**T.C.**
**PAMUKKALE ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

# SOSYAL MEDYA HESAPLARININ KURAL TABANLI PROFİL ÇIKARIMI: KULLANICI SİYASİ EĞİLİMLERİNİN SINIFLANDIRILMASI VE ARAŞTIRILMASI

**YÜKSEK LİSANS TEZİ**

**EMRE ŞAHİN**

**DENİZLİ, AĞUSTOS - 2018**

**T.C.**
**PAMUKKALE ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

# SOSYAL MEDYA HESAPLARININ KURAL TABANLI PROFİL ÇIKARIMI: KULLANICI SİYASİ EĞİLİMLERİNİN SINIFLANDIRILMASI VE ARAŞTIRILMASI

**YÜKSEK LİSANS TEZİ**

**EMRE ŞAHİN**

**DENİZLİ, AĞUSTOS - 2018**

## KABUL VE ONAY SAYFASI

**EMRE ŞAHİN** tarafından hazırlanan "**SOSYAL MEDYA HESAPLARININ KURAL TABANLI PROFİL ÇIKARIMI: KULLANICI SİYASİ EĞİLİMLERİNİN SINIFLANDIRILMASI VE ARAŞTIRILMASI**" adlı tez çalışmasının savunma sınavı 03.08.2018 tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği / ~~oy çokluğu~~ ile Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Tezi olarak kabul edilmiştir.

| Jüri Üyeleri | İmza |
|---|---|
| Danışman<br>Prof. Dr. Sezai TOKAT | ..... |
| Üye<br>Dr. Öğr. Üyesi Elif HAYTAOĞLU | ..... |
| Üye<br>Dr. Öğr. Üyesi Mahmut SİNECEN | ..... |

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 16/08/2018 tarih ve ..33/93.... sayılı kararıyla onaylanmıştır.

.....

Prof. Dr. Uğur YÜCEL

Fen Bilimleri Enstitüsü Müdürü

**Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğine beyan ederim.**

**EMRE ŞAHİN**

# ÖZET


**SOSYAL MEDYA HESAPLARININ KURAL TABANLI PROFİL ÇIKARIMI: KULLANICI SİYASİ EĞİLİMLERİNİN SINIFLANDIRILMASI VE ARAŞTIRILMASI**

**YÜKSEK LİSANS TEZİ**

**EMRE ŞAHİN**

**PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**

**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**(TEZ DANIŞMANI: PROF. DR. SEZAİ TOKAT)**

**DENİZLİ, AĞUSTOS - 2018**

İnsanların kişilik özelliklerini, düşüncelerini ve tercihlerini tahminlemek pazarlama ve reklamcılık gibi alanların ilgisini çeken konulardır. Geçmişte bunun için anket ve testler kullanılırken günümüzde sosyal medya kullanımının artmasıyla birlikte bu platformlar bu çalışmalar için daha uygun ortamlar haline gelmiştir. Ancak sosyal medya üzerindeki bilgiler test ve anketlerdeki gibi belirli bir amaca yönelik bilgiler olmadığı için işlemesi ve analiz edilmesi daha zordur. Bununla ilgili sürekli olarak yeni yöntemler, araçlar ve teknikler önerilmekte ve geliştirilmektedir.

Profil çıkarımı, sosyal medyada önemli bir tahminleme konusudur. Bu çalışmada sosyal medya kullanıcılarının profil çıkarım çalışmalarında kullanılabilecek, arkadaşlık benzerliğine dayanan 13 farklı özellik önerilmekte ve bu özelliklerden bir kural tabanı elde edilmektedir. Önerilen bu özelliklerin kullanılabilirliğini test etmek için Twitter kullanıcılarının siyasi parti eğilimlerini tahminlemeye yönelik bir uygulama gerçekleştirilmiştir. Uygulamada siyasi partilerin ve liderlerinin resmi Twitter hesaplarının arkadaş ve takipçi listesinden elde edilen benzerlik değerleri kullanılarak sınıflandırma ve kümeleme işlemleri gerçekleştirilmiştir. Sınıflandırma için farklı eğitim veri setleriyle sistemin başarımı test edildiği zaman minimum %70.81 *doğruluk*, %77.40 *kesinlik* ve %70.81 *f1* değeri elde edilirken genel olarak *k*-NN ile karar ağacına göre daha başarılı sonuçlar elde edilmiştir. Ancak karar ağacı yönteminin faydası, sınıflandırmayı görsel olarak ifade edebilmesi ve kural tabanı çıkarımına yardımcı olmasıdır. Kümeleme için de aynı özellikler *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle farklı veri setleri üzerinde test edilmiştir. Testler sonucunda *k*-Ortalamalar ile daha başarılı sonuçlar alınmasına rağmen Bulanık *c*-Ortalamalar örneklere küme üyelik dereceleri atadığı için yanlış kümelenen örneklerin gözlenmesine, karşılaştırılmasına olanak sağlamaktadır. *k*-Ortalamalar yönteminde örnekler birden fazla kümenin merkezine aynı uzaklıkta olsalar bile yalnızca bir tane kümeye dahil edilerek gösterilmekte ve diğer kümeler göz ardı edilmektedir. Halbuki Bulanık *c*-Ortalamalar'ın kullanıldığı durumda, eğilim analizinde katkı sağlayacak şekilde, örnekler üyelik derecesine bağlı olarak farklı kümelere farklı oranda dahil olabilmektedir.



**ANAHTAR KELİMELER:** Sosyal Medya, Profilleme, Kural Tabanı, Sosyal Medya Analizi, Profil Çıkarımı

# ABSTRACT


**RULE BASED PROFILE EXTRACTION OF SOCIAL MEDIA ACCOUNTS: CLASSIFICATION AND EXPLORATION OF USER POLITICAL TENDENCIES**

**MSC THESIS**

**EMRE ŞAHİN**

**PAMUKKALE UNIVERSITY INSTITUTE OF SCIENCE**

**COMPUTER ENGİNEERİNG**

**(SUPERVISOR: PROF. DR. SEZAİ TOKAT)**

**DENİZLİ, AUGUST 2018**

Marketing and advertising sectors are very interested in predicting people's personality, ideas and preferences. In the past, surveys and tests have been used for this aim, however with the increasing use of social media, these platforms have become more suitable environments. On the other hand, the information on social media is more difficult to process and analyze as there is no specific purpose-oriented information such in tests and surveys. Therefore new methods, tools and techniques have being explored and developed.

Profile extraction is an important prediction topic in social media. In this study, 13 different features based on the resemblance of fundamental friendship are proposed and a rule base is derived from these features, which can be used in the profile extraction of social media users. To test the usability of these features, an application was implemented to anticipate the political party tendencies of Twitter users. In this application classification and clustering was carried out using similarity values obtained from official Twitter accounts of friends and followers of political parties and leaders. When the system was tested for performance with different training data sets for classification, the results of *k*-NN were generally more successful than decision tree where the minimum performance scores are 70.81% accuracy, 77.40% precision and 70.81% *f1*. However, the advantage of the decision tree method is that it can visually express the classification and helps to extract the rule base. For clustering, the same features have been tested on different data sets with *k*-Means and Fuzzy *c*-Means methods. Even though more successful results are obtained with the *k*-Means as a result of the tests, Fuzzy *c*-Means allows to observe and compare the wrong clustered samples because they assign cluster membership values to the samples. In the *k*-Means method, even though the samples are at the same distance from the center of more than one cluster, this samples are included in only one cluster, other clusters are ignored. However in cases where Fuzzy *c*-Means is used, different samples can be included in different clusters depending on the degree of membership, as a contribution to the trend analysis.



**KEYWORDS:** Social Media, Profilling, Rule Base, Social Media Analysis, Profile Extraction

# İÇİNDEKİLER

**Sayfa**

# ŞEKİL LİSTESİ

**Sayfa**

# TABLO LİSTESİ

**Sayfa**

# SEMBOL LİSTESİ

$s(\boldsymbol{A})$ : **A** Kümesinin Eleman Sayısı

$d(\boldsymbol{a}, \boldsymbol{b})$ : **a** Noktasının **b** Noktasına Olan Uzaklığı

$\boldsymbol{KA}$ : Kullanıcının Arkadaş Listesinden Oluşan Küme

$\boldsymbol{PA}$ : Siyasi Partinin Arkadaş Listesinden Oluşan Küme

$\boldsymbol{PAA_i}$ : Siyasi Partinin Arkadaş Listesindeki $i$. Arkadaşının Arkadaş Listesinden Oluşan Küme

$\boldsymbol{PTA_i}$ : Siyasi Partinin Takipçi Listesindeki $i$. Takipçisinin Arkadaş Listesinden Oluşan Küme

$\boldsymbol{LA_i}$ : Siyasi Partinin Liderleri Arasında Yer Alan $i$. Liderin Arkadaş Listesinden Oluşan Küme

# KISALTMALAR LİSTESİ

**API** : Application Programming Interface, Uygulama Programlama Arayüzü
**BFPI** : Big Five Personality Inventory, Beş Büyük Kişilik Envanteri
**BLR** : Bayesian Logistic Regression, Bayes Lojistik Regresyon
**CART** : Classification and Regression Trees, Sınıflandırma ve Regresyon Ağaçları
**GBDT** : Gradient Boosted Decision Tree, Gradyan Güçlendirmeli Karar Ağacı
**GSM** : Global System for Mobile Communications, Mobil İletişim için Küresel Sistem
**ID3** : Iterative Dicholomiser 3, Tekrarlı Dicholomiser 3
**JCR** : Journal Citation Reports, Dergi Atıf Raporları
***k*-NN** : *k*-Nearest Neighbourhood, *k*-En Yakın Komşuluk
**LDA** : Latent Dirichlet Allocation, Gizli Dirichlet Ataması
**LIWC** : Linguistic Inquiry and Word Count, Dilbilimsel Araştırma ve Kelime Sayısı
**LSA** : Latent Semantic Analysis, Gizli Anlamsal Analiz
**LSE** : Least Squares Estimation, En Küçük Kareler Tahmini
**ME** : Maximum Entropy, Maksimum Entropi
**MLE** : Maximum Likehood Estimation, Maksimum Olabilirlik Tahmini
**MNB** : Multinomial Naive Bayes, Çok Terimli Naive Bayes
**MRR** : Mean Reciprocal Rank, Karşılıklı Sıralamaların Ortalaması
**MRC** : Machine Readable Dictionary, Makine Tarafından Okunabilir Sözlük
**NTIT** : News and Twitter Interaction Topic, Haber ve Twitter Konu Etkileşimi
**NodeXL** : Network Overview, Discovery and Exploration Add-in for Excel, Excel için Ağ Değerlendirme, Keşif ve Araştırma Eklentisi
**SMO** : Sequential Minimal Optimization, Sıralı Minimum Optimizasyon
**SSCI** : Social Sciences Citation Index, Sosyal Bilimler Atıf Dizini
**SVM** : Support Vector Machine, Destek Vektör Makinesi
**T2T** : Tag to Tag, Etikete karşı Etiket
**TF-IDF** : Term Frequency – Inverse Document Frequency, Terim Sıklığı, Ters Doküman Sıklığı
**ULAKBİM** : Ulusal Akademik Ağ ve Bilgi Merkezi
**URL** : Uniform Resource Locator, Standart Kaynak Bulucu
**Weka** : Waikato Environment for Knowledge Analysis, Bilgi Analizi için Waikato Ortamı
**WWW** : World Wide Web, Dünya Çapında Ağ

# ÖNSÖZ

Her zaman çalışmanın beraberinde başarıyı getireceğine inanan birisi olarak dünyanın ancak okuyarak, araştırarak ve sorgulayarak gelişebileceğini düşünmekteyim. Bu doğrultuda benim bu düşünce ve fikirlerimin oluşmasında emeği olan bugüne kadarki tüm öğretmenlerime, lisans ve yüksek lisans eğitimim sırasında üzerimde emeği geçen tüm hocalarıma, Pamukkale Üniversitesi Bilgisayar Mühendisliği ve Yönetim Bilişim Sistemleri bölümlerinin tüm öğretim üyelerine, iyi ve kötü günümde yanımda olan tüm arkadaşlarıma teşekkür ederim.

Tez çalışmamın her aşamasında yardımcı olan, değerli bilgilerini ve zamanını esirgemeyen tez danışmanım sayın Prof. Dr. Sezai TOKAT'a ve hayatımın her döneminde bana gerek maddi gerekse manevi olarak destek olan başta annem olmak üzere aileme teşekkürü bir borç bilirim.

# 1. GİRİŞ

Günümüzde içerisinde yaşadığımız toplum bilgi toplumu olarak adlandırılmaktadır ve bu toplumun unsurlarını hayatımızın her alanında görmek mümkündür. Artık herkes bir akıllı telefon kullanmakta, her evde bir bilgisayar ya da tablet ve tüm şirketlerde bilgi teknolojileri birimleri bulunmaktadır. Bilgisayarın insan hayatına girmesinden yaklaşık yarım asır sonra bu veriler, insan gözlemiyle mümkün olmayan bilgi çıkarımları yapılabilecek boyutlara ulaşmıştır. Bu süreçte sadece üretilen ve depolanan veri artmamış aynı zamanda bilgiye erişim de hızlanmıştır. 2012 yılında tüm dünyada günlük üretilen ve depolanan veri miktarının 2.5 kentrilyon byte olduğu hesaplanmıştır (Dülger 2015). Günümüzde üretilen büyük verinin önemli bir kısmını artık hayatımızın vazgeçilmezi haline gelen sosyal medya platformlarındaki veriler oluşturmaktadır. Bunların başında da Facebook, Twitter, Instagram ve Youtube gelmektedir.

Geniş bir kesim sosyal medyayı çeşitli topluluklar arasında dağıtık olarak bulunan günlük konuşma dilindeki içerikleri farklı bir ortamda kendine özel yöntem ve sembollerle üretmek, yaymak ve iletişim sağlamak amacıyla kullanmaktadır. Geleneksel ve endüstriyel medyanın aksine sosyal medya yazar ile okuyucu arasındaki sınırları kaldırır. İçeriğin üretilmesi, paylaşılması ve tüketilmesi birbirine bağlı olaylardır (Zeng ve diğ. 2010). Bu kadar büyük, değerli ve hızlı veri akışının olduğu sosyal medya üzerinde bu içeriklerin anlamlandırılması oldukça önem kazanmıştır.

Genel olarak büyük veriden ve doğrudan ele alındığı zaman bir anlam ifade etmeyen verilerden oluşan veri setlerinden belli aşamalardan geçirilerek anlamlı bilgi çıkarılması işlemine veri madenciliği denilmektedir. Bu çalışmaların sosyal medya üzerinde kullanıcıların sınıflandırılması, kümelenmesi vb. şeklinde olanları ise profilleme çalışmaları olarak adlandırılmaktadır. Literatürde profilleme ile ilgili yapılan çalışmalarda kullanıcının profil bilgilerinden, paylaşmış olduğu içerik bilgilerinden ve sosyal ağ bilgilerinden elde edilen özelliklerden yararlanılmıştır. Özellikler nitel ve nicel olarak ikiye ayrılmaktadır. Nitel özellikler duygu ve anlam

gibi sayısal olmayan verilerden oluşurken, nicel özellikler oran, uzunluk ve benzerlik gibi sayısal değerlerden oluşmaktadır.

Bu çalışmada sosyal medya platformlarında kullanıcı profil çıkarımını sağlayacak nicel özellikler kümesi ve kullanıcıların sınıflandırılmasını sağlacak bir kural tabanı sistemi önerilmektedir. Önerilen bu nicel özelliklerin profilleme çalışmalarındaki başarımını ölçmek için Twitter üzerinde 4 farklı uygulama gerçekleştirilmiştir. Bu uygulamalarda sınıflandırma ve kümeleme için ikişer farklı yöntemle bu özelliklerin farklı kombinasyonlarından yararlanılarak Twitter kullanıcılarının siyasi eğilimleri ve siyasi gruplaşmalar tespit edilmeye çalışılmıştır.

Çalışmanın genel akışı şu şekilde verilebilir: Bölüm 2'de sosyal medya ve veri madenciliği gibi kavramlar açıklanmaya çalışılmış ve bu alanda literatürde yapılmış çalışmalar özetlenmiştir. Bölüm 3'te çalışmada kullanılan yöntemler, verilerin elde edilmesi, hazırlanması, işlenmesi ve analizlerin yapılması için kullanılan teknolojiler ve sistemin başarımının ölçülmesi için kullanılan başarım ölçütleri açıklanmıştır. Bölüm 4'te verilerin toplanması, ön işleme adımları, yapılan analiz çalışmaları ve sonuçları anlatılmıştır. Son olarak Bölüm 5'te ise genel olarak önerilen sistemin değerlendirmesi yapılmış, yöntemler karşılaştırılmış ve gelecekte yapılması planlanan çalışmalar hakkında bilgiler verilmiştir.

# 2. SOSYAL MEDYA VE VERİ MADENCİLİĞİ

Veri madenciliği bilgi teknolojilerinin gelişiminin doğal bir sonucu olarak ortaya çıkmıştır. Dünya çapında yapılan bir araştırma sonuçlarına göre 2013-2020 yılları arasında üretilecek veri miktarının 44 zettabaytı aşması beklenmektedir, ki bu miktar 2013 yılına kadar üretilen veri miktarının yaklaşık 10 katıdır (Kashyap ve diğ. 2014). Büyük ve sayısız veri havuzunda toplanan, depolanan, hızla büyüyen çok sayıdaki veri güçlü araçlar olmaksızın insanların geleneksel yöntemlerle onları anlama, kavrama ve analiz etme yeteneğini aşmıştır. Çok miktardaki veri, güçlü veri analiz araçlarına duyulan ihtiyaç ile birleştiğinde, büyük veri havuzlarından toplanan veriler "veri mezarları" ya da nadiren ziyaret edilen veri arşivlerine dönüşürler ve böylece veri açısından zengin, bilgi açısından fakir bir durum ortaya çıkar (Han ve diğ. 2011). Böyle ortamlarda önemli kararlar genellikle veri havuzlarında depolanan bilgi açısından zengin verilere değil, karar vericinin sezgisine dayanır, çünkü karar vericinin çok miktarda ki veri içerisinde gömülü olarak bulunan bilgiyi ayıklayacak araçları yoktur (Han ve diğ. 2011). Bundan dolayı çoğunlukla kullanıcıların ya da uzmanların bilgi tabanına bilgiyi elle girdiği uzman sistemler ve bilgi tabanlı teknolojiler geliştirilmeye çalışılmaktadır, ancak bu sistemler genellikle yönlenmeye ve hataya eğilimlidir, ayrıca oldukça maliyetli ve zaman alıcıdırlar (Han ve diğ. 2011).

Veri ve bilgi arasındaki genişleyen uçurum veri mezarlarını bilginin "altın külçelerine" dönüştüren veri madenciliği araçlarının sistematik gelişimini gerektirmektedir (Han ve diğ. 2011). Veri madenciliği, büyük veriden temel bilgi ve öngörülerin elde edilmesini sağlayan yöntem ve algoritmaları içeren, veritabanı sistemleri, istatistik, makine öğrenmesi ve örüntü tanıma gibi bağlantılı alanlardan kavramları bir araya getiren disiplinler arası bir alandır (Zaki ve Jr. 2014). Aslında veri madenciliği, modellemek ve model uygulamak, hipotez doğrulamak ve genelleştirmek gibi işlem sonrası adımların yanı sıra veri çıkarımı, veri temizleme, veri birleştirme, veri indirgeme ve özellik çıkarımı gibi işlem öncesi görevleri içeren büyük bir bilgi keşif sürecidir (Zaki ve Jr. 2014). Kısaca veri madenciliği, geleneksel yöntemlerle anlamlandırılamayan veri yığınlarından bilgi çıkarım süreci olarak tanımlanabilir.

Veri madenciliği bilimsel ve mühendislik çalışmaları, bankacılık ve finans, müşteri ilişkileri yönetimi, sahtekarlık tespiti, güvenlik/istihbarat, eğitim, sağlık ve

biyomedikal, pazarlama ve reklamcılık gibi birçok farklı alana uygulanabilmektedir. Aslında bu çalışmaların tamamında amaç insanların doğru karar vermesini, mevcut durumdan en iyi şekilde yararlanmasını sağlamak ya da onları etkileyerek kendi düşünceleri etkisine almaya çalışmaktır. Örneğin; pazarlama ve reklamcılık açısından bakıldığı zaman günümüzde yalnızca bir ürünü tanıtmak yeterli olarak görülmüyor, doğru ürünün doğru kişiye doğru zamanda doğru şekilde tanıtılması ya da aynı ürünün kişiye özel farklı yöntemlerle tanıtılması hedefleniyor. Bunun için de insanların kişilik özelliklerinin, arkadaşlık ilişkilerinin, fikirlerinin ve düşüncelerinin olabildiğince doğru bir şekilde çıkarımsanması gerekmektedir.

Medeniyetin insan hayatına girmesinden sonra insanı tanımlayan özelliklerin en başında sosyal bir varlık oluşu gelmektedir. Çünkü yerleşik hayata geçilmesiyle birlikte insanoğlu hayatını sürdürebilmek için tek başına yaşamayı bırakarak gruplar halinde yaşamaya başlamıştır. Dolayısıyla bir toplum kavramı ortaya çıkmış ve bu toplum içerisinde bireyler birbirleriyle bir şekilde iletişim ve etkileşim içerisinde olmuşlardır. Sözlük karşılığına bakıldığı zaman sosyal kelimesi topluma ait, toplumsal, içtimai, insanların toplum içinde ve birlikte yaşamaları ile ilgili anlamlarına gelmektedir. Toplumdaki her bir bireyin davranışı, iletişim şekli veya biçimi aslında o bireyin kişiliği hakkında bilgiler vermektedir. Bu bağlamda toplumsal davranışları ve insan ilişkilerini inceleyen bilim dalına da sosyoloji denilmektedir (Wani 2017). Ayrıca sosyoloji için, toplum bilimi, sosyal organizasyon ve sosyal değişim bilimi, insan ilişikilerini inceleyen bilim dalı ve kolektif davranış bilimi gibi tanımlar da yapılmaktadır (Wani 2017).

Çoğu insana göre, Dünya Çapında Ağ (WWW, World Wide Web)'ın varoluş amacı bağlantı kuran bir ağ oluşturmaktır fakat sosyal ağlar, insan oğlunun avcı ve toplayıcı olduğu dönemden beri hayatımızdadır (Kadushin 2012). İnsanlar ilişkileriyle ve bağımlılıklarıyla birbirlerine bağlıdırlar. Akrabalık ve aile ilişkileri, kabileler, ongunlar, hiyerarşiler, mahallelerin, köylerin ve şehirlerin sorumluluk ve ilişki ağları hepsi birer sosyal ağ örneğidir. Akrabalık ilişkilerinin yanı sıra modern toplumda insanlar postalarının getirilmesi, çimlerinin biçilmesi ve iyi bir restoran tavsiyesi almak için birbirlerine ihtiyaç duyarlar (Kadushin 2012).

Sosyoloji açısından bakıldığı zaman sosyal ağ kavramı uzun bir tarihe ve çok geniş bir içeriğe sahiptir (Scott 2017). Sosyal ağlar genel olarak dijital ve çevrimiçi

ağlar olarak algılansa da aslında karşılıklı ilişkileri, siyasi işbirliği ve bağlantıları, işletmeler arasındaki ekonomik işlemleri, ülkeler ve uluslararası ajanslar arasındaki jeopolitik ilişkileri de içermektedir. Yıllar boyunca sosyologlar her türlü ilişkiyi incelemek ve yorumlamak için sistematik analiz biçimleri tasarlamışlardır (Scott 2017).

Dünyayı daraltan ve küçülten iletişim ve taşımacılık teknolojileriyle nitelendirilen 21. yüzyıl coğrafik olarak birbirinden uzakta olan bireylerin sosyal iletişim açısından yakın ilişkiler kurabilmesini sağlar (Prell 2012). Paradoksal olarak bizler küçük ve geniş bir dünyada yaşıyoruz: her birimiz yerel topluluklar içinde bulunurken aslında aynı zamanda dünyayı ve daha fazlasını kapsayan bağlantılar kurabiliyoruz ve daha da ilginç olanı bizler aslında dünyanın her hangi bir yerindeki birisine sadece birkaç bağlantı kurarak ulaşabiliyoruz (Milgram 1967). Bu karşılıklı bağımlılıkların hem olumlu hem de olumsuz sonuçları vardır. Dünyanın hem geniş hem de küçük olduğu hakkındaki bilgimiz bizi eylemlerimizin dünya çapında etkileri olabileceği ve yerel sosyal çevremizde bulunan birisinin uzak bir coğrafyadaki birisiyle çok iyi bağ kurabileceği hakkında daha duyarlı hale getirmektedir (Prell 2012). Ancak aynı bağımlılıklar asosyal ve terörist grupların nefret ve hoşgörüsüz mesajlarını yerel bağlantıları kullanarak dünyaya yayabileceğini de göstermektedir (Prell 2012).

1980'lerin ortalarında kişisel bilgisayaların halka tanıtılmasıyla başlayan dijitalleşme süreci günümüzde oldukça önemli bir noktaya ulaşmıştır. 1989 yılında icat edilip 2001 yılında halka açık hale getirilen WWW insanların hayatına bir devrim olarak girmiş ve hemen ardından iTunes ve Vikipedi faaliyete başlamıştır. Bundan sonraki gelişmeler çok hızlı ve kısa sürede gerçekleşmiştir. 2003'te LinkedIn, 2004'te Facebook, bir yıl sonra YouTube, Flickr, Reddit ve 2006'da Twitter kurulmuştur. 2007 yılında akıllı telefonların piyasaya sürülmesi ise insanları bu platformlara her an her yerden ulaşabilir hale getirerek farklı bir boyut kazandırmıştır. Aynı yıl içerisinde Tumblr tanıtılırken 2008 yılında Spotify yayın hayatına başlamıştır. Bu gelişmeleri 2010 yılında tablet bilgisayarların icadı ve Instagram'ın kurulması, 2011 yılında da Pinterest ve Google+'ın kurulması takip etmiştir. Facebook gibi çevrimiçi platformlar, akıllı telefonlar ve akıllı hesaplama sistemlerini içeren dijital teknolojiler, son on yıl içinde ulaşımdan eğitime, aile hayatından aktivizme, cezaevi yönetiminden vahşi

yaşamın korunmasına kadar geniş bir yelpazeye yayılmış olduğundan önemli bir toplumsal olguyu sunmaktadırlar (Marres 2017). Kısacası artık dijital bir toplumda yaşamaktayız. Yeni dijital teknolojiler günlük yaşamızı, sosyal ilişkilerimizi, siyasi, ticari, ekonomik olayları, bilgi üretim ve yayılım şekillerini etkilemektedir (Lupton 2014). Hemen hemen herkes gününün büyük bir kısmını çevrimiçi olarak geçirmekte, toplumun büyük bir kısmı akıllı telefon, tablet bilgisayar gibi cihazları sürekli olarak yanlarında taşımaktadır, hatta giyilebilir cihazlar sayesinde gece ve gündüz farketmeksizin sürekli olarak bedensel faaliyetler izlenebilmektedir (Lupton 2014). Yine benzer şekilde haber, müzik ve televizyon yayınlarına dijital platformlar ve cihazlar üzerinden erişilebilmekte, LinkedIn, Facebook ve Twitter gibi sosyal medya araçları üzerinden arkadaşlık ve kurumsal ağlar kurulabilmekte, fotoğraf ve videolar YouTube, Instagram ve Flickr üzerinden dünya ile paylaşılabilmekte, merak edilen bir konu hakkında Google ve Bing gibi bir arama motoru kullanılarak birkaç saniye içerisinde bilgi edinilebilmektedir (Lupton 2014). Bu gelişmelerin sosyoloji için önemli etkileri vardır. Fakat toplumun süregelen dijitalleşmesi sadece önemli bir araştırma konusu sunmaz, aynı zamanda sosyal araştırmaların bizzat toplumda oynadığı rolü değiştirme potansiyeline de sahiptir (Marres 2017). Bu durum son yıllarda "Dijital Sosyoloji" olarak adlandırılmaya başlanan sosyal araştırma biçimlerinin önemli bir ayırt edici özelliğini açıklığa kavuşturmaya yardımcı olur (Marres 2017).

Dijital sosyolojinin bir başka boyutu, sosyal araştırma yapmak için büyük dijital veri kümelerinin kullanılmasıdır. Sosyal medya platformları, dijital sosyolojinin en önemli veri kaynalarını oluşturmaktadır. Çünkü sosyal medya, yüz yüze soru sormak ya da anket yapmak gibi geleneksel yöntemlerin aksine kişilerin gerek paylaşımlarıyla, gerek beğenileriyle, gerekse diğer bireylerle (arkadaş, akraba vb.) olan ilişkileriyle kendi kimliklerini tanımlayabildikleri platformlardır. Bu platformlardan alınan büyük miktardaki veriler, veri madenciliği, makine öğrenmesi gibi bilgisayar bilimlerine ait ya da istatistiksel yaklaşımlar kullanılarak analiz edilerek çıkarımlar yapılabilir.

Sosyal medya üzerinde yapılan veri madenciliği genel olarak kullanıcıların yaş, cinsiyet, eğitim durumu, siyasi görüş gibi kişisel özelliklerinin tahminlenmesi, kişilik özelliklerinin tespit edilmesi, arkadaş ve içerik tavsiye sistemlerinin geliştirilmesi,

içeriklerin konu ve başlığa göre sınıflandırılması, ağ kavramı ve ağ teorisine dayalı olarak gerçekleştirilen sosyal ağ analizi çalışmalarını içermektedir. Bu çalışmalarda kullanıcıların profil bilgilerinden, paylaşmış olduğu içeriklerden ve sosyal ağ yapısından elde edilen özelliklerden yararlanılmaktadır.

## 2.1 Cinsiyet, Yaş, Eğitim Düzeyi, Siyasi Görüş ve Sahte Hesapların Tahminlenmesine Yönelik Çalışmalar

Sosyal medya kullanımının artmasıyla birlikte kötüye kullanım ve sahtekarlık olayları da bir hayli artmıştır. Genellikle bu tarz olaylar için sahte hesaplar açılmakta ve bu işlemler bu hesaplar üzerinden gerçekleştirilmektedir. Yanlış bilgi ve söylentiler içeren paylaşımların yapılması insanları yanlış yönlendirilebilmekte ve sonuçları hata ve kazalara neden olabilmektedir. Böyle hesapların kimlik bilgilerinin belirlenmesi hem suçluların tespit edilmesine hem de böyle olayların azalmasına yardımcı olacaktır. Bunun yanında günümüzde reklam ve pazarlama hizmetlerinin gelişmesiyle birlikte kişiye özel reklamcılık anlayışı ortaya çıkmıştır. Sosyal medya da reklam ve pazarlama için oldukça sık kullanılan platformlardır. Sosyal medya üzerinde kullanıcıların kişisel özelliklerinin daha iyi bilinmesi demek onların ilgi alanına giren reklamların gösterilmesi veya onlara özel reklamların hazırlanacağı dolayısıyla reklamların doğru hedef kitleye ulaşacağı anlamına gelmektedir. Ayrıca günümüzde sosyal medya insanların siyasi görüşleri üzerinde de oldukça büyük etkilere neden olmaktadır. Yakın zamanda gerçekleşen Facebook skandalı bunun en büyük örneklerinden birisidir (Tuttle 2018).

Bununla ilgili literatürde, Facebook yorumları incelenerek naive Bayes, *k*-En Yakın Komşuluk (*k*-NN, *k*-Nearest Neighbourhood) ve Destek Vektör Makinesi (SVM, Support Vector Machine) yöntemleriyle kullanıcıların yaş, cinsiyet ve eğitim düzeyleri tahminlenmiş, naive Bayes yöntemiyle yaş, cinsiyet ve eğitim düzeyi için sırasıyla %89.67, %90.85 ve %86.15 test *doğruluk* değerleri elde edilmiştir (Talebi ve Köse 2013). Belçika'nın sosyal medya platformu olan Netlog üzerindeki 1537283 adet paylaşımdan doğal dil işleme yöntemleriyle çıkarılan ve bu paylaşımları yapan kullanıcıların profil bilgilerinden elde edilen özellikler kullanılarak SVM yöntemiyle kullanıcıların yaş ve cinsiyetleri belirlenmeye çalışılmıştır (Peersman ve diğ. 2011).

Yine bir başka çalışmada Twitter üzerinde kullanıcılar, paylaşımların içerikleri ve profillerden elde edilen kişi adı, kullanıcı adı, lokasyon, Standart Kaynak Bulucu (URL, Uniform Resource Locator) ve açıklama gibi özelliklerden NGram modeliyle çıkartılan öznitelikler kullanılarak SVM, naive Bayes ve Dengeli Ayırma 2 (Balanced Winnow 2) yöntemleriyle erkek ve kadın şeklinde sınıflandırılmaya çalışılmıştır (Burger ve diğ. 2011). Çalışmada elde edilen en başarılı sınıflandırma için *doğruluk* değeri %92 iken yalnızca metinsel özellikler kullanılması durumunda elde edilen en yüksek *doğruluk* değeri %76'dır. Twitter üzerindeki kullanıcıların kuruluşlar, gazeteciler/medya bloggerları ve sıradan bireyler şeklinde sınıflandırıldığı bir başka çalışmada 8 farklı olay için 8 farklı yöntem ağ yapısal özellikleri, içeriklerinden elde edilen özellikler, konu dağılımlarından elde edilen özellikler ve etkinlik özellikleriyle test edilmiş ve en başarılı sonuçlar $k$=10 için $k$-NN ile elde edilmiştir (Choudhury ve diğ. 2012). Yapılan bu çalışmada elde edilen sonuçlar Amazon Mechanical Turk üzerinden etiketlenen sonuçlar ile karşılaştırılmış ve en başarılı durumda %88.73 doğru sınıflandırma elde edilmiştir.

Twitter üzerinden 3 farklı seçim olayına ait veriler çekilerek ve bir Mamdani tipi bulanık mantık sistemi kullanılarak kullanıcıların seçimlerdeki davranışları tespit edilmeye çalışılmıştır (Albornoz 2015). Çalışmada 3 öncül değişken ve 1 çıktı değişkeni kullanılarak kullanıcılar, içerik paylaşımlarına bağlı olarak saldırgan, karşıt, nötr, seçmen, propagandacı ve istenmeyen içerik paylaşıcı şeklinde sınıflandırılmıştır. Öncü değişkenlerin araştırma uzayı tweet frekansı olarak belirlenirken, çıktı değişkeni ise bir kullanıcının seçimdeki olası niyeti olarak belirlenmiştir. Ayrıca tweet frekans dağılımı, Kuvvet Yasası (Power Law)'na uyan tüm seçimlerde bu modelin uygulanabileceği belirtilmiştir.

Rao ve diğ. (2010), Twitter kullanıcılarının tweet ya da durum mesajlarını, sosyal ağ yapısını ve iletişim davranışlarını kullanarak kullanıcının cinsiyet, yaş, bölgesel köken ve politik eğilim gibi gizli özelliklerini otomatik olarak ortaya çıkarmaya çalışmışlardır. Kullanıcı özelliklerinin çıkarımı açısından durum mesajı ya da tweet içeriklerinin, sosyal ağ yapısı özellikleri ve iletişim davranışlarından daha değerli olduğu sonucuna varılmıştır. Çalışmada SVM yöntemi toplum dilbilimsel özellik modeli, NGram özellik modeli ve ikisinin birleşimi olan yığınsal model olmak üzere 3 farklı model için test edilmiş ve en yüksek *doğruluk* değerleri, cinsiyet ve yaş

için yığınsal modelde %72.33 ve %74.11 olarak, bölgesel köken için toplum dilbilimsel modelde %77.08 ve politik eğilim için NGram modelinde %82.84 olarak hesaplanmıştır.

Pennacchiotti ve Popescu (2011) Twitter kullanıcılarını, profil özellikleri, içerik paylaşım davranışlarına ait özellikler, sosyal ağ özellikleri ve içeriklerin dilsel özelliklerinden yararlanılarak politik eğilimi, etnik kökeni ve bir işletmeye bağlılıkları açısından ayrı ayrı ikili sınıflandırmışlardır. Her bir özellik grubunun etkisini ölçmek için Gradyan Güçlendirmeli Karar Ağacı (GBDT, Gradient Boosted Decision Tree)'nı farklı özellik gruplarıyla test etmişler ve dilsel özelliklerin bilhassa konu temelli durumlarda daha güvenilir olduğu sonucuna varmışlardır. Ayrıca sosyal ağ özelliklerinin toplanmasının oldukça zor olmasına rağmen hedef sınıfın aktif Twitter varlığı açısından zengin olan ünlü kişilerden oluşması durumunda oldukça değerli olduğunu belirtmişlerdir.

Twitter üzerinde gerçekleştirilen başka bir çalışmada 2010 A.B.D kongre ara dönem seçimleri sırasında Twitter kullanıcılarının sağ ve sol şeklindeki siyasi gruplaşmaları belirlenmeye çalışılmıştır (Conover ve diğ. 2011). Çalışmada SVM ile tweet içerikleri ve ağ özelliklerine göre iki farklı sınıflandırma gerçekleştirilmiştir. Her iki sınıf için de belirlenen etiketlerden en az bir tanesini içeren 252 bin tweet toplanmış ve bunların bir kısmı eğitim veri seti olarak kullanılmak üzere sağ, sol ve belirsiz şeklinde 2 kişi tarafından elle etiketlenmiştir. Metinsel içerikler üzerinde, bir sözcüğün metin içerisinde ne kadar önemli olduğunu bulmaya yarayan Terim Sıklığı, Ters Doküman Sıklığı (TF-IDF, Term Frequency-Inverse Document Frequency) ve metinler içerisindeki konunun belirlenmesi için kullanılan Gizli Anlamsal Analiz (LSA, Latent Semantic Analysis) algoritmalarıyla elde edilen skorlar kullanılarak yapılan sınıflandırmada %90.8 *doğruluk* değeri elde edilmiştir. Anma (mention) ve tekrar paylaşma (retweet) ağları çıkarılarak elde edilen ağlar 2 gruba ayrılacak şekilde kümelenmeye çalışılmıştır, daha sonra da hangi kümenin sağı hangi kümenin solu temsil ettiği belirlenmiştir. Bu durumda elde edilen *doğruluk* değeri ise %95 olarak hesaplanmıştır.

Facebook üzerindeki kullanıcıların cinsiyet ve etnik köken gibi gizli özelliklerinin çıkartılmasına yönelik bir çalışmada, isim özelliklerinden elde edilen isim modeli, yorumlardan elde edilen özelliklerden oluşan içerik modeli ve iki modelin

birleşimi olan toplam 3 model 2 farklı yöntemle test edilmiştir (Rao ve diğ. 2011) İçerik modeli yorumlardan Gizli Dirichlet Ataması (LDA, Latent Dirichlet Allocation) kullanılarak konu ve etiket çıkarımıyla oluşturulmuştur. Her bir model SVM, naive Bayes ve hiyerarşik naive Bayes yöntemleriyle çalıştırılarak sınıflandırma sonuçları karşılaştırılmıştır. En başarılı *doğruluk* sonuçları cinsiyet için %80.1 ve etnik köken için %81.1 olarak iki modelin birleşiminin yarı denetimli hiyerarşik naive Bayes ile çalıştırılması durumunda elde edilmiştir.

## 2.2 Kişilik Özelliklerinin Belirlenmesine Yönelik Çalışmalar

Kişilik bireyin etkileşimini ve tercihlerini etkileyen insan davranışlarının temel dayanağıdır. Kişilik aslında bireyin iş hayatını ve başarımını, özel hayatını ve ilişkilerini, arkadaş ilişkilerini, düşünce, istek ve tercihlerini etkileyen çok önemli bir kavramdır. Günümüzde bu kişilik özelliklerinin tespit edilmesi için en kabul görmüş yöntem kişilik testlerinin/anketlerinin yapılmasıdır. Ancak insanlar anket ve testler sırasında kaygı ve endişelerden dolayı bazı sorulara yanlış ya da eksik cevaplar verebilmektedir, bu da test ve analizlerin sonuçlarını etkileyebilmektedir.

Sosyal medya ve blogların oluşmasıyla, insanlar bu platformları sosyal, siyasi veya etnik olaylarla ilgili düşünce ve fikirlerini belirterek kendilerini ifade edebilecekleri mecralar olarak görmeye başlamışlardır. Böylece kişilik özelliklerinin tespiti için sosyal medya verilerinin kullanıldığı çalışmalar ortaya çıkmıştır (Rosen ve Kluemper 2008).

Literatürde kişilik özelliklerinin belirlenmesine yönelik birçok çalışma olmasına rağmen sosyal medya üzerinden kullanıcıların kişilik özelliklerinin belirlenmesine yönelik ilk çalışma kendilerine Beş Büyük Kişilik Envanteri (BFPI, Big Five Personality Inventory) testi uygulanmış Facebook kullanıcıları üzerinde gerçekleştirilmiştir (Golbeck ve diğ. 2011[a]). Çalışmada kullanıcıların profillerinden elde edilen özellikler, içerikler kullanılarak Dilbilimsel Araştırma ve Kelime Sayısı (LIWC, Linguistic Inquiry and Word Count) veritabanı üzerinden elde edilen istatistiksel özellikler, Makine Tarafından Okunabilir Sözlük (MRC, Machine Readable Dictionary) veritabanı üzerinden elde edilen davranışsal özellikler ve Genel Soruşturma (General Inquirer) veritabanı üzerinden elde edilen duygusal puan

özellikleri kullanılarak, bu özelliklerle BFPI'ın her bir kişilik özelliği arasındaki Pearson Korelasyonu hesaplanmaya çalışılmıştır. Buna bağlı olarak da Gauss Süreci (Gaussian Processes) ve ZeroR yöntemleriyle regresyon analizi yapılarak her bir kullanıcının kişilik özelliği değerleri tahminlenmeye çalışılmıştır. Twitter üzerinde ise kullanıcıların takipçi ve arkadaş sayısı gibi profillerinden elde edilen özellikler, anma, cevaplama (reply), etiket ve bağlantı sayısı gibi içeriklerinden doğrudan elde edilen özellikler, LIWC veritabanı üzerinden elde edilen içeriklerin istatistiksel özellikleri ve ağ yoğunluğu gibi yapısal ağ özellikleri kullanılarak Çoklu Doğrusal Regresyon Analizi (Multiple Linear Regression Analysis) yöntemiyle BFPI kişilik özellik değerleri tahminlenmeye çalışılmıştır (Golbeck ve diğ. 2011[b]).

Literatürde OMD, Sanders ve SemEvam2013 olarak geçen veri setleri üzerinde naive Bayes, SVM ve Çok Katmanlı Algılayıcı Sinir Ağı (Multi Layer Perceptron Neural Network) yöntemleriyle profil bilgilerinden elde edilen özellikler olmaksızın yalnızca içerik bilgilerinden yararlanılarak kişilik çıkarımının yapılmaya çalışıldığı çok etiketli bir sınıflandırma uygulaması mevcuttur (Lima ve de Castro 2014). Bu çalışmada BFPI'nin her bir kişilik özelliği için bir ikili sınıflandırma olacak şekilde toplamda 5 tane ikili sınıflandırma uygulanmış ve ortalama %83'lük bir başarı elde edilmiştir.

Pratama ve Sarno (2015) tarafından Twitter kullanıcılarının kişillikleri yalnızca içerik verileri kullanılarak belirlenmeye çalışılmıştır. Çalışmada daha önceden BFPI uygulanmış kullanıcılar metin içeriklerinde yer alan kelimelerle temsil edilmiş, *k*-NN, Çok Terimli naive Bayes (MNB, Multinomial naive Bayes) ve SVM yöntemleriyle her bir kelimenin her bir kişilik özelliği için ikili sınıflandırma skoru elde edilmiş ve bu skorlara bağlı olarak da kullanıcıların kişilikleri tahminlenmeye çalışılmıştır.

Facebook üzerinde gerçekleştirilen bir çalışmada, myPersonality veritabanına ait verilerle kullanıcıların profil ve demografik bilgilerinden elde edilen özellikler, paylaşmış olduğu içeriklerin LIWC veritabanı üzerinden elde edilen istatistiksel özellikler kullanılarak kişilik özellikleri tahminlenmeye çalışılmıştır (Ateş 2014). Çalışmada tahminleme işlemi için Sıralı Minimal Optimizasyon (SMO, Sequential Minimal Optimization), J48 ve Rastgele Orman (Random Forest) yöntemleri test edilmiş ve en başarılı sonuçlar SMO ile elde edilmiştir. Yine veri seti olarak myPersonality veritabanının kullanıldığı başka bir çalışmada, durum mesajlarından

Bag of Words ve NGram yöntemleri kullanılarak kelimeler çıkartılmış, TF-IDF kullanılarak bu kelimelerden sözcük vektörleri elde edilmiştir (Alam ve diğ. 2013). Daha sonra bu sözcük vektörleri SMO, MNB ve Bayes Lojistik Regresyon (BLR, Bayesian Logistic Regression) yöntemleri ile sınıflandırılarak BFPI özellikleri tahminlenmeye çalışılmıştır. Veri setinin %66'sının eğitim %34'ünün test verisi olarak kullanıldığı 10 katlı çapraz doğrulama sonucunda en başarılı sonuçlar MNB ile elde edilmiştir.

## 2.3 Arkadaş ve İçerik Öneri Sistemi Çalışmaları

Teknolojinin gelişmesiyle birlikte platformlar arası rekabet artmış ve çeşitlilik ortaya çıkmıştır. Bunun için de bir kullanıcının yalnız platforma üye olması platformun başarımı için yeterli görülmemekte bu üyeliğin kalıcılığının ve devamlılığının sağlanması önem kazanmaktadır. Bu yüzden de bu platformlar için arkadaş ve içerik öneri sistemleri oldukça önemlidir. Bu sayede kullanıcıların ilgisini çeken kişiler ya da içerikler önerilerek kullanıcıların platforma bağlılılığı arttırılmaya çalışılmaktadır.

Friendster, MySpace ve Orkut gibi sosyal ağ sitelerinden toplanan kullanıcı profilleri üzerinde insanların ilgi alanları dikkate alınarak yapılan bir öneri sistemi geliştirilmiştir. Bu çalışmada, geleneksel öneri sistemlerindeki gibi insanların geçmiş davranışlarını dikkate almak yerine, ilgi alanları ve kişilikler arasındaki ilişkileri görselleştiren bir ağ tarzı olan İlgi Haritası (Interest Map) yöntemi kullanılmıştır (Liu ve Maes 2005). Böylece İlgi Haritası kullanılarak geleneksel öneri sistemlerine göre daha doğru tavsiyeler üretildiği, gerçek hayattaki bir insana ait ilgi alanlarının ve tercihlerinin sezgisel ve görsel olarak daha doğru biçimde modellendiği belirtilmiştir.

Bir müzik topluluğu sitesi olan Last.fm üzerinde gerçekleştirilen bir çalışmada basit bir etiket analiz metodu olarak kullanıcının sahip olduğu bir parçayla ilişkili etiketleri ve ilişki puanını belirlemek için o parçaya ait genel etiketler ve etiketleme frekansı kullanılarak ve İşbirlikçi Filtreleme Öneri Sistemi (Collaborative Filtering Recommender System**)** fikri sürdürülerek, kullanıcı-etiket (user-tag) matrisinden bazı benzer kullanıcılar bulunmuş ve benzer kullanıcı etiketlerini içeren parçalar önerilmiştir (Firan ve diğ. 2007). Geleneksel Parça Tabanlı Öneri Yaklaşımı (Track Based Recommender Approach) temel alınarak sonuçlar karşılaştırıldığında, etiket

tabanlı kullanıcı profillerinin kullanımının başarımı önemli derecede arttırdığı belirtilmiştir.

Michlmayr ve Cayzer (2007) bir kullanıcıyı, kullanıcı tarafından kullanılan etiketlerin düğümleri ve bu etiketler arasındaki ilişkilerin de kenarları oluşturduğu bir profil çizgesi şeklinde temsil ederek, etiketleme verilerinden kullanıcı profil çıkarımını amaçlamışlardır. Ayrıca birlikte meydana gelen ve geçici olan bilgileri bir araya getirerek Etiket Ekleme (Add-Tag) algoritmasını geliştirmişler, bu sayede hem etiket çiftleri arasındaki kenar ağırlıklarını belirlemiş hem de dinamik kullanıcı profilleri için çizge görselleştiricisi sağlamışlardır.

Hung ve diğ. (2008) kullanıcı ve içerikleri Goldberg diğ. (1992)'nin yapmış olduğu gibi derece vektörleriyle ifade etmek yerine tanımlayıcı etiketlerle ifade etmişlerdir. Yapmış oldukları çalışmada del.icio.us kullanıcı profil etiketleri ve yer imi olarak eklenen URL etiketlerinden oluşan ve her bir hücrenin o satırdaki kullanıcı etiketine sahip bir kullanıcının içeriklerinde o sütundaki içerik etiketine sahip olma oranını gösterdiği Etikete karşı Etiket (T2T, Tag-to-Tag) matrisini kullanmışlardır. Yeni bir içerik veya kullanıcı geldiğinde bu matris kullanılarak öneri skorlarını hesaplamışlar ve bu değerin eşik değerinden büyük olması durumunda içeriği/kullanıcıyı önerilebilir olarak kabul etmişlerdir. Ayrıca bu çalışmada etiketler, kullanıcının kendisi tarafından eklenenler Kişisel Görünüm (Personal View), kendisi dışındaki kullanıcılar tarafından eklenenler Sosyal Görünüm (Social View) olarak değerlendirilmiştir.

Twitter üzerinden TV programlarının rating sıralamalarının tahminlenmeye çalışıldığı bir çalışmada programlarla ilgili önceden belirlenmiş etiketler için atılan tweetler toplanarak doğal dil işleme ve duygu analizi aşamalarından geçirilerek tahminleme işlemi gerçekleştirilmiştir (Akarsu ve Diri 2016). Çalışmada doğal dil işleme süreçleri için Türkçe için geliştirilmiş olan Zemberek kütüphanesinden yararlanılmış, tahminleme için Weka üzerinde SMO, J48, MNB ve Rastgele Orman yöntemleri test edilmiştir. İçlerinden en başarılı sınıflandırma yapan Rastgele Orman yöntemi kullanılarak sıra gözetmeksizin ilk 5 sıradaki programlar tahminlenmeye çalışılmış ve dizi programları için %68.5, ana haber programları için %59.7 ve yarışma programları için %92.1 başarım elde edilirken sırasıyla aynı program türleri için 0.367,

0.497 ve 0.628 Karşılıklı Sıralamaların Ortalaması (MRR, Mean Reciprocal Rank) değerleri elde edilmiştir.

Drobnjak (2012) yapmış olduğu yüksek lisans tez çalışmasında öğrencilerin birbirlerine ders çalışma arkadaşları aradığı bir sistem üzerinde çizge teorisine dayalı arkadaş öneri sistemiyle bulanık mantık teorisine dayalı arkadaş öneri sistemini karşılaştırmıştır. Çalışmada verilerin çok boyuttan indirgenmesi için Sammon Haritalama (Sammon Mapping) yöntemi kullanılırken, seçilmiş bir öğrencinin seçilmiş konulara aitlik sınırını belirlemek için Top-N yöntemi kullanılmıştır. Arkadaş önerilerinin yapılması sırasında da Bulanık *c*-Ortalamalar (Fuzzy *c*-Means) yönteminden yararlanılmıştır.

## 2.4 Konu ve Duygu Sınıflandırma Çalışmaları

Teknolojinin gelişmesi ve dijitalleşmeyle her geçen gün içerik üretim süresi kısalırken, üretilen içerik miktarı artmakta, içeriklere ulaşım süresi azalmaktadır. Bunun yanında bazı olumsuz durumlar da ortaya çıkmaktadır, bunların başında da üretilen içerik kalitesinin düşmesi, yalan veya yanlış haberlerin ortaya çıkması gelmektedir. Ayrıca çok fazla içerik üretimi ve kaynak olduğu için bunların konu ve içerik olarak sınıflandırılması geleneksel yöntemler kullanılarak mümkün olmamaktadır. Her fikir, düşünce ve söylem beraberinde destek ve tepkiyi getirmektedir, sosyal medya ve bloglar gibi dijital platformalarda yapılan içerik paylaşımları da aynı şekilde olumlu veya olumsuz dönütler almaktadır. Ancak bu platformlardaki dönütler milyonları bulabildiği için bunların analizi ve değerlendirilmesi geleneksel ve sıradan yöntemlerle yapılamamaktadır. Bu nedenlerden dolayı metin madenciliği, içerik sınıflandırılması ve duygu analizi olarak adlandırılan birçok çalışma alanı ortaya çıkmıştır.

Twitter üzerinde 14777 tweet kullanılarak yapılan bir çalışmada LDA algoritması kullanılarak her bir tweetin konu (topic) modeli çıkartılmış ve buna bağlı olarak da tweet içerikleri *k*-NN, naive Bayes, MNB, SVM ve Maksimum Entropi (ME, Maximum Entropy) yöntemleriyle 2 farklı konuya ayrılmaya çalışılmıştır (Çoban ve Özyer 2016). Çalışmada tweet içerikleri Bag of Words ve NGram modelleri kullanılarak elde edilen öznitelikler vektörü şeklinde temsil edilmiştir. Genel anlamda

en başarılı sonuç Bag of Words modeli ile elde edilen öznitelikler kullanılarak MNB yöntemiyle elde edilmiş ve %92 doğru sınıflandırma gerçekleştirilmiştir.

PHP tabanlı bir uygulamada Twitter kullanıcılarının istenen sayıdaki paylaşımlarına bakılarak bu paylaşımların haber, kültür ve siyaset kategorilerine ayrılmasını ve her bir kategoride duygu analizi yapılarak paylaşımların olumlu, olumsuz ve nötr şeklinde sınflandırılmasını sağlayan bir sistem Twitter platformu için geliştirilmiştir (Baykara ve Gürtürk 2017). Bu uygulama paylaşımların kategorilere ayrılması sırasında Libchart kütüphanesinden yararlanarak, her bir paylaşımın içerisindeki her bir kategoriye ait kelime sayısını belirlemekte, daha sonra ise naive Bayes algoritmasını kullanarak her bir paylaşımı en çok kelimeye sahip olduğu kategoriyle etiketlemektedir. Kategori bazında paylaşımları duygu analizi açısından olumlu, olumsuz ve nötr şeklinde sınıflandırmak için her bir kategoride yer alan paylaşımların içerisindeki olumlu, olumsuz ve nötr kelimelerin sayıları kullanılarak skorlar hesaplanmakta ve bu skorlar kullanılarak sınıflandırma işlemi gerçekleştirilmektedir.

Twitter üzerinde yapılan başka bir çalışmada da Türkiye'deki 3 Mobil İletişim için Küresel Sistem (GSM, Global System for Mobile Communications) firması hakkında atılan tweetler olumlu ve olumsuz şeklinde sınıflandırılmaya çalışılmış, böylece hakkında en çok olumlu ve olumsuz tweet atılan GSM firmaları belirlenmeye çalışılmıştır (İşeri ve diğ. 2017). Çalışmada dilbilgisi özelliklerinden, söz sanatı özelliklerinden, istatistiksel özelliklerden ve sözlük bilgisi özelliklerinden oluşan bir özellik vektörü ve NGram ile elde edilen özellik vektörleri kullanılarak iki farklı özellik vektörü için *k*-NN algoritması kullanılarak sınıflandırma işlemi gerçekleştirilmiştir.

Twitter ve haberlere ait konu dağılımlarının, bu konuların ortaya çıkışını ve yayılmasını etkileyen içsel ve dışsal faktörlerin incelendiği bir çalışmada sosyal medya ve haber konularını birlikte öğrenen ve konuların etkilerini dikkatle ele alan Haber ve Twitter Konu Etkileşimi (NTIT, News and Twitter Interaction Topic) modeli adında yeni bir konu model ve hem haber medyası hem de sosyal medya da yaygınlıklarını dikkate alarak konuların popülerliklerine göre sıralanmasını sağlayan Social Rank adında yeni bir sistem önerilmiştir (Jeya ve Bala 2018). Çalışmada aynı olay için bile haber ve Twitter konularının odak noktalarının farklı olabileceği, genellikle haber

konularının Twitter konularına göre daha etkili olacağı ve konular genellikle baskın veri kaynağında ortaya çıksa da bazen ilk olarak bir veri kaynağında görülen konuların başka bir veri kaynağında baskın konu olabileceği sonuçlarına varılmıştır.

Başka bir çalışmada ürün üretilmeden önce uzun vadede, ürün piyasaya sürüldükten sonra da kısa vadede kapalı ağlarda ürün konularının ortaya çıkışı tahminlenmeye çalışılmıştır (Peng ve diğ. 2018). Çalışmada içerisinde yazar çeşitliliği ve rekabet diye adlandırılan iki yeni özelliği barındıran Ortaya Çıkacak Konuyu Öngörücü (ETP, Emerging Topic Predictor) adında bir çerçeve (framework) önerilmiştir. Bu çerçeve hem uzun hem kısa vadeli tahminler için Reddit ve PTT (https:/www.ptt.cc)'den alınan film yorumları üzerinde Rastgele Orman, SVM, Lojistik Regresyon ve GBDT ile test edilmiştir. Uzun vadeli tahminler için 285 gün öncesinde %91, 104 gün öncesinde %95 *doğruluk* elde edilirken, ürün piyasaya sürüldükten sonraki ilk haftada %92, ikinci ve üçüncü haftalarda %96 başarı elde edilmiştir.

## 2.5 Sosyal Ağ Analizi Çalışmaları

Sosyal ağ kavramı temeli çizge teoremine dayanan varlıkların düğümlerle, varlıklar arasındaki ilişki ve etkileşimlerin de kenarlarla temsil edildiği bir yapıdır. Sosyal ağ kavramındaki düğümler genellikle insanlar ya da insanların etkileşimde olduğu diğer varlıklardan oluşmaktadır. Hemen hemen her şeyi sosyal ağ yapısı kullanarak tanımlamak mümkündür. Ağ üzerindeki en önemli düğümlerin belirlenmesi, en kısa yolun tespiti, bir düğüm eklenmesi ya da çıkarılması durumunda oluşacak ağ yapısının tespiti vb. durumlar için çap, merkezilik, yakınlık gibi kavramlar ve çeşitli analiz yöntemleri geliştirilmiştir. Sosyal medya günümüzde sosyal ağ kavramının en önemli örneklerinden birisi olarak kabul edilmektedir. Sosyal ağ analizi için kullanılan kavramlar ve yöntemlerin bir sınıftaki arkadaşlık ilişkileri, ders–öğrenci ilişkileri gibi küçük ağ yapıları üzerinde uygulanması mümkünken sosyal medya ve e-posta trafiği gibi milyonlarca düğüm ve kenara sahip ağlarda bu analizlerin geleneksel yöntemlerle uygulanması mümkün değildir. Bu sebepten dolayı bu alanda yapılmış ve yapılmakta olan birçok çalışma bulunmaktadır.

Disiplinlerarası bilimsel dergilerin bir göstergesinin belirlenmesi için yapılan çalışmada 7379 adet Dergi Atıf Raporları (JCR, Journal Citation Reports) bilimsel dergisine ait arasındalık merkeziliği, yakınlık merkeziliği, girdi derece merkeziliği, çıktı derece merkeziliği gibi sosyal ağ özellikleri ve atıf sayısı, referans sayısı, etki vb. istatistiksel özellikler incelenmiş ve arasındalık merkeziliğinin bunun için en uygun özellik olduğu belirtilmiştir (Leydesdorff 2007).

Türkiye'nin bilimsel yayınlarının ağ yapısı ile ilgili yapılmış olan ilk çalışmada Sosyal Bilimler Atıf Dizini (SSCI, Social Sciences Citation Index) ve Ulusal Akademik Ağ ve Bilgi Merkezi (ULAKBİM) veri tabanlarındaki veriler üzerinde Türkiye'nin ortak yazarlık ağları analiz edilmiştir (Gossart ve Özman 2009). Ayrıca bu çalışmada Ankara ve İstanbul'daki üniversitelerin yayın üretiminin yüksek olduğuna değinilirken aynı zamanda SSCI ve ULAKBİM veri tabanlarındaki ortak yazarlık ağlarının farklılık gösterdiğine dikkat çekilmiştir.

Bilimsel yayınların ağ yapısıyla ilgili yapılmış başka bir çalışmada 1968 ve 2009 yılları arasında Hacettepe Üniversitesi kaynaklı olarak yapılan yayınlar sosyal ağ analizi yöntemleriyle yazarların/kurumların birliktelik analizi ve bu birlikteliklerin zamansal değişimi incelenmiştir (Al ve diğ. 2012). Çalışmada özellik olarak arasındalık merkeziliği kullanılırken bu değerlerin hesaplanması ve verilerin görselleştirilmesi için CiteSpace'den yararlanılmıştır.

Sert ve diğ. (2014) NodeXL kullanarak Twitter üzerinden her birisi ayrı önem taşıyan 5 farklı tarih için #akademikzam etiketiyle atılmış tweetleri incelemişlerdir. Her bir tarih için ayrı sosyal ağ oluşturmuşlar, Fuchterman-Reingold ve Harel-Koren algoritmalarından yararlanarak bu ağları görselleştirmişlerdir. Sosyal ağ analizinin temel kavramları olan girdi derecesi, çıktı derecesi, girdi/çıktı yakınlık merkeziliği, yoğunluk, ortalama derece, çap ve kümeleme katsayısı kavramları kullanılarak ağlardaki en çok tweeti atan başka bir deyişle en yoğun kullanıcılar ya da kendisinin anılarak tweete dahil edildiği düğümler (ki bunlar genellikle milletvekilleri ve bakanlardır) tespit edilmiş ve yine bu çalışmada kümeleme analizi yapılarak ağlardaki gruplaşmalar tespit edilmiştir. Ayrıca bu çalışmanın başka bir önemli sonucu ise oluşturulan 5 farklı ağın da çaplarının 6 ya da 7 birim olmasıdır. Bu da oluşan ağların Küçük Dünya Hipotezini (Small World Phenomenon) desteklediğini göstermektedir.

Twitter üzerindeki içerik verilerinin girdi derece dağılımının Kuvvet Yasası olasılık dağılımına uygun olup olmadığını anlamak için 5 farklı dönemde #akademikzam etiketiyle atılmış tweet, anma ve cevaplamalar incelenmiştir (Gürsakal ve diğ. 2014). Çalışmada grafiksel yöntemin yanında D istatistikleri ve Kolmogorov Smirnov Testi adında 2 farklı matematiksel yöntem kullanılmıştır. Ölçekleme parametresine ait En Küçük Kareler Tahmini (LSE, Least Squares Estimation) değerleri 4. dönemde sapmalı olduğu için Maksimum Olabilirlik Tahmini (MLE, Maximum Likehood Estimation) değerleri hesaplanmıştır. Çalışmanın sonucunda deneysel verilerin Kuvvet Yasası olasılık dağılımına uyduğu sonucuna varılmıştır.

NodeXL üzerinde yapılan sosyal ağ analizi çalışmaları arasında Twitter takipçi ve arkadaş ağının görselleştirilerek toplulukların iletişim sürecinin keşfedilmeye çalışıldığı bir çalışma (Choi ve diğ. 2012), ağdaki önemli içerikleri bulmak ve ağ içeriğini analiz etmek için Youtube'daki ameliyat videoları ağının incelendiği bir çalışma (Hansen 2011) ve işbirlikçi ortamlarda etkileşimin incelendiği ve sonucunda kişilerin profesyonel geçmişleriyle benzerlik gösteren kişilerle iletişim kurma eğiliminde oldukları sonucuna varılan bir çalışma (Doran ve diğ. 2011) yer almaktadır.

# 3. YÖNTEMLER, KULLANILAN TEKNOLOJİLER VE BAŞARIM ÖLÇÜTLERİ

## 3.1 Yöntemler

Bu çalışmada Twitter üzerinden elde edilen veriler üzerinde *k*-NN ve karar ağacı yöntemleri kullanılarak verilerin sınıflandırılmasına yönelik ve *k*-Ortalamalar (*k*-Means) yöntemiyle verilerin kümelenmesine yönelik çalışmalar yapılmıştır. Ayrıca *k*-Ortalamalar yöntemiyle kümeleme sonuçlarının karşılaştırılabilmesi için bir bulanık mantık kümeleme yöntemi olan ve veri madenciliği uygulamalarında da sıklıkla kullanılan Bulanık *c*-Ortalamalar yöntemi kullanılarak kümeleme yapılmış ve elde edilen sonuçlar *k*-Ortalamalar yöntemiyle karşılaştırılmıştır.

Yapılan tüm analiz uygulamalarında yöntemin, parametrelerin, özelliklerin ve veri setinin başarımını ölçmeye yönelik birçok test yapılmıştır. Bu testler sırasında *doğruluk, kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü ve ayrıca sınıflandırma uygulamalarında her bir sınıfın başarısı için *duyarlılık* ölçütü kullanılmıştır. Bunun yanında elde edilen sonuçların anlaşılabilirliğini kolaylaştırmak için karışıklık matrisinden yararlanılmıştır.

### 3.1.1 *k*-En Yakın Komşuluk

Veri madenciliği, makine öğrenmesi gibi alanlarda en sık kullanılan örnek tabanlı öğrenme yapan sınıflandırma algoritmalarından birisidir (Taşçı ve Onan. 2016). Örnek tabanlı öğrenme algoritmalarında başlangıçta verilen eğitim setiyle sistem eğitilir ve belirlenen özellikler için yeni gelen verinin eğitim setindeki verilerle benzerliklerine bakılarak sınıflandırma yapılır (Mitchell 1997).

*k*-NN algoritmasında sınıflandırılmak üzere yeni bir örnek geldiği zaman daha önceden eğitim setinde bulunan verilerin her birisiyle belirlenen özelliklerin uzaklıkları hesaplanır ve bu uzaklık değerleri içerisinde en düşük uzaklığa sahip *k* tanesine bakılarak yeni örneğin sınıfı belirlenir (Özkan 2016).

*k*-NN algoritmasında uzaklık hesabı için birçok uzaklık ölçütü kullanılsa da en sık kullanılan uzaklık ölçütü Öklid uzaklığıdır. Bu çalışmada da *k*-NN algoritmasındaki uzaklıklar Öklid uzaklığına göre hesaplanmıştır.

*n* boyutlu bir düzlemde bulunan $\boldsymbol{a}$ ve $\boldsymbol{b}$ noktalarının arasındaki Öklid uzaklığı şu şekilde hesaplanır:

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{1}$$

### 3.1.2 Karar Ağaçları

Sınıflandırma amaçlı kullanılan bir başka makine öğrenmesi ve veri madenciliği yöntemi karar ağaçlardır. *k*-NN algoritması gibi karar ağaçları da örnek tabanlı öğrenme yapan bir sınıflandırma yöntemidir. Ağaç yapısı kök, dallar ve yapraklardan meydana gelir. En uçtaki çocukları olmayan düğümlere yaprak, en üst kısımdaki ebeveyni olmayan düğüme kök, aralardakilere ise dal adı verilir (Özkan 2016). Karar ağacında önceden öğrenme verisi olarak verilen örneklerden yararlanarak sınıf ayrımı sağlayan özelliklere göre kökten yaprağa doğru dallanmalar meydana gelir ve yaprakta sınıf etiketi yer alır (Şeker 2013).

Karar ağacı oluştururken en önemli konulardan birisi hangi özelliklere göre dallanmanın gerçekleşeceğidir. Bu konuda Twoing, Gini ve regresyon ağaçları gibi daha farklı algoritmalar olmasına rağmen en sık kullanılan algoritmalar C4.5, C5.0, Sınıflandırma ve Regresyon Ağaçları (CART, Classification and Regression Trees) ve Tekrarlı Dichotomiser 3 (ID3, Iterative Dichotomiser 3) gibi entropi tabanlı algoritmalardır.

ID3, Ross Quinlan tarafından 1986 yılında geliştirilmiştir. ID3 algoritması her bir düğümde sınıflar için en büyük bilgi kazancını veren özellikleri bularak çok yönlü bir ağaç oluşturur. Ağaç, olabilecek maksimum boyuta kadar büyür ve daha sonra ağacın gücünü arttırmak için belirsiz verilere budama işlemi uygulanır (Quinlan 1986).

C4.5, özelliklerin kategorik olması kısıtını ortadan kaldıran ID3'ün gelişmiş bir versiyonudur. C4.5 eğitilmiş ağaçları, eğer-ise yapısındaki kurallara dönüştürür. Eğer herhangi bir kural çıkarıldığı zaman *doğruluk* artıyorsa, o kuralın ön koşulu çıkarılarak budama gerçekleştirilir (Decision_Tree 2017).

C5.0 ise Quinlan tarafından geliştirilmiş ve lisansı kendisine ait olan son sürümdür. C5.0, C4.5'e göre daha az bellek kullanmakta ve daha küçük kural setleri oluşturmaktadır ancak C4.5 ile daha doğru sonuçlar elde edilebilmektedir (Decision_Tree 2017).

CART, C4.5'e oldukça benzer bir algoritmadır ancak yalnızca numerik hedef değişkenleri desteklemesi ve kural kümelerini hesaplamaması bakımından C4.5'ten ayrılır. CART, her düğümde en büyük bilgi kazanımını sağlayan özellik ve eşik değerini kullanarak ikili ağaçlar oluşturur (Decision_Tree 2017).

Bu tez çalışmasında karar ağacının dallanması için CART algoritması kullanılmıştır.

### 3.1.3 *k*-Ortalamalar

*k*-Ortalamalar algoritması, kümeleme için en sık kullanılan geleneksel algoritmalardan birisidir. *k*-Ortalamalar yönteminde başlangıçta *k* adet küme için merkez noktaları belirlenir. Bu merkez noktalarını belirlemeyle ilgili çok farklı yaklaşımlar olmasına rağmen iki yöntem çok sık kullanılmaktadır. Bunlardan birincisi başlangıç durumunda her bir kümenin merkez noktasının belirtilmesi, diğeri veri seti içerisinden *k* adet verinin her birisinin bir küme merkezi olarak seçildiği kmeans++'dır (KMeans 2018). Bu çalışmada da başlangıç merkezlerinin seçimi için kmeans++ yöntemi kullanılmıştır.

*k*-Ortalamalar algoritmasına, yeni bir veri geldiği zaman bu verinin her bir küme merkezine olan Öklid uzaklığı hesaplanır ve veri en düşük uzaklığa sahip olan kümeye dahil edilir. Kümeye yeni veri eklendiği zaman da küme merkezleri yeniden hesaplanır (Qi ve diğ. 2017).

Her yeni kümeleme işleminden sonra her bir kümenin merkez noktasının o kümeye ait verilere olan Öklid uzaklıklarının kareleri toplanarak küme içi değişim değerleri hesaplanır (Özkan 2016). Bir $i$ kümesine ait $m$ adet veri örneği, $x_j$'nin bu örneklerden $j$.si ve $M_k$ nin de $k$ kümesinin merkez noktası olduğu kabul edilirse, $k$ kümesi için küme içi değişim değeri şu şekilde hesaplanır:

$$e_k^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ji} - M_{ki})^2 \tag{2}$$

Son olarak da tüm kümelere ait küme içi değişim değerleri toplanarak karesel hata değeri hesaplanır.

$$E_K^2 = \sum_{k=1}^{K} e_k^2 \tag{3}$$

Bu algoritmanın amacı belirlenen $k$ adet küme için her adımda karesel hata değerini minimize eden kümeler oluşturmaktır. Bu doğrultuda karesel hata değerinin her iterasyonda bir öncekine göre düşük olması beklenir (Özkan 2016).

### 3.1.4 Bulanık *c*-Ortalamalar

Bulanık mantık, sadece bir kümeye üyeliğin olduğu Aristo mantığının genelleştirilmesi ile elde edilmiştir ve gerçek dünyada kesin sınırların olmadığı durumlarda faydalı sonuçlara ulaşılmasını sağlar (Elmas 2016).

*k*-Ortalamalar algoritması, bilinen en eski kümeleme yöntemlerinden birisi olmasının yanında bazı sorunları da beraberinde getirmektedir. Örneğin; *k*-Ortalamalar algoritmasında bir veri örneği bir kümeye ait ise 1, değilse 0 ile gösterilmektedir ancak gerçek dünya için bu durum pek mümkün değildir. Çünkü gerçek dünya örneklerinde birden fazla kümeye ya da gruba farklı derecelerde bağlılıklar söz konusu olabilmektedir. Bulanık *c*-Ortalamalar yönteminde gerçek dünyaya uygun olarak bir veri örneği tüm kümelere farklı üyelik dereceleriyle dahil olabilmektedir (Ross 2004). Bulanık *c*-Ortalamalar yöntemi, ilk olarak 1981 yılında

Jim Bezdek tarafından daha önceki kümeleme yöntemlerinin geliştirilmesiyle ortaya atılmıştır.

Yine benzer şekilde *k*-Ortalamalar algoritması bir veri örneğinin küme merkezlerine uzaklığına göre kümeleme yapmaktadır ancak bir veri örneğinin birden fazla küme merkezine aynı uzaklıkta olduğu durumda bu veri örneğinin ait olduğu küme her iterasyonda sürekli olarak rastgele değişir. Fakat Bulanık *c*-Ortalamalar algoritmasında bu kümeler eşit üyelik derecesine sahip olacak şekilde temsil edilirler (Stetco ve diğ. 2015).

## 3.2 Kullanılan Teknolojiler

Bu tez çalışmasında verilerin elde edilmesi, depolanması, temizlenmesi ve analizlerin gerçekleştirilmesi sırasında birçok teknolojiden yararlanılmıştır. Bu bölümde bu teknolojiler kısaca açıklanmaya çalışılmıştır.

### 3.2.1 Veri Kaynağı: Twitter

Twitter, kullanıcıların fikir, düşünce ve haberlerini en fazla 280 karakterden oluşan metinlerle paylaştığı elektronik platformdur (Rigolin 2018). Twitter'da kullanıcılar diğer bireyler ya da gruplarla iletişim kurmak ve onların güncellemelerinden haberdar olmak için birbirlerini takip eder ve etkileşimde bulunurlar.

Jack Dorsey, Noah Glass, Biz Stone ve Evan Williams tarafından 2006 yılında kurulan Twitter'ın kuruluş amacı, bir grup arkadaşın birbiriyle iletişim kurmasını sağlamak ve bunların kaydını tutmaktı. 2006'dan sonra sağladığı haber ve sosyal ağ hizmetlerinin gelişmesiyle birlikte popülerliği sürekli olarak artmış ve 2017'nin üçüncü çeyreği itibariyle 330 milyondan fazla aktif kullanıcıya ulaşmıştır (Rigolin 2018).

### 3.2.2 Programlama Dilleri ve Platformlar

Çalışmanın gerçekleştirilmesi sırasında 2 farklı programlama dilinden yararlanılmıştır. 4. bölümde yer alan Uygulama 1, Uygulama 2, verilerin toplanması ve ön işlenmesi gibi işlemler Python programlama dili kullanılarak PyCharm Editörü üzerinde gerçekleştirilmiştir. Çalışma kapsamında gerçekleştirilen diğer uygulamalar olan Uygulama 3 ve Uygulama 4 ise Matlab üzerinde gerçekleştirilmiştir.

#### 3.2.2.1 Python

Geniş söz dizimine sahip, yorumlanabilir, nesne tabanlı ve açık kaynak kodlu yüksek seviye bir programlama dilidir. Python'ın yüksek seviyeli veri yapılarının, hızlı yazım özelliğiyle birleşmesi onu hızlı uygulama geliştirme ve farklı ortamlarda yazılmış kodların birleştirilmesi için uygun bir programlama dili yapmaktadır (Python 2018).

Python'un basitliği ve öğrenim kolaylığı okunabilirliği arttırdığı için programın bakım maliyetlerini düşürür. Python paket ve modül kullanımını destekleyerek modüler ve tekrar kullanılabilir kodlara sahip programlar yazılmasını sağlar.

#### 3.2.2.2 Matlab

Lineer cebir, istatistik, optimizasyon, nümerik analiz gibi matematiksel hesaplamalarda, iki ve üç boyutlu çizimlerde karmaşık hesaplamalarda oldukça başarılı olan MathWorks adlı bir firma tarafından geliştirilen çok paradigmalı bir bilimsel hesaplama ortamıdır. Adını Matrix Laboratuvarı (Matrix Laboratory)'nın kısaltmasından almıştır. Ayrıca Matlab'da C veya Fortran'da yazılmış fonksiyonlar doğrudan kullanılabilir (Matlab 2018).

İlk başlardaki geliştirilme amacı problemlerin çözümünde bilim adamlarına matris tabanlı çözümler sunarak yardımcı olmakken, günümüzde işaret işleme, kontrol, bulanık mantık, yapay sinir ağları gibi birçok alana özgü geliştirilmiş kütüphane ve araçlara sahip olduğu için hem akademik hem de sanayi alanında araştırma, geliştirme ve analiz aracı olarak yaygın bir şekilde kullanılmaktadır.

### 3.2.3 Verilerin Depolanması: MongoDB

NoSql, verilerin geleneksel yöntemlerde olduğu gibi ilişkisel veritabanlarında saklamak yerine büyük miktardaki verilerin dağıtık ve ilişkisel olmayan bir şekilde saklanmasını sağlayan veri yönetim sistemidir. NoSql veritabanları anahtar-değer (key-value), grafik tabanlı (graph DB's) ve döküman tabanlı olmak üzere 3'e ayrılır (Şavklı 2016).

MongoDB, C++ programlama diliyle geliştirilmiş döküman tabanlı bir NoSql veritabanıdır. Döküman tabanlı NoSql veritabanları verileri genellikle JSON formatında saklarlar. MongoDB ise verileri JSON'ın özel bir formatı olan BSON (Binary JSON) şeklinde saklar (Şavklı 2016).

### 3.2.4 Python Çerçeve, Kütüphane ve Modülleri

#### Anaconda

Python üzerinde kolay bir şekilde makine öğrenmesi ve veri bilimiyle ilgili çalışmalar yapmak için geliştirilmiş açık kaynak bir dağıtımdır. Haziran 2018 itibariyle Anaconda'nın kullanıcı sayısı 6 milyonu geçerken içerisinde barındırdığı paket sayısı da 1400'ü aşmıştır (Anaconda 2018). Ayrıca Anaconda, Scikit-learn, TensorFlow ve Scipy gibi karmaşık veri bilimi ve makine öğrenmesi ortamlarının kurulması, çalıştırılması ve yükseltilmesini kolaylaştırır.

#### tweepy

Github üzerinde barındırılan, Twitter platformuyla Uygulama Programlama Arayüzlerini (API, Application Programming Interface) kullanarak iletişim kurmayı sağlayan açık kaynaklı bir Python kütüphanesidir (Tweepy 2013). Mevcut birçok metodu sayesinde Twitter üzerinden kullanıcı bilgilerine, arkadaş listelerine, takipçi listelerine, paylaşım bilgilerine, beğenilere, listelere, trendlere vb. birçok bilgiye ulaşılmasını sağlamaktadır.

#### pymongo

İçerisinde MongoDB üzerinde çalışmayı sağlayan araçları barındıran Python dağıtımıdır. Ayrıca pymongo Python üzerinde MongoDB için doğal bir sürücü içerir, bu sayede ekleme silme, güncelleme vb. veritabanı işlemleri pymongo metodları sayesinde kolaylıkla gerçekleştirilebilir (PyMongo 2008).

**graphviz**

Yapısal bilgilerin soyut grafikler ve ağlarla temsil edilmesini sağlayan açık kaynak bir Python grafik görselleştirme kütüphanesidir (Ellson ve diğ. 2002). Ağ oluşturma, biyoenformatik, yazılım müdendisliği, veritabanı, web tasarımı, makine öğrenmesi ve diğer teknik alanlarda görselleştirme açısından önemli uygulamaları bulunmaktadır.

**matplotlib**

Dizilerin 2 boyutlu çizimlerinin oluşturulabilmesini sağlayan Python kütüphanesidir. Köken olarak Matlab grafik komutlarını taklit etmektedir. Ancak Matlab'dan bağımsız olarak nesne yönelimli olarak kullanıma olanak sağlamaktadır. matplotlib temelde saf Python ile yazılmasına rağmen, yoğun bir şekilde numpy ve diğer harici kodların kullanımı büyük dizilerde bile iyi başarım elde edilmesini sağlamaktadır (Hunter 2007).

**numpy**

Bilimsel hesaplamalar yapmaya yarayan temel Python kütüphanesi olan numpy, içerisinde güçlü bir çok boyutlu dizi nesnesini, gelişmiş ve sık kullanılan fonksiyonları, C/C++ ve Fortran kodlarının entegrasyonunu sağlayan araçları, kullanışlı lineer cebir, Fourier dönüşümü ve rastgele sayı fonksiyonlarını barındırır (Van Der Walt ve diğ. 2011).

numpy bilimsel kullanımının yanında genel veriler için çok boyutlu bir kapsayıcı olarak da kullanılabilir ve keyfi veri tipleri tanımlanmasına izin verir. Bu özellikler numpy'nin birçok veritabanıyla hızlı ve sorunsuz bir şekilde entegre olmasını sağlar.

**itertools.combinations**

itertools, APL, Haskell ve SML gibi yapılardan esinlenilerek yapı taşlarının oluşturulduğu bir dizi yineleyici uygulayan bir Python modülüdür (Bernard 2016).

Bu modülün combinations metodu, parametre olarak aldığı bir dizi elemanın yine parametre olarak aldığı bir değere göre tüm kombinasyonlarını döndürür (Nanjekye 2017). Matlab'ın nchoosek fonksiyonuyla benzerlik göstermektedir.

**sklearn**

numpy, scipy ve matplotlib üzerinde geliştirilmiş herkes tarafından erişilebilen, veri madenciliği, veri analizi ve makine öğrenmesi için basit ve etkili araçlar içeren açık kaynak bir Python kütüphanesidir (Buitinck ve diğ. 2013).

**sklearn.metrics**

İçerisinde skor fonksiyonları, başarım ölçütleri, ikili ölçütler ve uzaklık hesaplamaları barındıran sklearn kütüphanesinin bir modülüdür (Pedregosa ve diğ. 2011). metrics modülünün barındırdığı fonksiyonlardan bazıları şunlardır:

- *doğruluk* değerini hesaplayan accuracy_score fonksiyonu,
- *kesinlik* değerini hesaplayan precision_score fonksiyonu,
- *duyarlılık* değerini hesaplayan recall_score fonksiyonu,
- *kesinlik* ve *duyarlılık* değerinin harmonik ortalaması olan *f1* değerini hesaplayan *f1*_score fonksiyonu,
- ortalama *kesinlik* değerini hesaplayan average_precision_score fonksiyonu,
- sınıflandırma ölçütlerinin metinsel raporunu görüntüleyen classification_report fonksiyonu,
- karışıklık matrisini döndüren confusion_matrix fonksiyonudur.

**sklearn.neighbors**

sklearn kütüphanesinin *k*-NN algoritmasını uygulamak için geliştirilmiş modülüdür (Pedregosa ve diğ. 2011). İçerisinde genel olarak KNeighborsClassifier,

RadiusNeighborsClassifier, NearestNeighbors sınıflandırıcılarını ve uzaklık metriklerinin hesaplanması için kullanılan DistanceMetric sınıfını barındırır.

**sklearn.tree**

sklearn kütüphanesi içerisinde regresyon ve sınıflandırma amaçlı karar ağacı modellerini içeren modüldür (Pedregosa ve diğ. 2011). İçerisinde DecisionTreeClassifier ve ExtraTreeClasssifier gibi sınıflandıma amaçlı kullanılan sınıfları, DecisionTreeRegressor ve ExtraTreeRegressor gibi regresyon amaçlı kullanılan sınıfları ve karar ağaçlarının DOT formatında dışa aktarılmasını sağlayan export_graphviz fonksiyonunu barındırır.

### 3.2.5 Matlab Fonksiyonları ve Paralelleştirme

Yukarıda da bahsedildiği üzere çalışmanın 4. bölümünde yer alan Uygulama 3 ve Uygulama 4 Matlab platformu üzerinde gerçekleştirilmiştir. Uygulama 3'te kullanılan $k$-Ortalamalar yöntemi için kmeans, Uygulama 4'te kullanılan Bulanık $c$-Ortalamalar yöntemi için fcm fonksiyonları kullanılmıştır. Ayrıca özellik ve veri seti testleri sırasında tüm kombinasyonların oluşturulması için nchoosek fonksiyonundan ve bu kombinasyonların aynı anda çalıştırılmasını sağlamak için de parfor'dan yararlanılmıştır.

**kmeans**

Matlab'ın $k$-Ortalamalar algoritması için kullanılan fonksiyonudur. Genel tanımı $[idx,\ C] = kmeans(X,\ k)$ şeklinde olan kmeans fonksiyonu her birisi $p$ tane özelliğe sahip $n$ örneği temsil eden $n * p$'lik $X$ matrisini ve oluşturulacak küme sayısını belirten $k$'yı giriş parametresi olarak alır ve çıktı olarak da $n$ tane örneğin her birisinin ait olduğu kümenin indisini içeren $n * 1$'lik $idx$ matrisini ve $k$ adet kümenin her birisinin orta noktalarını gösteren $k * p$'lik $C$ matrisini geriye döndürür.

**fcm**

Matlab'ın Bulanık $c$-Ortalamalar yöntemi için geliştirilmiş fonksiyonudur. Genel tanımı $[centers,\ U] = fcm(data,\ Nc)$ şeklinde olan fcm fonksiyonu her

birisi $p$ tane özelliğe sahip $n$ örneği temsil eden $n * p$ 'lik $data$ matrisi ve oluşturulacak küme sayısını belirten $Nc$'yi giriş parametresi olarak alır ve çıktı olarak da $Nc$ adet kümenin her birisinin orta noktalarını gösteren $Nc * p$ 'lik $centers$ matrisini ve $n$ tane örneğin her bir küme için üyelik derecelerini içeren $n * Nc$'lik $U$ matrisini döndürür.

**nchoosek**

Matlab'ın binom katsayısı ve tüm kombinasyonların hesaplanması için geliştirilmiş fonksiyonudur. Bu fonksiyon iki farklı şekilde kullanılır. Birinci kullanım şekli olan $b = nchoosek(n,\ k)$ .şeklinde kullanıldığı zaman $n$ örneğin $k$ 'lı kombinasyonlarının sayısı olan $n!/\big((n-k)!\,k!\big)$ değerini döndürür. Diğer kullanım şekli olan $C = nchoosek(v,\ k)$ şeklinde kullanıldığı zaman da $n$ boyutlu $v$ vektörünü ve $k$ değerini giriş parametresi olarak alır ve çıktı olarak $v$ vektörünün $k$'lı tüm kombinasyonlarını içeren $k$ sütundan ve $n!\,\big((n-k)!\,k!\big)$ satırdan oluşan $C$ matrisini döndürür.

**parfor**

Matlab'ın geleneksel for döngüsünün yapmış olduğu işlemlerin paralel olarak gerçekleştirilmesini sağlar Paralelleştirme işlemi parfor tarafından otomatik olarak gerçekleştirilir. Her işçi geleneksel for döngüsünün her bir iterasyonunda gerçekleştirilen işlemlerin bir kısmını gerçekleştirir. parfor kullanılırken dikkat edilmesi gereken konu her iterasyonun birbirinden bağımsız olması gerektiğidir, bir iterasyonun sonucunun diğer iterasyonu etkilediği durumlarda parfor kullanılamaz.

## 3.3 Başarım Ölçütleri

Veri madenciliği, sosyal ağ analizi, finansal analiz, pazarlama araştırmaları vb. alanlarda yapılan çalışmalar sonucunda bazı tahminlerde bulunulmaktadır. Bu tahminlerin başarımının ve doğruluğunun ölçülebilmesi için bazı ölçütler kullanılmaktadır. Bunlar arasında en çok bilinenleri *doğruluk*, *hata oranı*, *kesinlik*, *duyarlılık*, *f1* ölçütü, *hata karelerinin ortalaması*, *mutlak hata* ve *özgünlük* ölçütleridir. Bu çalışmada, veri madenciliği uygulamalarında en çok kullanılan ölçütler olan

*doğruluk*, *kesinlik*, *duyarlılık* ve *f1* ölçütü kullanılmıştır. Ayrıca bu tür uygulamalarda sınıflandırma sonuçlarında gerçek değer ile tahmin edilen değerin karşılaştırılması için karışıklık matrisinden yararlanılır. Tablo 3.1'de $S_1$, $S_2$, $S_3$, …, $S_n$ olmak üzere $n$ adet sınıf için bir karışıklık matrisi örneği bulunmaktadır. Bu tabloda $X_{ij}$ gerçekte ait olduğu sınıf $i$. sınıfken, tahminleme sonucu elde edilen sınıf etiketinin $j$. sınıf olduğu örnek sayısıdır.

### 3.3.1 Karışıklık Matrisi

Karışıklık matrisi, tahminleme yapılan çalışmalarda gerçek sınıf değerleriyle tahminleme sonucu elde edilen sınıf değerlerinin karşılaştırılması için kullanılır (Özkan 2016). Karışıklık matrisinin satırlarında gerçek sınıflara ait örnekler, sütunların da tahmin edilen sınıflara ait örnek sayıları bulunur. Köşegenler de ise gerçek ve tahmin edilen sınıf etiketleri aynı olan örnek sayıları yer alır.

Tablo 3.1: $\boldsymbol{n}$ sınıf için karışıklık matrisi

| | tahmin edilen sınıf | | | | | |
|---|---|---|---|---|---|---|
| gerçek sınıf | | $\boldsymbol{S_1}$ | $\boldsymbol{S_2}$ | $\boldsymbol{S_3}$ | … | $\boldsymbol{S_n}$ |
| | $\boldsymbol{S_1}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | … | $X_{1n}$ |
| | $\boldsymbol{S_2}$ | $X_{21}$ | $X_{22}$ | $X_{23}$ | … | $X_{2n}$ |
| | $\boldsymbol{S_3}$ | $X_{31}$ | $X_{32}$ | $X_{33}$ | … | $X_{3n}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | $\boldsymbol{S_n}$ | $X_{n1}$ | $X_{n2}$ | $X_{n3}$ | … | $X_{nn}$ |

### *3.3.2 doğruluk*

*doğruluk (accuracy)*, gerçek sınıf değerlerinin tahmin edilen sınıf değerleriyle hangi oranda aynı olduğunu gösterir (Özkan 2016).

Tablo 3.1'de verilen karışıklık matrisi dikkate alındığında *doğruluk* ölçütü şu şekilde hesaplanabilir:

$$doğruluk = \frac{\sum_{i=1}^{n} X_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}} \quad (4)$$

Yani kısaca *doğruluk* ölçütü karışıklık matrisinin köşegenlerinde yer alan örnek sayısının toplam örnek sayısına bölünmesiyle bulunmaktadır. En iyi durumda alacağı değer 1, en kötü durumda alacağı değer 0'dır. Ayrıca bir sınıflandırmanın hata oranı da şu şekilde bulunabilir:

$$hata\ oranı = 1 - doğruluk \quad (5)$$

### 3.3.3 *kesinlik*

*kesinlik* (precision), gerçekte bir sınıfa ait örnek sayısının, tahminleme sonunu yine aynı sınıfla etiketlenen örnek sayısına yani o sütunda yer alan tüm değerlerin toplamına oranıdır. Başka bir deyişle, bir sınıfla etiketlenen örneklerin gerçekte o sınıfa ait olma ihtimalidir (McCarthy ve Lehnert 1995).

*kesinlik* ölçütü her bir sınıf için ayrı ayrı hesaplanabildiği gibi tüm sınıfların *kesinlik* ölçütünün ağırlıklı ortalaması alınarak ortalama *kesinlik* ölçütü de hesaplanabilmektedir. En iyi durumda alacağı değer 1, en kötü durumda alacağı değer 0'dır. Tablo 3.1'de verilen karışıklık matrisi dikkate alındığında, $S_k$ sınıfının *kesinlik* ölçütü şu şekilde hesaplanabilir:

$$kesinlik_{S_k} = \frac{X_{kk}}{\sum_{i=1}^{n} X_{ki}} \quad (6)$$

Benzer şekilde toplam ortalama *kesinlik* ölçütü şu şekilde hesaplanabilir:

$$\frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{n} X_{ij} * kesinlik_{S_i} \right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}} \quad (7)$$

### 3.3.4 *duyarlılık*

*duyarlılık* bir sınıfa ait örneklerin, tahminleme sonucunda ne kadarının yine aynı sınıfla etiketlendiğinin ölçütüdür. Başka bir ifadeyle sınıflandırıcının tüm doğru sınıflandırmaları bulma kabiliyetidir (Caruana ve Niculescu-Mizil 2004).

*duyarlılık* ölçütünün en iyi durumda alacağı değer 1, en kötü durumda alacağı değer 0'dır. Tablo 3.1'de verilen karışıklık matrisi dikkate alındığında, $S_k$ sınıfının *duyarlılık* ölçütü şu şekilde hesaplanabilir:

$$duyarlılık_{S_k} = \frac{X_{kk}}{\sum_{i=1}^{n} X_{ki}} \tag{8}$$

### 3.3.5 *f1* ölçütü

*f1* ölçütü, *duyarlılık* ve *kesinlik* değerlerinin harmonik ortalamasıdır (Coşkun 2011). Bu ölçüt sayesinde *duyarlılık* ve *kesinlik* ölçütleri birlikte değerlendirilebilme imkanı bulmaktadır (Özkan 2016).

$f1_{S_k}$ ölçütü her bir $S_k$ sınıfı için ayrı ayrı hesaplanabildiği gibi tüm sınıfların *f1* ölçütlerinin ağırlıklı ortalaması alınarak ortalama *f1* ölçütü de hesaplanabilmektedir. En iyi durumda alacağı değer 1, en kötü durumda alacağı değer 0'dır. Tablo 3.1'de verilen karışıklık matrisi dikkate alındığında ve *duyarlılık* ile *kesinlik* ölçütlerinin *f1* ölçütüne katkısının eşit olduğu kabul edildiğinde, $s_k$ sınıfının *f1* ölçütü şu şekilde hesaplanabilir:

$$f1_{S_k} = 2 * \frac{duyarlılık_{S_k} * kesinlik_{S_k}}{duyarlılık_{S_k} + kesinlik_{S_k}} \tag{9}$$

Benzer şekilde ortalama *f1* ölçütü şu şekilde hesaplanabilir:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij} * f1_{S_i}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}} \tag{10}$$

# 4. UYGULAMALAR

Bu çalışmanın amacı kullanıcıların yalnızca arkadaşlık ilişkilerinden yararlanarak profilleri hakkında çıkarımda bulunabilecek nicel bir sistem geliştirmek ve bu profilleme işlemi için kullanılabilecek bir kural tabanı oluşturmaktır.

Literatürde bu konuda yapılmış birçok çalışma bulunmasına rağmen bu çalışmayı tüm bu çalışmalardan ayıran en büyük özellik profil çıkarımı için geçmişteki çalışmalardan farklı olarak yalnızca kullanıcıların arkadaşlık ilişkilerinden yararlanılmasıdır. Bu doğrultuda önerilen sistemin kullanılabilirliğini test etmek için kullanıcıların siyasi görüşlerini tahmin edebilmek adına farklı yöntemlerle çeşitli uygulama örnekleri yapılmıştır. Yapılan uygulamalar ve yapılış amaçları kısaca şöyledir:

1. *k*-NN yöntemi kullanılarak Twitter kullanıcılarının siyasi görüşlerini tahmin etmek ve en yüksek başarımı sağlayan *k* değerinin, eğitim setinin ve özelliklerinin belirlenmesi,
2. Karar ağacı yöntemi kullanarak Twitter kullanıcılarının siyasi görüşlerini tahmin etmek, bu karar ağacına bağlı olarak kurallar oluşturmak ve en yüksek başarımı sağlayan karar ağacı için eğitim setinin ve özelliklerinin belirlenmesi,
3. *k*-Ortalamalar yöntemi kullanılarak Twitter kullanıcılarının siyasi parti sayısı (bu çalışmada 3 adet) kadar kümeye ayrılmasını sağlamak, en başarılı kümelenmeleri sağlayan özelliklerin ve en başarılı kümelenmenin sağlandığı veri setinin belirlenmesi,
4. *k*-Ortalamalar yönteminin bulanıklaştırılmış bir versiyonu olan Bulanık *c*-Ortalamalar yöntemini kullanılarak Twitter kullanıcılarının siyasi parti sayısı kadar kümeye ayrılmasını sağlamak, en başarılı kümelenmeleri sağlayan özelliklerin, en başarılı kümelenmenin sağlandığı veri setinin belirlenmesi ve yanlış kümelenen örneklerin üyelik derecelerinin ve özelliklerinin incelenmesidir.

Twitter diğer sosyal medyaların aksine genellikle insanların siyasi, ekonomik ve toplumsal olaylara dahil olmak, bu olaylardan haberdar olmak, bu olaylara destek olmak ya da tepkilerini göstermek amacıyla duygu, düşünce ve fikirlerini belirttiği bir

araç olarak kullanılır. Ayrıca sosyolojik açıdan bakıldığında Twitter topluluk ve topluluklar arasındaki etkileşimlerle sosyal ağ kavramının teknolojik bir yansıması olarak görülebilir. Bu doğrultuda bu tez çalışmasında ele alınan uygulamalarda veri kaynağı olarak Twitter seçilmiş, analizler ve uygulamalar sırasında Twitter kullanıcı verileri kullanılmıştır.

## 4.1 Verilerin Toplanması ve Ön İşleme

Twitter, verilerine ulaşımı API'leri üzerinden sağlamaktadır. Twitter'ın Stream API ve Rest API olmak üzere mevcut 2 API'si bulunmaktadır. Stream API canlı olarak akan veriye ulaşmayı sağlarken, Rest API kullanıcı bilgileri, paylaşım bilgileri gibi verilere ulaşmayı sağlamaktadır. Bu çalışmada kullanılan nicel verilere bağlı olarak bu verilerin çekilmesi için Rest API'den faydalanılmıştır. Bu tez çalışmasında Rest API'yi doğrudan kullanmak yerine çalışmada verilerin çekilmesi, temizlenmesi, hazırlanması ve bazı analizlerin yapılması için kullanılan Python programlama dili için geliştirilmiş ve en sık kullanılan kütüphanelerden birisi olan tweepy'den yararlanılmıştır.

tweepy kütüphanesi üzerinden yapılan Rest API istekleri sonuç olarak json formatında ya da nesne türünde veriler döndürmektedir. Bundan dolayı çekilen veriler ilişkisel bir veritabanında tutulmak yerine NoSql bir veritabanı olan MongoDB üzerinde tutulmuştur. Özellikle MongoDB'nin seçilmesinin nedeni MongoDB'nin döküman adı verilen her bir kaydı json formatına benzeyen ve bson adı verilen bir formatta saklamasıdır. Ayrıca MongoDB üzerinde verilere erişim, sorgulama ve analiz işlemleri javascript ile doğrudan yapılabildiği için oldukça basit olması da diğer tercih sebepleri arasındadır. Verilerin çekilmesi Python ortamında gerçekleştiği için çekilen verilerin MongoDB'ye eklenmesi, daha sonra gerçekleştirilen silme ve güncelleme gibi işlemler yine bir Python kütüphanesi olan pymongo kullanılarak gerçekleştirilmiştir.

Öncelikle analizin gerçekleştireceği 3 siyasi partinin resmi Twitter hesaplarının kullanıcı adları belirlenmiştir. Çalışmada kişisel bilgilerin güvenliliği ve gizliliği açısından bu siyasi partiler gerçek isimleri yerine P1, P2 ve P3 şeklindeki etiketlerle temsil edilmişlerdir. Daha sonra her bir siyasi parti için, bir Python kütüphanesi olan

tweepy'nin get_user metodu kullanılarak geliştirilen bir robot vasıtasıyla adı, kullanıcı adı, takipçi sayısı, arkadaş sayısı ve paylaşım sayısı gibi profil bilgileri çekilmiş, çekilen her kullanıcı için yine tweepy kütüphanesinin bir metodu olan friends_ids çalıştırılarak o partinin resmi Twitter hesabının arkadaş listesinde yer alan Twitter kullanıcılarının ID'leri çekilerek C4 koleksiyonunda depolanmıştır. Benzer şekilde her bir parti için liderinin resmi Twitter hesabının kullanıcı adı belirlenmiş ve yine get_user metodu kullanılarak parti liderlerinin profil bilgileri ve friends_ids metoduyla da arkadaş listesindeki kullanıcıların ID'leri çekilerek C5 koleksiyonunda depolanmıştır. Daha sonra get_friends ve get_followers metodları kullanılarak her bir partinin arkadaş ve takipçi listesi elde edilmiş, listelerdeki Twitter kullanıcıların user_id'leriyle bir döngü halinde tweepy kütüphanesinin get_user metodu çağırılarak profil bilgileri toplanmış, arkadaş listesinde yer alan P1 için 48, P2 için 426 ve P3 için 72 kullanıcı bilgisi C6, takipçi listesinde yer alan kullanıcıların bilgileri de C7 koleksiyonuna kaydedilmiştir.

Siyasi partilerin arkadaş listesi, genelde o partinin milletvekilleri, bakanları ya da o partinin önde gelen liderlerinden oluşmaktadır ve bundan dolayı bu kullanıcıların siyasi görüşü doğrudan o partiyle etiketlenmiştir. Ancak takipçilerde böyle bir şey söz konusu değildir. Bir kullanıcı birden fazla partinin takipçi listesinde yer alabilmektedir. Bu durumda bu kullanıcıların etiketlenmesi mümkün olmamaktadır, bundan dolayı takipçi listelerindeki kullanıcılar çekilirken yalnız bir siyasi partiyi takip eden kullanıcıların bilgileri çekilmiş ve sonrasında siyasi görüş olarak da takip ettiği partiyle etiketlenmişlerdir.

Bu verileri çekmek için tweepy kütüphanesinden yararlandığımızı daha önce belirtmiştik ancak tweepy, aslında arka planda Twitter API'sini kullanan ve Python programlama dili için düzenlenmiş fonksiyon ve metotlar barındıran bir kütüphanedir. Twitter API'sinin sahip olduğu ve bu tez çalışmasını kısıtlayan bazı sınırlamalar bulunmaktadır, bunlardan birincisi kullanıcıların profil bilgilerine, arkadaş listesine ve takipçi listesine vb. ulaşmak için kullandığımız metotlar yalnızca profili gizli olmayan yani profili herkes tarafından görülebilen kullanıcıların ya da Twitter API için erişim izni olan hesabın arkadaş veya takipçi listesinde olan kullanıcıların bilgilerine erişime izin vermesidir. Bu sınırlama siyasi partilerin, siyasi parti liderlerinin ve siyasi partilerin arkadaşlarının bilgilerini çekerken herhangi bir sorun oluşturmamıştır.

Çünkü bu hesaplar, siyasi partilere, siyasi parti liderlerine ve arkadaş listelerinde yer alanlar da bakan, milletvekili ya da siyasetçilere ait olduğu için profilleri herkes tarafından erişilebilir durumdadır. Ancak takipçi listesinde yer alan kullanıcılar sıradan bireyler de olabileceği için içlerinde profilleri gizli olan hesaplar da bulunmaktadır. Ayrıca Twitter API'nin sınırlamalarından bir diğeri de belli bir süre içerisinde belli sayıda istek gönderilebilmesidir. Örneğin; 15 dakika içerisinde sadece 15 tane GET friends/ids isteğinde bulunulabilmektedir. Bu siyasi partilerin en az takipçiye sahip olanının 646859, en çok takipçiye sahip olanının da 1414588 adet takipçisi olduğunu düşünürsek tüm takipçilerin bilgilerini çekmek bu sınırlar doğrultusunda mümkün olmamaktadır. Bu sınırlandırma doğrultusunda her bir parti için takipçi listesinde yer alan 500 kullanıcı bilgisi çekilmiş daha sonra bunların içerisinde birden fazla siyasi partiyi takip eden kullanıcılar veri setinden çıkarılmıştır. Son durumda P1 için 196, P2 için 262 ve P3 için 243 adet kullanıcı bilgisi C7 koleksiyonuna kaydedilmiştir.

Buraya kadar çekilen veriler eğitim verisi olarak kullanılması hedeflenen verilerdir, test amaçlı kullanılmak istenen veriler de @pauedutr kullanıcı adına sahip Pamukkale Üniversitesi'nin resmi Twitter hesabının takipçi listesindeki kullanıcılar çekilerek elde edilmiştir. Bu hesabın kullanıcı listesinde 17734 tane kullanıcı yer almaktadır ancak bunların büyük bir kısmının profili gizli olduğu için kullanıcı bilgilerine ulaşılamamakta ve aynı zamanda bu 17734 kullanıcının içerisinde sahte ya da robot ve yeni açılmış hesaplar bulunmaktadır. Bu doğrultuda bu 17734 kullanıcıdan profili gizli olmayanlar, en az 200 en fazla 2500 tweet atmış olanlar, arkadaş ve takipçi sayısı 50 ile 2500 arasında olanlar seçilerek bu sayı 3357'ye indirilmiş, daha sonra da bunların içerisinden 500 tanesi rastgele seçilerek gözlemciler tarafından etiketlenmiştir. Etiketleme işlemi kullanıcıların profillerine bakılarak yapılmıştır. Örneğin; profil resmi, profilinin açıklama kısmında herhangi bir siyasi partiyle ilgili bir bilginin olup olmaması, kullanıcının atmış olduğu tweetlerin içerikleri, favori olarak seçmiş olduğu paylaşımlar etiketleme sırasında dikkate alınmıştır. Bu etiketleme işleminin sonucunda 500 adet kullanıcının yalnız 170 tanesi siyasi görüşü açısından etiketlenebilmiş ve bunların da 9 tanesi tez çalışmasındaki analizler için seçilmiş olan siyasi partilerin dışında kalan partilerdir. Test veri seti olarak elde edilen kullanıcı bilgileri de C8 koleksiyonuna kaydedilmiştir.

Her bir siyasi partinin takipçileri ile arkadaşlarının kullanıcı bilgileri ve Pamukkale Üniversitesi resmi Twitter hesabının takipçileri arasından rastgele seçilen kullanıcı bilgileri elde edildikten sonra bunların analizde kullanılabilmesi için bazı özelliklerin belirlenmesi gerekmektedir. Bu çalışmada kullanılması hedeflenen özelliklerin listesi ve açıklamaları Tablo 4.2'de yer almaktadır.

Sırasıyla C6, C7 ve C8 koleksiyonlarında yer alan her bir kullanıcının her bir siyasi parti için Tablo 4.2'de yer alan 13 adet özelliği hesaplanmış ve sırasıyla C1, C2 ve C3 koleksiyonlarına kaydedilmiştir.

Tablo 4.1: Veri seti ve koleksiyonlar

| **no** | **kısaltmalar** | **koleksiyonlar** |
|---|---|---|
| 1 | C1 | PoliticalPartiesFriendsSimilarities |
| 2 | C2 | PoliticalPartiesFollowersSimilarities |
| 3 | C3 | UsersPoliticalSimilarities |
| 4 | C4 | PoliticalParties |
| 5 | C5 | PoliticalPartiesLeaders |
| 6 | C6 | PoliticalPartiesFriends |
| 7 | C7 | PoliticalPartiesFollowers |
| 8 | C8 | Users |

Sırasıyla **KA** bir Twitter kullanıcısının arkadaş listesinden oluşan kümeyi, **PA** siyasi partinin arkadaş listesinden oluşan kümeyi, $\mathbf{PAA}_i$ siyasi partinin arkadaş listesinde yer alan $i$. arkadaşının arkadaş listesinden oluşan kümeyi, $\mathbf{PTA}_i$ siyasi partinin takipçi listesinde yer alan $i$. takipçisinin arkadaş listesinden oluşan kümeyi, $\mathbf{LA}_i$ siyasi partinin liderleri arasında yer alan $i$. liderin arkadaş listesinden oluşan kümeyi temsil etmek üzere A1'den A13'e kadar olan özellikler Tablo 4.2'de tanımlanmıştır.

Tablo 4.2: Analiz için kullanılan özellikler ve kısaltmaları

| **no** | **özellik** | **formül** |
|---|---|---|
| A1 | party_similarity | $\frac{s(KA \cap PA)^2}{s(KA) * s(PA)}$ |

| | | |
|---|---|---|
| A2 | party_similarity2 | $\frac{s(KA \cap PA)}{s(KA)}$ |
| A3 | party_similarity3 | $\frac{s(KA \cap PA)}{s(PA)}$ |
| A4 | party_is_being_followed | $\begin{cases} 1, parti\ takip\ ediliyorsa \\ 0, parti\ takip\ edilmiyorsa \end{cases}$ |
| A5 | party_leaders_similarity | $\frac{\left(\sum_{i=1}^{n} s(KA \cap LA_i)^2 \Big/ s(LA_i)\right)}{\left(n * s(KA)\right)}$ |
| A6 | party_leaders_similarity2 | $\frac{\sum_{i=1}^{n} s(KA \cap LA_i)}{\left(n * s(KA)\right)}$ |
| A7 | user_intersection_political_party_leaders_total_friends_count_division_political_party_leaders_total_friends_count | $\frac{\sum_{i=1}^{n} s(KA \cap LA_i)}{\sum_{i=1}^{n} s(LA_i)}$ |
| A8 | party_friends_similarity | $\frac{\left(\sum_{i=1}^{n} s(KA \cap PAA_i)^2 \Big/ s(PAA_i)\right)}{\left(n * s(KA)\right)}$ |
| A9 | party_friends_similarity2 | $\frac{\sum_{i=1}^{n} s(KA \cap PAA_i)}{\left(n * s(KA)\right)}$ |
| A10 | user_total_intersection_friends_count_of_political_party_friends_division_total_friends_count_of_political_party_friends | $\frac{\sum_{i=1}^{n} s(KA \cap PAA_i)}{\sum_{i=1}^{n} s(PAA_i)}$ |
| A11 | party_followers_similarity | $\frac{\left(\sum_{i=1}^{n} s(KA \cap PTA_i)^2 \Big/ s(PTA_i)\right)}{\left(n * s(KA)\right)}$ |
| A12 | party_followers_similarity2 | $\frac{\sum_{i=1}^{n} s(KA \cap PTA_i)}{\left(n * s(KA)\right)}$ |
| A13 | user_total_intersection_friends_count_of_political_party_followers_division_total_friends_count_of_political_party_followers | $\frac{\sum_{i=1}^{n} s(KA \cap PTA_i)}{\sum_{i=1}^{n} s(PTA_i)}$ |

Bu 13 özelliğin her birisi analizde kullanılan her bir parti için ayrı ayrı hesaplanmaktadır. Örneğin bu tez çalışmasın da P1, P2 ve P3 olmak üzere 3 parti kullanılmıştır, bundan dolayı bir kullanıcı A1 için aslında P1_A1, P2_A1 ve P3_A1

şeklinde 3 özelliğe sahiptir. Bundan dolayı bu tez çalışmasında her bir kullanıcı aslında 13 * 3‘den 39 tane özellik ile temsil edilmektedir.

## 4.2 Sınıflandırma Uygulamaları

Bu bölümde Tablo 4.1’de yer alan C1, C2 ve C3 veri setleri üzerinde Tablo 4.2’de yer alan özellikler kullanılarak *k*-NN ve karar ağacı yöntemlerine göre sınıflandırma uygulamaları yapılmıştır. Ayrıca karar ağacı yönteminden elde edilen sonuçlara bağlı olarak da kural tabanları çıkartılmıştır.

### 4.2.1 Uygulama 1: *k*-En Yakın Komşuluk Yöntemiyle Kullanıcıların Siyasi Görüşlerinin Tahmin Edilmesi

Bu uygulamadaki temel amaç, Twitter üzerinden elde edilen C1, C2 ve C3 veri setlerini *k*-NN algoritmasıyla birlikte kullanarak Twitter kullanıcılarının siyasi görüşlerini tahminlemeye çalışmaktır. Bunu yaparken yukarıda Tablo 4.1’de yer alan toplam 8 adet koleksiyon arasından C1, C2 ve C3’ten ve Tablo 4.2’de yer alan 13 adet özelliğin tamamından yararlanılmıştır.

Eğitim için seçilen veri seti, veri setinde yer alan özellikler ve seçilen *k* değeri *k*-NN algoritmasının başarımını etkileyen faktörlerdir. Buna bağlı olarak bu uygulamada 4 temel soruya cevap aranmıştır:

1. En başarılı tahminleme sonucu nedir?
2. En başarılı tahminleme için en uygun eğitim veri seti hangisi veya hangileridir?
3. En başarılı tahminleme için en uygun özellik hangisi veya hangileridir?
4. En başarılı tahminleme için en uygun *k* değeri nedir?

Yapılan uygulamada Tablo 4.1’de yer alan C1, C2 ve C3 koleksiyonlarından, C1 ve C2 *k*-NN algoritması için eğitim, C3 ise test veri seti olarak kullanılmıştır. Ancak yalnız C1, yalnız C2 ve C1 ile C2’nin birlikte eğitim veri seti olarak kullanılabileceği 3 durum söz konusudur ve yukarıda bahsedilen 2 nolu sorunun

cevabının bulunabilmesi için bu 3 durumun ayrı ayrı test edilmesi gerekmektedir. Yine aynı şekilde soru 3'ün cevabını bulabilmek için, Tablo 4.2'de yer alan 13 adet özellikten hangisi veya hangilerinin kullanılması gerektiğini bulabilmek için tüm kombinasyonların test edilmesi, bu yüzden de $\binom{13}{1}, \binom{13}{2}, \dots, \binom{13}{13}$ şeklinde tüm kombinasyonlar için toplamda 8191 durumun ayrı ayrı test edilmesi gerekmektedir. *k*-NN algoritmasının başarımını etkileyen en önemli faktörlerden birisi de *k* değerinin seçimidir, ki bu 4 nolu soru ile ilişkilidir. Bu çalışmada *k*'nın 1-10 arasında değer alabileceği kabul edilmiştir. Yani bu durumda da sistemin toplamda 10 farklı *k* değeri için ayrı ayrı test edilmesi gerekmektedir.

Özetlemek gerekirse 3 farklı eğitim veri seti, 8191 farklı özellik seçimi ve 10 farklı *k* değeri seçimi yapılabilmektedir ve bu faktörlerin hepsi birbirini etkilemektedir, dolayısıyla toplamda $3 * 8191 * 10$'dan 245730 farklı kombinasyon bulunmaktadır.

1 nolu soruda da yer alan algoritmanın başarımını ölçmek için Bölüm 3.3'te açıklanan *doğruluk, kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü kullanılmıştır. 245730 farklı kombinasyon için bu 3 ölçüt hesaplanmış ve her bir ölçüt için en yüksek değeri sağlayan, eğitim veri seti veya setleri, özellik veya özellikler ve *k* değeri bulunmaya çalışılmıştır.

Bu uygulama Python ortamında sklearn kütüphanesinin neighbors modülü kullanılarak gerçekleştirilmiştir. Öncelikle neighbors modülünün KNeighborsClassifier sınflandırıcısından bir nesne türetilmiştir. Daha sonra bu nesnenin fit metodu parametre olarak eğitim verilerini alarak sistem eğitilmiş ve son olarak da predict metoduna test verileri parametre olarak gönderilerek tahminleme yapılmıştır. 1 ile 10 arasında değerler alabilen *k*'nın her bir değeri için KNeigborsClassifier sınıflandırıcısı, n_neighbors parametresi *k*'nın aldığı değer, p parametresi 2 ve metric parametresi "minkowski" olacak şekilde kullanılmıştır. n_neighbors parametresi bakılacak en yakın komşuluk sayısını temsil ederken, p'nin 2, metric'inse "minkowski" olması bakılacak uzaklık metriğinin standart Öklid uzaklığı olduğunu .gösterir.

Yapılan tahminlerin başarımını ölçmek için sırasıyla:

- Tüm sınıfların ortalama *doğruluk* ölçütü için sklearn kütüphanesinin metrics modülünün accuracy_score fonksiyonu,
- Hem tüm sınıfların ortalama *kesinlik* ölçütü hem de sınıf bazında tek tek *kesinlik* ölçütü için sklearn kütüphanesinin metrics modülünün precision_score fonksiyonu,
- Tüm sınıfların sınıf bazında tek tek *duyarlılık* ölçütü için sklearn kütüphanesinin metrics modülünün recall_score fonksiyonu,
- Hem tüm sınıfların ortalama *f1* ölçütü hem de sınıf bazında tek tek *f1* ölçütü için sklearn kütüphanesinin metrics modülünün *f1*_score fonksiyonu,
- Sınıflandırmanın karışıklık matrisini elde etmek için sklearn kütüphanesinin metrics modülünün confusion_matrix fonksiyonu,
- Sınıflandırmanın raporunu elde etmek için sklearn kütüphanesinin metrics modülünün classification_reports fonksiyonu kullanılmıştır.

Elde edilen sonuçların grafiksel olarak gösterimi için, çizim için geliştirilmiş bir Python kütüphanesi olan matplotlib'in pyplot modülünden yararlanılmıştır. Bu modülün plot fonksiyonu kullanılarak, 1 ile 10 arasında değer alabilen *k*'nın her bir değerinde hesaplanan başarım ölçütü değerleri görselleştirilmiştir.

Tablo 4.3: Eğitim veri setinin C1 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 | | | | |
|---|---|---|---|---|---|
| **maksimum** | ***k*** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | 4 | A3, A10 | **0.7516** | 0.8141 | 0.7379 |
| ***kesinlik*** | 1 | A2, A3 | 0.4224 | **0.8700** | 0.4425 |
| ***f1*** | 9 | A10, A13 | 0.7391 | 0.7681 | **0.7454** |
| **süre** | 391.6998 saniye | | | | |

Tablo 4.3'te *k*-NN algoritması için eğitim veri seti olarak C1'in kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.3'te de görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.7516 olarak hesaplanmış, bu değer, özellik olarak A3 ve A10'un, *k*'nın ise 4 olarak seçildiği durumda elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.87 olarak hesaplanmış, bu değer özellik olarak A2 ve A3'ün, *k*'nın ise 1 olarak seçildiği durumda elde edilirken,

en yüksek *f1* değeri özellik olarak A10 ve A13'ün, *k*'nın ise 9 seçilmesi durumunda 0.7454 olarak hesaplanmıştır.

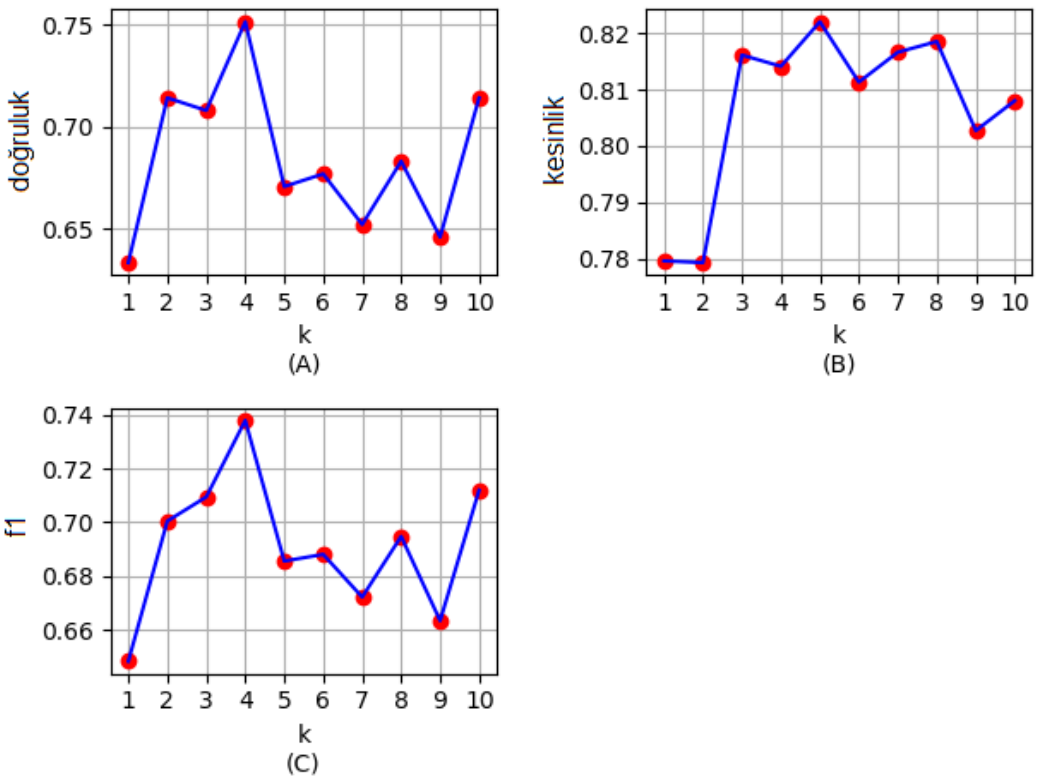
Bu değerler aynı zamanda şu anlama da gelmektedir;

- Eğitim veri seti olarak C1, özellik olarak A3 ve A10 kullanıldığında ve *k* değeri 4 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, ortalama %75.16'sı doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C1, özellik olarak A3 ve A10 kullanıldığında ve *k* değeri 4 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %87'dir.
- Eğitim veri seti olarak C1, özellik olarak A10 ve A13 kullanıldığında ve *k* değeri 9 olarak kabul edildiğinde ortalama *f1* değeri %74.54'tür.

Tablo 4.4: Eğitim veri setinin C1, özelliklerin A3 ve A10 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A3, A10 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| 1 | 0.6335 | 0.7796 | 0.6482 |
| 2 | 0.7143 | 0.7793 | 0.7004 |
| 3 | 0.7081 | 0.8161 | 0.7094 |
| **4** | **0.7516** | **0.8141** | **0.7379** |
| 5 | 0.6708 | 0.8220 | 0.6855 |
| 6 | 0.6770 | 0.8113 | 0.6880 |
| 7 | 0.6522 | 0.8166 | 0.6721 |
| 8 | 0.6832 | 0.8185 | 0.6947 |
| 9 | 0.6460 | 0.8027 | 0.6633 |
| 10 | 0.7143 | 0.8080 | 0.7119 |

Tablo 4.4'te eğitim veri setinin C1, özelliklerin A3 ve A10 olması durumunda *doğruluk, kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.1 ise Tablo 4.4'teki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.1: Eğitim veri setinin C1, özelliklerin A3 ve A10 olması durumunda tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.1'deki (A) ve (C) grafiklerinde *k* değeri 4 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.7516, ortalama *f1* puanı 0.7379 iken, ortalama *kesinlik* 0.8141'dir. Şekil 4.1 (B) grafiğinde ise *k* değeri 5 için en yüksek değere ulaşılmıştır. Şekil 4.1 (B) grafiğinde *k* değeri 5 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.6708, ortalama *kesinlik* 0.8220 ve ortalama *f1* puanı 0.7379 olarak hesaplanmıştır.

Tablo 4.5: Eğitim veri setinin C1, özelliklerin A3 ve A10 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.86 | 0.89 | 0.87 | 93 |
| **P2** | 0.49 | 0.89 | 0.63 | 27 |
| **P3** | 0.93 | 0.34 | 0.50 | 41 |
| **ort./toplam** | 0.81 | 0.75 | 0.74 | 161 |

Tablo 4.5'te eğitim veri setinin C1, özelliklerin A3 ve A10, *k*'nın ise 4 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir. Tablolardaki ortalama başarım ölçütleri, ölçütlerinin örnek sayısına göre ağırlıklı ortalamasıdır.

Tablo 4.6: Eğitim veri setinin C1, özelliklerin A3 ve A10 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **83** | 10 | 0 | 93 |
| **P2** | 2 | **24** | 1 | 27 |
| **P3** | 12 | 15 | **14** | 41 |
| **toplam** | 97 | 49 | 15 | 161 |

Tablo 4.6'da ise eğitim veri setinin C1, özelliklerin A3 ve A10, *k*'nın ise 4 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Bu tabloda da görüldüğü gibi 161 örnekten oluşan test veri setimiz 93 tane P1, 27 tane P2 ve 41 tane P3 sınıfına ait örnekten oluşmaktadır. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 83 tanesi, 27 tane P2 sınıfına ait örnekten 24 tanesi ve 41 tane P3 sınıfına ait örnekten 14 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 83/93'ten 0.89, 24/27'den 0.89 ve 14/41'den 0.34 olarak hesaplanmıştır.

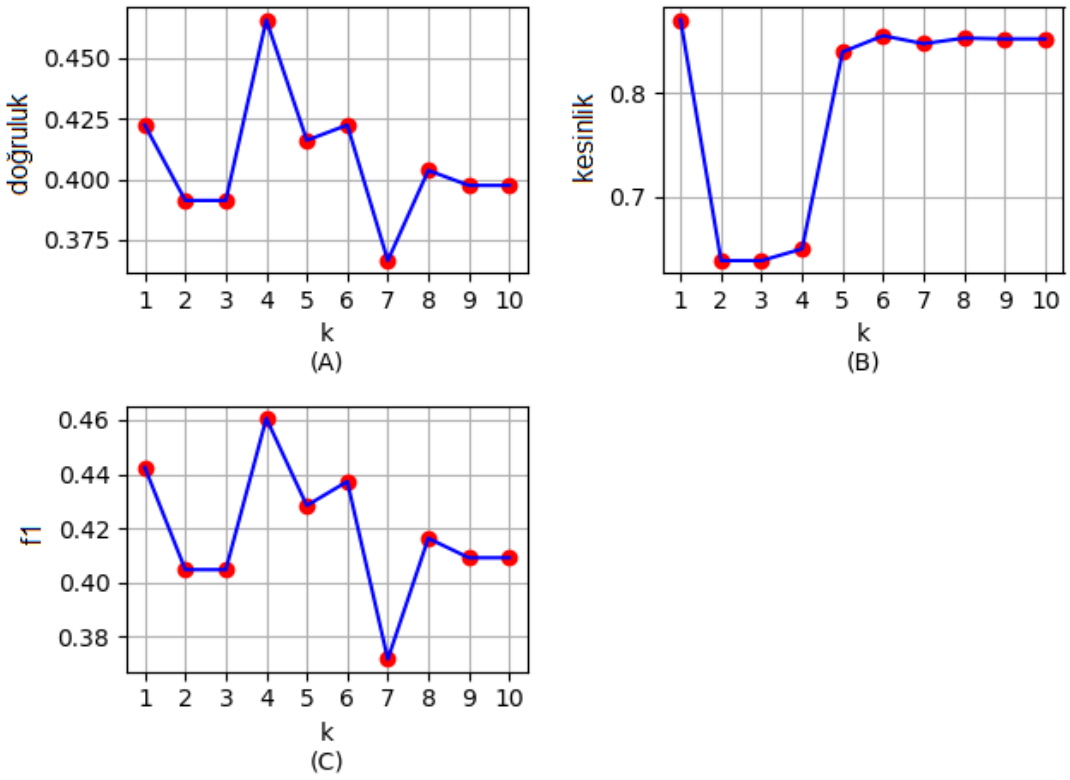
Bu tez çalışmasında yapılan analizler sonucunda toplamda 97 örnek P1 sınıfıyla etiketlenmiş ancak gerçekte bunun 83 tanesi bu sınıfa aittir, yine benzer şekilde 49 örnek P2 sınıfıyla, 15 örnek P3 sınıfıyla etiketlenirken gerçekte 49'dan 24 tanesi, 15'ten ise 14 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri

P1 için, 83/97'den 0.86, P2 için 24/49'dan 0.49 ve P3 için 14/15'ten 0.93 olarak hesaplanmıştır.

Tablo 4.7: Eğitim veri setinin C1, özelliklerin A2 ve A3 olması durumda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A2, A3 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| **1** | **0.4224** | **0.8700** | **0.4425** |
| 2 | 0.3913 | 0.6387 | 0.4047 |
| 3 | 0.3913 | 0.6387 | 0.4047 |
| 4 | 0.4658 | 0.6500 | 0.4606 |
| 5 | 0.4161 | 0.8391 | 0.4284 |
| 6 | 0.4224 | 0.8547 | 0.4373 |
| 7 | 0.3665 | 0.8470 | 0.3716 |
| 8 | 0.4037 | 0.8524 | 0.4163 |
| 9 | 0.3975 | 0.8516 | 0.4091 |
| 10 | 0.3975 | 0.8516 | 0.4091 |

Tablo 4.7'de eğitim veri setinin C1, özelliklerin A2 ve A3 olması durumunda *doğruluk*, *kesinlik*, *f1* ve duyarlılığın *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.2 ise Tablo 4.7'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.2: Eğitim veri setinin C1, özelliklerin A2 ve A3 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.2'deki (A) ve (C) grafiklerinde *k* değeri 4 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.4658, ortalama *f1* 0.4606 iken, ortalama *kesinlik* 0.65'tir. (B) grafiğinde ise *k* değeri 1 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 1 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.4224, ortalama *kesinlik* 0.87 ve ortalama *f1* 0.4425 olarak hesaplanmıştır.

Tablo 4.8: Eğitim veri setinin C1, özelliklerin A2 ve A3 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 1.00 | 0.38 | 0.55 | 93 |
| **P2** | 0.23 | 1.00 | 0.37 | 27 |
| **P3** | 1.00 | 0.15 | 0.26 | 41 |
| **ort./toplam** | 0.87 | 0.42 | 0.44 | 161 |

Tablo 4.8'de eğitim veri setinin C1, özelliklerin A2 ve A3, *k*'nın ise 1 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.9: Eğitim veri setinin C1, özelliklerin A2 ve A3 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **35** | 58 | 0 | 93 |
| **P2** | 0 | **27** | 0 | 27 |
| **P3** | 0 | 35 | **6** | 41 |
| **toplam** | 35 | 120 | 6 | 161 |

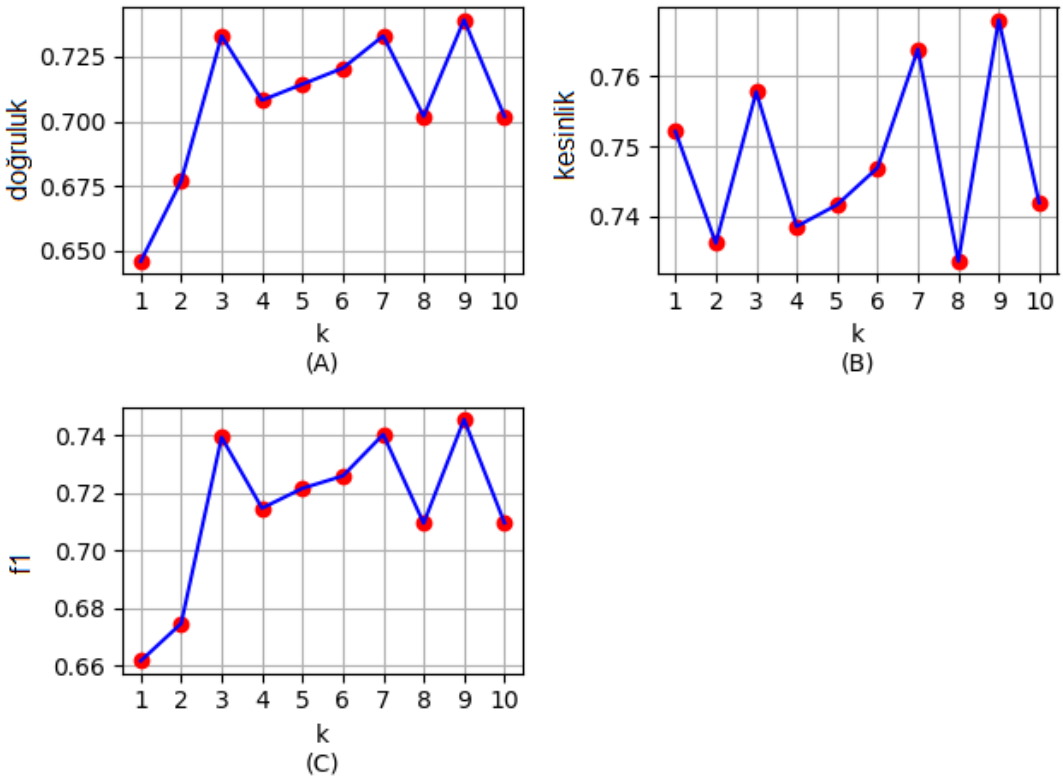
Tablo 4.9'da ise eğitim veri setinin C1, özelliklerin A2 ve A3, *k*'nın ise 1 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 35 tanesi, 27 tane P2 sınıfına ait örnekten 27 tanesi ve 41 tane P3 sınıfına ait örnekten 6 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 35/93'ten 0.38, 27/27'den 1.00 ve 6/41'den 0.15 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 35 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun tamamı bu sınıfa aittir, yine benzer şekilde 120 örnek P2 sınıfıyla, 27 örnek P3 sınıfıyla etiketlenirken gerçekte 120'den 27 tanesi, 6'nın ise tamamı bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 35/35'ten 1.00, P2 için 27/120'den 0.23 ve P3 için 6/6'dan 1.00 olarak hesaplanmıştır.

Tablo 4.10: Eğitim veri setinin C1, özelliklerin A10 ve A13 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A10, A13 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| 1 | 0.6460 | 0.7521 | 0.6616 |
| 2 | 0.6770 | 0.7361 | 0.6746 |
| 3 | 0.7329 | 0.7578 | 0.7392 |
| 4 | 0.7081 | 0.7385 | 0.7146 |
| 5 | 0.7143 | 0.7416 | 0.7215 |
| 6 | 0.7205 | 0.7469 | 0.7258 |
| 7 | 0.7329 | 0.7639 | 0.7402 |
| 8 | 0.7019 | 0.7335 | 0.7094 |
| **9** | **0.7391** | **0.7681** | **0.7454** |
| 10 | 0.7019 | 0.7419 | 0.7095 |

Tablo 4.10'da eğitim veri setinin C1, özelliklerin A10 ve A13 olması durumunda *doğruluk*, *kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.3 ise Tablo 4.10'daki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.3: Eğitim veri setinin C1, özelliklerin A10 ve A13 olması durumunda tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.3'teki (A) ve (C) grafiklerinde *k* değeri 9 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.7391, ortalama *f1* 0.7454 iken, ortalama *kesinlik* 0.7681'dir. (B) grafiğinde ise *k* değeri 9 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 9 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.7391, ortalama *kesinlik* 0.7454 ve ortalama *f1* 0.7681 olarak hesaplanmıştır.

Tablo 4.11: Eğitim veri setinin C1, özelliklerin A10 ve A13 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.88 | 0.72 | 0.79 | 93 |
| **P2** | 0.70 | 0.85 | 0.77 | 27 |
| **P3** | 0.56 | 0.71 | 0.62 | 41 |
| **ort./toplam** | 0.77 | 0.74 | 0.75 | 161 |

Tablo 4.11'de eğitim veri setinin C1, özelliklerin A10 ve A13, *k*'nın ise 9 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.12: Eğitim veri setinin C1, özelliklerin A10 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **67** | 5 | 21 | 93 |
| **P2** | 2 | **23** | 2 | 27 |
| **P3** | 7 | 5 | **29** | 41 |
| **toplam** | 76 | 33 | 52 | 161 |

Tablo 4.12'de ise eğitim veri setinin C1, özelliklerin A10 ve A13, *k*'nın ise 9 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 67 tanesi, 27 tane P2 sınıfına ait örnekten 23 tanesi ve 41 tane P3 sınıfına ait örnekten 29 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 67/93'ten 0.72, 23/27'den 0.85 ve 29/41'den 0.71 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 76 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 67 tanesi bu sınıfa aittir, yine benzer şekilde 33 örnek P2 sınıfıyla, 52 örnek P3 sınıfıyla etiketlenirken gerçekte 33'den 23 tanesi, 52'nin de 29 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 67/73'ten 0.88, P2 için 23/33'ten 0.70 ve P3 için 29/52'den 0.56 olarak hesaplanmıştır.

Tablo 4.13: Eğitim veri setinin C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 | | | | |
|---|---|---|---|---|---|
| **maksimum** | ***k*** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | 6 | A9 | **0.7391** | 0.7550 | 0.7429 |
| ***kesinlik*** | 1 | A4, A9, A12, A13 | 0.6087 | **0.7740** | 0.6295 |
| ***f1*** | 10 | A9, A11 | 0.7391 | 0.7652 | **0.7442** |
| **süre** | 387.4336 saniye | | | | |

Tablo 4.13'te *k*-NN algoritması için eğitim veri seti olarak C2'in kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.13'te de görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.7391 olarak hesaplanmış, bu değer, özellik olarak A9'un, *k*'nın ise 6 olarak seçildiği durumda elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.7740 olarak hesaplanmış, bu değer özellik olarak A4, A9, A12 ve A13'ün, *k*'nın ise 1 olarak seçildiği durumda elde edilirken, en

yüksek *f1* değeri özellik olarak A9 ve A11'ün, *k*'nın ise 10 olduğu durumda 0.7442 olarak hesaplanmıştır.

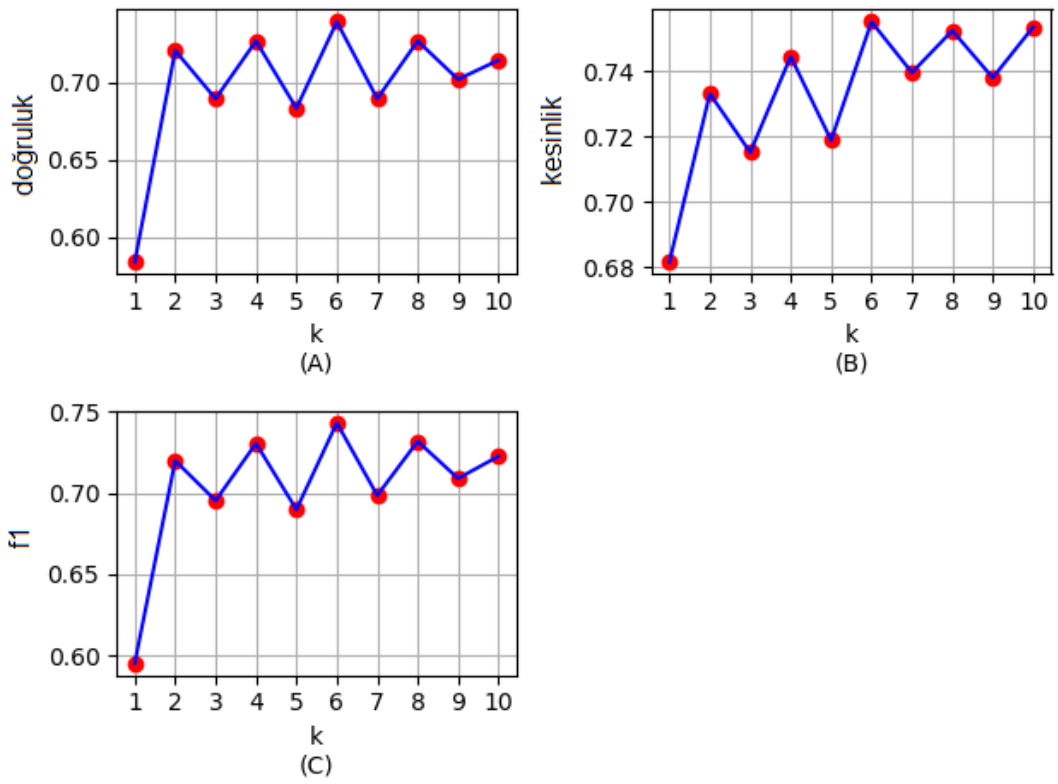
Bu değerler aynı zamanda şu anlama da gelmektedir;

- Eğitim veri seti olarak C2, özellik olarak A9 kullanıldığında ve *k* değeri 6 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, ortalama %73.91'sı doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C2, özellik olarak A4, A9, A12 ve A13 kullanıldığında ve *k* değeri 1 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %77.40'tır.
- Eğitim veri seti olarak C2, özellik olarak A9 ve A11 kullanıldığında ve *k* değeri 10 olarak kabul edildiğinde *kesinlik* ve duyarlılığın ağırlıklı ortalaması %74.42'dir.

Tablo 4.14: Eğitim veri setinin C2, özelliğin A9 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A9 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| 1 | 0.5839 | 0.6816 | 0.5955 |
| 2 | 0.7205 | 0.7330 | 0.7197 |
| 3 | 0.6894 | 0.7151 | 0.6952 |
| 4 | 0.7267 | 0.7444 | 0.7303 |
| 5 | 0.6832 | 0.7188 | 0.6897 |
| **6** | **0.7391** | **0.7550** | **0.7429** |
| 7 | 0.6894 | 0.7395 | 0.6984 |
| 8 | 0.7267 | 0.7523 | 0.7316 |
| 9 | 0.7019 | 0.7379 | 0.7088 |
| 10 | 0.7143 | 0.7534 | 0.7224 |

Tablo 4.14'de eğitim veri setinin C2, özelliğin A9 olması durumunda *doğruluk, kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.4 ise Tablo 4.14'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.4: Eğitim setinin C2, özelliğin A9 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.4'deki (A) ve (C) grafiklerinde *k* değeri 6 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.7391, ortalama *f1* 0.7429 iken, ortalama *kesinlik* 0.7550'dir. (B) grafiğinde ise yine *k* değeri 6 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 6 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.7391, ortalama *kesinlik* 0.7550 ve ortalama *f1* 0.7429 olarak hesaplanmıştır.

Tablo 4.15: Eğitim veri setinin C2, özelliğin A9 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***f1*** | ***duyarlılık*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.87 | 0.81 | 0.76 | 93 |
| **P2** | 0.66 | 0.74 | 0.85 | 27 |
| **P3** | 0.57 | 0.59 | 0.61 | 41 |
| **ort./toplam** | 0.76 | 0.74 | 0.74 | 161 |

Tablo 4.15'te eğitim veri setinin C2, özelliğin A9, *k*'nın ise 6 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.16: Eğitim veri setinin C2, özelliğin A9 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **71** | 4 | 18 | 93 |
| **P2** | 3 | **23** | 1 | 27 |
| **P3** | 8 | 8 | **25** | 41 |
| **toplam** | 82 | 35 | 44 | 161 |

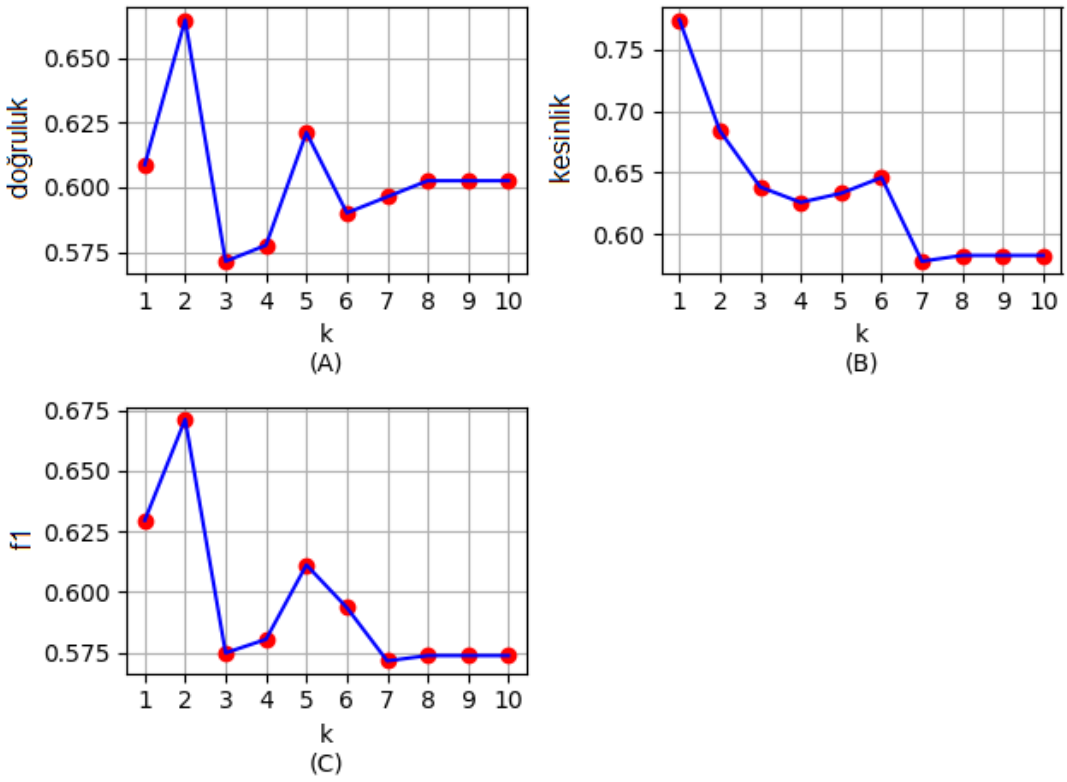
Tablo 4.16'da ise eğitim veri setinin C2, özelliğin A9, *k*'nın ise 6 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 71 tanesi, 27 tane P2 sınıfına ait örnekten 23 tanesi ve 41 tane P3 sınıfına ait örnekten 25 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 71/93'ten 0.76, 23/27'den 0.85 ve 25/41'den 0.61 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 82 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 71 tanesi bu sınıfa aittir, yine benzer şekilde 35 örnek P2 sınıfıyla, 44 örnek P3 sınıfıyla etiketlenirken gerçekte 35'ten 23 tanesi, 44'ün de 25 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 71/82'den 0.87, P2 için 23/35'ten 0.66 ve P3 için 25/44'ten 0.57 olarak hesaplanmıştır.

Tablo 4.17: Eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A4, A9, A12, A13 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| **1** | **0.6087** | **0.7740** | **0.6295** |
| 2 | 0.6646 | 0.6840 | 0.6711 |
| 3 | 0.5714 | 0.6384 | 0.5748 |
| 4 | 0.5776 | 0.6259 | 0.5804 |
| 5 | 0.6211 | 0.6333 | 0.6111 |
| 6 | 0.5901 | 0.6461 | 0.5935 |
| 7 | 0.5963 | 0.5781 | 0.5715 |
| 8 | 0.6025 | 0.5829 | 0.5737 |
| 9 | 0.6025 | 0.5829 | 0.5737 |
| 10 | 0.6025 | 0.5829 | 0.5737 |

Tablo 4.17'de eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13 olması durumunda *doğruluk*, *kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.5 ise Tablo 4.17'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.5: Eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.5'teki (A) ve (C) grafiklerinde *k* değeri 2 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.6646, ortalama *f1* 0.6711 iken, ortalama *kesinlik* 0.6840'tır. (B) grafiğinde ise *k* değeri 1 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 1 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.6087, ortalama *kesinlik* 0.7740 ve ortalama *f1* 0.6295 olarak hesaplanmıştır.

Tablo 4.18: Eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.94 | 0.52 | 0.67 | 93 |
| **P2** | 0.79 | 0.56 | 0.65 | 27 |
| **P3** | 0.38 | 0.85 | 0.53 | 41 |
| **ort/toplam** | 0.77 | 0.61 | 0.63 | 161 |

Tablo 4.18'de eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13, *k*'nın ise 1 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.19: Eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **48** | 1 | 44 | 93 |
| **P2** | 0 | **15** | 12 | 27 |
| **P3** | 3 | 3 | **35** | 41 |
| **toplam** | 51 | 19 | 91 | 161 |

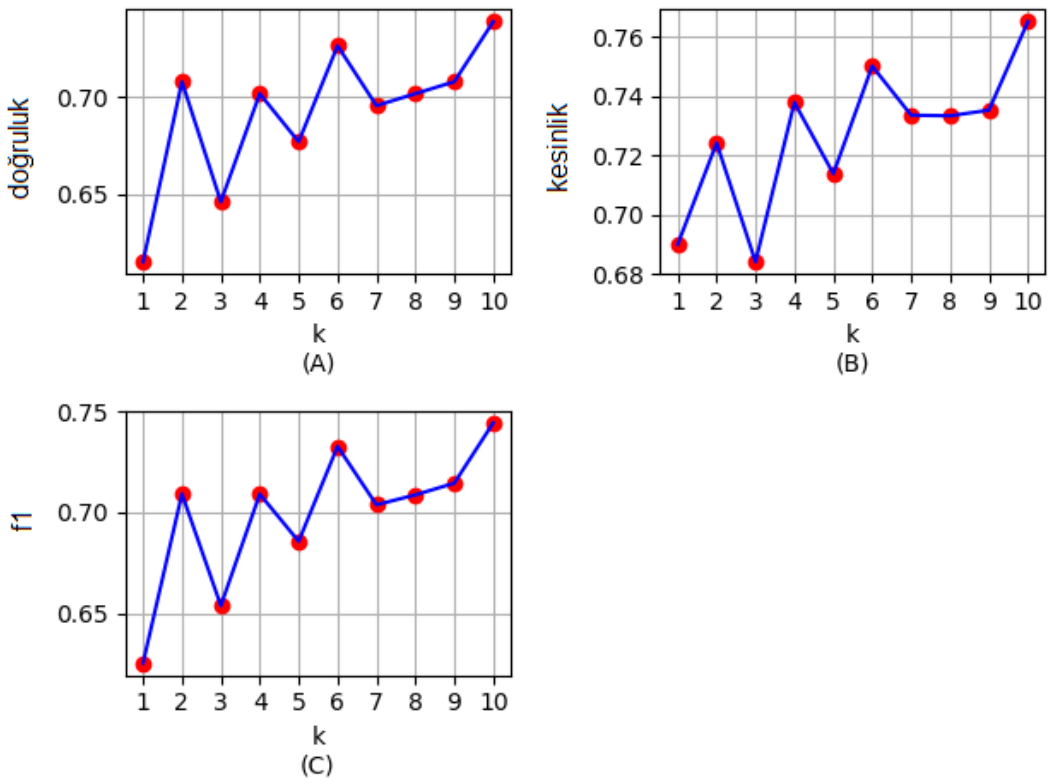
Tablo 4.19'da ise eğitim veri setinin C2, özelliklerin A4, A9, A12 ve A13, *k*'nın ise 1 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 48 tanesi, 27 tane P2 sınıfına ait örnekten 15 tanesi ve 41 tane P3 sınıfına ait örnekten 35 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 48/93'ten 0.52, 15/27'den 0.56 ve 35/41'den 0.85 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 51 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 48 tanesi bu sınıfa aittir, yine benzer şekilde 19 örnek P2 sınıfıyla, 91 örnek P3 sınıfıyla etiketlenirken gerçekte 19'dan 15 tanesi, 91'in de 35 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 48/51'den 0.94, P2 için 15/19'dan 0.79 ve P3 için 35/91'den 0.38 olarak hesaplanmıştır.

Tablo 4.20: Eğitim veri setinin C2, özelliklerin A9 ve A11 olduğu durumda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A9, A11 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| 1 | 0.6149 | 0.6899 | 0.6254 |
| 2 | 0.7081 | 0.7242 | 0.7091 |
| 3 | 0.6460 | 0.6840 | 0.6538 |
| 4 | 0.7019 | 0.7379 | 0.7090 |
| 5 | 0.6770 | 0.7138 | 0.6856 |
| 6 | 0.7267 | 0.7502 | 0.7326 |
| 7 | 0.6957 | 0.7336 | 0.7037 |
| 8 | 0.7019 | 0.7334 | 0.7086 |
| 9 | 0.7081 | 0.7353 | 0.7144 |
| **10** | **0.7391** | **0.7652** | **0.7442** |

Tablo 4.20'de eğitim veri setinin C2, özelliklerin A9 ve A11 olması durumunda *doğruluk*, *kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 7.2.6 ise Tablo 4.20'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.6: Eğitim veri setinin C2, özelliklerin A9 ve A11 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.6'daki (A) ve (C) grafiklerinde *k* değeri 10 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.7391, ortalama *f1* 0.7442 iken, ortalama *kesinlik* 0.7652'dir. (B) grafiğinde ise yine *k* değeri 10 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 10 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.7391, ortalama *kesinlik* 0.7652 ve ortalama *f1* 0.7442 olarak hesaplanmıştır.

Tablo 4.21: Eğitim veri setinin C2, özelliklerin A9 ve A11 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***f1*** | ***duyarlılık*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.88 | 0.73 | 0.80 | 93 |
| **P2** | 0.65 | 0.89 | 0.75 | 27 |
| **P3** | 0.57 | 0.66 | 0.61 | 41 |
| **ort./toplam** | 0.77 | 0.74 | 0.74 | 161 |

Tablo 4.21'de eğitim veri setinin C2, özelliklerin A9 ve A11, *k*'nın ise 10 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.22: Eğitim veri setinin C2, özelliklerin A9 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **68** | 6 | 19 | 93 |
| **P2** | 2 | **24** | 1 | 27 |
| **P3** | 7 | 7 | **27** | 41 |
| **toplam** | 77 | 37 | 47 | 161 |

Tablo 4.22'de ise eğitim veri setinin C2, özelliklerin A9 ve A11, *k*'nın ise 10 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 68 tanesi, 27 tane P2 sınıfına ait örnekten 24 tanesi ve 41 tane P3 sınıfına ait örnekten 27 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 68/93'ten 0.73, 24/27'den 0.89 ve 27/41'den 0.66 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 77 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 68 tanesi bu sınıfa aittir, yine benzer şekilde 37 örnek P2 sınıfıyla, 47 örnek P3 sınıfıyla etiketlenirken gerçekte 37'den 24 tanesi, 47'nin de 27 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *duyarlılık* değerleri P1 için, 68/77'den 0.88, P2 için 24/37'den 0.65 ve P3 için 27/47'den 0.57 olarak hesaplanmıştır.

Tablo 4.23: Eğitim veri setinin C1 ve C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C2 | | | | |
|---|---|---|---|---|---|
| **maksimum** | ***k*** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | 10 | A9, A11 | **0.7702** | 0.7907 | 0.7744 |
| ***kesinlik*** | 1 | A4 | 0.5031 | **0.8070** | 0.5024 |
| ***f1*** | 10 | A9, A11 | 0.7702 | 0.7907 | **0.7744** |
| **süre** | 538.9182 saniye | | | | |

Tablo 4.23'te *k*-NN algoritması için eğitim veri seti olarak C1 ve C2'nin kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.23'te de görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.7702 olarak hesaplanmış, bu değer, özellik olarak A9 ve A11'in, *k*'nın ise 10 olarak seçilmesiyle elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.8070 olarak hesaplanmış, bu değer özellik olarak A4'ün, *k*'nın ise 1 olarak seçildiği durumda elde edilirken, en

yüksek *f1* değeri özellik olarak A9 ve A11'in, *k*'nın ise 10 seçildiği durumda 0.7744 olarak hesaplanmıştır.

Bu değerler aynı zamanda şu anlama da gelmektedir;

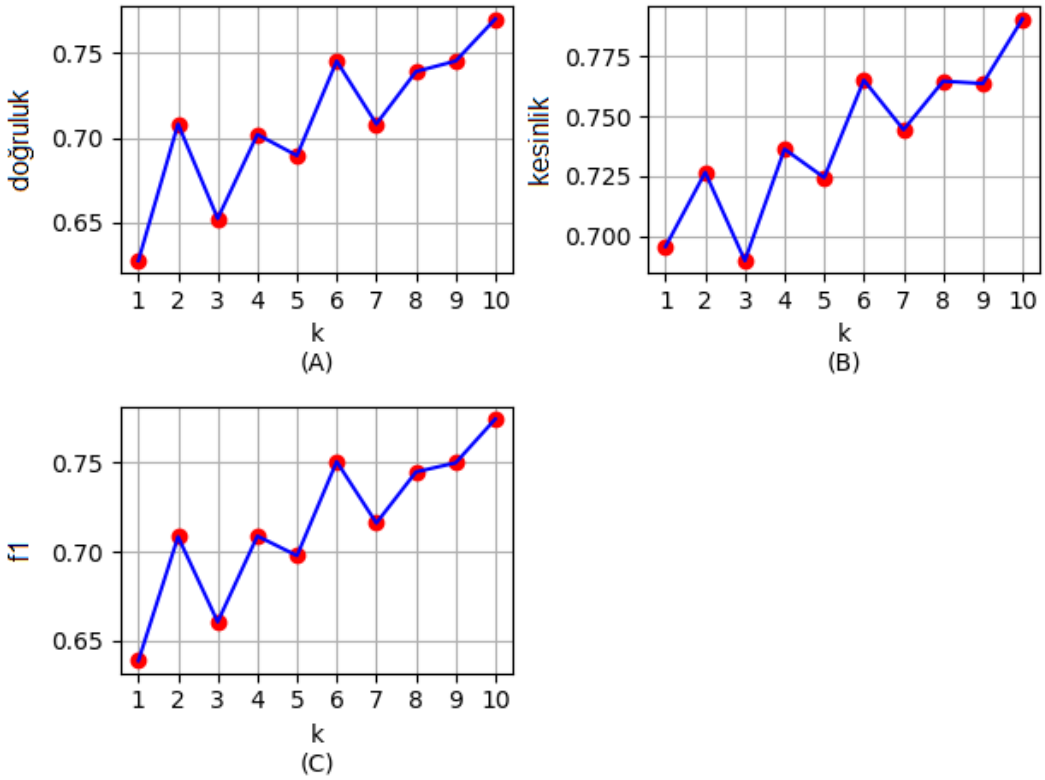
- Eğitim veri seti olarak C1 ve C2, özellik olarak A9 ve A11 kullanıldığında ve *k* değeri 10 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, ortalama %77.02'si doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C1 ve C2, özellik olarak A4 kullanıldığında ve *k* değeri 1 olarak kabul edildiğinde, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %80.70'tir.
- Eğitim veri seti olarak C1 ve C2, özellik olarak A9 ve A11 kullanıldığında ve *k* değeri 10 olarak kabul edildiğinde ortalama *f1* değeri %74.44'tür.

Tablo 4.24: Eğitim setinin C1 ve C2, özelliklerin A9 ve A11 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A9, A11 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| 1 | 0.6273 | 0.6956 | 0.6387 |
| 2 | 0.7081 | 0.7266 | 0.7083 |
| 3 | 0.6522 | 0.6897 | 0.6605 |
| 4 | 0.7019 | 0.7363 | 0.7087 |
| 5 | 0.6894 | 0.7244 | 0.6977 |
| 6 | 0.7453 | 0.7651 | 0.7505 |
| 7 | 0.7081 | 0.7444 | 0.7160 |
| 8 | 0.7391 | 0.7647 | 0.7447 |
| 9 | 0.7453 | 0.7637 | 0.7498 |
| **10** | **0.7702** | **0.7907** | **0.7744** |

Tablo 4.24'de eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11 olması durumunda *doğruluk*, *kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı

ortalama değerler gösterilmektedir. Şekil 4.7 ise Tablo 4.24'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.7: Eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.7'deki (A) ve (C) grafiklerinde *k* değeri 10 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.7702, ortalama *f1* 0.7744 iken, ortalama *kesinlik* 0.7907'dir. (B) grafiğinde ise yine *k* değeri 10 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 10 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.7702, ortalama *kesinlik* 0.7907 ve ortalama *f1* 0.7744 olarak hesaplanmıştır.

Tablo 4.25: Eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.90 | 0.77 | 0.83 | 93 |
| **P2** | 0.65 | 0.89 | 0.75 | 27 |
| **P3** | 0.64 | 0.68 | 0.66 | 41 |
| **ort./toplam** | 0.79 | 0.77 | 0.77 | 161 |

Tablo 4.25'te eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11, *k*'nın ise 10 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.26: Eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **72** | 6 | 15 | 93 |
| **P2** | 2 | **24** | 1 | 27 |
| **P3** | 6 | 7 | **28** | 41 |
| **toplam** | 80 | 37 | 44 | 161 |

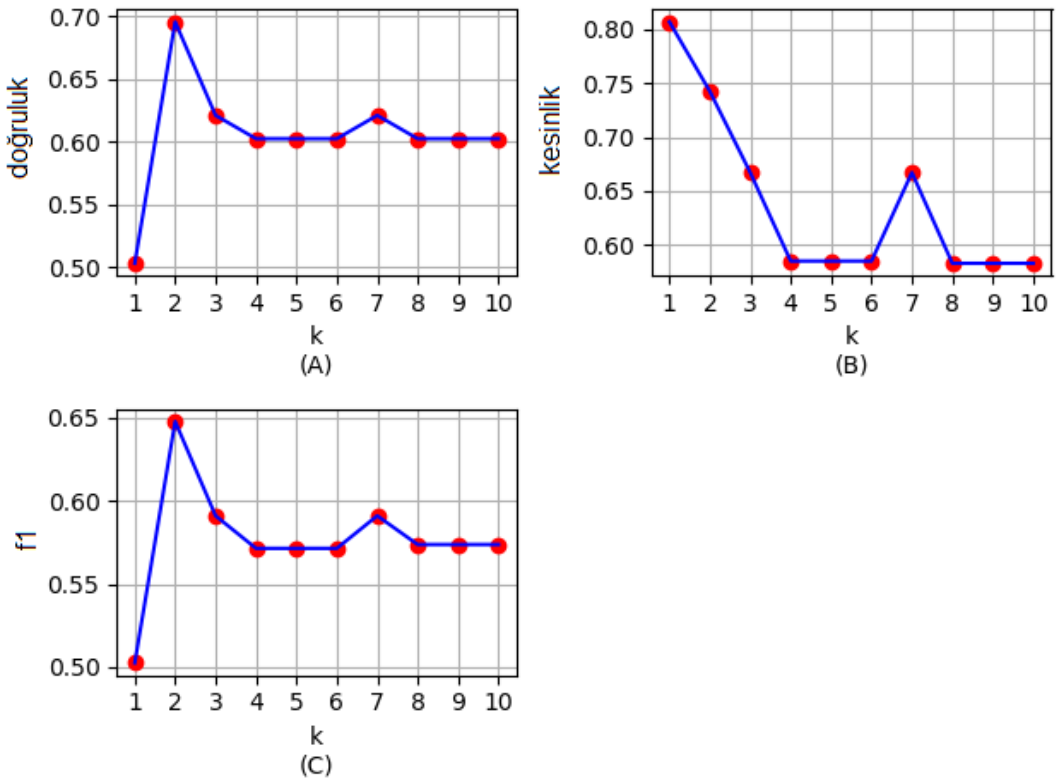
Tablo 4.26'da ise eğitim veri setinin C1 ve C2, özelliklerin A9 ve A11, *k*'nın ise 10 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 72 tanesi, 27 tane P2 sınıfına ait örnekten 24 tanesi ve 41 tane P3 sınıfına ait örnekten 28 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 72/93'ten 0.77, 24/27'den 0.89 ve 28/41'den 0.68 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 80 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 72 tanesi bu sınıfa aittir, yine benzer şekilde 37 örnek P2 sınıfıyla, 44 örnek P3 sınıfıyla etiketlenirken gerçekte 37'den 24 tanesi, 44'ün de 28 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 72/80'den 0.90, P2 için 24/37'den 0.65 ve P3 için 28/44'den 0.64 olarak hesaplanmıştır.

Tablo 4.27: Eğitim ve setinin C1 ve C2, özelliğin A4 olması durumunda başarım ölçütlerinin *k* değerine göre değişimi

| **özellik** | A4 | | |
|---|---|---|---|
| ***k*** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| **1** | **0.5031** | **0.8070** | **0.5024** |
| 2 | 0.6957 | 0.7425 | 0.6478 |
| 3 | 0.6211 | 0.6671 | 0.5910 |
| 4 | 0.6025 | 0.5850 | 0.5717 |
| 5 | 0.6025 | 0.5850 | 0.5714 |
| 6 | 0.6025 | 0.5850 | 0.5714 |
| 7 | 0.6211 | 0.6671 | 0.5910 |
| 8 | 0.6025 | 0.5829 | 0.5737 |
| 9 | 0.6025 | 0.5829 | 0.5737 |
| 10 | 0.6025 | 0.5829 | 0.5737 |

Tablo 4.27'de eğitim veri setinin C1 ve C2, özelliğin A4 olması durumunda *doğruluk*, *kesinlik* ve *f1*'in *k*'nın 1 ile 10 arasındaki değişimine göre aldığı ortalama değerler gösterilmektedir. Şekil 4.8 ise Tablo 4.27'deki bu değerlerin grafiksel olarak dağılımını göstermektedir.




Şekil 4.8: Eğitim veri setinin C1 ve C2, özelliğin A4 olması durumunda, tahminleme başarım ölçütlerinin ortalamasının *k* değerine göre değişimi, (A): *doğruluk*, (B): *kesinlik*, (C): *f1*

Şekil 4.8'deki (A) ve (C) grafiklerinde *k* değeri 2 için en yüksek değerlere ulaşılmıştır. Bu noktada elde edilen ortalama *doğruluk* 0.6957, ortalama *f1* 0.6478 iken, ortalama *kesinlik* 0.7425'tir. (B) grafiğinde ise *k* değeri 1 için en yüksek değere ulaşılmıştır. (B) grafiğinde *k* değeri 1 için elde edilen başarım ölçütleri sırasıyla, ortalama *doğruluk* 0.5031, ortalama *kesinlik* 0.8070 ve ortalama *f1* 0.5024 olarak hesaplanmıştır.

Tablo 4.28: Eğitim veri setinin C1 ve C2, özelliğin A4 olması durumunda her bir sınıfın başarım ölçütleri

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 1.00 | 0.38 | 0.55 | 93 |
| **P2** | 0.86 | 0.22 | 0.35 | 27 |
| **P3** | 0.34 | 0.98 | 0.50 | 41 |
| **ort./toplam** | 0.81 | 0.50 | 0.50 | 161 |

Tablo 4.28'de eğitim veri setinin C1 ve C2, özelliğin A4, *k*'nın ise 1 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.29: Eğitim veri setinin C1 ve C2, özelliğin A4 olması durumunda her bir sınıf için tahminleme sonuçları

| | **P1** | **P2** | **P3** | **toplam** |
|---|---|---|---|---|
| **P1** | **35** | 0 | 58 | 93 |
| **P2** | 0 | **6** | 21 | 27 |
| **P3** | 0 | 1 | **40** | 41 |
| **toplam** | 35 | 7 | 119 | 161 |

Tablo 4.29'da ise eğitim veri setinin C1 ve C2, özelliğin A4, *k*'nın ise 1 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 35 tanesi, 27 tane P2 sınıfına ait örnekten 6 tanesi ve 41 tane P3 sınıfına ait örnekten 40 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 35/93'ten 0.38, 6/27'den 0.22 ve 40/41'den 0.98 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 35 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 35 tanesi bu sınıfa aittir, yine benzer şekilde 7 örnek P2 sınıfıyla, 119 örnek P3 sınıfıyla etiketlenirken gerçekte 7 den 6 tanesi, 119'un da 40 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 35/35'ten 1.00, P2 için 6/7'den 0.86 ve P3 için 40/119'dan 0.34 olarak hesaplanmıştır.

Yukarıda da bahsedildiği üzere toplamda 245730 farklı kombinasyon için *k*-NN algoritması çalıştırılmış ve çıktı olarak 3 farklı eğitim veri setinde, tüm sınıfların ortalaması bazında *doğruluk, kesinlik* ve *f1*, her bir sınıfın sınıf bazında başarımını ölçmek için *kesinlik, duyarlılık* ve *f1* başarım ölçütleri kullanılarak en başarılı sonuçları sağlayan *k* değeri ve özellikler belirlenmiştir. Elde edilen bu en başarılı sonuçlar içinde en sık kullanılan özelliklerden A9 5 kez, A11 3 kez, A3 2 kez, A13 2 kez, A10 2 kez, A4 2 kez ve A2 1 kez kullanılmıştır.

*doğruluk* ölçütü açısından en başarılı sonuç, eğitim veri setinin C1 ve C2, *k* değerinin 10, özelliklerin A9 ve A11 kabul edildiği durumda elde edilmiştir. Bu durumda elde edilen *doğruluk* değeri 0.7702, *kesinlik* değeri 0.7907 ve *f1* değeri 0.7744'tür.

*kesinlik* ölçütü açısından en başarılı sonuçsa, eğitim veri setinin C1, *k* değerinin 1, özelliklerin A2 ve A3 olduğu durumda elde edilmiştir. Bu durumda elde edilen

*doğruluk* değeri 0.4224, *kesinlik* değeri 0.87 ve *f1* değeri 0.4425'tir. Bu koşullarda *kesinlik* değeri oldukça yüksek olmasına rağmen *doğruluk* değeri oldukça düşüktür. Bu yüzden tek başına *doğruluk* ya da *kesinlik* ölçütüne bakmak yerine *duyarlılık* ve *kesinlik* ölçütünün harmonik ortalaması olan *f1* değerine göre değerlendirme yapmanın daha iyi sonuçlar vereceği düşünülmektedir. *f1* ölçütünün en yüksek değeri ise, *doğruluk* ölçütünün en başarılı sonuçlarının elde edildiği koşullarda elde edilmiştir.

Tek tek sınıf bazında incelediğimiz zamansa;

P1 sınıfı için *duyarlılık* açısından en iyi sonuç olan 0.89 değeri eğitim veri setinin C1, özelliklerin A3 ve A10 ve *k*'nın ise 4 seçilmesi durumunda elde edilmiştir. Bu koşullar altında elde edilen *kesinlik* değeri 0.86 iken, *f1* değeri 0.87'dir. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 1 değeri iki farklı durumda elde edilmiştir. Bu durumlardan birisi eğitim veri setinin C1, *k* değerinin 1 ve özelliklerin A3 ve A10 olmasıyken, diğeri eğitim veri setinin C3, *k* değerinin 1 ve özelliğin A4 olması durumudur. Bu koşullar altında elde edilen diğer ölçüt değerleri *duyarlılık* için 0.38, *f1* için 0.55'tir. Görüldüğü gibi bu koşullarda elde edilen *kesinlik* değeri 1 olmasına rağmen *duyarlılık* değeri oldukça düşüktür, bu yüzden daha öncede belirtildiği gibi en başarılı sonucun belirlenebilmesi için *f1* değerine bakılması daha uygundur. Bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1, *k* değerinin 4, özelliklerin A3 ve A10 olması durumunda *duyarlılık* 0.89, *kesinlik* 0.86 ve *f1* 0.87 olarak hesaplanmıştır.
- Eğitim veri setinin C3, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.77, *kesinlik* 0.90 ve *f1* 0.83 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.76, *kesinlik* 0.87 ve *f1* 0.81 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.88, *kesinlik* 0.73 ve *f1* 0.80 olarak hesaplanmıştır.

- Eğitim veri setinin C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.72, *kesinlik* 0.88 ve *f1* 0.79 olarak hesaplanmıştır.

P2 sınıfı için *duyarlılık* açısından en iyi sonuç olan 1 değeri eğitim veri setinin C1, özelliklerin A2 ve A3 ve *k*'nın ise 1 seçilmesi durumunda elde edilmiştir. Bu koşullar altında elde edilen *kesinlik* değeri 0.23 iken, *f1* değeri 0.37'dir. Yani bu şartlar altında elde edilen *duyarlılık* değeri en yüksek değer olan 1 olmasına rağmen *kesinlik* değeri oldukça düşüktür. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 0.86 değeri eğitim veri setinin C1 ve C2, *k* değerinin 1, özelliğin A4 olması durumunda elde edilmiştir. Bu durumda elde edilen diğer ölçüt değerleri *duyarlılık* için 0.22, *f1* için 0.35'tir. Görüldüğü gibi bu koşullarda elde edilen *kesinlik* değeri 0.86 gibi yüksek bir değer olmasına rağmen *duyarlılık* değeri oldukça düşüktür, bu yüzden en başarılı sonucu belirlemek için *f1* değerine bakılması daha uygundur. Bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1, *k* değerinin 9, özelliklerin A10 ve A13 olması durumunda *duyarlılık* 0.85, *kesinlik* 0.70 ve *f1* 0.77 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.89, *kesinlik* 0.65 ve *f1* 0.75 olarak hesaplanmıştır.
- Eğitim veri setinin C1 ve C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.89, *kesinlik* 0.65 ve *f1* 0.75 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 6, özelliğin A9 olması durumunda *duyarlılık* 0.85, *kesinlik* 0.66 ve *f1* 0.74 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 1, özelliklerin A4, A9, A12 ve A13 olması durumunda *duyarlılık* 0.56, *kesinlik* 0.79 ve *f1* 0.65 olarak hesaplanmıştır.

P3 sınıfı için *duyarlılık* açısından en iyi sonuç olan 0.98 değeri eğitim veri setinin C1 ve C2, özelliğin A4 ve *k*'nın ise 1 seçilmesi durumunda elde edilmiştir. Bu koşullar altında elde edilen *kesinlik* değeri 0.34 iken, *f1* değeri 0.50'dir. Yani bu şartlar

altında elde edilen *duyarlılık* değeri oldukça yüksek bir değer olmasına rağmen *kesinlik* değeri oldukça düşüktür. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 0.93 değeri eğitim veri setinin C1, *k* değerinin 4, özelliklerin A3 ve A10 olması durumunda elde edilmiştir. Bu koşullar altında elde edilen diğer ölçüt değerleri *duyarlılık* için 0.34, *f1* için 0.50'dir. Görüldüğü gibi bu koşullarda elde edilen *kesinlik* değeri 0.93 gibi yüksek bir değer olmasına rağmen *duyarlılık* değeri oldukça düşüktür, bu yüzden en başarılı sonucun belirlenebilmesi için *f1* değerine bakılmıştır ve bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1 ve C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.68, *kesinlik* 0.64 ve *f1* 0.66 olarak hesaplanmıştır.
- Eğitim veri setinin C1, *k* değerinin 9, özelliklerin A10 ve A13 olması durumunda *duyarlılık* 0.71, *kesinlik* 0.56 ve *f1* 0.62 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 10, özelliklerin A9 ve A11 olması durumunda *duyarlılık* 0.66, *kesinlik* 0.57 ve *f1* 0.61 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 6, özelliğin A9 olması durumunda *duyarlılık* 0.61, *kesinlik* 0.57 ve *f1* 0.59 olarak hesaplanmıştır.
- Eğitim veri setinin C2, *k* değerinin 1, özelliklerin A4, A9, A12 ve A13 olması durumunda *duyarlılık* 0.85, *kesinlik* 0.38 ve *f1* 0.53 olarak hesaplanmıştır.

### 4.2.2 Uygulama 2: Karar Ağacı Yöntemiyle Kullanıcıların Siyasi Görüşlerinin Tahmin Edilmesi ve Kural Tabanı Çıkarımı

Bu uygulamadaki temel amacımız, Twitter üzerinden elde ettiğimiz C1, C2 ve C3 veri setlerini CART karar ağacı algoritmasıyla birlikte kullanarak Twitter kullanıcılarının siyasi görüşlerini tahminlemeye çalışmak ve kullanıcıları sınıflandırabilmek için bir kural tabanı oluşturmaktır. Bunu yaparken yukarıda Tablo 4.1'de yer alan 8 adet koleksiyon arasından C1, C2 ve C3'den ve Tablo 4.2'de yer alan 13 adet özelliğin tamamından yararlanılmıştır.

Eğitim için seçilen veri seti ve veri setinde yer alan özellikler karar ağacını, kural tabanını ve karar ağacının başarımını etkileyen faktörlerdir. Buna bağlı olarak bu uygulamada 3 temel soruya cevap aranmaya çalışılmıştır:

1. En başarılı tahminleme sonucunu sağlayan karar ağacı ve kural tabanı hangisidir?
2. En başarılı tahminleme için en uygun eğitim veri seti hangisi veya hangileridir?
3. En başarılı tahminleme için en uygun özellik hangisi veya hangileridir?

Yapılan uygulamada Tablo 4.1'de yer alan C1, C2 ve C3 koleksiyonlarından C1 ve C2 karar ağacı için eğitim, C3 ise test veri seti olarak kullanılmıştır. Ancak yalnız C1, yalnız C2 ve C1 ile C2'nin birlikte eğitim seti olarak kullanılabileceği 3 durum söz konusudur ve yukarıda bahsedilen 2 nolu sorunun cevabının bulunabilmesi için bu 3 durumun ayrı ayrı test edilmesi gerekmektedir. Yine aynı şekilde soru 3'ün cevabını bulabilmek için Tablo 4.2'de yer alan 13 adet özellikten hangisi veya hangilerinin kullanılması gerektiğini bulabilmek için tüm kombinasyonların test edilmesi, bu yüzden de $\binom{13}{1}, \binom{13}{2}, \dots, \binom{13}{13}$ şeklinde tüm kombinasyonlar için toplamda 8191 durumun ayrı ayrı test edilmesi gerekmektedir.

Özetlemek gerekirse 3 farklı eğitim veri seti ve 8191 farklı özellik seçimi yapılabilmektedir ve bu faktörlerin hepsi birbirini etkilemektedir, dolayısıyla toplamda $3 * 8191$'den 24573 farklı kombinasyon bulunmaktadır.

1 nolu soruda da yer alan algoritmanın başarımını ölçmek için Bölüm 3.3'te açıklanan *doğruluk*, *kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü kullanılmıştır. 24573 farklı kombinasyon için bu 4 ölçüt hesaplanmış ve her bir ölçüt için en yüksek değeri sağlayan, eğitim veri seti veya setleri ve özellik veya özellikler bulunmaya çalışılmıştır. Ancak bu uygulamada sonuçların daha da iyileştirilmesi için bu 24573 kombinasyon 10 kez çalıştırarak, sonunda en başarılı sonucu elde eden değerler çıktı olarak kabul edilmiştir.

Bu uygulama Python ortamında sklearn kütüphanesinin tree modülü kullanılarak gerçekleştirilmiştir. Öncelikle tree modülünün DecisionTreeClassifier sınıflandırıcısından bir nesne türetilerek bir karar ağacı oluşturulmuştur. Daha sonra

bu nesnenin fit metodu parametre olarak eğitim verilerini alarak ağaç eğitilmiş ve son olarak da predict metoduna test verileri parametre olarak gönderilerek tahminleme yapılmıştır. Yapılan analizler sırasında DecisionTreeClassifier sınıflandırıcısı criterion parametresi "gini", max_depth değeri 6 olacak şekilde kullanılmıştır. Bu, ağaç oluşturulurken dallanmaların gini indeksine bağlı olarak belirlendiği ve ağacın maksimum derinlik değerinin 6 olarak kabul edildiği anlamına gelmektedir. Bunun nedeni Şekil 4.10, Şekil 4.12, Şekil 4.13, Şekil 4.14 ve Şekil 4.15 gibi büyük ağaçlarda çok fazla dallanma meydana geldiği için ağaçların görselleştirilmesinin ve görüntülenmesinin zor olmasıdır.

Oluşturulan karar ağaçlarına göre gerçekleştirilen tahminlerin başarımını ölçmek için *k*-NN uygulamasında olduğu gibi yine Python sklearn kütüphanesinin metrics modülünden yararlanılmıştır.

Elde edilen karar ağaçlarının görselleştirilmesi Python için geliştirilmiş grapviz kütüphanesi kullanılarak gerçekleştirilmiştir. DecisionTreeClassifier sınıflandırıcısından türetilen nesnenin export_graphviz fonksiyonu kullanılarak karar ağacının GraphViz gösterimi olan bir çıktı dosyası oluşturulmuş, daha sonra graphviz'in Source fonksiyonuna oluşturulan bu dosya parametre olarak verilerek ağacın graf gösterimi olan graph nesnesi oluşturulmuş ve son olarak da bu graph nesnesinin view metodu kullanılarak karar ağacı görüntülenmiştir.

Tablo 4.30: Eğitim veri setinin C1 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A5, A13 | **0.7578** | 0.7838 | 0.7591 |
| ***kesinlik*** | A1, A7 | 0.5093 | **0.8750** | 0.5357 |
| ***f1*** | A2, A6, A7, A13 | 0.7578 | 0.7829 | **0.7597** |
| **süre** | 41.5880 saniye | | | |

Tablo 4.30'da CART algoritması için eğitim veri seti olarak C1'in kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.30'da da görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.7578 olarak hesaplanmış, bu değer, özellik olarak A2, A5 ve A13'ün seçilmesiyle elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.8750 olarak hesaplanmış, bu değer özellik olarak A1 ve A7'nin seçildiği durumda elde edilirken, en yüksek *f1* değeri özellik olarak A2, A6, A7 ve A13'ün seçildiği durumda 0.7597 olarak hesaplanmıştır.

Bu değerler aynı zamanda şu anlama da gelmektedir;

- Eğitim veri seti olarak C1, özellik olarak A2, A5 ve A13 kullanıldığında, test veri setimizde yer alan 161 örneğin, ortalama %75.78'i doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C1, özellik olarak A1 ve A7 kullanıldığında, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %87.50'dir.
- Eğitim veri seti olarak C1, özellik olarak A2, A6, A7 ve 13 kullanıldığında ortalama *f1* değeri %75.97'dir.

Tablo 4.31: Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A2, A5, A13 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 0.88 | 0.85 | 0.86 | 93 |
| **P2** | 0.51 | 0.81 | 0.63 | 27 |
| **P3** | 0.75 | 0.51 | 0.61 | 41 |
| **ort./toplam** | 0.78 | 0.76 | 0.76 | 161 |

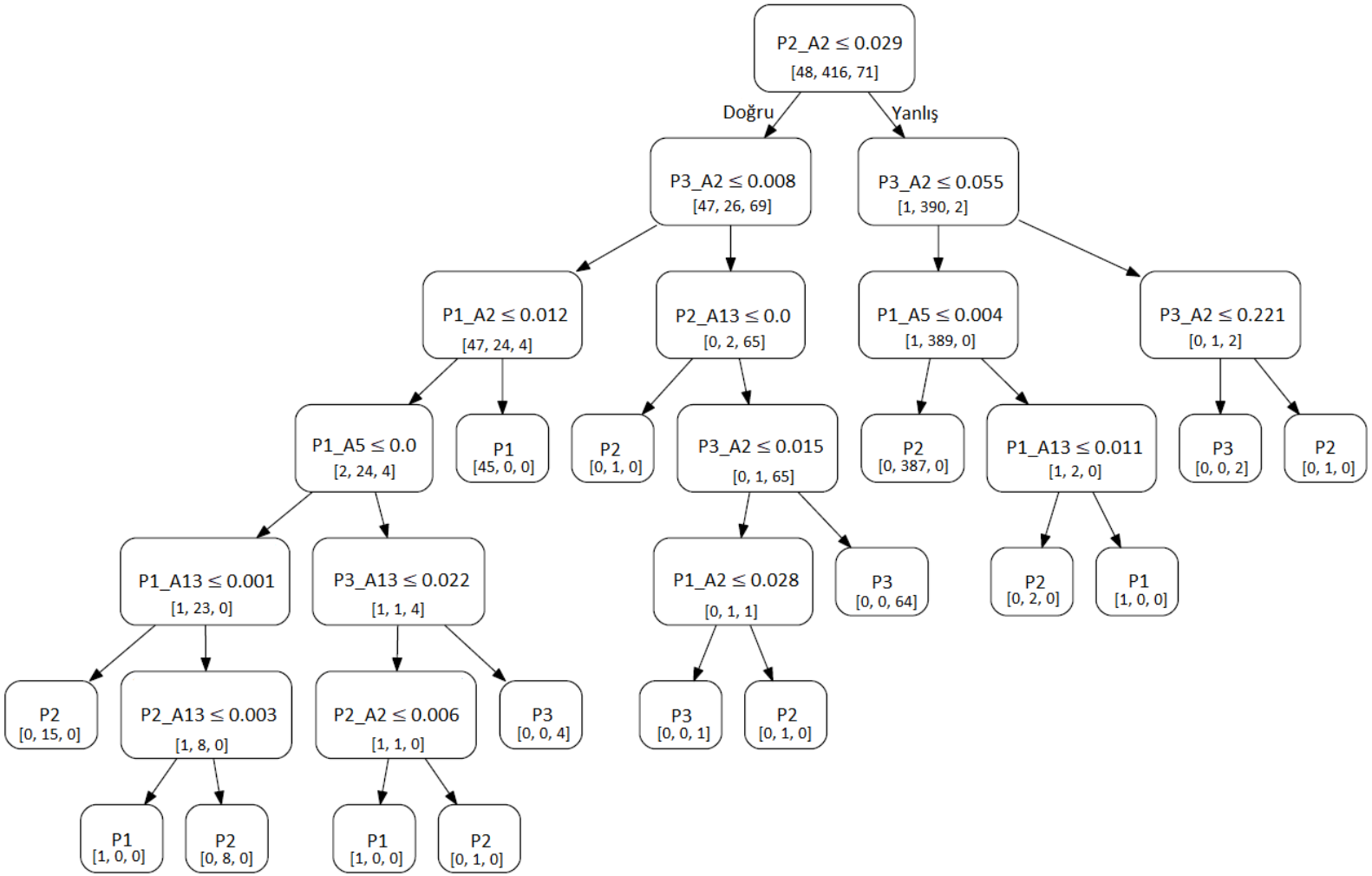
Tablo 4.31'de eğitim veri setinin C1, özelliğin A2, A5 ve A13 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.32: Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları

| **özellik** | A2, A5, A13 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **79** | 10 | 4 | 93 |
| **P2** | 2 | **22** | 3 | 27 |
| **P3** | 9 | 11 | **21** | 41 |
| **toplam** | 90 | 43 | 28 | 161 |

Tablo 4.32'de ise eğitim veri setinin C1, özelliğin A4, A5 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 79 tanesi, 27 tane P2 sınıfına ait örnekten 22 tanesi ve 41 tane P3 sınıfına ait örnekten 21 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değeri 79/93'ten 0.85, 22/27'den 0.81 ve 21/41'den 0.51 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 90 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 79 tanesi bu sınıfa aittir, yine benzer şekilde 43 örnek P2 sınıfıyla, 28 örnek P3 sınıfıyla etiketlenirken gerçekte 43'ten 22 tanesi, 28'in de 21 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 79/90'dan 0.88, P2 için 22/43'ten 0.51 ve P3 için 21/28'den 0.75 olarak hesaplanmıştır.




Şekil 4.9: Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.33: Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K1 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 ≤ 0.0) VE (P1_A13 ≤ 0.001) İSE | P2 |
| K2 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 ≤ 0.0) VE (P1_A13 > 0.001) VE (P2_13 ≤ 0.003) İSE | P1 |
| K3 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 ≤ 0.0) VE (P1_A13 > 0.001) VE (P2_13 > 0.003) İSE | P2 |
| K4 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 > 0.0) VE (P3_A13 ≤ 0.022) VE (P2_A2 ≤ 0.006) İSE | P1 |
| K5 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 > 0.0) VE (P3_A13 ≤ 0.022) VE (P2_A2 > 0.006) İSE | P2 |
| K6 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A5 > 0.0) VE (P3_A13 > 0.022) İSE | P3 |
| K7 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 > 0.012) İSE | P1 |
| K8 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 ≤ 0.0) İSE | P2 |
| K9 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P3_A2 ≤ 0.015) VE (P1_A2 ≤ 0.028) İSE | P3 |
| K10 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P3_A2 ≤ 0.015) VE (P1_A2 > 0.028) İSE | P2 |
| K11 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P3_A2 > 0.015) İSE | P3 |
| K12 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A5 ≤ 0.004) İSE | P2 |
| K13 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A5 > 0.004) VE (P1_A13 ≤ 0.011) İSE | P2 |

Tablo 4.33 (devam): Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K14 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A5 > 0.004) VE (P1_A13 > 0.011) İSE | P1 |

| | | |
|---|---|---|
| K15 | EĞER (P2_A2 > 0.029) VE (P3_A2 > 0.055) VE (P3_A2 ≤ 0.221) İSE | P3 |
| K16 | EĞER (P2_A2 > 0.029) VE (P3_A2 > 0.055) VE (P3_A2 > 0.221) İSE | P2 |

Şekil 4.9'da CART algoritması tarafından eğitim veri seti olarak C1, özellik olarak da A2, A5 ve A13'ün kullanılması durumunda elde edilen karar ağacı gösterilmektedir. Tablo 4.33'te ise bu karar ağacına bağlı olarak elde edilen 16 adet kuraldan oluşan kural tabanı yer almaktadır.

Tablo 4.34: Eğitim veri setinin C1, özelliklerin A1 ve A7 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A1, A7 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 1.00 | 0.52 | 0.68 | 93 |
| **P2** | 0.25 | 1.00 | 0.41 | 27 |
| **P3** | 1.00 | 0.17 | 0.29 | 41 |
| **ort./toplam** | 0.88 | 0.51 | 0.54 | 161 |

Tablo 4.34'de eğitim veri setinin C1, özelliğin A1 ve A7 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

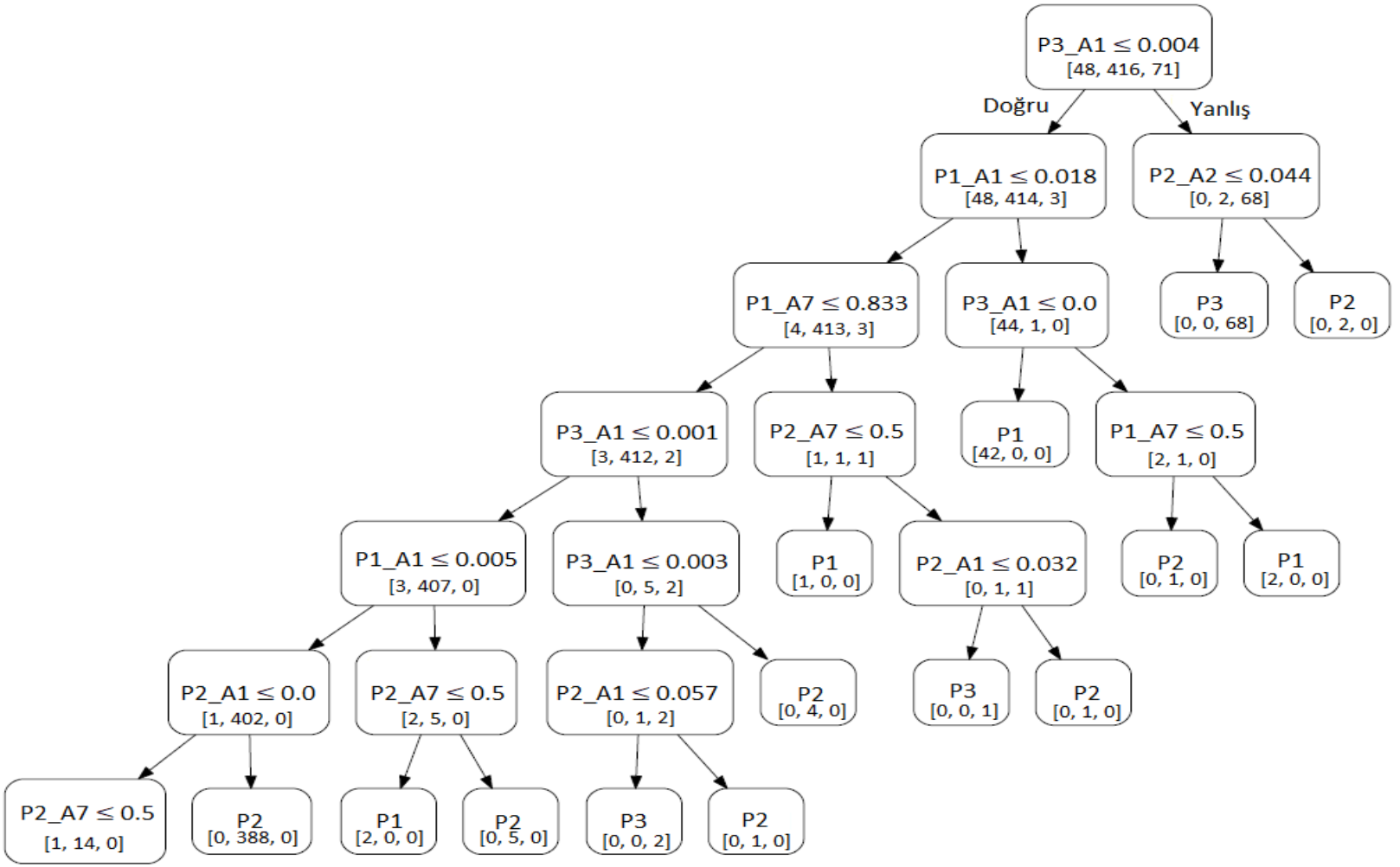
Tablo 4.35: Eğitim veri setinin C1, özelliklerin A1 ve A7 olması durumunda her bir sınıf için tahminleme sonuçları

| **özellik** | A1, A7 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **48** | 45 | 0 | 93 |
| **P2** | 0 | **27** | 0 | 27 |
| **P3** | 0 | 34 | **7** | 41 |
| **toplam** | 48 | 106 | 7 | 161 |

Tablo 4.35'te ise eğitim veri setinin C1, özelliğin A1 ve A7 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 48 tanesi, 27 tane P2 sınıfına ait örneğin tamamı ve 41 tane P3 sınıfına ait örnekten 7 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da

sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 48/93'ten 0.52, 27/27'den 1.00 ve 7/41'den 0.17 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 48 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun tamamı bu sınıfa aittir, yine benzer şekilde 106 örnek P2 sınıfıyla, 7 örnek P3 sınıfıyla etiketlenirken gerçekte 106'dan 27 tanesi, 7'nin de tamamı bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 48/48'den 1.00, P2 için 27/106'dan 0.25 ve P3 için 7/7'den 1.00 olarak hesaplanmıştır.




Şekil 4.10: Eğitim veri setinin C1, özelliklerin A1 ve A7 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.36: Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K1 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 ≤ 0.001) ve (P1_A1 ≤ 0.005) İSE | P2 |
| K2 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 ≤ 0.001) VE (P1_A1 > 0.005) VE (P2_A7 ≤ 0.5) İSE | P1 |
| K3 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 ≤ 0.001) VE (P1_A1 > 0.005) VE (P2_A7 > 0.5) İSE | P2 |
| K4 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 > 0.001) VE (P3_A1 ≤ 0.003) VE (P2_A1 ≤ 0.057) İSE | P3 |
| K5 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 > 0.001) VE (P3_A1 ≤ 0.003) VE (P2_A1 > 0.057) İSE | P2 |
| K6 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 ≤ 0.833) VE (P3_A1 > 0.001) VE (P3_A1 > 0.003) | P2 |
| K7 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 > 0.833) VE (P2_A7 ≤ 0.5) İSE | P1 |
| K8 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 > 0.833) VE (P2_A1 ≤ 0.032) İSE | P3 |
| K9 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 ≤ 0.018) VE (P1_A7 > 0.833) VE (P2_A1 > 0.032) İSE | P2 |
| K10 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 > 0.018) VE (P3_A1 ≤ 0.0) İSE | P1 |
| K11 | EĞER (P3_A1 ≤ 0.004) VE (P1_A1 > 0.018) VE (P3_A1 > 0.0) VE (P1_A7 ≤ 0.5) İSE | P2 |
| K12 | EĞER (P3_A1 ≤ 0.004 ) VE (P1_A1 > 0.018) VE (P3_A1 > 0.0) VE (P1_A7 > 0.5) İSE | P1 |
| K13 | EĞER (P3_A1 > 0.004) VE (P2_A1 ≤ 0.044) İSE | P3 |
| K14 | EĞER (P3_A1 > 0.004) VE (P2_A1 > 0.044) İSE | P2 |

Şekil 4.10'da CART algoritması tarafından eğitim veri seti olarak C1, özellik olarak da A1 ve A7'nin kullanılması durumunda elde edilen karar ağacı

gösterilmektedir. Tablo 4.36'da ise bu karar ağacına bağlı olarak elde edilen 14 adet kuraldan oluşan kural tabanı yer almaktadır. Şekil 4.10'daki ağacın en sol yaprağında görüldüğü üzere tam olarak ayrışma sağlanmamıştır. Bundan dolayı normalde bu yapraktan aşağıya doğru dallanma devam etmektedir ancak burada bir sınıftan 1, diğer sınıftan ise 14 tane örnek olduğu yani 14 tane örneği olan sınıf, 1 tane örneği olan sınıfa oldukça baskın olduğu, ağacın daha fazla dallanarak görselleştirilmesini zorlaştırmamak ve maksimum derinlik 6 olarak kabul edildiği için buradaki dallanma ihmal edilerek sınıf etiketi örnek sayısı fazla olan sınıfın etiketi olarak kabul edilmiştir. Aynı durum Tablo 5.43'te K9, K11 ve K12 nolu kurallarda ve Tablo 5.46'da K9, K10, K19, K20, K21, K29 ve K30 nolu kurallarda da mevcuttur.

Yukarıda açıklandığı üzere ihmal edilen dallanma sonucunda oluşan ağacın en solundaki yaprağın bir üst seviyesinde yer alan düğümdeki koşul P2_A1 ≤ 0.0'dır. Ancak bu koşulun doğru veya yanlış olması durumunda elde edilen sınıf etiketi P2'dir. Yani bu koşulun kuraldan çıkarılması herhangi bir şeyi değiştirmemekte ve aynı zamanda daha az kural oluşturulmasını sağlamaktadır. Bu nedenlerden dolayı bu koşul kural tabanında ihmal edilerek bunun yerine K1 nolu kural oluşturulmuştur. Benzer durum Tablo 5.43'te K22 nolu kuralda, Tablo 5.46'da K25 nolu kuralda, Tablo 5.50'de K19 ve K23 nolu kurallarda ve Tablo 5.53'te ise K1 ve K20 nolu kurallarda da mevcuttur.

Tablo 4.37: Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A2, A6, A7, A13 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 0.89 | 0.85 | 0.87 | 93 |
| **P2** | 0.51 | 0.81 | 0.63 | 27 |
| **P3** | 0.72 | 0.51 | 0.60 | 41 |
| **ort./toplam** | 0.78 | 0.76 | 0.76 | 161 |

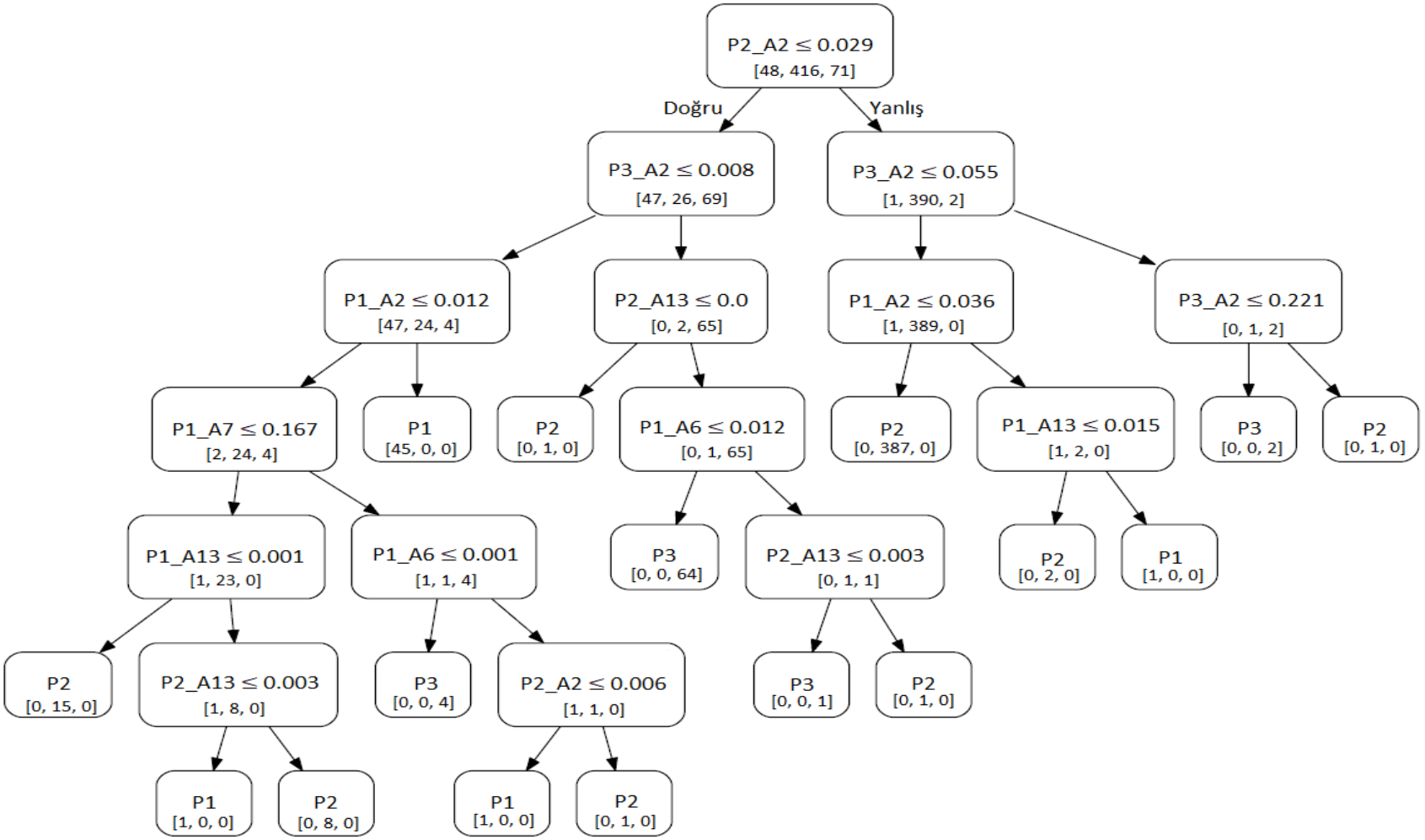
Tablo 4.37'de eğitim veri setinin C1, özelliğin A2, A6, A7 ve A13 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.38: Eğitim veri setinin C1, özelliğin A2, A6, A7 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları

| **özellik** | A2, A6, A7, A13 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **79** | 10 | 4 | 93 |
| **P2** | 1 | **22** | 4 | 27 |
| **P3** | 9 | 11 | **21** | 41 |
| **toplam** | 89 | 43 | 29 | 161 |

Tablo 4.38'de ise eğitim veri setinin C1, özelliğin A2, A6, A7 ve A13 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 79 tanesi, 27 tane P2 sınıfına ait örneğin 22 tanesi ve 4 tane P3 sınıfına ait örnekten 21 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 79/93'ten 0.85, 22/27'den 0.81 ve 21/41'den 0.51 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 89 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 79 tanesi bu sınıfa aittir, yine benzer şekilde 43 örnek P2 sınıfıyla, 29 örnek P3 sınıfıyla etiketlenirken gerçekte 43'ten 22 tanesi, 29'un da 21 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 79/89'dan 0.89, P2 için 22/43'ten 0.51 ve P3 için 21/29'dan 0.72 olarak hesaplanmıştır.




Şekil 4.11: Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.39: Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K1 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A7 ≤ 0.167) VE (P1_A13 ≤ 0.001) | P2 |
| K2 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A7 ≤ 0.167) VE (P1_A13 > 0.001) VE (P2_A13 ≤ 0.003) İSE | P1 |
| K3 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A7 ≤ 0.167) VE (P1_A13 > 0.001) VE (P2_A13 > 0.003) İSE | P2 |
| K4 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A7 > 0.167) VE (P1_A6 ≤ 0.001) İSE | P3 |
| K5 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A6 > 0.167) VE (P1_A6 > 0.001) VE (P2_A2 ≤ 0.006) İSE | P1 |
| K6 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 ≤ 0.012) VE (P1_A6 > 0.167) VE (P1_A6 > 0.001) VE (P2_A2 > 0.006) İSE | P2 |
| K7 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 ≤ 0.008) VE (P1_A2 > 0.012) İSE | P1 |
| K8 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 ≤ 0.0) İSE | P2 |
| K9 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P1_A6 ≤ 0.012) İSE | P3 |
| K10 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P1_A6 > 0.012) VE (P2_A13 ≤ 0.003) İSE | P3 |
| K11 | EĞER (P2_A2 ≤ 0.029) VE (P3_A2 > 0.008) VE (P2_A13 > 0.0) VE (P1_A6 > 0.012) VE (P2_A13 > 0.003) İSE | P2 |
| K12 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A2 ≤ 0.036) İSE | P2 |
| K13 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A2 > 0.036) VE (P1_A13 ≤ 0.015) İSE | P2 |

Tablo 4.39 (devam): Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K14 | EĞER (P2_A2 > 0.029) VE (P3_A2 ≤ 0.055) VE (P1_A2 > 0.036) VE (P1_A13 > 0.015) İSE | P1 |
| K15 | EĞER (P2_A2 > 0.029) VE (P3_A2 > 0.055) VE (P3_A2 ≤ 0.221) İSE | P3 |
| K16 | EĞER (P2_A2 > 0.029) VE (P3_A2 > 0.055) VE (P3_A2 > 0.221) İSE | P2 |

Şekil 4.11'de CART algoritması tarafından eğitim veri seti olarak C1, özellik olarak da A2, A6, A7 ve A13'ün kullanılması durumunda elde edilen karar ağacı gösterilmektedir. Tablo 4.39'da ise bu karar ağacına bağlı olarak elde edilen 16 adet kuraldan oluşan kural tabanı yer almaktadır.

Tablo 4.40: Eğitim veri setinink C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A1, A4, A7, A8, 10, A11 | **0.7267** | 0.7277 | 0.7263 |
| ***kesinlik*** | A2, A4, A5, A6, A8, A9, A10 | 0.5776 | **0.7993** | 0.5860 |
| ***f1*** | A1, A4, A7, A8, 10, A11 | 0.7267 | 0.7277 | **0.7263** |
| **süre** | 61.3802 saniye | | | |

Tablo 4.40'da CART algoritması için eğitim veri seti olarak C2'nin kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.40'da da görüldüğü gibi en yüksek ortalama *doğruluk* ve ortalama *duyarlılık* değeri 0.7267 olarak hesaplanmış, bu değer, özellik olarak A1, A4, A7, A8, A10 ve A11'in seçilmesiyle elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.7993 olarak hesaplanmış, bu değer özellik olarak A2, A4, A5, A6, A8, A9 ve A10'un seçilmesi durumunda elde edilirken, en yüksek *f1* değeri özellik olarak A1, A4, A7, A8, A10 ve A11'in seçilmesi durumunda 0.7263 olarak hesaplanmıştır.

Bu değerler aynı zamanda şu anlama da gelmektedir;

- Eğitim veri seti olarak C2, özellik olarak A1, A4, A7, A8, A10 ve A11 kullanıldığında, test veri setimizde yer alan 161 örneğin, ortalama %72.67'si doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C2, özellik olarak A2, A4, A5, A6, A8, A9 ve A10 kullanıldığında, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %79.93'tür.
- Eğitim veri seti olarak C2, özellik olarak A1, A4, A7, A8, A10 ve A11 kullanıldığında, ortalama *f1* değeri %72.63'tür.

Tablo 4.41: Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A1, A4, A7, A8, A10, A11 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 0.82 | 0.82 | 0.82 | 93 |
| **P2** | 0.58 | 0.67 | 0.62 | 27 |
| **P3** | 0.62 | 0.56 | 0.59 | 41 |
| **ort./toplam** | 0.73 | 0.73 | 0.73 | 161 |

Tablo 4.41'de eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir.

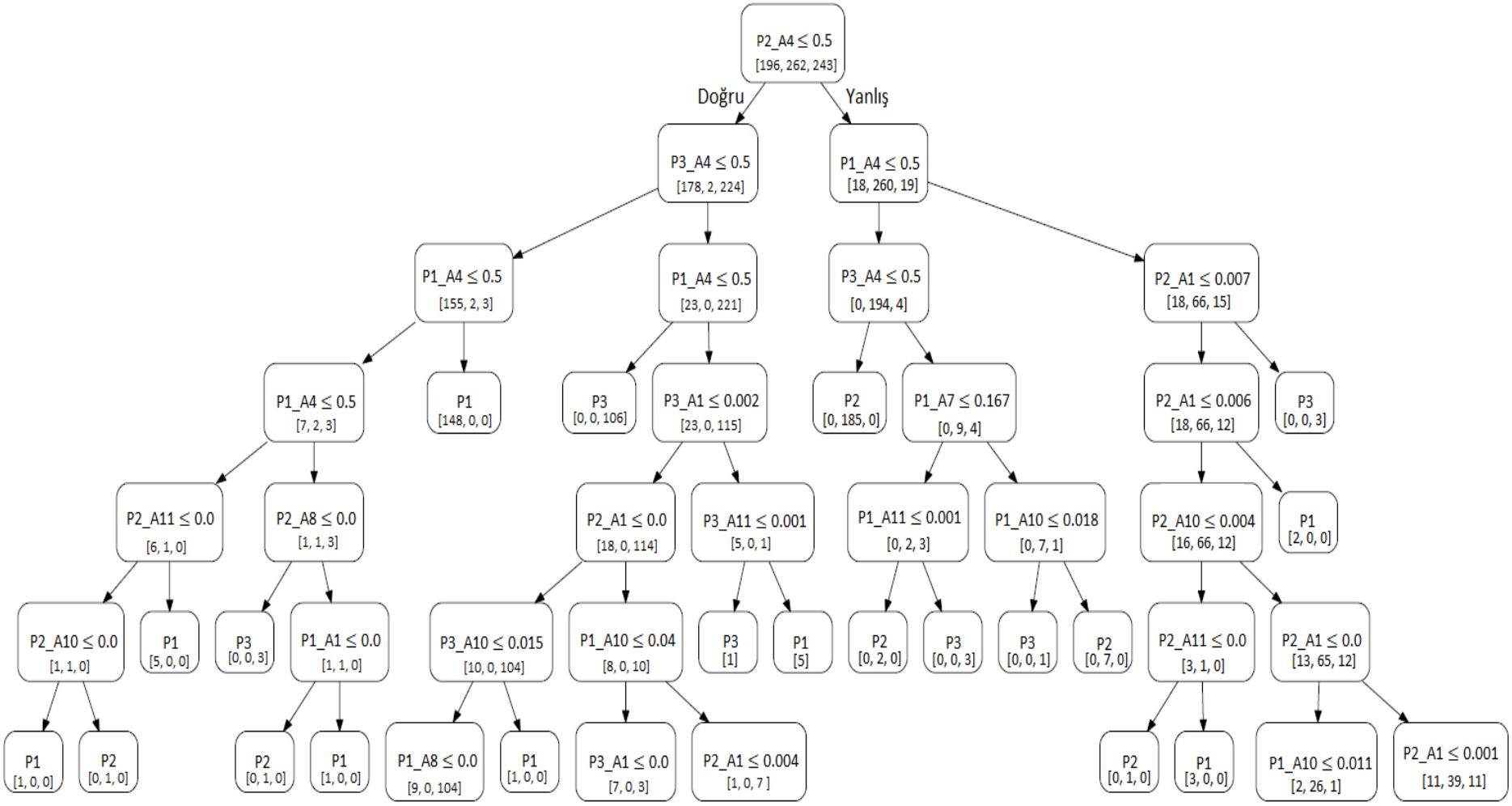
Tablo 4.42: Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları

| **özellik** | A1, A4, A7, A8, A10, A11 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **76** | 5 | 12 | 93 |
| **P2** | 7 | **18** | 2 | 27 |
| **P3** | 10 | 8 | **23** | 41 |
| **toplam** | 93 | 31 | 37 | 161 |

Tablo 4.42'de ise eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 76 tanesi, 27 tane P2 sınıfına ait

örneğin 23 tanesi ve 41 tane P3 sınıfına ait örnekten 23 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 76/93'ten 0.82, 18/27'den 0.67 ve 23/41'den 0.56 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 93 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 76 tanesi bu sınıfa aittir, yine benzer şekilde 31 örnek P2 sınıfıyla, 37 örnek P3 sınıfıyla etiketlenirken gerçekte 31'den 18 tanesi, 37'nin de 23 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 76/93'ten 0.82, P2 için 18/31'den 0.58 ve P3 için 23/37'den 0.62 olarak hesaplanmıştır.




Şekil 4.12: Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.43: Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K1 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 ≤ 0.0) VE (P2_A11 ≤ 0.0) VE (P2_A10 ≤ 0.0) İSE | P1 |
| K2 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 ≤ 0.0) VE (P2_A11 ≤ 0.0) VE A(P2_A10 > 0.0) İSE | P2 |
| K3 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 ≤ 0.0) VE (P2_A11 > 0.0) İSE | P1 |
| K4 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 > 0.0) VE (P2_A8 ≤ 0.0) İSE | P3 |
| K5 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 > 0.0) VE (P2_A8 > 0.0) VE (P1_A1 ≤ 0.0) İSE | P2 |
| K6 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 > 0.0) VE (P2_A8 > 0.0) VE (P1_A1 > 0.0) İSE | P1 |
| K7 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) İSE | P1 |
| K8 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 ≤ 0.5) İSE | P3 |
| K9 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 ≤ 0.002) VE (P2_A1 ≤ 0.0) VE (P3_A10 ≤ 0.015) İSE | P3 |
| K10 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 ≤ 0.002) VE (P2_A1 ≤ 0.0) VE (P3_A10 > 0.015) İSE | P1 |
| K11 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 ≤ 0.002) VE (P2_A1 > 0.0) VE (P1_A10 ≤ 0.04) İSE | P1 |
| K12 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 ≤ 0.002) VE (P2_A1 > 0.0) VE (P1_A10 > 0.04) İSE | P3 |
| K13 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 > 0.002) VE (P3_A11 ≤ 0.001) İSE | P3 |
| K14 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A1 > 0.002) VE (P3_A11 > 0.001) İSE | P1 |
| K15 | EĞER (P2_A4 > 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) İSE | P2 |
| K16 | EĞER (P2_A4 > 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A7 ≤ 0.167) VE (P1_A11 ≤ 0.001) İSE | P2 |

Tablo 4.43 (devam): Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K17 | EĞER (P2_A4 > 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A7 ≤ 0.167) VE (P1_A11 > 0.001) İSE | P3 |
| K18 | EĞER (P2_A4 > 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A7 > 0.167) VE (P1_A10 ≤ 0.018) İSE | P3 |
| K19 | EĞER (P2_A4 > 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A7 > 0.167) VE (P1_A10 > 0.018) İSE | P2 |
| K20 | EĞER (P2_A4 > 0.5) VE (P1_A4 > 0.5) VE (P2_A1 ≤ 0.007) VE (P2_A1 ≤ 0.006) VE (P2_A10 ≤ 0.004) VE (P2_A11 ≤ 0.0) İSE | P2 |
| K21 | EĞER (P2_A4 > 0.5) VE (P1_A4 > 0.5) VE (P2_A1 ≤ 0.007) VE (P2_A1 ≤ 0.006) VE (P2_A10 ≤ 0.004) VE (P2_A11 > 0.0) İSE | P1 |
| K22 | EĞER (P2_A4 > 0.5) VE (P1_A4 > 0.5) VE (P2_A1 ≤ 0.007) VE (P2_A1 ≤ 0.006) VE (P2_A10 > 0.004) İSE | P2 |
| K23 | EĞER (P2_A4 > 0.5) VE (P1_A4 > 0.5) VE (P2_A1 ≤ 0.007) VE (P2_A1 > 0.006) İSE | P1 |
| K24 | EĞER (P2_A4 > 0.5) VE (P1_A4 > 0.5) VE (P2_A1 > 0.007) İSE | P3 |

Şekil 4.12'de CART algoritması tarafından eğitim veri seti olarak C2, özellik olarak da A1, A4, A7, A8, A10 ve A11'in kullanılması durumunda elde edilen karar ağacı gösterilmektedir. Tablo 4.43'te ise bu karar ağacına bağlı olarak elde edilen 24 adet kuraldan oluşan kural tabanı yer almaktadır.

Tablo 4.44: Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A2, A4, A5, A6, A8, A9, A10 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 0.97 | 0.42 | 0.59 | 93 |
| **P2** | 0.82 | 0.52 | 0.64 | 27 |
| **P3** | 0.38 | 0.98 | 0.55 | 41 |
| **ort./toplam** | 0.80 | 0.58 | 0.59 | 161 |

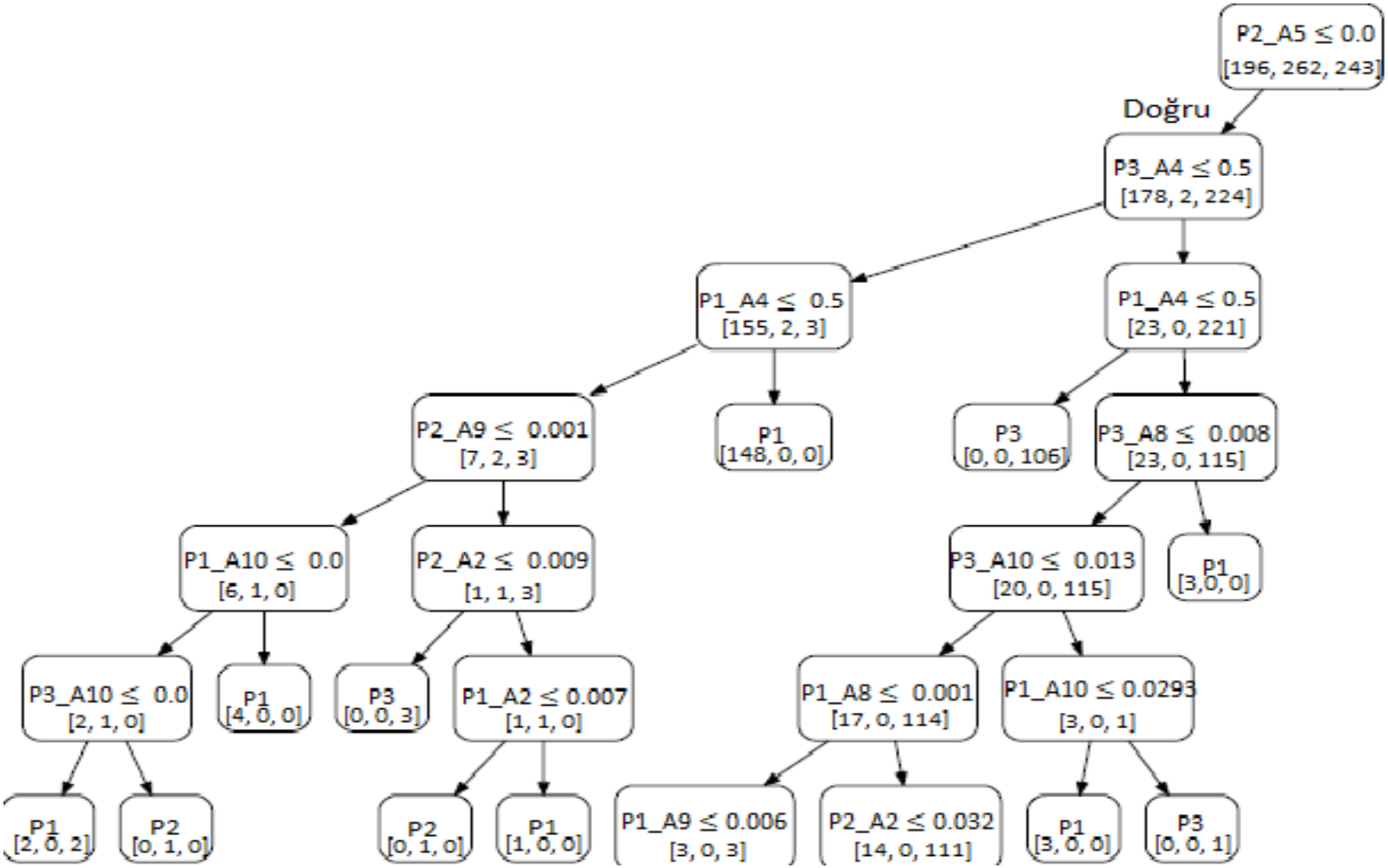
Tablo 4.44'de eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.45: Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda her bir sınıf için tahminleme sonuçları

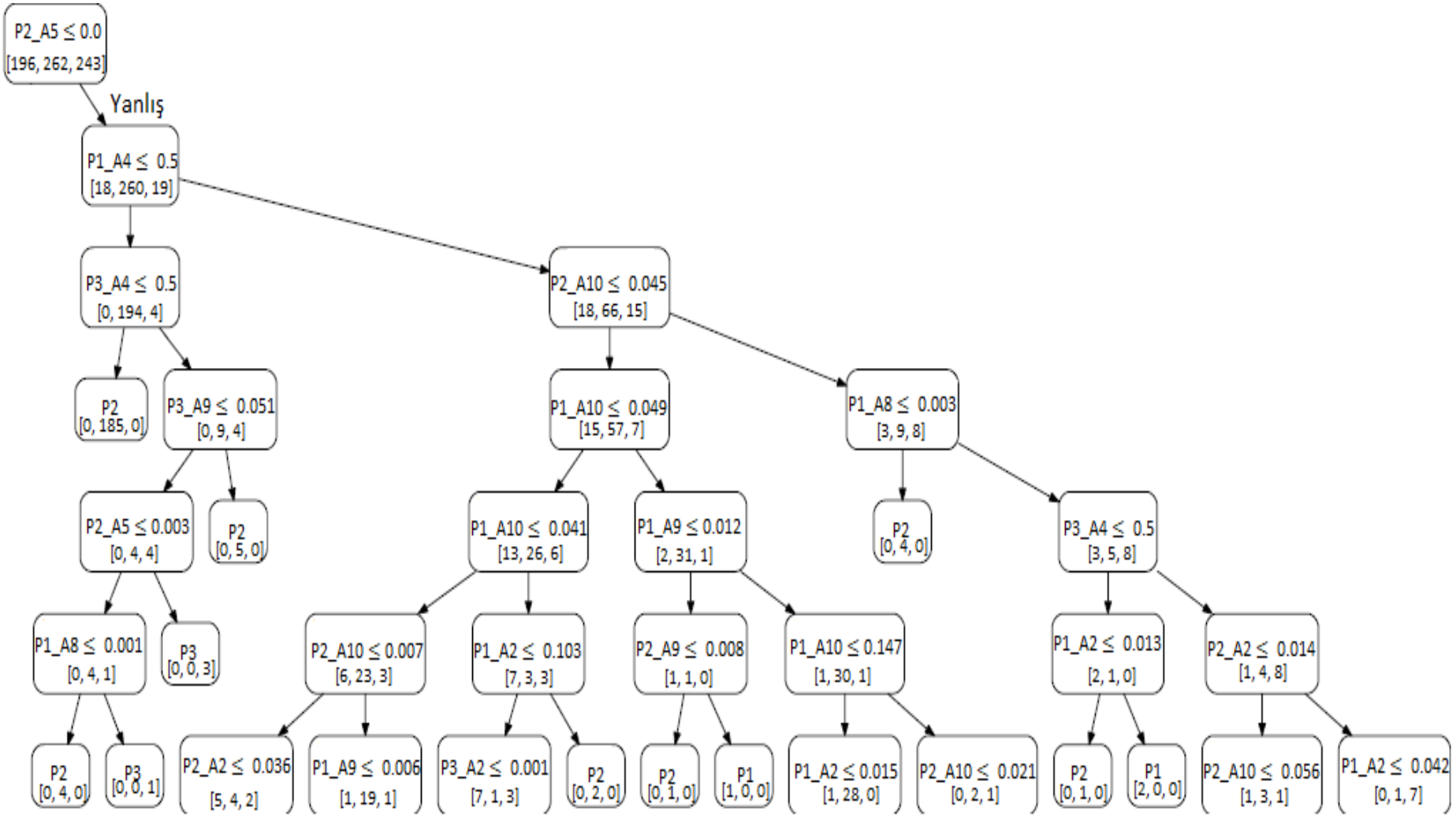
| **özellikler** | A2, A4, A5, A6, A8, A9, A10 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **39** | 2 | 52 | 93 |
| **P2** | 1 | **14** | 12 | 27 |
| **P3** | 0 | 1 | **40** | 41 |
| **toplam** | 40 | 17 | 104 | 161 |

Tablo 4.45'te ise eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 39 tanesi, 27 tane P2 sınıfına ait örneğin 14 tanesi ve 41 tane P3 sınıfına ait örnekten 40 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 39/93'ten 0.42, 14/27'den 0.52 ve 40/41'den 0.98 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 40 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 39 tanesi bu sınıfa aittir, yine benzer şekilde 17 örnek P2 sınıfıyla, 104 örnek P3 sınıfıyla etiketlenirken gerçekte 17'den 14 tanesi, 104'ün de 40 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 39/40'dan 0.97, P2 için 14/17'den 0.82 ve P3 için 40/104'den 0.38 olarak hesaplanmıştır.




Şekil 4.13: Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8 ve A10 olması durumunda CART algoritmasına göre oluşan karar ağacı




Şekil 4.13 (devam): Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8 ve A10 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.46: Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8 ve A10 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K1 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 ≤ 0.001) VE (P1_A10 ≤ 0.0) VE (P3_A10 ≤ 0.0) İSE | P1 |
| K2 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 ≤ 0.001) VE (P1_A10 ≤ 0.0) VE (P3_A10 > 0.0) İSE | P2 |
| K3 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 ≤ 0.001) VE (P1_A10 > 0.0) İSE | P1 |
| K4 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 > 0.001) VE (P2_A2 ≤ 0.009) İSE | P3 |
| K5 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 > 0.001) VE (P2_A2 > 0.009) VE (P1_A2 ≤ 0.007) İSE | P2 |
| K6 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P2_A9 > 0.001) VE (P2_A2 > 0.009) VE (P1_A2 > 0.007) İSE | P1 |
| K7 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) İSE | P1 |
| K8 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 ≤ 0.5) İSE | P3 |
| K9 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P3_A10 ≤ 0.013) VE (P1_A8 ≤ 0.001) İSE | P1 |
| K10 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P3_A10 ≤ 0.013) VE (P1_A8 > 0.001) İSE | P3 |
| K11 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P3_A10 > 0.013) VE (P1_A10 ≤ 0.293) İSE | P1 |
| K12 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P3_A10 > 0.013) VE (P1_A10 > 0.293) İSE | P3 |
| K13 | EĞER (P2_A5 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 > 0.008) İSE | P1 |
| K14 | EĞER (P2_A5 > 0.0) VE (P1_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) İSE | P2 |
| K15 | EĞER (P2_A5 > 0.0) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P3_A9 ≤ 0.051) VE (P2_A5 ≤ 0.003) VE (P1_A8 ≤ 0.001) İSE | P2 |
| K16 | EĞER (P2_A5 > 0.0) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P3_A9 ≤ 0.051) VE (P2_A5 ≤ 0.003) VE (P1_A8 > 0.001) İSE | P3 |

Tablo 4.46 (devam): Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8 ve A10 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K17 | EĞER (P2_A5 > 0.0) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P3_A9 ≤ 0.051) VE (P2_A5 > 0.003) İSE | P3 |
| K18 | EĞER (P2_A5 > 0.0) VE (P1_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P3_A9 > 0.051) İSE | P2 |
| K19 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 ≤ 0.049) VE (P1_A10 ≤ 0.041) VE (P2_A10 ≤ 0.007) İSE | P1 |
| K20 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 ≤ 0.049) VE (P1_A10 ≤ 0.041) VE (P2_A10 > 0.007) İSE | P2 |
| K21 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 ≤ 0.049) VE (P1_A10 > 0.041) VE (P1_A2 ≤ 0.103) İSE | P1 |
| K22 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 ≤ 0.049) VE (P1_A10 > 0.041) VE (P1_A2 > 0.103) İSE | P2 |
| K23 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 > 0.049) VE (P1_A9 ≤ 0.012) VE (P2_A9 ≤ 0.008) İSE | P2 |
| K24 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 > 0.049) VE (P1_A9 ≤ 0.012) VE (P2_A9 > 0.008) İSE | P1 |
| K25 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A10 > 0.049) VE (P1_A9 > 0.012) İSE | P2 |
| K26 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A8 ≤ 0.003) İSE | P2 |
| K27 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A8 > 0.003) VE (P3_A4 ≤ 0.5) VE (P1_A2 ≤ 0.013) İSE | P2 |
| K28 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A8 > 0.003) VE (P3_A4 ≤ 0.5) VE (P1_A2 > 0.013) İSE | P1 |
| K29 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A8 > 0.003) VE (P3_A4 > 0.5) VE (P2_A2 ≤ 0.014) İSE | P2 |
| K30 | EĞER (P2_A5 > 0.0) VE (P1_A4 > 0.5) VE (P2_A10 ≤ 0.045) VE (P1_A8 > 0.003) VE (P3_A4 > 0.5) VE (P2_A2 > 0.014) İSE | P3 |

Şekil 4.13'te CART algoritması tarafından eğitim veri seti olarak C2, özellik olarak da A2, A4, A5, A6, A8, A9 ve A10'un kullanılması durumunda elde edilen

karar ağacı gösterilmektedir. Tablo 4.46'da ise bu karar ağacına bağlı olarak elde edilen 30 adet kuraldan oluşan kural tabanı yer almaktadır.

Tablo 4.47: Eğitim veri setinin C1 ve C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C2 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A4, A10, A11 | **0.7081** | 0.7164 | 0.7106 |
| ***kesinlik*** | A4, A6, A7, A8, A10 | 0.6273 | **0.7775** | 0.6558 |
| ***f1*** | A2, A4, A10, A11 | 0.7081 | 0.7164 | **0.7106** |
| **süre** | 90.5425 saniye | | | |

Tablo 4.47'de CART algoritması için eğitim veri seti olarak C1 ve C2'nin kullanılması durumunda elde edilen en başarılı sonuçlar gösterilmektedir. Tablo 4.47'de de görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.7081 olarak hesaplanmış, bu değer, özellik olarak A2, A4, A10 ve A11'in seçilmesiyle elde edilmiştir. Yine aynı şekilde en yüksek *kesinlik* değeri 0.7775 olarak hesaplanmış, bu değer özellik olarak A4, A6, A7, A8 ve A10'un seçilmesi durumunda elde edilirken, en yüksek *f1* değeri özellik olarak A2, A4, A10 ve A11'in seçilmesi durumunda 0.7089 olarak hesaplanmıştır.

Bu değerler aynı zamanda şu anlama da gelmektedir;

- Eğitim veri seti olarak C1 ve C2, özellik olarak A2, A4, A10 ve A11 kullanıldığında, test veri setimizde yer alan 161 örneğin, ortalama %70.81'i doğru olarak sınıflandırılmaktadır.
- Eğitim veri seti olarak C1 ve C2, özellik olarak A4, A6, A7, A8 ve A10 kullanıldığında, test veri setimizde yer alan 161 örneğin, herhangi bir sınıf ile etiketlendiği zaman gerçekte o sınıfa ait olma ihtimali ortalama %77.75'dir.
- Eğitim veri seti olarak C1 ve C2, özellik olarak A2, A4, A10 ve A11 kullanıldığında, ortalama *f1* değeri %70.89'dur.

Tablo 4.48: Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A2, A4, A10, A11 | | | |
|---|---|---|---|---|
| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
| **P1** | 0.81 | 0.75 | 0.78 | 93 |
| **P2** | 0.66 | 0.78 | 0.71 | 27 |
| **P3** | 0.53 | 0.56 | 0.55 | 41 |
| **ort./toplam** | 0.72 | 0.71 | 0.71 | 161 |

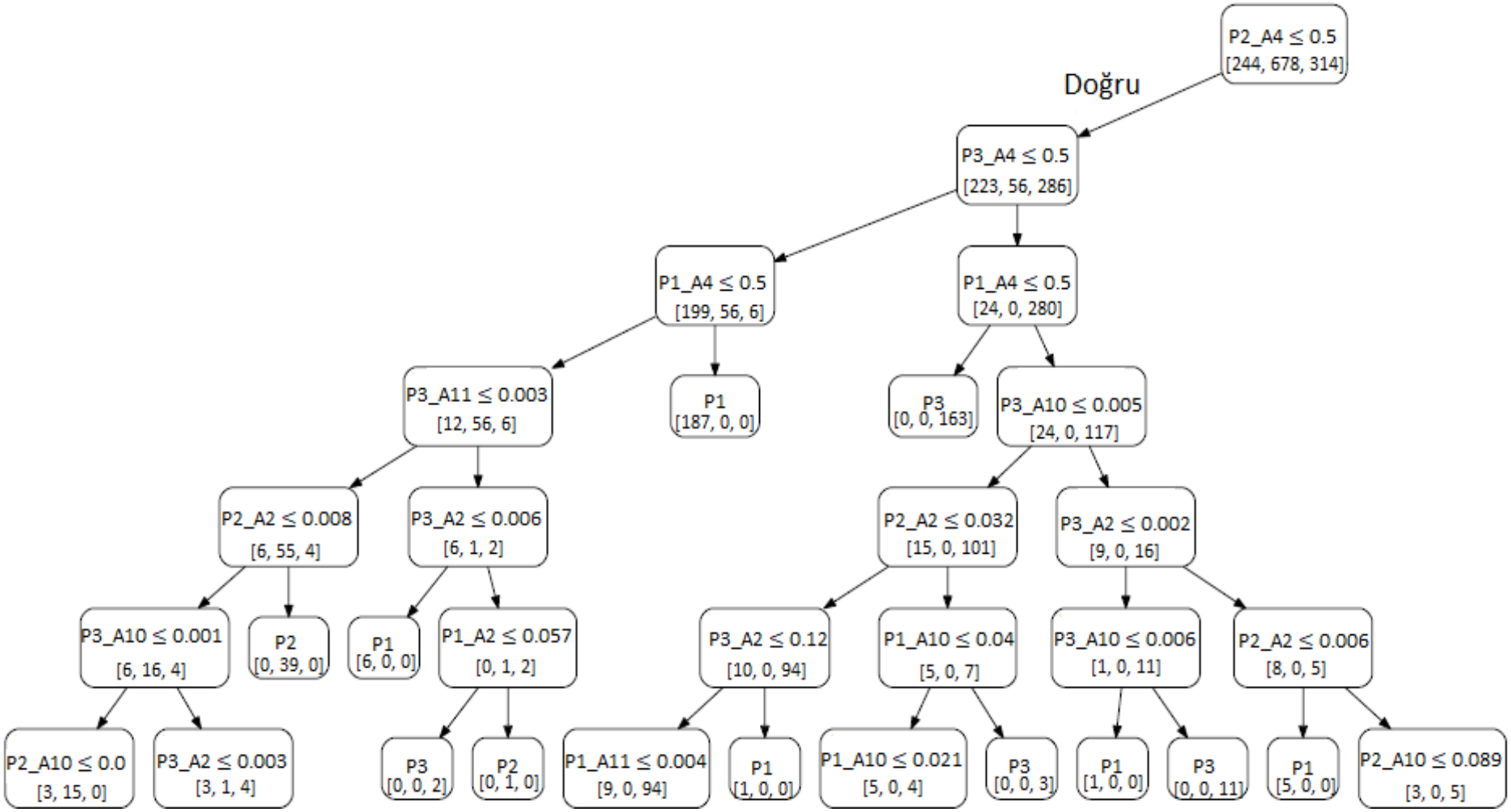
Tablo 4.48'de eğitim veri setinin C2, özelliklerin A2, A4, A10 ve A11 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir

Tablo 4.49: Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları

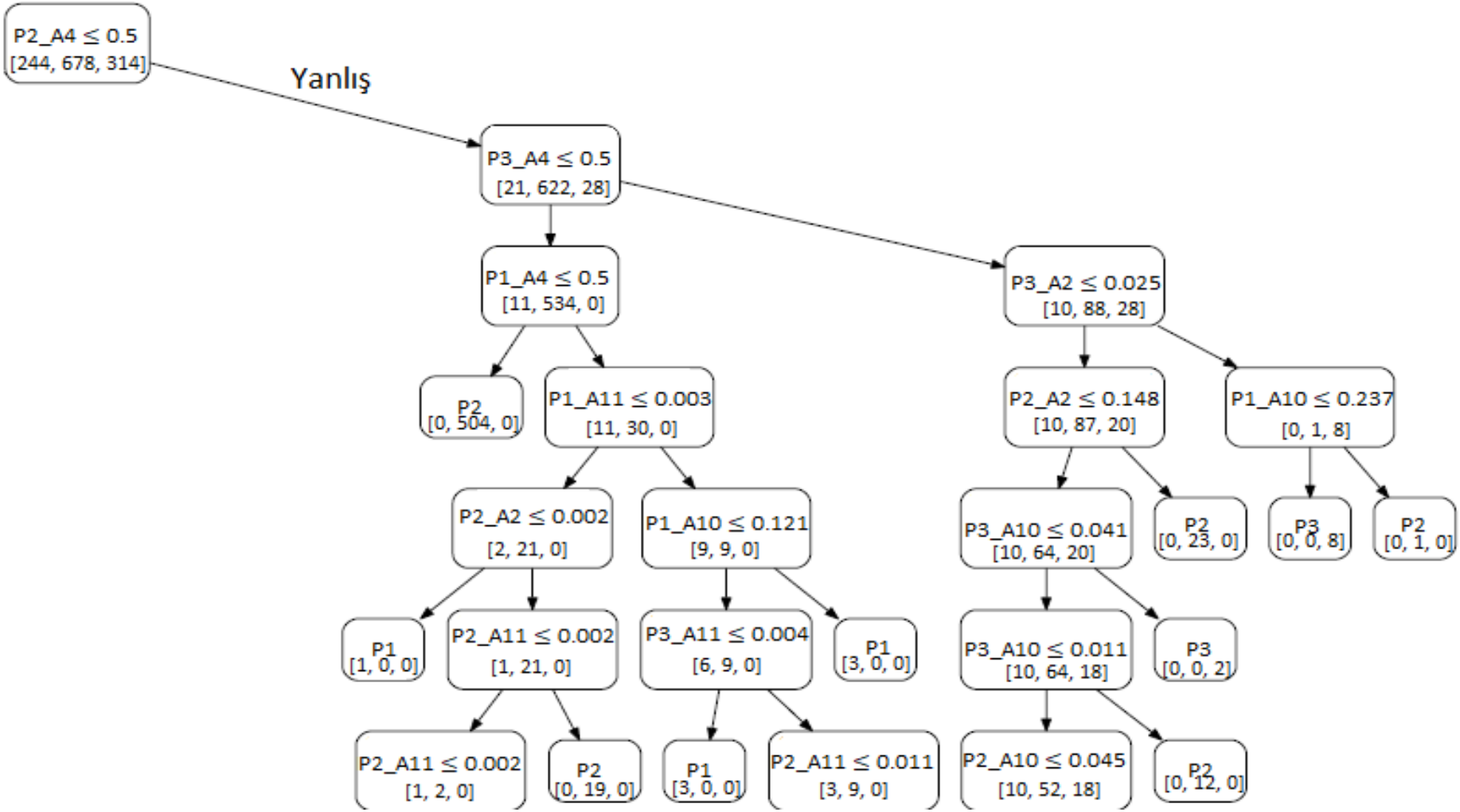
| **özellik** | A2, A4, A10, A11 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **70** | 6 | 17 | 93 |
| **P2** | 3 | **21** | 3 | 27 |
| **P3** | 13 | 5 | **23** | 41 |
| **toplam** | 86 | 32 | 43 | 161 |

Tablo 4.49'da ise eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 70 tanesi, 27 tane P2 sınıfına ait örneğin 21 tanesi ve 41 tane P3 sınıfına ait örnekten 23 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 70/93'ten 0.75, 21/27'den 0.78 ve 23/41'den 0.56 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 86 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 70 tanesi bu sınıfa aittir, yine benzer şekilde 32 örnek P2 sınıfıyla, 43 örnek P3 sınıfıyla etiketlenirken gerçekte 32'den 21 tanesi, 43'ün de 23 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 70/86'dan 0.81, P2 için 21/32'den 0.66 ve P3 için 23/43'den 0.53 olarak hesaplanmıştır.




Şekil 4.14: Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda CART algoritmasına göre oluşan karar ağacı




Şekil 4.14 (devam): Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.50: Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A5, A10 ve A11 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K1 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 ≤ 0.003) VE (P2_A2 ≤ 0.008) VE (P3_A10 ≤ 0.001) İSE | P2 |
| K2 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 ≤ 0.003) VE (P2_A2 ≤ 0.008) VE (P3_A10 > 0.001) İSE | P3 |
| K3 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 ≤ 0.003) VE (P2_A2 > 0.008) İSE | P2 |
| K4 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 > 0.003) VE (P3_A2 ≤ 0.006) İSE | P1 |
| K5 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 > 0.003) VE (P3_A2 > 0.006) VE (P1_A2 ≤ 0.057) İSE | P3 |
| K6 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A11 > 0.003) VE (P3_A2 > 0.006) VE (P1_A2 > 0.057) İSE | P2 |
| K7 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) İSE | P1 |
| K8 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 ≤ 0.5) İSE | P3 |
| K9 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 ≤ 0.005) VE (P2_A2 ≤ 0.032) VE (P3_A2 ≤ 0.12) İSE | P3 |
| K10 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 ≤ 0.005) VE (P2_A2 ≤ 0.032) VE (P3_A2 > 0.12) İSE | P1 |
| K11 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 ≤ 0.005) VE (P2_A2 > 0.032) VE (P1_A10 ≤ 0.04) İSE | P1 |
| K12 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 ≤ 0.005) VE (P2_A2 > 0.032) VE (P1_A10 > 0.04) İSE | P3 |
| K13 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 > 0.005) VE (P3_A2 ≤ 0.002) VE (P3_A10 ≤ 0.006) İSE | P1 |
| K14 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 > 0.005) VE (P3_A2 ≤ 0.002) VE (P3_A10 > 0.006) İSE | P3 |

Tablo 4.50 (devam): Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A5, A10 ve A11 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K15 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 > 0.005) VE (P3_A2 ≤ 0.002) VE (P2_A2 ≤ 0.006) İSE | P1 |

| | | |
|---|---|---|
| K16 | EĞER (P2_A4 ≤ 0.5) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A10 > 0.005) VE (P3_A2 ≤ 0.002) VE (P2_A2 > 0.006) İSE | P3 |
| K17 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) İSE | P2 |
| K18 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A11 ≤ 0.003) VE (P2_A2 ≤ 0.002) İSE | P1 |
| K19 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A11 ≤ 0.003) VE (P2_A2 > 0.002) İSE | P2 |
| K20 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A11 > 0.003) VE (P1_A10 ≤ 0.121) VE (P3_A11 ≤ 0.004) İSE | P1 |
| K21 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A11 > 0.003) VE (P1_A10 ≤ 0.121) VE (P3_A11 > 0.004) İSE | P2 |
| K22 | EĞER (P2_A4) > 0.5 VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A11 > 0.003) VE (P1_A10 > 0.121) İSE | P1 |
| K23 | EĞER (P2_A4 > 0.5) VE (P3_A4 > 0.5) VE (P3_A2 ≤ 0.025) VE (P2_A2 ≤ 0.148) VE (P3_A10 ≤ 0.041) İSE | P2 |
| K24 | EĞER (P2_A4 > 0.5) VE (P3_A4 > 0.5) VE (P3_A2 ≤ 0.025) VE (P2_A2 ≤ 0.148) VE (P3_A10 > 0.041) İSE | P3 |
| K25 | EĞER (P2_A4 > 0.5) VE (P3_A4 > 0.5) VE (P3_A2 ≤ 0.025) VE (P2_A2 > 0.148) İSE | P2 |
| K26 | EĞER (P2_A4 > 0.5) VE (P3_A4 > 0.5) VE (P3_A2 > 0.025) VE (P1_A10 ≤ 0.237) İSE | P3 |
| K27 | EĞER (P2_A4 > 0.5) VE (P3_A4 > 0.5) VE (P3_A2 > 0.025) VE (P1_A10 > 0.237) İSE | P2 |

Şekil 4.14'de CART algoritması tarafından eğitim veri seti olarak C1 ve C2, özellik olarak da A2, A4, A10 ve A11'in kullanılması durumunda elde edilen karar ağacı gösterilmektedir. Tablo 4.50'de ise bu karar ağacına bağlı olarak elde edilen 27 adet kuraldan oluşan kural tabanı yer almaktadır.

Tablo 4.51: Eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda her bir sınıfın başarım ölçütleri

| **özellik** | A4, A6, A7, A8, A10 |
|---|---|

| | ***kesinlik*** | ***duyarlılık*** | ***f1*** | **örnek sayısı** |
|---|---|---|---|---|
| **P1** | 0.93 | 0.58 | 0.72 | 93 |
| **P2** | 0.34 | 0.89 | 0.49 | 27 |
| **P3** | 0.72 | 0.56 | 0.63 | 41 |
| **ort./toplam** | 0.78 | 0.63 | 0.66 | 161 |

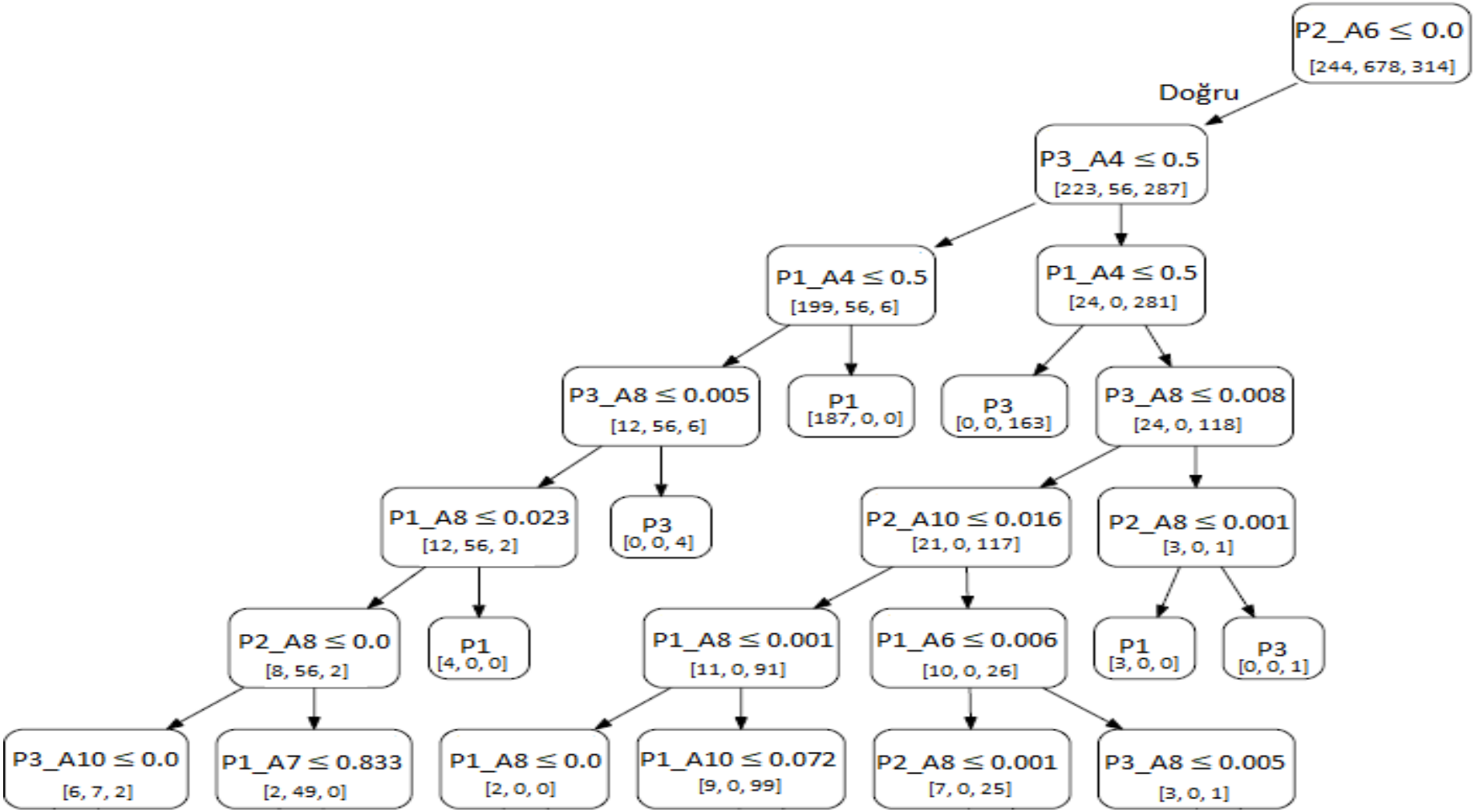
Tablo 4.51'de eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda her bir sınıf için başarım ölçütlerinin aldığı değerler gösterilmiştir.

Tablo 4.52: Eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A11 olması durumunda her bir sınıf için tahminleme sonuçları

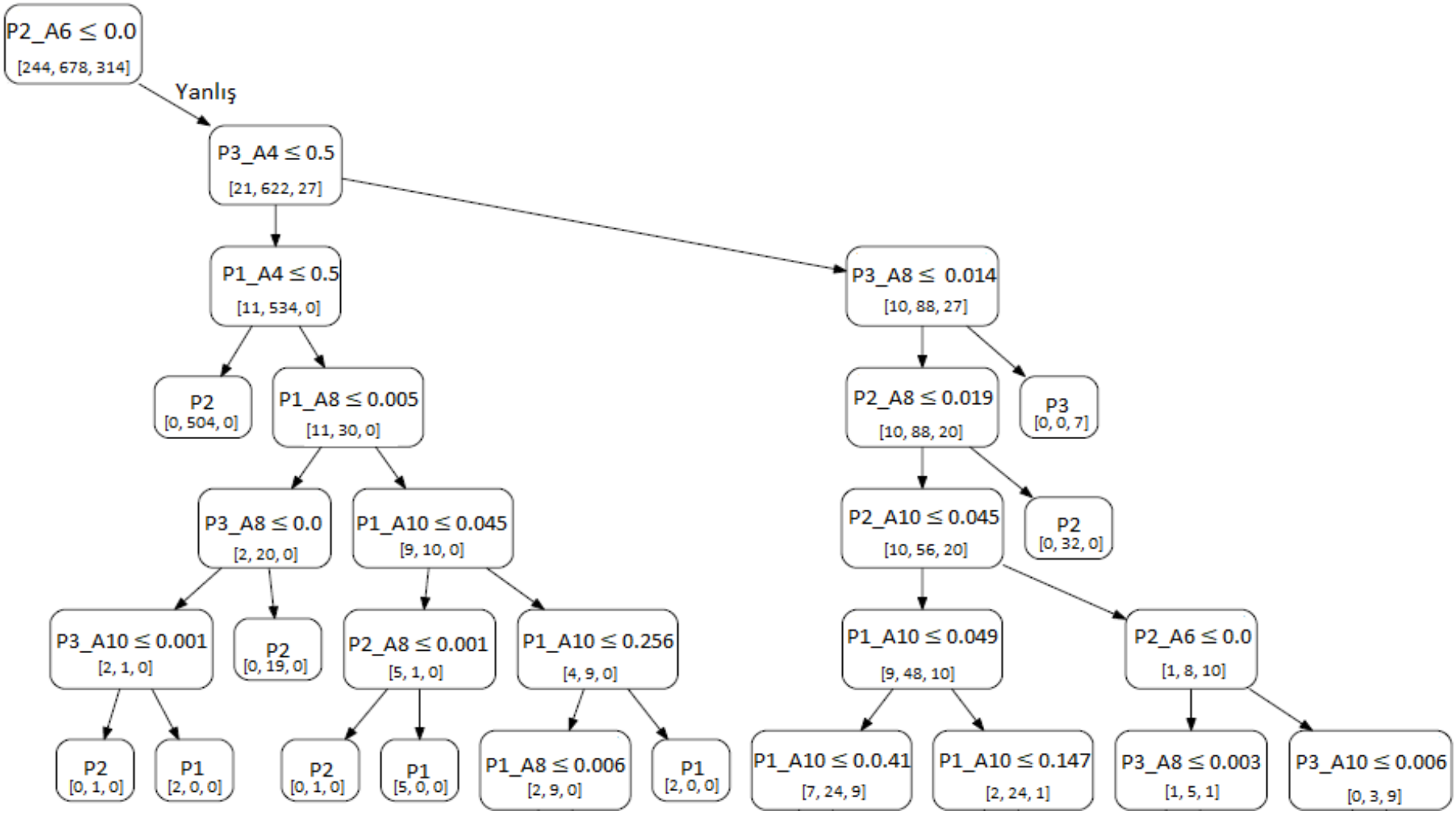
| **özellik** | A4, A6, A7, A8, A10 | | | |
|---|---|---|---|---|
| | **P1** | **P2** | **P3** | **toplam** |
| **P1** | **54** | 32 | 7 | 93 |
| **P2** | 1 | **24** | 2 | 27 |
| **P3** | 3 | 15 | **23** | 41 |
| **toplam** | 58 | 71 | 32 | 161 |

Tablo 4.52'de ise eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda her bir sınıf için tahminleme sonuçları gösterilmiştir. Test veri setimizde yer alan 93 tane P1 sınıfına ait örnekten 54 tanesi, 27 tane P2 sınıfına ait örneğin 24 tanesi ve 41 tane P3 sınıfına ait örnekten 23 tanesi doğru tahmin edilmiştir ve buna bağlı olarak da sırasıyla P1, P2 ve P3 sınıflarının *duyarlılık* değerleri 54/93'ten 0.58, 24/27'den 0.89 ve 23/41'den 0.56 olarak hesaplanmıştır.

Yapılan analizler sonucunda toplamda 58 örnek P1 sınıfıyla etiketlenmiş ve gerçekte bunun 54 tanesi bu sınıfa aittir, yine benzer şekilde 71 örnek P2 sınıfıyla, 32 örnek P3 sınıfıyla etiketlenirken gerçekte 71'den 24 tanesi, 32'nin de 23 tanesi bu sınıfa aittir. Bu durumda sınıflara ait *kesinlik* değerleri P1 için, 54/58'den 0.93, P2 için 24/71'den 0.34 ve P3 için 23/32'den 0.72 olarak hesaplanmıştır.




Şekil 4.15: Eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda CART algoritmasına göre oluşan karar ağacı




Şekil 4.15 (devam): Eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda CART algoritmasına göre oluşan karar ağacı

Tablo 4.53: Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A6, A7, A8 ve A10 olması durumunda elde edilen kural tabanı

| **no** | **kural** | **sonuç** |
|---|---|---|
| K1 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 ≤ 0.005) VE (P1_A8 ≤ 0.023) İSE | P2 |
| K2 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 ≤ 0.005) VE (P1_A8 > 0.023) İSE | P1 |
| K3 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) VE (P3_A8 > 0.005) İSE | P3 |
| K4 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) İSE | P1 |
| K5 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 ≤ 0.5) İSE | P3 |
| K6 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P2_A10 ≤ 0.016) VE (P1_A8 ≤ 0.001) İSE | P1 |
| K7 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P2_A10 ≤ 0.016) VE (P1_A8 > 0.001) İSE | P3 |
| K8 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P2_A10 > 0.006) VE (P1_A6 ≤ 0.006) İSE | P3 |
| K9 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 ≤ 0.008) VE (P2_A10 > 0.006) VE (P1_A6 > 0.006) İSE | P1 |
| K10 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 > 0.008) VE (P2_A8 ≤ 0.001) İSE | P1 |
| K11 | EĞER (P2_A6 ≤ 0.0) VE (P3_A4 > 0.5) VE (P1_A4 > 0.5) VE (P3_A8 > 0.008) VE (P2_A8 > 0.001) İSE | P3 |
| K12 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 ≤ 0.5) İSE | P2 |
| K13 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 ≤ 0.005) VE (P3_A8 ≤ 0.0) VE (P3_A10 ≤ 0.001) | P2 |
| K14 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 ≤ 0.005) VE (P3_A8 ≤ 0.0) VE (P3_A10 > 0.001) | P1 |
| K15 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 ≤ 0.005) VE (P3_A8 > 0.0) | P2 |
| K16 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 > 0.005) VE (P1_A10 ≤ 0.045) VE (P2_A8 ≤ 0.001) | P2 |

Tablo 4.53 (devam): Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A6, A7, A8 ve A10 olması durumunda elde edilen kural tabanı

| no | kural | sonuç |
|---|---|---|
| K17 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 > 0.005) VE (P1_A10 ≤ 0.045) VE (P2_A8 > 0.001) | P1 |
| K18 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 > 0.005) VE (P1_A10 > 0.045) VE (P1_A10 ≤ 0.256) | P2 |
| K19 | EĞER (P2_A6 > 0.0) VE (P3_A4 ≤ 0.5) VE (P1_A4 > 0.5) VE (P1_A8 > 0.005) VE (P1_A10 > 0.045) VE (P1_A10 > 0.256) | P1 |
| K20 | EĞER (P2_A6 > 0.0) VE (P3_A4 > 0.5) VE (P3_A8 ≤ 0.014) VE (P2_A8 ≤ 0.019) VE (P2_A10 ≤ 0.045) İSE | P2 |
| K21 | EĞER (P2_A6 > 0.0) VE (P3_A4 > 0.5) VE (P3_A8 ≤ 0.014) VE (P2_A8 ≤ 0.019) VE (P2_A10 > 0.045) VE (P2_A6 ≤ 0.0) İSE | P2 |
| K22 | EĞER (P2_A6 > 0.0) VE (P3_A4 > 0.5) VE (P3_A8 ≤ 0.014) VE (P2_A8 ≤ 0.019) VE (P2_A10 > 0.045) VE (P2_A6 > 0.0) İSE | P3 |
| K23 | EĞER (P2_A6 > 0.0) VE (P3_A4 > 0.5) VE (P3_A8 ≤ 0.014) VE (P2_A8 > 0.019) İSE | P2 |
| K24 | EĞER (P2_A6 > 0.0) VE (P3_A4 > 0.5) VE (P3_A8 > 0.014) İSE | P3 |

Şekil 4.15'te CART algoritması tarafından eğitim veri seti olarak C1 ve C2, özellik olarak da A4, A6, A7, A8 ve A10'un kullanılması durumunda elde edilen karar ağacı gösterilmektedir. Tablo 4.53'te ise bu karar ağacına bağlı olarak elde edilen 24 adet kuraldan oluşan kural tabanı yer almaktadır.

Yukarıda da bahsedildiği üzere toplamda 245730 farklı kombinasyonun her birisi için 10 kez CART algoritması çalıştırılarak karar ağacı oluşturulmuş ve çıktı olarak 3 farklı eğitim veri setinde, tüm sınıfların ortalaması bazında *doğruluk*, *kesinlik* ve *f1*, her bir sınıfın sınıf bazında başarımını ölçmek için *kesinlik*, *duyarlılık* ve *f1* başarım ölçütleri kullanılarak en başarılı sonuçları sağlayan özellikler belirlenmiştir. Elde edilen bu en başarılı sonuçlar içinde en sık kullanılan özelliklerden A4 6 kez, A10 6 kez, A2 5 kez, A7 5 kez, A11 4 kez, A8 4 kez, A6 3 kez ve A1 3 kez kullanılmıştır.

*doğruluk* ölçütü açısından en başarılı sonuç, eğitim veri setinin C1, özelliklerin A2, A5 ve A13 kabul edildiği durumda elde edilmiştir. Bu durumda elde edilen *doğruluk* ve *duyarlılık* değeri 0.7578, *kesinlik* değeri 0.7838 ve *f1* değeri 0.7591 dir.

*kesinlik* ölçütü açısından en başarılı sonuç, eğitim veri setinin C1, özelliklerin A1 ve A7 olduğu durumda elde edilmiştir. Bu durumda elde edilen *doğruluk* değeri 0.5093, *kesinlik* değeri 0.8750 ve *f1* değeri 0.5357'dir.

*f1* ölçütü açısından en yüksek değere sahip sonuçlar, *doğruluk* ölçütünün en başarılı sonuçlarının elde edildiği koşullarda elde edilmiştir.

Tek tek sınıf bazında incelediğimiz zamansa;

P1 sınıfı için *duyarlılık* açısından en iyi sonuç olan 0.85 iki farklı durumda elde edilmiştir. Bunların her ikisi de eğitim veri setinin C1 olduğu durumda elde edilirken, seçilen özellikler birinde A2, A5 ve A13 iken diğerinde A2, A6, A7 ve A13'dir. Bu koşullarda elde edilen elde edilen *kesinlik* değerleri sırasıyla 0.88 ve 0.89 iken, *f1* değerleri 0.86 ve 0.87'dir. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 1.00 eğitim veri setinin C1, özelliklerin A1 ve A7 olduğu durumda elde edilmiştir. Bu koşullar altında elde edilen diğer ölçüt değerleri *duyarlılık* için 0.52, *f1* için 0.68 tir. Daha öncede belirtildiği gibi bu uygulamada başarılı sonuçların belirlenmesinde daha doğru sonuçlar elde edebilmek için *duyarlılık* ya da *kesinlik* yerine *f1* ölçütüne bakılmaktadır Bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda *duyarlılık* 0.85, *kesinlik* 0.89 ve *f1* 0.87 olarak hesaplanmıştır.
- Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda *duyarlılık* 0.75, *kesinlik* 0.88 ve *f1* 0.87 olarak hesaplanmıştır.
- Eğitim veri setinin C2, özellikleri A1, A4, A7, A8, A10 ve A11 olması durumunda *duyarlılık* 0.82, *kesinlik* 0.82 ve *f1* 0.82 olarak hesaplanmıştır.
- Eğitim veri setinin C1 ve C2, özellikleri A2, A4, A10 ve A11 olması durumunda *duyarlılık* 0.81, *kesinlik* 0.75 ve *f1* 0.78 olarak hesaplanmıştır.
- Eğitim veri setinin C1 ve C2, özellikleri A4, A6, A7, A8 ve A10 olması durumunda *duyarlılık* 0.58, *kesinlik* 0.93 ve *f1* 0.72 olarak hesaplanmıştır.

P2 sınıfı için *duyarlılık* açısından en iyi sonuç olan 1.00 değeri eğitim veri setinin C1, özelliklerin A2, A5 ve A13 seçilmesi durumunda elde edilmiştir. Bu koşullar altında elde edilen *kesinlik* değeri 0.25 iken, *f1* değeri 0.41'dir. Yani bu şartlar altında elde edilen *duyarlılık* değeri en yüksek değer olan 1 olmasına rağmen *kesinlik* değeri oldukça düşüktür. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 0.82 değeri eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda elde edilmiştir. Bu durumda elde edilen diğer ölçüt değerleri *duyarlılık* için 0.52, *f1* için 0.64'tür. Görüldüğü gibi bu koşullarda elde edilen *kesinlik* değeri 0.82 gibi yüksek bir değer olmasına rağmen *duyarlılık* değeri oldukça düşüktür, bu yüzden en başarılı sonucu belirlemek için *f1* değerine bakılması daha uygundur. Bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda *duyarlılık* 0.78, *kesinlik* 0.66 ve *f1* 0.71 olarak hesaplanmıştır.
- Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda *duyarlılık* 0.52, *kesinlik* 0.82 ve *f1* 0.64 olarak hesaplanmıştır.
- Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda *duyarlılık* 0.81, *kesinlik* 0.51 ve *f1* 0.63 olarak hesaplanmıştır.
- Eğitim veri setinin C2, özelliklerin A2, A6, A7 ve A13 olması durumunda *duyarlılık* 0.81, *kesinlik* 0.51 ve *f1* 0.63 olarak hesaplanmıştır.
- Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda *duyarlılık* 0.67, *kesinlik* 0.58 ve *f1* 0.62 olarak hesaplanmıştır.

P3 sınıfı için *duyarlılık* açısından en iyi sonuç olan 0.98 değeri eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 seçilmesi durumunda elde edilmiştir. Bu koşullar altında elde edilen *kesinlik* değeri 0.38 iken, *f1* değeri 0.55'tir. Yani bu şartlar altında elde edilen *duyarlılık* değeri oldukça yüksek bir değer olmasına rağmen *kesinlik* değeri oldukça düşüktür. Yine benzer şekilde *kesinlik* açısından en iyi sonuç olan 1.00 değeri eğitim veri setinin C1, özelliklerin A1 ve A7 olması durumunda elde edilmiştir. Bu koşullar altında elde edilen diğer ölçüt değerleri *duyarlılık* için 0.17,

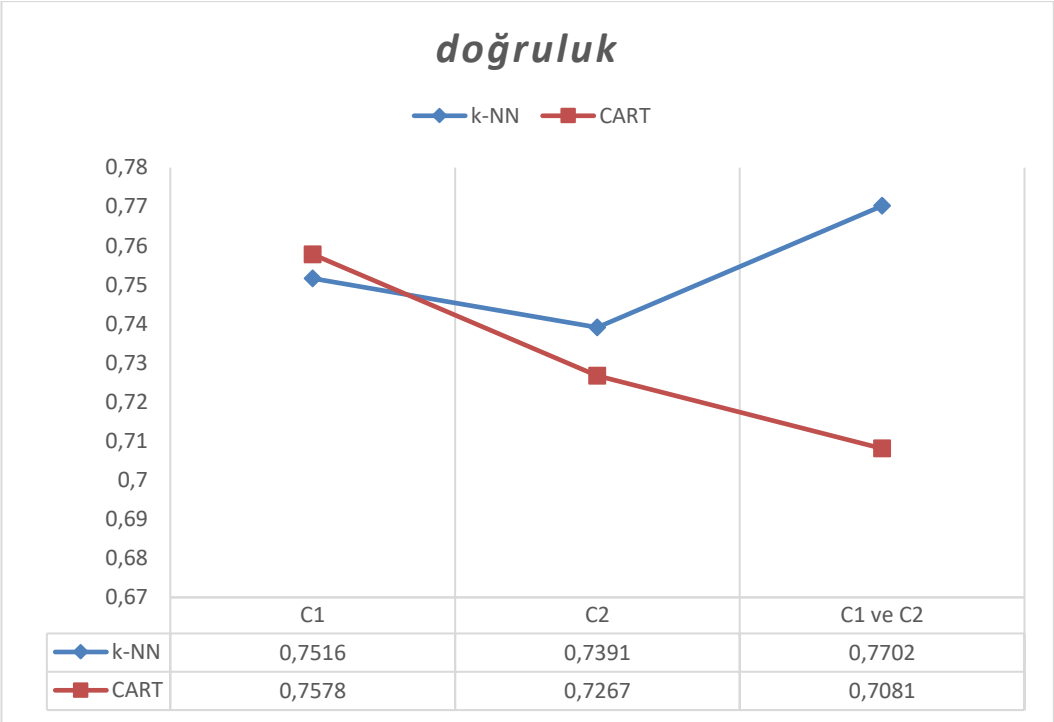
*f1* için 0.29'dur. Görüldüğü gibi bu koşullarda elde edilen *kesinlik* değeri 1.00 olmasına rağmen *duyarlılık* değeri oldukça düşüktür, bu yüzden en başarılı sonucun belirlenebilmesi için *f1* değerine bakılmıştır ve bu durumda *f1* değerine göre sıralanan en yüksek 5 sonuç şunlardır:

- Eğitim veri setinin C1 ve C2, özelliklerin A4, A6, A7, A8 ve A10 olması durumunda *duyarlılık* 0.56, *kesinlik* 0.72 ve *f1* 0.63 olarak hesaplanmıştır.
- Eğitim veri setinin C1, özelliklerin A2, A5 ve A13 olması durumunda *duyarlılık* 0.51, *kesinlik* 0.75 ve *f1* 0.61 olarak hesaplanmıştır.
- Eğitim veri setinin C1, özelliklerin A2, A6, A7 ve A13 olması durumunda *duyarlılık* 0.51, *kesinlik* 0.72 ve *f1* 0.60 olarak hesaplanmıştır.
- Eğitim veri setinin C2, özelliklerin A1, A4, A7, A8, A10 ve A11 olması durumunda *duyarlılık* 0.56, *kesinlik* 0.62 ve *f1* 0.59 olarak hesaplanmıştır.

Eğitim veri setinin C2, özelliklerin A2, A4, A5, A6, A8, A9 ve A10 olması durumunda *duyarlılık* 0.98, *kesinlik* 0.38 ve *f1* 0.55 olarak ve yine aynı şekilde eğitim veri setinin C1 ve C2, özelliklerin A2, A4, A10 ve A11 olması durumunda *duyarlılık* 0.56, *kesinlik* 0.53 ve *f1* 0.55 olarak hesaplanmıştır.

### 4.2.3 Sınıflandırma Sonuçları

Çalışma kapsamında önerilen özelliklerin sınıflandırma çalışmalarındaki başarımını test etmek için *k*-NN ve CART algoritmaları kullanılarak Bölüm 4.2.1 ve Bölüm 4.2.2'de yer alan uygulamalar gerçekleştirilmiştir. Şekil 4.16'da Uygulama 1 ve Uygulama 2'de kullanılan *k*-NN algoritması ve bir karar ağacı yöntemi olan CART algoritmasının ortalama *doğruluk* değerine göre karşılaştırıldığı grafik görülmektedir.

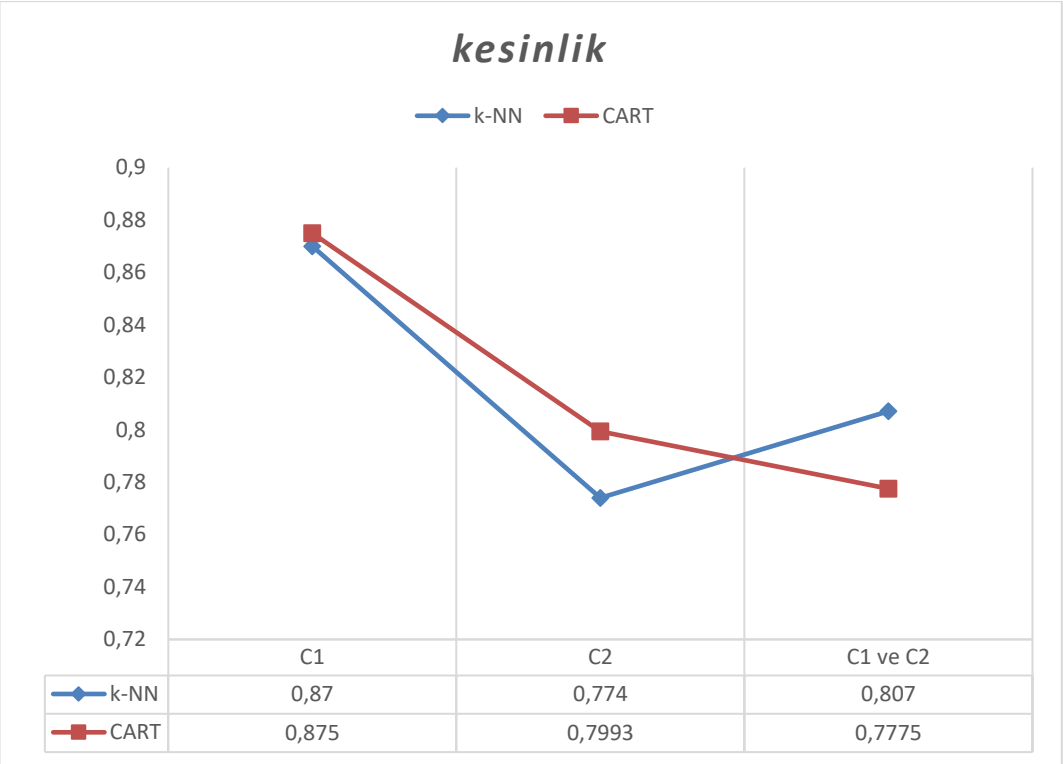



Şekil 4.16: Eğitim veri seti olarak yalnız C1, yalnız C2, C1 ve C2 seçilmesi durumunda *k*-NN ve karar ağacı yöntemleriyle elde edilen sınıflandırmaların karşılaştırmalı ortalama *doğruluk* değerleri

Şekil 4.16'da da görüldüğü gibi eğitim veri setinin C1 olarak seçilmesi durumunda *k*-NN algoritmasıyla elde edilen ortalama *doğruluk* değeri 0.7573 iken karar ağacı ile elde edilen ortalama *doğruluk* değeri 0.7578'dir. C2'nin eğitim veri seti olarak seçilmesi durumunda *k*-NN ve karar ağacıyla elde edilen ortalama *doğruluk* değerleri sırasıyla 0.7391 ve 0.7267 iken C1 ve C2'nin birlikte eğitim seti olarak seçilmesi durumunda 0.7702 ve 0.7081 olarak hesaplanmıştır.

Yöntem açısında bakıldığı zaman eğitim veri setinin C1 olarak seçilmesi durumunda ortalama *doğruluk* değerine göre karar ağacı ile *k*-NN'den daha başarılı sonuçlar elde edilmesine rağmen arada çok büyük bir fark bulunmamaktadır. Diğer eğitim veri seti kombinasyonlarında da *k*-NN, karar ağacına göre daha başarılı sınıflandırmalar gerçekleştirmiştir.

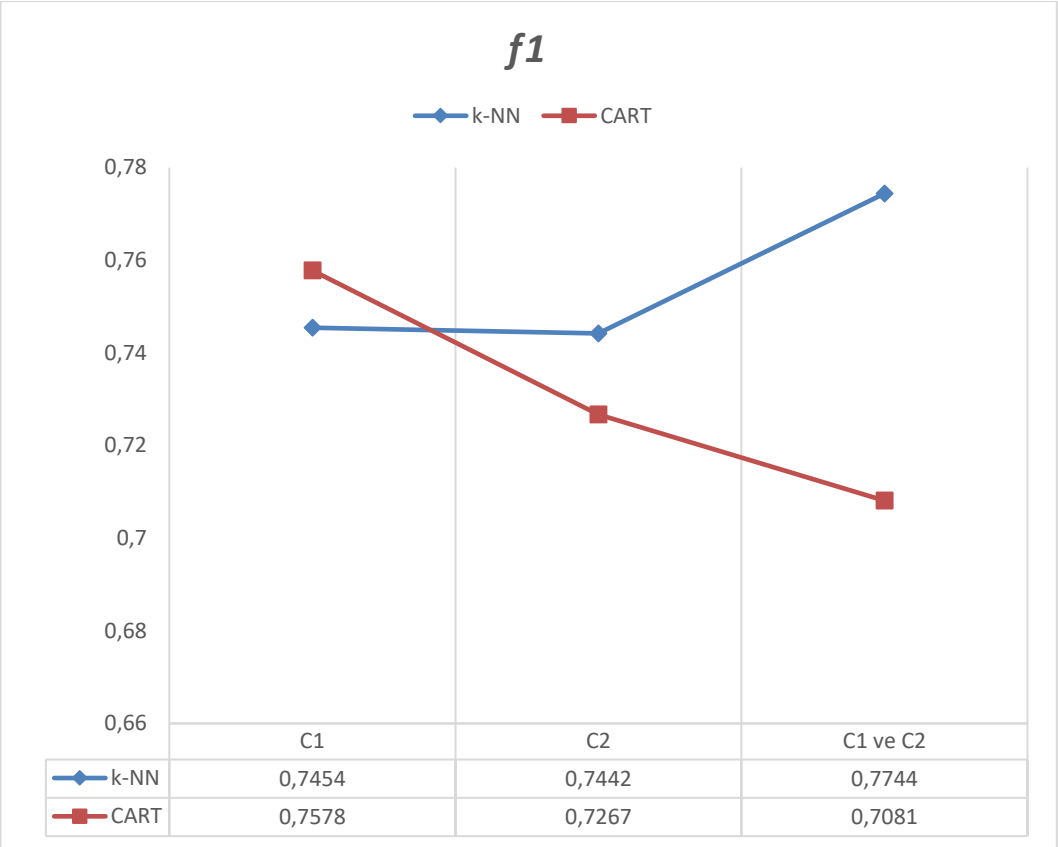
Eğitim veri seti açısından bakıldığı zaman *k*-NN için en yüksek ortalama *doğruluk* değeri C1 ve C2'nin birlikte kullanılması durumunda elde edilirken ki bu aynı zamanda genel anlamda da en başarılı sınıflandırmadır, karar ağacı için en yüksek ortalama *doğruluk* değeri de C1'in kullanılması durumunda elde edilmiştir.




Şekil 4.17: Eğitim veri seti olarak yalnız C1, yalnız C2, C1 ve C2 seçilmesi durumunda *k*-NN ve karar ağacı yöntemleriyle elde edilen sınıflandırmaların karşılaştırmalı ortalama *kesinlik* değerleri

*k*-NN ve CART algoritmalarının farklı eğitim veri setlerine göre gösterdikleri performasın ortalama *kesinlik* değeri açısından karşılaştırılması Şekil 4.17'de görülmektedir. Her iki algoritma içinde en yüksek ortalama *kesinlik* değerleri C1'in eğitim veri seti olarak kullanılması durumunda elde edilmiştir. Bu durumda *k*-NN ile elde edilen ortalama *kesinlik* değeri 0.87 iken karar ağacı ile elde edilen değer 0.875'tir, aralarında oldukça küçük bir fark olmamasına rağmen karar ağacıyla daha başarılı sonuçlar elde edilmiştir.

Yöntem açısından bakıldığı zaman C1 ve C2'nin birlikte eğitim veri seti olarak seçildiği durum dışında karar ağacı ile *k*-NN'e göre daha yüksek ortalama *kesinlik* değeri elde edilmiştir.




Şekil 4.18: Eğitim veri seti olarak yalnız C1, yalnız C2, C1 ve C2 seçilmesi durumunda *k*-NN ve karar ağacı yöntemleriyle elde edilen sınıflandırmaların karşılaştırmalı ortalama *f1* değerleri

*doğruluk* ve *kesinlik* değerlerinin her birisi açısından en başarılı sınıflandırmalar farklı koşullar altında elde edilmiştir ve bu yüzden bu değerlerin karşılaştırması oldukça zor olmaktadır. *duyarlılık* ve *kesinlik* değerlerinin ağırlıklı ortalaması olan *f1* değeri, *duyarlılık* (ortalamaları alındığı zaman *doğruluk* ile aynı değere sahiptir) ve *kesinlik* değerinin birlikte değerlendirilmesine imkan vermektedir. Buna bağlı olarak her iki yöntem ile farklı eğitim veri setleri kullanılarak elde edilen ortalama *f1* değerleri Şekil 4.18'de gösterilmiştir.

*k*-NN için en yüksek ortalama *f1* değeri C1 ve C2'nin eğitim veri seti olarak birlikte kullanıldığı durumda elde edilirken karar ağacı için C1'in kullanıldığı durumda elde edilmiştir.

Yöntem açısında bakıldığı zaman *k*-NN ile, eğitim veri setinin C1 seçilmesi durum dışındaki diğer tüm veri setlerinde daha yüksek ortalama *f1* değeri elde edilmiştir.

Sonuç olarak sınıflandırma için *k*-NN algoritmasının kullanılması durumunda en yüksek başarım değerleri, ortalama *doğruluk* ve ortalama *f1* için eğitim veri seti olarak C1 ile C2'nin birlikte seçilmesi, ortalama *kesinlik* için C1'in seçilmesi önerilmektedir. Karar ağacı yöntemi kullanılması durumunda en yüksek başarım değerlerinin elde edilebilmesi için de eğitim veri seti olarak C1'in seçilmesi önerilmektedir.

Eğitim veri seti açısından bakıldığı zaman, C1' in seçilmesi durumunda karar ağacının, C2'nin seçilmesi durumunda ortalama *doğruluk* ve ortalama *f1* için *k*-NN'in, ortalama *kesinlik* için karar ağacının kullanılması önerilmektedir. Ayrıca C1 ve C2'nin birlikte seçilmesi durumunda *k*-NN ile daha başarılı sonuçlar elde edileceği ön görülmektedir.

## 4.3 Kümeleme Uygulamaları

Bu bölümde Tablo 4.1'de yer alan C1, C2 ve C3 veri setleri üzerinde Tablo 4.2'de yar alan özellikler kullanılarak *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemlerine göre kümeleme uygulamaları yapılmıştır.

### 4.3.1 Uygulama 3: *k*-Ortalamalar Yöntemiyle Kullanıcıların Kümelenmesi

Bu uygulamada amaç, Uygulama 1 ve Uygulama 2'de olduğu gibi kullanıcıların siyasi görüşlerini tahmin ederek onları sınıflandırmak değildir. Buradaki amaç *k*-Ortalamalar algoritmasını kullanarak kullanıcılar arasındaki siyasi görüşe bağlı gruplanmaların belirlenmesidir. Yani burada önemli olan bir grupta yer alan kişilerin aynı siyasi görüşe sahip olmasıdır ancak hangi siyasi görüşe sahip oldukları önemli değildir. Örneğin bu çalışmada 3 tane siyasi parti bulunmaktadır ve buna bağlı olarak veri setlerindeki örnekler *k*-Ortalamalar algoritmasıyla 3 kümeye ayrılmaya çalışıldığında, en iyi durum her bir kümede yalnızca bir partiye ait kullanıcıların bulunmasıdır. Ancak burada dikkat edilmesi gereken nokta hangi kümenin hangi siyasi partiye ait kullanıcıları barındırdığının bilinmemesidir.

Yapılan uygulamada Tablo 4.1'de yer alan 8 adet koleksiyon arasından C1, C2 ve C3'ten ve Tablo 4.2'de yer alan 13 adet özelliğin tamamından yararlanılmış ve şu 3 temel sorunun cevabı aranmıştır:

1. En başarılı kümeleme sonucu nedir?
2. En başarılı kümeleme sonucunu sağlayan özellik hangisi ya da hangileridir?
3. En başarılı kümeleme sonucu hangi veri setinde elde edilmiştir?

Buradaki 2. sorunun cevabını bulabilmek için Tablo 4.2'deki 13 özelliğin tüm kombinasyonlarının denenmesi, bu yüzden de $\binom{13}{1}, \binom{13}{2}, \dots, \binom{13}{13}$ şeklinde tüm kombinasyonlar için toplamda 8191 durumun ayrı ayrı test edilmesi gerekmektedir. Benzer şekilde 3. sorunun cevabının bulunabilmesi için Tablo 4.1'de yer alan C1, C2 ve C3 koleksiyonlarının, $\binom{3}{1}, \binom{3}{2}, \binom{3}{3}$ şeklindeki toplamda 7 farklı kombinasyonunun test edilmesi gerekmektedir.

Kısaca 7 farklı veri seti ve 8191 farklı özellik seçimi yapılabilmektedir ve bu faktörlerin hepsi birbirini etkilemektedir, dolayısıyla toplamda $7 * 8191$'den 57337 farklı kombinasyon bulunmaktadır.

Bu uygulamada yöntemin başarımını ölçebilmek için veri setlerinde yer alan 3 siyasi parti verisinin her birisinin bir küme olduğu düşünülmüştür ve kümeleme sonucunda elde edilen kümeler, bunlarla karşılaştırılarak başarım ölçütleri hesaplanmıştır.

1 nolu soruda da yer alan algoritmanın başarımını ölçmek için Bölüm 3.3'te açıklanan *doğruluk, kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü kullanılmıştır. 57337 farklı kombinasyon için bu 3 ölçüt hesaplanmış ve her bir ölçüt için en yüksek değeri sağlayan, eğitim veri seti veya setleri ve özellik veya özellikler bulunmaya çalışılmıştır. Ancak bu uygulamada sonuçların daha da iyileştirilmesi için bu 57337 kombinasyon 10 kez çalıştırarak, sonunda en başarılı sonucu elde eden değerler çıktı olarak kabul edilmiştir.

Uygulama MATLAB üzerinde kmeans fonksiyonu kullanılarak gerçekleştirilmiştir. 13 farklı özelliğin tüm kombinasyonları olan 8191 farklı durumun

ve C1, C2 ve C3 veri setlerinin tüm kombinasyonları olan 7 farklı durumun oluşturulması için nchoosek fonksiyonundan yararlanılmıştır. Bu fonksiyon bir diziyi ve bir integer değeri parametre olarak almakta ve bu integer sayıya göre dizinin kombinasyonlarını döndürmektedir. Ayrıca tüm bu kombinasyonların 10 kez tekrarlanarak toplam 573370 kombinasyonun test edilmesi sırasında işlemlerin daha kısa sürede gerçekleştirilmesi için paralelleştirme gerçekleştirilmiş ve bunun içinde parfor döngüsü kullanılmıştır.

Tablo 4.54: Veri setinin C1 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A3, A4, A8, A9, A12 | **0.9925** | 0.9926 | 0.9925 |
| ***kesinlik*** | A2, A3, A4, A8, A9, A12 | 0.9925 | **0.9926** | 0.9925 |
| ***f1*** | A2, A3, A4, A8, A9, A12 | 0.9925 | 0.9926 | **0.9925** |
| **süre** | 7873.6988 saniye | | | |

Tablo 4.54'te *k*-Ortalamalar algoritmasının C1 veri seti üzerinde çalıştırılması sonucunda elde edilen en başarılı sonuçlar gösterilmektedir. Tabloda da görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.9925, ortalama *kesinlik* değeri 0.9926 ve *f1* değeri 0.9925 olarak hesaplanmıştır. Bu değerlerin tamamı özellik olarak A2, A3, A4, A8, A9 ve A12'nin seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi için tüm özellik kombinasyonları 10 iterasyon boyunca denenerek en iyi sonuçlar tespit edilmiştir. Bu işlem sırasında her iterasyon ortalama 7873.6988 saniye o da yaklaşık 2 saat 11 dakika sürmüştür.

Tablo 4.55'te oluşan kümelerin her birisi, kümenin içerisinde en fazla bulunan sınıfa ait örnek sayıları gösterilmiştir. Bu tabloya göre Küme1, Küme2 ve Küme3 için şunları söyleyebiliriz:

- Küme 1 içerisinde toplam 72 tane örnek bulunmakta ve bunun 70 tanesini P3 sınıfına ait örnekler oluşturmaktadır. Bu, kümede yer alan örneklerden sadece 2 tanesinin farklı sınıfa ya da sınıflara ait olduğunu göstermektedir. Yine aynı şekilde bu veri seti içerisinde bulunan P3 sınıfına ait 71 örnekten 70 tanesi Küme 1 içerisinde yer alırken yalnız 1 tanesi farklı bir kümede yer almıştır.

- Küme 2 içerisinde toplam 414 tane örnek bulunurken bu örneklerin 413 tanesinin P2 sınıfına, 1 tanesinin farklı bir sınıfa ait olduğu görülmektedir. Ayrıca P2 sınıfının bu veri setinde bulunan toplam 416 örneğinin 413 tanesi Küme 2 içerisinde yer alırken 3 tanesi farklı küme ya da kümelerde yer almıştır.
- Küme 3 içerisinde toplam 49 tane örnek bulunmakta ve bunun 48 tanesini P1 sınıfına ait örnekler oluşturmaktadır. Bu, kümede yer alan örneklerden sadece 1 tanesinin farklı sınıfa ait olduğunu göstermektedir. Bununla beraber bu veri seti içerisinde bulunan P1 sınıfına ait 48 örneğin tamamı Küme 3 içerisinde yer almaktadır.

Tablo 4.55: Veri setinin C1, özelliklerin A2, A3, A4, A8, A9 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 | | |
|---|---|---|---|---|
| **özellik** | | A2, A3, A4, A8, A9, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 70 | 72 | 71 |
| Küme 2 | P2 | 413 | 414 | 416 |
| Küme 3 | P1 | 48 | 49 | 48 |
| **toplam** | | 531 | 535 | 535 |

Tablo 4.55'ten de anlaşıldığı gibi C1 veri seti içerisinde yer alan toplam 535 örnekten 531 tanesi doğru kümelenirken yalnız 4 tanesi yanlış kümeler içerisinde yer almıştır. Yine benzer şekilde Tablo 4.54'te bulunan başarım ölçütlerine bakılarakta kümelemenin ne kadar başarılı olduğu görülebilir.

Tablo 4.56: Veri setinin C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A4, A7 | **0.9073** | 0.9120 | 0.9058 |
| ***kesinlik*** | A2, A4, A7 | 0.9073 | **0.9120** | 0.9058 |
| ***f1*** | A2, A4, A7 | 0.9073 | 0.9120 | **0.9058** |
| **süre** | 11189.4778 saniye | | | |

Tablo 4.56, C2 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.9073, ortalama *kesinlik* değeri 0.9058 ve ortalama *f1* değeri 0.9120 olarak hesaplanmıştır. Bu değerlerin tamamı özellik olarak A2, A4 ve A7'nin seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 11189.4778 saniye o da yaklaşık 3 saat 6 dakika sürmüştür.

Tablo 4.57: Veri setinin C2, özelliklerin A2, A4 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 | | |
|---|---|---|---|---|
| **özellik** | | A2, A4, A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 260 | 297 | 262 |
| Küme 2 | P1 | 155 | 160 | 196 |
| Küme 3 | P3 | 221 | 244 | 243 |
| **toplam** | | 636 | 701 | 701 |

Tablo 4.57'den de anlaşıldığı gibi C2 veri seti içerisinde yer alan toplam 701 örnekten 636 tanesi doğru kümelenirken 65 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.58: Veri setinin C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A7 | **0.6584** | 0.7109 | 0.6596 |
| ***kesinlik*** | A2, A5, A6, A8, A11, A13 | 0.4534 | **0.8217** | 0.4642 |
| ***f1*** | A7 | 0.6584 | 0.7109 | **0.6596** |
| **süre** | 5538.5731 saniye | | | |

Tablo 4.58, C3 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* ve ortalama *f1* değeri özellik olarak A7'nin seçilmesi durumunda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* ve ortalama *duyarlılık* değeri 0.6584, ortalama *f1* değeri 0.6596'dır. Benzer şekilde en yüksek ortalama *kesinlik* değeri özellik olarak A2, A5,

A6, A8, A11 ve A13'ün seçilmesi durumunda 0.8217 olarak elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 5538.5731 saniye o da yaklaşık 1 saat 32 dakika sürmüştür.

Tablo 4.59: Veri setinin C3, özelliklerin A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C3 | | |
|---|---|---|---|---|
| **özellik** | | A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 70 | 79 | 93 |
| Küme 2 | P2 | 6 | 11 | 27 |
| Küme 3 | P3 | 30 | 71 | 41 |
| **toplam** | | 106 | 161 | 161 |

Tablo 4.59'dan da anlaşıldığı gibi özellik olarak A7'nin seçilmesi durumunda C3 veri seti içerisinde yer alan toplam 161 örnekten 106 tanesi doğru kümelenirken 55 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.60: Veri setinin C3, özelliklerin A2, A5, A6, A8, A11 ve A13 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A5, A6, A8, A11, A13 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 41 | 45 | 93 |
| Küme 2 | P2 | 27 | 111 | 27 |
| Küme 3 | P3 | 5 | 5 | 41 |
| **toplam** | | 73 | 161 | 161 |

Tablo 4.60'dan da anlaşıldığı gibi özellik olarak A2, A5, A6, A8, A11 ve A13'ün seçilmesi durumunda C3 veri seti içerisinde yer alan toplam 161 örnekten 73 tanesi doğru kümelenirken 88 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.61: Veri setinin C1 ve C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C2 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A4, A7 | **0.8908** | 0.8903 | 0.8916 |
| ***kesinlik*** | A2, A4, A9, A10 | 0.8827 | **0.9072** | 0.8861 |
| ***f1*** | A4, A7 | 0.8908 | 0.8903 | **0.8916** |
| **süre** | 25718.6625 saniye | | | |

Tablo 4.61, C1 ve C2 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* ve ortalama *f1* değeri özellik olarak A4 ve A7'nin seçilmesi durumunda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8908 ve ortalama *f1* değeri 0.8916'dır. Benzer şekilde en yüksek ortalama *kesinlik* değeri ise özellik olarak A2, A4, A9 ve A10'un seçilmesi durumunda 0.9072 olarak elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 25718.6625 saniye o da yaklaşık 7 saat 9 dakika sürmüştür.

Tablo 4.62: Veri setinin C1 ve C2, özelliklerin A4 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C2 | | |
|---|---|---|---|---|
| **özellik** | | A4, A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 622 | 671 | 678 |
| Küme 2 | P3 | 280 | 304 | 314 |
| Küme 3 | P1 | 199 | 261 | 244 |
| **toplam** | | 1101 | 1236 | 1236 |

Tablo 4.62'den de anlaşıldığı gibi özellik olarak A4 ve A7'nin seçilmesi durumunda C1 ve C2 veri seti içerisinde yer alan toplam 1236 örnekten 1101 tanesi doğru kümelenirken 135 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.63: Veri setinin C1 ve C2, özelliklerin A2, A4, A9, A10 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C2 | | |
|---|---|---|---|---|
| **özellik** | | A2, A4, A9, A10 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 194 | 195 | 244 |
| Küme 2 | P3 | 308 | 430 | 314 |
| Küme 3 | P2 | 589 | 611 | 678 |
| **toplam** | | 1091 | 1236 | 1236 |

Tablo 4.63'ten de anlaşıldığı gibi özellik olarak A2, A4, A9 ve A10'un seçilmesi durumunda C1 ve C2 veri seti içerisinde yer alan toplam 1236 örnekten 1091 tanesi doğru kümelenirken 145 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.64: Veri setinin C1 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A3, A8, A9, A12, A13 | **0.8664** | 0.8822 | 0.8582 |
| ***kesinlik*** | A4, A5, A12 | 0.7457 | **0.8981** | 0.7747 |
| ***f1*** | A2, A3, A8, A9, A12, A13 | 0.8664 | 0.8822 | **0.8582** |
| **süre** | 11550.2767 saniye | | | |

Tablo 4.64, C1 ve C3 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* ve ortalama *f1* değeri özellik olarak A2, A3, A8, A9, A12 ve A13'ün seçilmesi durumunda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8664, ortalama *f1* değeri de 0.8582'dir. Benzer şekilde en yüksek ortalama *kesinlik* değeri özellik olarak A4, A5 ve A12'nin seçilmesi durumunda 0.8981 olarak elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 11550.2767 saniye o da yaklaşık 3 saat 12 dakika sürmüştür.

Tablo 4.65: Veri setinin C1 ve C3, özelliklerin A2, A3, A8, A9, A12 ve A13 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A3, A8, A9, A12, A13 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 440 | 527 | 443 |
| Küme 2 | P1 | 92 | 97 | 141 |
| Küme 3 | P3 | 71 | 72 | 112 |
| **toplam** | | 603 | 696 | 696 |

Tablo 4.65'ten de anlaşıldığı gibi özellik olarak A2, A3, A8, A9, A12 ve A13'ün seçilmesi durumunda C1 ve C3 veri seti içerisinde yer alan toplam 696 örnekten 603 tanesi doğru kümelenirken 93 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.66: Veri setinin C1 ve C3, özelliklerin A4, A5 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A4, A5, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 112 | 287 | 112 |
| Küme 2 | P2 | 333 | 335 | 443 |
| Küme 3 | P1 | 74 | 74 | 141 |
| **toplam** | | 519 | 696 | 696 |

Tablo 4.66'dan da anlaşıldığı gibi özellik olarak A4, A5 ve A12'nin seçilmesi durumunda C1 ve C3 veri seti içerisinde yer alan toplam 696 örnekten 519 tanesi doğru kümelenirken 177 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.67: Veri setinin C2 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A4, A7 | **0.8503** | 0.8499 | 0.8496 |
| ***kesinlik*** | A2, A4, A10, A12, A13 | 0.7865 | **0.8607** | 0.7898 |
| ***f1*** | A4, A7 | 0.8503 | 0.8499 | **0.8496** |
| **süre** | 21530.2972 saniye | | | |

Tablo 4.67, C2 ve C3 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* ve ortalama *f1* değeri özellik olarak A4 ve A7'nin seçilmesi durumunda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8503 ve ortalama *f1* değeri 0.8496'dır. Benzer şekilde en yüksek ortalama *kesinlik* değeri de özellik olarak A2, A4, A10, A12 ve A13'ün seçilmesi durumunda 0.8496 olarak elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 21530.2972 saniye o da yaklaşık 5 saat 59 dakika sürmüştür.

Tablo 4.68: Veri setinin C2 ve C3, özelliklerin A4 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A4, A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 231 | 278 | 289 |
| Küme 2 | P2 | 266 | 308 | 289 |
| Küme 3 | P3 | 236 | 276 | 284 |
| **toplam** | | 733 | 862 | 862 |

Tablo 4.68'den de anlaşıldığı gibi özellik olarak A4 ve A7'nin seçilmesi durumunda C2 ve C3 veri seti içerisinde yer alan toplam 862 örnekten 733 tanesi doğru kümelenirken 129 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.69: Veri setinin C2 ve C3, özelliklerin A2, A4, A10, A12 ve A13 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A4, A10, A12, A13 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 284 | 459 | 284 |
| Küme 2 | P1 | 211 | 220 | 289 |
| Küme 3 | P2 | 183 | 183 | 289 |
| **toplam** | | 678 | 862 | 862 |

Tablo 4.69'dan da anlaşıldığı gibi özellik olarak A2, A4, A10, A12 ve A13'ün seçilmesi durumunda C2 ve C3 veri seti içerisinde yer alan toplam 862 örnekten 678 tanesi doğru kümelenirken 184 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.70: Veri setinin C1, C2 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1, C2 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A4, A6, A7 | **0.8576** | 0.8628 | 0.8593 |
| ***kesinlik*** | A2, A4, A6, A8, A9, A10 | 0.8003 | **0.8804** | 0.8105 |
| ***f1*** | A4, A6, A7 | 0.8576 | 0.8628 | **0.8593** |
| **süre** | 11550.2767 saniye | | | |

Tablo 4.70, C1, C2 ve C3 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* ve ortalama *f1* değeri özellik olarak A4, A6 ve A7'nin seçilmesi durumunda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8576 ve ortalama *f1* değeri 0.8593'tür. Benzer şekilde en yüksek ortalama *kesinlik* değeri de özellik olarak A2, A4, A6, A8, A9 ve A10'un seçilmesi durumunda 0.8804 olarak elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 11550.2767 saniye o da yaklaşık 3 saat 12 dakika sürmüştür.

Tablo 4.71: Veri setinin C1, C2 ve C3, özelliklerin A4, A6 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | C1, C2 ve C3 |
|---|---|

| **özellik** | | A4, A6, A7 | | |
|---|---|---|---|---|
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 295 | 336 | 355 |
| Küme 2 | P1 | 275 | 379 | 337 |
| Küme 3 | P2 | 628 | 682 | 705 |
| **toplam** | | 1198 | 1397 | 1397 |

Tablo 4.71'den de anlaşıldığı gibi özellik olarak A4, A6 ve A7'nin seçilmesi durumunda C1, C2 ve C3 veri seti içerisinde yer alan toplam 1397 örnekten 1198 tanesi doğru kümelenirken 199 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.72: Veri setinin C1, C2 ve C3, özelliklerin A2, A4, A6, A8, A9 ve A10 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1, C2 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A4, A6, A8, A9, A10 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 540 | 550 | 705 |
| Küme 2 | P3 | 355 | 623 | 355 |
| Küme 3 | P1 | 223 | 224 | 337 |
| **toplam** | | 1118 | 1397 | 1397 |

Tablo 4.72'den de anlaşıldığı gibi özellik olarak A2, A4, A6, A8, A9 ve A10'un seçilmesi durumunda C1, C2 ve C3 veri seti içerisinde yer alan toplam 1397 örnekten 1118 tanesi doğru kümelenirken 279 tanesi yanlış kümeler içerisinde yer almıştır.

Yukarıda da bahsedildiği üzere toplamda 57737 farklı kombinasyonun her birisi için 10 kez *k*-Ortalamalar algoritması çalıştırılarak kümeler oluşturulmuş ve çıktı olarak 7 farklı veri setinin *doğruluk*, *kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü için en başarılı sonuçlarının sağlayan özelliklerin belirlendiği 21 sonuç elde edilmiştir. Elde edilen bu en başarılı 21 sonuç içinde en sık kullanılan özelliklerden A4: 16 kez, A2: 12 kez, A7: 11 kez, A8, A9 ve A12: 7 kez, A3: 5 kez ve A6 ve A13: 4 kez kullanılmıştır.

Genel olarak bakıldığı zaman tüm başarım ölçütleri için en başarılı sonuçlar C1 koleksiyonunda elde edilmiştir. 535 adet örneğin 531 tanesi doğru kümelenerek yalnızca 4 tanesi yanlış kümelenmiştir. Buna bağlı olarak da ortalama *doğruluk* ve ortalama *f1* için 0.9925, ortalama *kesinlik* için 0.9926 değeri elde edilmiştir. Bu değerler aynı zamanda şu anlama gelmektedir:

Bu yöntemle A2, A3, A4, A8, A9 ve A12 özellikleri kullanılarak C1 koleksiyonuna ait örneklerin hemen hemen tamamı doğru kümelere ayrılmıştır. Bunun sebebi C1 koleksiyonun özelliğinden kaynaklanmaktadır. Bu veri seti siyasi partilerin resmi Twitter hesaplarının arkadaş listesinde yer alan kullanıcılardan oluşmaktadır, başka bir deyişle bu kullanıcılar genellikle partinin millet vekilleri, bakanları ve diğer siyasetçilerinden oluşmaktadır, bu yüzden de partiye benzerlikleri oldukça yüksektir. Diğer tekli veri setleri arasında ise sırasıyla C2 için özelliklerin A2, A4 ve A7 seçilmesi durumunda ortalama *doğruluk* değeri 0.9073, ortalama *kesinlik* değeri 0.9120 ve ortalama *f1* değeri 0.9058 olarak, C3 için özelliğin A7 olması durumunda ortalama *doğruluk* değeri 0.6584, ortalama *kesinlik* değeri 0.7109 ve ortalama *f1* değeri 0.6596 olarak ve yine C3 için özelliklerin A2, A5, A6, A8, A11 ve A13 olması durumunda ortalama *doğruluk* değeri 0.4534, ortalama *kesinlik* değeri 0.8217 ve ortalama *f1* değeri 0.4642 olarak hesaplanmıştır. Buradan da görüldüğü üzere tekli veri setleri için C1'den sonraki en başarılı sonuç C2'ye, en başarısız sonuç da C3'e aittir. Bunun nedeni yine koleksiyonların özelliklerinden kaynaklanmaktadır. C2 veri koleksiyonu siyasi partilerin takipçi listesinde yer alan ve yalnızca tek bir partiyi takip eden kullanıcı verilerinden oluşan bir veri setiyken, C3 rastgele seçilen kullanıcı verilerinden oluşan bir veri setidir.

İkili veri setleri arasında elde edilen en başarılı sonuçlar da C1 ve C2 veri seti kullanılarak elde edilmiştir. Bu veri setlerinin birlikte ele alındığı durumda ortalama *doğruluk* ve ortalama *f1* değeri açısından en başarılı sonuç A4 ve A7 özelliklerinin seçildiği durumda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8908, ortalama *kesinlik* değeri 0.8903 ve ortalama *f1* değeri 0.8916'dır. Benzer şekilde A2, A4, A9 ve A10 özelliklerinin seçilmesi durumunda *kesinlik* değeri için en başarılı sonuç elde edilmiştir ve bu durumda elde edilen ortalama *doğruluk* değeri 0.8827, ortalama *kesinlik* değeri 0.9022 ve ortalama *f1* değeri 0.8861'dir.

C1, C2 ve C3 veri setlerinin üçünün birlikte kullanılmasıyla elde edilen en başarılı ortalama *doğruluk* ve ortalama *f1* değerleri özellik olarak A4, A6 ve A7'nin seçildiği durumda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8576, ortalama *kesinlik* değeri 0.8628 ve ortalama *f1* değeri 0.8593'tür. Yine benzer şekilde *kesinlik* değeri için en başarılı sonuç da özelliklerin A2, A4, A6, A8, A9 ve A10 seçilmesi durumunda elde edilmiştir ve elde edilen değer sırasıyla ortalama *doğruluk* 0.8003, ortalama *kesinlik* 0.8804 ve ortalama *f1* 0.8105'tir.

### 4.3.2 Uygulama 4: Bulanık *c*-Ortalamalar Yöntemiyle Kullanıcıların Kümelenmesi

Bu uygulamada veri setlerinde yer alan örnekler Uygulama 1 ve Uygulama 2'de olduğu gibi sınıflandırılmak yerine Uygulama 3'e benzer şekilde kümelenmeye çalışılmıştır. Yani burada önemli olan bir grupta yer alan kişilerin aynı siyasi görüşe sahip olmasıdır ancak hangi siyasi görüşe sahip oldukları önemli değildir. Bu doğrultuda çalışmada 3 siyasi parti yer aldığı için veri setlerindeki örnekler Bulanık *c*-Ortalamalar algoritması kullanılarak 3 kümeye ayrılmaya çalışılmıştır.

Yapılan uygulamada Tablo 4.1'de yer alan 8 adet koleksiyon arasından C1, C2 ve C3'ten ve Tablo 4.2'de yer alan 13 adet özelliğin tamamından yararlanılmış ve şu 4 temel sorunun cevabı aranmıştır:

1. En başarılı kümeleme sonucu nedir?
2. En başarılı kümeleme sonucunu sağlayan özellik hangisi ya da hangileridir?
3. En başarılı kümeleme sonucu hangi veri setinde elde edilmiştir?
4. Yanlış kümelenen örneklerin üyelik derecelerinin dağılımı nasıldır?

Uygulamada Matlab üzerinde fcm fonksiyonu kullanılarak her bir örneğin her bir küme için .üyelik derecesi hesaplanmış ve örneğin en yüksek üyelik derecesine sahip kümeye ait olduğu kabul edilmiştir. Tüm veri seti ve özellik kombinasyonlarının test edilebilmesi için yine Matlab'ın nchoosek fonksiyonunundan yararlanılmıştır. Her kombinasyon 10 kez çalıştırılarak en başarılı sonuçlar çıktı olarak kabul edilmiştir.

Toplam 573370 kombinasyonun daha hızlı bir şekilde hesaplanabilmesi için Matlab'ın parfor fonksiyonundan yararlanılarak işlemlerin paralelleştirilmesi sağlanmıştır.

1 nolu soruda da yer alan, yöntemin başarımını ölçmek için Bölüm 3.3'te açıklanan ve diğer tüm uygulamalarda olduğu gibi yine *doğruluk*, *kesinlik* ve *f1* olmak üzere 3 farklı ölçütten yararlanılmıştır. En iyi kümelenme durumunun her bir kümede yalnız bir siyasi partiye ait kullanıcı örneğinin olduğu durum olarak kabul edilmiş ve başarım ölçütleri buna göre hesaplanmıştır.

2 ve 3 nolu soruların cevaplanabilmesi için Tablo 4.2'deki 13 farklı özelliğin 8191 kombinasyonunun ve Tablo 4.1'deki C1, C2 ve C3 veri setlerinin 7 farklı kombinasyonun oluşturduğu 57337 farklı durum ayrı ayrı test edilmiştir. Bu işlem 10 kez tekrarlanarak en başarılı sonuçlar çıktı olarak kabul edilmiştir.

Bulanık kümelemenin geleneksel kümelemeden farkı örneklerin bir kümeye ait olup olmamak yerine onların kümelerin üyelik dereceleriyle ifade edilmesidir. Yani Uygulama 3'te görüldüğü gibi geleneksel kümeleme yöntemlerinden birisi olan *k*-Ortalamalar yönteminde kümeleme sonucunda örnekler ya bir kümeye aittir ya da değildir şeklinde ifade edilmektedir. Yani bir örnek bir kümeye ait ise o küme 1 ile ait olmadığı diğer kümeler de 0 ile temsil edilmektedir. Ancak Bulanık *c*-Ortalamalar algoritması kullanılarak elde edilen sonuçlarda her bir örneğin her bir kümeye olan yakınlık ve benzerliklerine bağlı olarak 0 ve 1 aralığında üyelik dereceleri hesaplanmaktadır.

"Bulanık mantık için, matematiğin gerçek dünyaya uygulanması denilebilir. Çünkü gerçek dünyada her an değişen durumlarda değişik sonuçlar çıkarılabilir"(Elmas 2016).

Yukarıda da ifade edildiği gibi gerçek dünyadaki olaylarda yalnızca bir gruba, bir sınıfa ya da bir kümeye üyelik söz konusu değildir. Örneğin bu çalışmadan yola çıkarsak insanlar oy kullanıken yalnızca bir siyasi partiyi destekleyebilirler ancak aynı anda birden fazla siyasi partiyi takip edebilirler ve aynı şekilde birden fazla siyasi partiyle ilgilenebilirler. Ancak bu siyasi partilerin her birisine yakınlıkları farklılık göstermektedir. Bu uygulamada da örneklerin üyelik derecesi en yüksek olan kümeye ait olduğu kabul edilmiştir.

Bu uygulamanın Uygulama 3'ten farkı yanlış kümelenen örneklerin üyelik derecelerinin incelenebiliyor olmasıdır. Ayrıca bu yöntemi diğerlerinden ayıran en önemli özelliklerden birisi de bir kümeden diğerine geçebilecek olan örneklerin üyelik derecelerine bakılarak kolaylıkla tespit edilebilmesidir. Yine benzer şekilde bir kümede yer almasına rağmen diğer kümelere de aynı derecede yakın olan ve aslında tüm kümelere ait olabilecek örnekler bu değerler sayesinde tespit edilebilmektedir.

Tablo 4.73: Veri setinin C1 olması durumunda elde edilen en yüksek başarım ölçütleri sonuçlar

| **koleksiyon** | C1 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A1, A2, A3, A4, A6, A7, A9, A10, A12 | **0.8879** | 0.9444 | 0.9042 |
| ***kesinlik*** | A2, A3, A4, A5, A9, A10 | 0.8673 | **0.9465** | 0.8896 |
| ***f1*** | A1, A2, A3, A4, A6, A7, A9, A10, A12 | 0.8879 | 0.9444 | **0.9042** |
| **Süre** | 1806.2290 saniye | | | |

Tablo 4.73'te Bulanık *c*-Ortalamalar algoritmasının C1 veri seti üzerinde çalıştırılması sonucunda elde edilen en başarılı sonuçlar gösterilmektedir. Tabloda da görüldüğü gibi en yüksek ortalama *doğruluk* değeri 0.8879 ve *f1* değeri 0.9042 olarak hesaplanmıştır. Bu değerler özellik olarak A1, A2, A3, A4, A6, A7, A9, A10 ve A12'nin seçildiği durumda elde edilmiştir. Bu koşullarda elde edilen ortalama *kesinlik* değeri de 0.9444 olarak hesaplanmıştır. Ortalama *kesinlik* değerinin en yüksek olduğu durumda özelliklerin A2, A3, A4, A5, A9 ve A10 olması durumudur ve bu durumda elde edilen ortalama *kesinlik* değeri 0.9465 iken, ortalama *doğruluk* değeri 0.8673, ortalama *f1* değeri de 0.8896'dır. Bu değerler tüm kombinasyonların, her bir iterasyonun ortalama 1806.2290 saniye o da yaklaşık 30 dakika süren 10 iterasyon boyunca çalıştırılması sonucunda elde edilen en iyi sonuçlardır.

Tablo 4.74'te ve Tablo 4.76'da oluşan kümelerin her birisi, kümenin içerisinde en fazla bulunan sınıfa ait örnek sayıları gösterilmiştir. Bu tablolara göre Küme1, Küme2 ve Küme3 için şunları söyleyebiliriz:

Tablo 4.74'e göre:

- Küme 1 içerisinde toplam 104 tane örnek bulunurken bu örneklerin 47 tanesinin P1 sınıfına, 57 tanesinin farklı bir sınıfa ait olduğu görülmektedir. Ayrıca P1 sınıfının bu veri setinde bulunan toplam 48 örneğinin 47 tanesi Küme 1 içerisinde yer alırken yalnız 1 tanesi Küme 2 içerisinde yer almıştır.
- Küme 2 içerisinde toplam 365 tane örnek bulunmakta ve bunun 362 tanesini P2 sınıfına ait örnekler oluşturmaktadır. Bu kümede yer alan örneklerden sadece 1 tanesinin P1, 2 tanesinin de P3 sınıfına ait olduğunu göstermektedir. Yine aynı şekilde bu veri seti içerisinde bulunan P2 sınıfına ait 416 örnekten 362 tanesi Küme 2 içerisinde yer alırken 54 tanesi Küme 1 içerisinde yer almıştır.

- Küme 3 içerisinde toplam 66 tane örnek bulunmakta ve bunun tamamı P3 sınıfına ait örneklerden oluşmaktadır. Benzer şekilde P3 sınıfının bu veri setinde yer alan toplam 71 tane örneğinin 66 tanesi Küme 3 içerisinde yer alırken 5 tanesi Küme 1 ya da Küme 2 içerisinde yer almıştır.

Tablo 4.76'ya göre:

- Küme 1 içerisinde toplam 69 tane örnek bulunmakta ve bunun tamamını P3 sınıfına ait örnekler oluşturmaktadır. Yine aynı şekilde bu veri seti içerisinde bulunan P3 sınıfına ait 71 örnekten 69 tanesi Küme 1 içerisinde yer alırken geriye kalan 2 tanesi Küme 2 içerisinde yer almıştır.
- Küme 2 içerisinde toplam 119 tane örnek bulunmakta ve bunun 48 tanesi P1 sınıfına aitken kalan toplam 71 örneğin 2 tanesi P3 sınıfına, 69 tanesi de P2 sınıfına ait örneklerden oluşmaktadır. Benzer şekilde P1 sınıfının bu veri setinde yer alan toplam 48 tane örneğinin tamamı Küme 2 içerisinde yer almıştır.

- Küme 3 içerisinde toplam 347 tane örnek bulunurken bu örneklerinde tamamı P2 sınıfına ait olduğu görülmektedir. Ayrıca P2 sınıfının bu veri setinde bulunan toplam 416 örneğinin 347 tanesi Küme 3 içerisinde yer alırken 69 tanesi Küme 2 içerisinde yer almıştır.

Tablo 4.74: Veri setinin C1, özelliklerin A1, A2, A3, A4, A6, A7, A9, A10 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 | | |
|---|---|---|---|---|
| **özellik** | | A1, A2, A3, A4, A6, A7, A9, A10, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 47 | 104 | 48 |
| Küme 2 | P2 | 362 | 365 | 416 |
| Küme 3 | P3 | 66 | 66 | 71 |
| **toplam** | | 475 | 535 | 535 |

Tablo 4.74'te ortalama *doğruluk* ve ortalama *f1* değerlerinin en yüksek değerlerine sahip olduğu, özelliklerin A1, A2, A3, A4, A6, A7, A9, A10 ve A12 olması durumunda C1 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C1 veri setinde yer alan toplam 535 örnekten 475 tanesi doğru kümelenirken 60 tanesi yanlış kümeler içerisinde yer almıştır. Ancak Tablo 4.75'e baktığımız zaman yanlış kümelenen bu 60 örnekten 3 tanesinin bulunduğu kümenin üyelik dereceleriyle aslında olması gereken kümenin üyelik derecesi arasında çok büyük bir farkın olmadığı görülmektedir. Örneğin 2 nolu örneğin Küme 1, Küme 2 ve Küme 3 üyelik dereceleri sırasıyla 0.3439, 0.3463 ve 0.3098'dir. Burada en yüksek üyelik derecesi 0.3463 olduğu için sistem tarafından bu örnek Küme 2 içerisine dahil edilmiştir. Ancak aynı örneğin Küme 1 üyelik derecesi 0.3439 yani 0.3463'e çok yakın bir değerdir ki bu değer de 2 nolu örneğin aslında içerisinde olması gereken Küme 1'in üyelik derecesidir.

Tablo 4.75: Veri setinin C1, özelliklerin A1, A2, A3, A4, A6, A7, A9, A10 ve A12 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|
| 2 | 0.3439 | 0.3463 | 0.3098 | K2 | K1 |
| 100 | 0.3316 | 0.3367 | 0.3317 | K2 | K3 |
| 119 | 0.2618 | 0.3909 | 0.3473 | K2 | K3 |

Tablo 4.75, Tablo 4.77, Tablo 4.80, Tablo 4.82, Tablo 4.85, Tablo 4.87, Tablo 4.90, Tablo 4.93, Tablo 4.95, Tablo 4.98 ve Tablo 4.101'de Bulanık *c*-Ortalamalar algoritması kullanılarak elde edilen en başarılı sonuçlarda yanlış kümelenen örnekler arasından yanlış kümelenmesine rağmen içerisinde bulunduğu kümenin üyelik derecesiyle aslında olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten az olan örnekler gösterilmiştir. Bu tabloların ilk sütunu olan no örneklerin veri setindeki indekslerini, K1 sütunu Küme 1'in üyelik derecesini, K2 sütunu Küme2'nin üyelik derecesini, K3 sütunu Küme3'ün üyelik derecesini, A sütunu örneğin içerisinde bulunduğu kümeyi, B sütunu ise örneğin gerçekte olması gereken kümeyi ifade etmektedir. Ayrıca A ve B sütunlarında yer alan K1 Küme 1, K2 Küme 2 ve K3 ise Küme 3 anlamına gelmektedir.

Tablo 4.76: Veri setinin C1, özelliklerin A2, A3, A4, A5, A9 ve A10 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 | | |
|---|---|---|---|---|
| **özellik** | | A2, A3, A4, A5,A9,A10 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 69 | 69 | 71 |
| Küme 2 | P1 | 48 | 119 | 48 |
| Küme 3 | P2 | 347 | 347 | 416 |
| **toplam** | | 464 | 535 | 535 |

Tablo 4.76'da ortalama *kesinlik* değerinin en yüksek değerine sahip olduğu, özelliklerin A2, A3, A4, A5, A9 ve A10 olması durumunda C1 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C1 veri setinde yer alan toplam 535 örnekten 464 tanesi doğru kümelenirken 71 tanesi yanlış kümeler içerisinde yer almıştır. Tablo 4.77'de de bu 71 örneğin 21 tanesine ait küme üyelik dereceleri, bulunduğu kümeler ve gerçekte bulunması gereken kümeler yer almaktadır. Bu 21 örnek rastgele seçilmiş örnekler değildir, bu örnekler bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten az olan örneklerdir.

Tablo 4.77: Veri setinin C1, özelliklerin A2,A3, A4, A5, A9 ve A10 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki

fark 0.05’ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| no | (K1) | (K2) | (K3) | (A) | (B) |
|---|---|---|---|---|---|
| 53 | 0.3464 | 0.3778 | 0.2758 | K2 | K1 |
| 128 | 0.2867 | 0.3752 | 0.3380 | K2 | K3 |
| 167 | 0.2998 | 0.3654 | 0.3348 | K2 | K3 |
| 169 | 0.3038 | 0.3562 | 0.3400 | K2 | K3 |
| 175 | 0.3055 | 0.3590 | 0.3355 | K2 | K3 |
| 186 | 0.2989 | 0.3698 | 0.3313 | K2 | K3 |
| 196 | 0.2832 | 0.3825 | 0.3342 | K2 | K3 |
| 223 | 0.3149 | 0.3658 | 0.3193 | K2 | K3 |
| 240 | 0.2843 | 0.3621 | 0.3536 | K2 | K3 |
| 259 | 0.2812 | 0.3788 | 0.3400 | K2 | K3 |
| 284 | 0.2953 | 0.3546 | 0.3501 | K2 | K3 |
| 402 | 0.2946 | 0.3745 | 0.3310 | K2 | K3 |
| 425 | 0.3013 | 0.3568 | 0.3419 | K2 | K3 |
| 450 | 0.3120 | 0.3523 | 0.3356 | K2 | K3 |
| 479 | 0.2821 | 0.3811 | 0.3368 | K2 | K3 |
| 496 | 0.2837 | 0.3669 | 0.3494 | K2 | K3 |
| 517 | 0.2844 | 0.3687 | 0.3468 | K2 | K3 |

Tablo 4.78: Veri setinin C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A4, A7 | **0.8845** | 0.8920 | 0.8843 |
| ***kesinlik*** | A3, A4, A7, A10 | 0.8845 | **0.8923** | 0.8843 |
| ***f1*** | A4, A7 | 0.8845 | 0.8920 | **0.8843** |
| **Süre** | 3579.8265 saniye | | | |

Tablo 4.78, C2 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.8845, en yüksek ortalama *kesinlik* değeri 0.8923 ve en yüksek ortalama *f1* değeri 0.8843 olarak hesaplanmıştır. Bu en yüksek değerlerden ortalama *doğruluk* ve ortalama *f1* özellik olarak A2, A4 ve A7'nin seçildiği durumda, ortalama *kesinlik* özellik olarak A3, A4, A7 ve A10'un seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 3579.8265 saniye o da yaklaşık 1 saat sürmüştür.

Tablo 4.79: Veri setinin C2, özelliklerin A4 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 | | |
|---|---|---|---|---|
| **özellik** | | A4, A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 228 | 278 | 243 |
| Küme 2 | P2 | 237 | 263 | 262 |
| Küme 3 | P1 | 155 | 160 | 196 |
| **toplam** | | 620 | 701 | 701 |

Tablo 4.79'da ortalama *doğruluk* ve ortalama *f1* değerlerinin en yüksek değerlerine sahip olduğu, özelliklerin A4 ve A7 olması durumunda C2 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C2 veri setinde yer alan toplam 701 örnekten 620 tanesi doğru kümelenirken 81 tanesi yanlış kümeler içerisinde yer almıştır.

Yanlış kümelenen 81 tane örnek arasından bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten az olan 29 örnek Tablo 4.80'de yer almaktadır.

Tablo 4.80: Veri setinin C2, özelliklerin A4 ve A7 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 123 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 331 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 433 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 |
| 157 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 344 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 454 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 |
| 167 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 376 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 458 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 |
| 212 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 380 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 28 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 224 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 385 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 61 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 238 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 388 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 77 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 260 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 395 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 107 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 292 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 411 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 175 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 298 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 421 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | 409 | 0.3616 | 0.3867 | 0.2517 | K2 | K1 |
| 321 | 0.3787 | 0.3481 | 0.2732 | K1 | K2 | 422 | 0.3734 | 0.3660 | 0.2606 | K1 | K2 | | | | | | |

Tablo 4.81: Veri setinin C2, özelliklerin A3, A4, A7 ve A10 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 | | |
|---|---|---|---|---|
| **özellik** | | A3, A4, A7, A10 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P3 | 229 | 280 | 243 |
| Küme 2 | P1 | 155 | 160 | 196 |
| Küme 3 | P2 | 236 | 261 | 262 |
| **toplam** | | 620 | 701 | 701 |

Tablo 4.81'de ise ortalama *kesinlik* değerinin en yüksek değerine sahip olduğu, özelliklerin A3, A4, A7 ve A10 olması durumunda C2 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C2 veri setinde yer alan toplam 701 örnekten yine 620 tanesi doğru kümelenirken 81 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.82: Veri setinin C2, özelliklerin A3, A4, A7 ve A10 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 123 | 0.3772 | 0.2878 | 0.3350 | K1 | K3 | 321 | 0.3773 | 0.2789 | 0.3438 | K1 | K3 | 422 | 0.3752 | 0.2742 | 0.3506 | K1 | K3 |
| 157 | 0.3715 | 0.2869 | 0.3416 | K1 | K3 | 331 | 0.3766 | 0.2784 | 0.3450 | K1 | K3 | 433 | 0.3774 | 0.2866 | 0.3360 | K1 | K3 |
| 167 | 0.3782 | 0.2831 | 0.3387 | K1 | K3 | 344 | 0.3782 | 0.2838 | 0.3379 | K1 | K3 | 454 | 0.3780 | 0.2799 | 0.3421 | K1 | K3 |
| 197 | 0.3672 | 0.2690 | 0.3638 | K1 | K3 | 376 | 0.3748 | 0.2930 | 0.3322 | K1 | K3 | 458 | 0.3755 | 0.2918 | 0.3326 | K1 | K3 |
| 212 | 0.3727 | 0.2661 | 0.3612 | K1 | K3 | 380 | 0.3738 | 0.2670 | 0.3591 | K1 | K3 | 61 | 0.3636 | 0.2579 | 0.3784 | K3 | K1 |
| 224 | 0.3752 | 0.2694 | 0.3554 | K1 | K3 | 385 | 0.3779 | 0.2799 | 0.3422 | K1 | K3 | 77 | 0.3594 | 0.2557 | 0.3849 | K3 | K1 |
| 238 | 0.3733 | 0.2670 | 0.3597 | K1 | K3 | 388 | 0.3777 | 0.2792 | 0.3431 | K1 | K3 | 107 | 0.3661 | 0.2640 | 0.3698 | K3 | K1 |
| 260 | 0.3698 | 0.2651 | 0.3651 | K1 | K3 | 395 | 0.3775 | 0.2794 | 0.3431 | K1 | K3 | 175 | 0.3596 | 0.2556 | 0.3848 | K3 | K1 |
| 292 | 0.3705 | 0.2652 | 0.3643 | K1 | K3 | 411 | 0.3731 | 0.2664 | 0.3605 | K1 | K3 | 409 | 0.3604 | 0.2558 | 0.3838 | K3 | K1 |
| 298 | 0.3784 | 0.2816 | 0.3400 | K1 | K3 | 421 | 0.3698 | 0.2651 | 0.3650 | K1 | K3 | | | | | | |

C2 veri seti için özelliklerin A4, A7 veya A3, A4, A7 ve A10 seçilmesi durumunda her iki durumda da 701 örnekten 620 tanesi doğru kümelenirken 81 tanesi yanlış kümelenmiştir. Ayrıca her iki durumda da bu 81 örnekten 29 tanesinin bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten azdır. A4 ve A7 özellikleri için bu 29 örnek Tablo 4.80'de gösterilirken A3, A4, A7 ve A10 özellikleri için Tablo 4.82'de gösterilmiştir.

Tablo 4.83: Veri setinin C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A6, A7, A9, A11, A12 | **0.5839** | 0.7419 | 0.5977 |
| ***kesinlik*** | A4, A7, A10, A12 | 0.4969 | **0.7574** | 0.5151 |
| ***f1*** | A2, A6, A7, A9, A11, A12 | 0.5839 | 0.7419 | **0.5977** |
| **süre** | 1573.8381 saniye | | | |

Tablo 4.83, C3 veri seti için elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.5839, en yüksek ortalama *kesinlik* değeri 0.7574 ve en yüksek ortalama *f1* değeri 0.5977 olarak hesaplanmıştır. Bu en yüksek değerlerden ortalama *doğruluk* özellik olarak A2, A6, A7, A9, A11 ve A12'nin seçildiği durumda, ortalama *kesinlik* özellik olarak A4, A7, A10 ve A12'un seçildiği durumda ve ortalama *f1* ise özellik olarak A2, A6, A7 ve A9'un seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 1573.8381 saniye o da yaklaşık 1 saat sürmüştür.

Tablo 4.84: Veri setinin C3, özelliklerin A2, A6, A7, A9, A11 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A6, A7, A9, A11, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 53 | 55 | 93 |
| Küme 2 | P3 | 35 | 95 | 41 |
| Küme 3 | P2 | 6 | 11 | 27 |
| **toplam** | | 94 | 161 | 161 |

Tablo 4.84'te ortalama *doğruluk* ve ortama *f1* değerlerinin en yüksek değerlerine sahip olduğu, özelliklerin A2, A6, A7, A9, A11 ve A12 olması durumunda C3 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C3 veri setinde yer alan toplam 161 örnekten 94 tanesi doğru kümelenirken 67 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.85'te ise yanlış kümelenen 67 tane örnek arasından bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten az olan 18 örnek yer almaktadır.

Tablo 4.85: Veri setinin C3, özelliklerin A2, A6, A7, A9, A11 ve A12 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 0.4892 | 0.4690 | 0.0418 | K1 | K2 | 28 | 0.4720 | 0.4870 | 0.0410 | K2 | K1 | 71 | 0.4633 | 0.4947 | 0.0420 | K2 | K1 |
| 9 | 0.4740 | 0.4849 | 0.0411 | K2 | K1 | 32 | 0.4670 | 0.4910 | 0.0420 | K2 | K1 | 134 | 0.4678 | 0.4906 | 0.0416 | K2 | K1 |
| 15 | 0.4688 | 0.4897 | 0.0415 | K2 | K1 | 34 | 0.4720 | 0.4870 | 0.0410 | K2 | K1 | 141 | 0.4750 | 0.4839 | 0.0410 | K2 | K1 |
| 20 | 0.4714 | 0.4873 | 0.0413 | K2 | K1 | 35 | 0.4700 | 0.4890 | 0.0410 | K2 | K1 | 144 | 0.4588 | 0.4987 | 0.0425 | K2 | K1 |
| 21 | 0.4768 | 0.4819 | 0.0413 | K2 | K1 | 41 | 0.4577 | 0.4995 | 0.0428 | K2 | K1 | 156 | 0.4628 | 0.4949 | 0.0423 | K2 | K1 |
| 24 | 0.4728 | 0.4861 | 0.0411 | K2 | K1 | 55 | 0.4778 | 0.4811 | 0.0411 | K2 | K1 | 159 | 0.4665 | 0.4918 | 0.0417 | K2 | K1 |

Tablo 4.86: Veri setinin C3, özelliklerin A4, A7, A10 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C3 | | |
|---|---|---|---|---|
| **özellik** | | A4, A7, A10, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 26 | 89 | 27 |
| Küme 2 | P1 | 35 | 35 | 93 |
| Küme 3 | P3 | 19 | 37 | 41 |
| **toplam** | | 80 | 161 | 161 |

Tablo 4.86'da ortalama *kesinlik* değerinin en yüksek değerine sahip olduğu, özelliklerin A4, A7, A10 ve A12 olması durumunda C3 veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. C3 veri setinde yer alan toplam 161 örnekten 80 tanesi doğru kümelenirken, 81 örnek yanlış kümeler içerisinde yer almıştır.

Tablo 4.87: Veri setinin C3, özelliklerin A4, A7, A10 ve A12 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|
| 57 | 0.3522 | 0.2556 | 0.3920 | K3 | K1 |

Tablo 4.87'de, veri setinin C3, özelliklerin A4, A7, A10 ve A12 seçilmesi durumunda Bulanık *c*-Ortalamalar algoritmasıyla kümelenmesi sonucunda elde edilen yanlış kümelenmiş 81 örnekten 57 nolu örneğin üyelik dereceleri gösterilmiştir. Kalan 80 örneğin bulunduğu küme ile olması gereken kümenin üyelik derecesi farkı 0.05'ten büyük olduğu için bu listede yer almamıştır.

Tablo 4.88: Veri setinin C1 ve C2 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C2 | | | |
|---|---|---|---|---|
| **maksimum** | **Özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A3, A4, A9, A10, A12 | **0.8568** | 0.8860 | 0.8594 |
| ***kesinlik*** | A2, A3, A4, A9, A10, A12 | 0.8568 | **0.8860** | 0.8594 |
| ***f1*** | A2, A3, A4, A9, A10, A12 | 0.8568 | 0.8860 | **0.8594** |
| **süre** | 7283.9235 saniye | | | |

Tablo 4.88, C1 ve C2 koleksiyonlarının veri seti olarak birlikte kullanılması durumunda elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.8568, ortalama *kesinlik* değeri 0.8860 ve ortalama *f1* değeri 0.8594 olarak hesaplanmıştır. Bu değerlerin tamamı özellik olarak A2, A3, A4, A9, A10 ve A12'nin seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 7283.9235 saniye o da yaklaşık 2 saat sürmüştür.

Tablo 4.89: Veri setinin C1 ve C2, özelliklerin A2, A3, A4, A9, A10 ve A12 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C2 | | |
|---|---|---|---|---|
| **özellik** | | A2, A3, A4, A9, A10, A12 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P1 | 210 | 277 | 244 |
| Küme 2 | P3 | 312 | 422 | 314 |
| Küme 3 | P2 | 537 | 537 | 678 |
| **toplam** | | 1059 | 1236 | 1236 |

Tablo 4.89'dan da anlaşıldığı gibi C1 ve C2 koleksiyonlarının birlikte veri seti olarak kullanılması, özellik olarak da A2, A3, A4, A9, A10 ve A12'nin seçilmesi durumunda toplam 1236 örnekten 1059 tanesi doğru kümelenirken 177 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.90: Veri setinin C1 ve C2, özelliklerin A2, A3, A4, A9, A10 ve A12 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 0.3484 | 0.3338 | 0.3179 | K1 | K2 | 974 | 0.3488 | 0.3319 | 0.3192 | K1 | K3 | 1081 | 0.3407 | 0.3249 | 0.3344 | K1 | K3 |
| 138 | 0.3493 | 0.3322 | 0.3185 | K1 | K2 | 976 | 0.3492 | 0.3320 | 0.3188 | K1 | K3 | 1082 | 0.3434 | 0.3271 | 0.3295 | K1 | K3 |
| 154 | 0.3490 | 0.3320 | 0.3190 | K1 | K3 | 995 | 0.3481 | 0.3312 | 0.3207 | K1 | K3 | 1083 | 0.3519 | 0.3301 | 0.3180 | K1 | K3 |
| 190 | 0.4336 | 0.1823 | 0.3841 | K1 | K3 | 996 | 0.3389 | 0.3235 | 0.3377 | K1 | K3 | 1084 | 0.3419 | 0.3259 | 0.3322 | K1 | K3 |
| 313 | 0.3456 | 0.3291 | 0.3253 | K1 | K3 | 998 | 0.3465 | 0.3348 | 0.3186 | K1 | K3 | 1085 | 0.3451 | 0.3286 | 0.3262 | K1 | K3 |
| 830 | 0.3482 | 0.3314 | 0.3204 | K1 | K3 | 1000 | 0.3431 | 0.3271 | 0.3298 | K1 | K3 | 1086 | 0.3430 | 0.3268 | 0.3302 | K1 | K3 |
| 831 | 0.3449 | 0.3285 | 0.3266 | K1 | K3 | 1002 | 0.3489 | 0.3320 | 0.3192 | K1 | K3 | 1087 | 0.3451 | 0.3284 | 0.3265 | K1 | K3 |
| 832 | 0.3466 | 0.3299 | 0.3234 | K1 | K3 | 1003 | 0.3461 | 0.3296 | 0.3243 | K1 | K3 | 1089 | 0.3439 | 0.3276 | 0.3285 | K1 | K3 |
| 833 | 0.3437 | 0.3275 | 0.3288 | K1 | K3 | 1004 | 0.3435 | 0.3274 | 0.3290 | K1 | K3 | 1091 | 0.3442 | 0.3296 | 0.3262 | K1 | K3 |
| 848 | 0.3486 | 0.3318 | 0.3196 | K1 | K3 | 1051 | 0.3487 | 0.3324 | 0.3189 | K1 | K3 | 1125 | 0.4151 | 0.1884 | 0.3965 | K1 | K3 |
| 851 | 0.3557 | 0.3318 | 0.3125 | K1 | K3 | 1056 | 0.3463 | 0.3298 | 0.3239 | K1 | K3 | 1209 | 0.3393 | 0.3239 | 0.3368 | K1 | K3 |
| 940 | 0.3944 | 0.2140 | 0.3916 | K1 | K3 | 1063 | 0.3463 | 0.3301 | 0.3236 | K1 | K3 | 541 | 0.3501 | 0.4000 | 0.2499 | K2 | K1 |
| 947 | 0.3392 | 0.3237 | 0.3371 | K1 | K3 | 1065 | 0.3415 | 0.3236 | 0.3349 | K1 | K3 | 695 | 0.3593 | 0.3869 | 0.2538 | K2 | K1 |
| 957 | 0.3422 | 0.3265 | 0.3313 | K1 | K3 | 1072 | 0.3485 | 0.3317 | 0.3198 | K1 | K3 | 703 | 0.3594 | 0.3884 | 0.2523 | K2 | K1 |
| 967 | 0.3406 | 0.3248 | 0.3346 | K1 | K3 | 1073 | 0.3505 | 0.3319 | 0.3176 | K1 | K3 | 745 | 0.4239 | 0.4293 | 0.1468 | K2 | K1 |
| 968 | 0.3478 | 0.3309 | 0.3213 | K1 | K3 | 1077 | 0.3481 | 0.3313 | 0.3206 | K1 | K3 | 927 | 0.3237 | 0.3610 | 0.3153 | K2 | K3 |

Tablo 4.90 (devam): Veri setinin C1 ve C2, özelliklerin A2, A3, A4, A9, A10 ve A12 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 973 | 0.3489 | 0.3320 | 0.3191 | K1 | K3 | 1078 | 0.3439 | 0.3275 | 0.3286 | K1 | K3 | | | | | | |

Bulanık *c*-Ortalamalar algoritması için veri seti olarak C1 ve C2'nin, özellik olarak da A2, A3, A4, A9, A10 ve A12'nin seçilmesi durumunda elde edilen en iyi durumda toplam 1236 örnekten 1059 tanesi doğru kümelenirken 177 tanesi yanlış kümelenmiştir. Ancak bu 177 örnekten 50 tanesinin bulunduğu kümenin üyelik derecesi ile olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten azdır. Bu 50 örnek Tablo 4.90'da gösterilmiştir.

Tablo 4.91: Veri setinin C1 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **Özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A3, A4, A7 | **0.7974** | 0.8720 | 0.8177 |
| ***kesinlik*** | A8 | 0.5920 | **0.8846** | 0.6359 |
| ***f1*** | A3, A4, A7 | 0.7974 | 0.8720 | **0.8177** |
| **süre** | 4640.1559 saniye | | | |

Tablo 4.91, C1 ve C3 koleksiyonlarının birlikte veri seti olarak kullanılması durumunda elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.7974, en yüksek ortalama *kesinlik* değeri 0.8846 ve en yüksek ortalama *f1* değeri 0.8177 olarak hesaplanmıştır. Bu en yüksek değerlerden ortalama *doğruluk* ve ortalama *f1* özellik olarak A3, A4 ve A7'nin seçildiği durumda, ortalama *kesinlik* de özellik olarak A8'in seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 4640.1559 saniye o da yaklaşık 1 saat 17 dakika sürmüştür.

Tablo 4.92: Veri setinin C1 ve C3, özelliklerin A3, A4 ve A7 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A3, A4, A7 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 365 | 375 | 443 |
| Küme 2 | P3 | 94 | 214 | 112 |
| Küme 3 | P1 | 96 | 107 | 141 |
| **toplam** | | 555 | 696 | 696 |

Tablo 4.92'de ortalama *doğruluk* ve ortalama *f1* değerlerinin en yüksek değerlerine sahip olduğu, özelliklerin A3, A4 ve A7 olması durumunda C1 ve C3'in birlikte kullanıldığı veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. Bu veri setinde yer alan toplam 696 örnekten 555 tanesi doğru kümelenirken 141 tanesi yanlış kümeler içerisinde yer almıştır. Bu 141 örnekten 140 tanesinin bulunduğu küme ile olması gereken kümenin üyelik dereceleri farkı 0.05'ten büyükken, 223 nolu örneğinki 0.05'ten küçüktür. Tablo 4.93'te bu örneğe ait küme üyelik dereceleri, bulunduğu küme ve olması gereken küme bilgileri gösterilmiştir.

Tablo 4.93: Veri setinin C1 ve C3, özelliklerin A3, A4 ve A7 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|
| 223 | 0.3644 | 0.2296 | 0.4060 | K3 | K1 |

Tablo 4.94: Veri setinin C1 ve C3, özelliklerin A8 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A8 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | AHP | 49 | 49 | 141 |
| Küme 2 | P3 | 112 | 396 | 112 |
| Küme 3 | P2 | 251 | 251 | 443 |
| **toplam** | | 412 | 696 | 696 |

Tablo 4.94'te ise ortalama *kesinlik* değerinin en yüksek değerine sahip olduğu, özelliğin A8 olması durumunda C1 ve C3'in birlikte kullanıldığı veri seti içerisindeki her bir sınıfın kümeler içerisindeki dağılımı gösterilmiştir. Bu veri setinde yer alan toplam 696 örnekten 412 tanesi doğru kümelenirken 284 tanesi yanlış kümeler içerisinde yer almıştır.

Tablo 4.95'te ise yanlış kümelenmiş 284 örnekten 296, 468 ve 531 nolu 3 örneklerin üyelik dereceleri gösterilmiştir. Kalan 281 örneğin bulunduğu küme ile

olması gereken kümenin üyelik dereceleri farkı 0.05'ten büyük olduğu için bu listede yer almamıştır.

Tablo 4.95: Veri setinin C1 ve C3, özelliğin A8 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|
| 296 | 0.0573 | 0.4731 | 0.4696 | K2 | K3 |
| 468 | 0.0576 | 0.4896 | 0.4528 | K2 | K3 |
| 531 | 0.0596 | 0.4857 | 0.4547 | K2 | K3 |

Tablo 4.96: Veri setinin C2 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C2 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A1, A2, A3, A4, A5, A6, A7, A11 | **0.8457** | 0.8454 | 0.8454 |
| ***kesinlik*** | A1, A2, A3, A4, A5, A6, A7, A11 | 0.8457 | **0.8454** | 0.8454 |
| ***f1*** | A1, A2, A3, A4, A5, A6, A7, A11 | 0.8457 | 0.8454 | **0.8454** |
| **süre** | 6974.8316 | | | |

Tablo 4.96, C2 ve C3 koleksiyonlarının veri seti olarak birlikte kullanılması durumunda elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.8457, ortalama *kesinlik* değeri 0.8454 ve ortalama *f1* değeri 0.8454 olarak hesaplanmıştır. Bu değerlerin tamamı özellik olarak A1, A2, A3, A4, A5, A6, A7 ve A11'in seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 6974.8316 saniye o da yaklaşık 1 saat 56 dakika sürmüştür.

Tablo 4.97: Veri setinin C2 ve C3, özelliklerin A1, A2, A3, A4, A5, A6, A7 ve A11 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C2 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A1, A2, A3, A4, A5, A6, A7, A11 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 258 | 296 | 289 |
| Küme 2 | P1 | 231 | 278 | 289 |
| Küme 3 | P3 | 240 | 288 | 284 |
| **toplam** | | 729 | 862 | 862 |

Tablo 4.97'den de anlaşıldığı gibi C2 ve C3 koleksiyonlarının birlikte veri seti olarak kullanılması, özellik olarak da A1, A2, A3, A4, A5, A6, A7 ve A11'in seçilmesi durumunda toplam 862 örnekten 729 tanesi doğru kümelenirken 133 tanesi yanlış kümeler içerisinde yer almıştır. Ancak yanlış kümelen bu 133 örnekten 13 tanesinin bulunduğu kümenin üyelik dereceleriyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten azdır. Bu 13 örneğe ait üyelik dereceleriyle, bulunduğu küme ve olması gereken küme Tablo 4.98'de gösterilmiştir.

Tablo 4.98: Veri setinin C2 ve C3, özelliklerin A1, A2, A3, A4, A5, A6, A7 ve A11 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|
| 28 | 0.3794 | 0.2725 | 0.3481 | K1 | K3 |
| 48 | 0.3757 | 0.2731 | 0.3511 | K1 | K3 |
| 99 | 0.3808 | 0.2703 | 0.3490 | K1 | K3 |
| 107 | 0.3854 | 0.2684 | 0.3462 | K1 | K3 |
| 561 | 0.3559 | 0.2955 | 0.3487 | K1 | K3 |
| 123 | 0.3471 | 0.2962 | 0.3567 | K3 | K1 |
| 157 | 0.3541 | 0.2916 | 0.3543 | K3 | K1 |
| 167 | 0.3526 | 0.2901 | 0.3573 | K3 | K1 |
| 298 | 0.3547 | 0.2890 | 0.3563 | K3 | K1 |
| 344 | 0.3519 | 0.2915 | 0.3567 | K3 | K1 |
| 376 | 0.3436 | 0.3003 | 0.3561 | K3 | K1 |
| 433 | 0.3496 | 0.2936 | 0.3568 | K3 | K1 |
| 458 | 0.3441 | 0.2994 | 0.3564 | K3 | K1 |

Tablo 4.99: Veri setinin C1, C2 ve C3 olması durumunda elde edilen en yüksek başarım ölçütleri

| **koleksiyon** | C1, C2 ve C3 | | | |
|---|---|---|---|---|
| **maksimum** | **Özellik** | ***doğruluk*** | ***kesinlik*** | ***f1*** |
| ***doğruluk*** | A2, A3, A4, A9, A10 | **0.8282** | 0.8614 | 0.8321 |
| ***kesinlik*** | A2, A3, A4, A9, A10 | 0.8282 | **0.8614** | 0.8321 |
| ***f1*** | A2, A3, A4, A9, A10 | 0.8282 | 0.8614 | **0.8321** |
| **süre** | 12348.9174 saniye | | | |

Tablo 4.99, C1, C2 ve C3 koleksiyonlarının veri seti olarak birlikte kullanılması durumunda elde edilen en yüksek ortalama başarım ölçütlerini göstermektedir. Buna göre elde edilen en yüksek ortalama *doğruluk* değeri 0.8282, ortalama *kesinlik* değeri 0.8614 ve ortalama *f1* değeri 0.8282 olarak hesaplanmıştır. Bu değerlerin tamamı özellik olarak A2, A3, A4, A9 ve A10'un seçildiği durumda elde edilmiştir. Bu değerlerin tespit edilmesi sırasında her iterasyon ortalama 12348.9174 saniye o da yaklaşık 3 saat 25 dakika sürmüştür.

Tablo 4.100: Veri setinin C1, C2 ve C3, özelliklerin A2, A3, A4, A9 ve A10 olması durumunda elde edilen kümeleme sonuçları

| **koleksiyon** | | C1, C2 ve C3 | | |
|---|---|---|---|---|
| **özellik** | | A2, A3, A4, A9, A10 | | |
| **küme** | **sınıf** | **doğru kümelenen örnek sayısı** | **kümedeki örnek sayısı** | **sınıftaki örnek sayısı** |
| Küme 1 | P2 | 539 | 539 | 705 |
| Küme 2 | P3 | 332 | 461 | 355 |
| Küme 3 | P1 | 286 | 397 | 337 |
| **toplam** | | 1157 | 1397 | 1397 |

Tablo 4.100'den de anlaşıldığı gibi C1, C2 ve C3 koleksiyonlarının birlikte veri seti olarak kullanılması, özellik olarak da A2, A3, A4, A9 ve A10'un seçilmesi durumunda toplam 1397 örnekten 1157 tanesi doğru kümelenirken 240 tanesi yanlış kümeler içerisinde yer almıştır. Yanlış kümelenen bu 240 örnekten 110 tanesinin bulunduğu kümenin üyelik derecesi ile olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten azdır. Bu 110 örnek Tablo 4.101'de gösterilmiştir.

Tablo 4.101: Veri setinin C1, C2 ve C3, özelliklerin A2, A3, A4, A9 ve A10 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| no | (K1) | (K2) | (K3) | (A) | (B) | no | (K1) | (K2) | (K3) | (A) | (B) | no | (K1) | (K2) | (K3) | (A) | (B) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 0.3061 | 0.3499 | 0.3440 | K2 | K3 | 1081 | 0.3229 | 0.3421 | 0.3350 | K2 | K1 | 1322 | 0.3086 | 0.3507 | 0.3407 | K2 | K1 |
| 138 | 0.3076 | 0.3503 | 0.3422 | K2 | K3 | 1082 | 0.3176 | 0.3451 | 0.3373 | K2 | K1 | 1323 | 0.3070 | 0.3508 | 0.3422 | K2 | K3 |
| 154 | 0.3070 | 0.3507 | 0.3422 | K2 | K1 | 1083 | 0.3057 | 0.3542 | 0.3400 | K2 | K1 | 1328 | 0.3074 | 0.3506 | 0.3420 | K2 | K1 |
| 190 | 0.3935 | 0.4200 | 0.1865 | K2 | K1 | 1084 | 0.3206 | 0.3434 | 0.3360 | K2 | K1 | 1329 | 0.3071 | 0.3501 | 0.3429 | K2 | K3 |
| 291 | 0.3891 | 0.4257 | 0.1853 | K2 | K1 | 1085 | 0.3143 | 0.3468 | 0.3389 | K2 | K1 | 1333 | 0.3075 | 0.3503 | 0.3421 | K2 | K1 |
| 313 | 0.3145 | 0.3464 | 0.3390 | K2 | K1 | 1086 | 0.3186 | 0.3445 | 0.3370 | K2 | K1 | 1334 | 0.3043 | 0.3517 | 0.3441 | K2 | K1 |
| 329 | 0.3903 | 0.4223 | 0.1875 | K2 | K1 | 1087 | 0.3147 | 0.3467 | 0.3386 | K2 | K1 | 1340 | 0.3072 | 0.3506 | 0.3422 | K2 | K3 |
| 349 | 0.3854 | 0.4255 | 0.1891 | K2 | K1 | 1088 | 0.3307 | 0.3378 | 0.3314 | K2 | K1 | 1343 | 0.3083 | 0.3499 | 0.3418 | K2 | K1 |
| 382 | 0.3827 | 0.4306 | 0.1867 | K2 | K1 | 1089 | 0.3167 | 0.3455 | 0.3378 | K2 | K1 | 1345 | 0.3193 | 0.3448 | 0.3359 | K2 | K1 |
| 830 | 0.3086 | 0.3498 | 0.3416 | K2 | K1 | 1091 | 0.3151 | 0.3454 | 0.3396 | K2 | K1 | 1363 | 0.3069 | 0.3506 | 0.3424 | K2 | K3 |
| 831 | 0.3148 | 0.3465 | 0.3387 | K2 | K1 | 1110 | 0.3275 | 0.3395 | 0.3330 | K2 | K1 | 1364 | 0.3092 | 0.3493 | 0.3414 | K2 | K1 |
| 832 | 0.3114 | 0.3484 | 0.3402 | K2 | K1 | 1125 | 0.4014 | 0.4063 | 0.1923 | K2 | K1 | 1368 | 0.3009 | 0.3576 | 0.3415 | K2 | K3 |
| 833 | 0.3169 | 0.3455 | 0.3376 | K2 | K1 | 1193 | 0.3338 | 0.3351 | 0.3311 | K2 | K1 | 1369 | 0.3085 | 0.3490 | 0.3425 | K2 | K1 |
| 848 | 0.3077 | 0.3503 | 0.3420 | K2 | K1 | 1209 | 0.3251 | 0.3408 | 0.3340 | K2 | K1 | 1371 | 0.3112 | 0.3483 | 0.3405 | K2 | K1 |
| 947 | 0.3255 | 0.3408 | 0.3337 | K2 | K1 | 1242 | 0.3068 | 0.3500 | 0.3433 | K2 | K3 | 1381 | 0.3085 | 0.3499 | 0.3416 | K2 | K1 |
| 957 | 0.3202 | 0.3433 | 0.3365 | K2 | K1 | 1244 | 0.3105 | 0.3487 | 0.3408 | K2 | K1 | 1383 | 0.3063 | 0.3544 | 0.3393 | K2 | K1 |
| 967 | 0.3230 | 0.3422 | 0.3348 | K2 | K1 | 1249 | 0.3066 | 0.3508 | 0.3426 | K2 | K3 | 1384 | 0.3131 | 0.3467 | 0.3403 | K2 | K1 |
| 968 | 0.3093 | 0.3495 | 0.3412 | K2 | K1 | 1252 | 0.3083 | 0.3499 | 0.3418 | K2 | K3 | 1393 | 0.3062 | 0.3519 | 0.3419 | K2 | K3 |
| 973 | 0.3072 | 0.3506 | 0.3422 | K2 | K1 | 1258 | 0.3078 | 0.3501 | 0.3421 | K2 | K1 | 294 | 0.3934 | 0.1682 | 0.4383 | K3 | K1 |
| 974 | 0.3073 | 0.3506 | 0.3421 | K2 | K1 | 1262 | 0.3072 | 0.3503 | 0.3425 | K2 | K3 | 695 | 0.2624 | 0.3538 | 0.3839 | K3 | K2 |
| 976 | 0.3069 | 0.3508 | 0.3422 | K2 | K1 | 1265 | 0.3070 | 0.3507 | 0.3422 | K2 | K1 | 703 | 0.2619 | 0.3531 | 0.3850 | K3 | K2 |
| 995 | 0.3088 | 0.3498 | 0.3414 | K2 | K1 | 1266 | 0.3069 | 0.3504 | 0.3426 | K2 | K3 | 745 | 0.1506 | 0.4228 | 0.4266 | K3 | K2 |

Tablo 4.101 (devam): Veri setinin C1, C2 ve C3, özelliklerin A2, A3, A4, A9 ve A10 olması durumunda yanlış kümelenmesine rağmen bulunduğu küme ile olması gerektiği kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnekler, (K1): Küme 1, (K2): Küme 2, (K3): Küme 3, (A): Bulunduğu küme, (B): Olması gereken küme

| **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** | **no** | **(K1)** | **(K2)** | **(K3)** | **(A)** | **(B)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 996 | 0.3262 | 0.3403 | 0.3335 | K2 | K1 | 1275 | 0.3038 | 0.3512 | 0.3450 | K2 | K3 | 916 | 0.3187 | 0.3214 | 0.3599 | K3 | K1 |
| 998 | 0.3074 | 0.3479 | 0.3447 | K2 | K1 | 1282 | 0.3052 | 0.3526 | 0.3421 | K2 | K3 | 927 | 0.3226 | 0.3188 | 0.3586 | K3 | K1 |
| 1000 | 0.3183 | 0.3445 | 0.3372 | K2 | K1 | 1284 | 0.3072 | 0.3505 | 0.3422 | K2 | K1 | 1099 | 0.3150 | 0.3201 | 0.3648 | K3 | K1 |
| 1002 | 0.3071 | 0.3507 | 0.3422 | K2 | K1 | 1289 | 0.3143 | 0.3465 | 0.3392 | K2 | K1 | 1126 | 0.3135 | 0.3265 | 0.3600 | K3 | K1 |
| 1003 | 0.3125 | 0.3477 | 0.3398 | K2 | K1 | 1290 | 0.3042 | 0.3540 | 0.3418 | K2 | K1 | 1296 | 0.1132 | 0.4213 | 0.4654 | K3 | K2 |
| 1004 | 0.3172 | 0.3451 | 0.3376 | K2 | K1 | 1294 | 0.3041 | 0.3534 | 0.3425 | K2 | K3 | 1303 | 0.1116 | 0.4215 | 0.4669 | K3 | K2 |
| 1034 | 0.3327 | 0.3364 | 0.3309 | K2 | K1 | 1300 | 0.3300 | 0.3373 | 0.3327 | K2 | K1 | 1320 | 0.1371 | 0.4270 | 0.4359 | K3 | K2 |
| 1051 | 0.3069 | 0.3505 | 0.3426 | K2 | K1 | 1305 | 0.3058 | 0.3487 | 0.3455 | K2 | K3 | 1348 | 0.1287 | 0.4271 | 0.4442 | K3 | K2 |
| 1056 | 0.3119 | 0.3481 | 0.3400 | K2 | K1 | 1310 | 0.3071 | 0.3507 | 0.3422 | K2 | K3 | 1351 | 0.1427 | 0.4267 | 0.4306 | K3 | K2 |
| 1063 | 0.3120 | 0.3478 | 0.3403 | K2 | K1 | 1311 | 0.3031 | 0.3509 | 0.3460 | K2 | K3 | 1373 | 0.1410 | 0.4280 | 0.4310 | K3 | K2 |
| 1065 | 0.3229 | 0.3436 | 0.3336 | K2 | K1 | 1312 | 0.3074 | 0.3505 | 0.3421 | K2 | K3 | 1387 | 0.2583 | 0.3474 | 0.3943 | K3 | K2 |
| 1072 | 0.3081 | 0.3500 | 0.3419 | K2 | K1 | 1313 | 0.3279 | 0.3384 | 0.3337 | K2 | K1 | 1390 | 0.1238 | 0.4244 | 0.4518 | K3 | K2 |
| 1073 | 0.3058 | 0.3522 | 0.3420 | K2 | K1 | 1314 | 0.3047 | 0.3511 | 0.3442 | K2 | K3 | 1391 | 0.1252 | 0.4246 | 0.4502 | K3 | K2 |
| 1077 | 0.3087 | 0.3499 | 0.3415 | K2 | K1 | 1315 | 0.2997 | 0.3546 | 0.3456 | K2 | K3 | 1394 | 0.1381 | 0.4259 | 0.4360 | K3 | K2 |
| 1078 | 0.3167 | 0.3455 | 0.3377 | K2 | K1 | 1316 | 0.2962 | 0.3631 | 0.3407 | K2 | K3 | | | | | | |

Bu uygulamada 13 özelliğin ve 3 farklı veri setinin toplam 57737 farklı kombinasyonun her birisi için 10 kez Bulanık *c*-Ortalamalar algoritması çalıştırılarak her bir örneğin her bir küme için üyelik dereceleri hesaplanmıştır ve bu değerlere bağlı olarak da kümeler oluşturulmuştur. Daha sonra çıktı olarak 7 farklı veri setinin *doğruluk, kesinlik* ve *f1* olmak üzere 3 farklı başarım ölçütü için en başarılı sonuçları sağlayan özelliklerin belirlendiği 21 sonuç elde edilmiştir. Elde edilen bu en başarılı 21 sonuç içinde en sık kullanılan özelliklerden A4: 18 kez, A3: 15 kez, A2: 14 kez, A7: 13 kez, A9: 11 kez, A10: 11 kez, A12: 8 kez, A6: 7 kez, A1 ve A11: 5 kez ve A5: 4 kez kullanılmıştır.

Uygulama 3'te de olduğu gibi genel olarak bakıldığı zaman tüm başarım ölçütleri için en başarılı sonuçlar C1 koleksiyonunda elde edilmiştir. Özelliklerin A1, A2, A3, A4, A6, A7, A9, A10 ve A12 seçilmesi durumunda 535 adet örneğin 475 tanesi doğru kümelenirken 60 tanesi yanlış kümelenmiştir. Ortalama *doğruluk* ve ortalama *f1* ölçütlerinin en yüksek değerlerini aldığı bu durumda elde edilen ortalama *doğruluk* değeri 0.8879, ortalama *kesinlik* değeri 0.9444 ve ortalama *f1* değeri 0.9042 olarak hesaplanmıştır. Ayrıca yanlış kümelenen 60 örneğin üyelik derecelerine bakıldığı zaman da 3 tane örneğin bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki farkın 0.05'ten daha az olduğu gözlenmiştir. Ortalama *kesinlik* değerinin en yüksek değerini aldığı durum olan özelliklerin A2, A3, A4, A5, A9 ve A10 seçilmesi durumunda toplam 535 örnekten 464 tanesi doğru kümelerinirken, 71 tanesi yanlış kümelenmiştir. Bu koşullarda elde edilen ortalama *doğruluk* değeri 0.8673, ortalama *kesinlik* değeri 0.9465 ve ortalama *f1* değeri 0.8896'dır. Bu koşullarda 71 örnek yanlış kümelenmiş olsada bu 71 örneğin 21 tanesinin bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten daha azdır. Tablo 4.75'te yer alan 3 örnek ve Tablo 4.77'de yer alan 21 örnek aslında yanlış kümelenmiş olsalarda bu örneklerin bulunduğu kümenin üyelik derecesiyle olması gereken kümenin üyelik derecesi arasındaki fark 0.05'ten daha azdır. Başka bir deyişle bu örnekler her iki kümeye de yaklaşık aynı derecede yakınlıktadırlar.

Uygulama 3'te olduğu gibi bu uygulamada da en başarılı sonuçların C1 koleksiyonundaki örnekler üzerinde elde edilmesinin nedeni bu koleksiyonun özelliğinden kaynaklanmaktadır. Bu veri seti siyasi partilerin resmi Twitter

hesaplarının arkadaş listesinde yer alan kullanıcılardan oluşmaktadır, başka bir deyişle bu kullanıcılar genellikle partinin millet vekilleri, bakanları ve diğer siyasetçilerinden oluşmaktadır, bu yüzden de partiye benzerlikleri oldukça yüksektir. Ancak burada dikkat çeken durum Tablo 4.75'teki 3 örnekte ve Tablo 4.77'deki 21 örnekte de görüldüğü gibi bu örnekler diğer kümelere ait üyelik dereceleri yüksek olduğu için yanlış kümelenmişlerdir. Aslında bu şu anlama gelmektedir, bu veri setinde yer alan örnekler siyasi partilerin temsilcileri, bakanlar, milletvekilleri vs. olmasına rağmen bazı örnekler diğer kümelere de oldukça yakındırlar. Diğer tekli veri setleri arasında ise sırasıyla C2 için özelliklerin A4 ve A7 seçilmesi durumunda ortalama *doğruluk* değeri 0.8845, ortalama *kesinlik* değeri 0.8920 ve ortalama *f1* değeri 0.8843 olarak, özelliklerin A3, A4, A7 ve A10 seçilmesi durumunda diğer ölçütler aynı kalırken, ortalama *kesinlik* değeri 0.8923 olarak hesaplanmıştır.

C3 için özelliklerin A2, A6, A7, A9, A11 ve A12 olması durumunda ortalama *doğruluk* değeri 0.5839, ortalama *kesinlik* değeri 0.7419 ve ortalama *f1* değeri 0.5977 olarak ve yine C3 için özelliklerin A4, A7, A10 ve A12 olması durumunda ortalama *doğruluk* değeri 0.4969, ortalama *kesinlik* değeri 0.7574 ve ortalama *f1* değeri 0.5151 olarak hesaplanmıştır. Burada da Uygulama 3'te görüldüğü gibi tekli veri setleri için C1'den sonraki en başarılı sonuç C2'ye, en başarısız sonuç C3'e aittir. Bunun nedeni yine koleksiyonların özelliklerinden kaynaklanmaktadır. C2 veri koleksiyonu siyasi partilerin takipçi listesinde yer alan ve yalnızca tek bir partiyi takip eden kullanıcı verilerinden oluşan bir veri setiyken, C3 rastgele seçilen kullanıcı verilerinden oluşan bir veri setidir. C2 veri seti için iki durumda da toplam 701 örnekten 81 tanesi yanlış kümelenirken bu 81 örneğin içinden Tablo 4.80 ve Tablo 4.82'de de görüldüğü gibi 29 örneğin buluduğu kümenin üyelik dereceleriyle olması gereken kümenin üyelik derecesi birbirlerine oldukça yakındır.C2 veri seti oluşturulurken sadece bir siyasi partiyi takip eden kullanıcı verileri toplanmıştır ancak bu 29 örneğin diğer kümelere de oldukça yakın oldukları görülmektedir.

İkili veri setleri arasında elde edilen en başarılı sonuçlar yine Uygulama 3'te olduğu gibi C1 ve C2 veri seti kullanılarak elde edilmiştir. Bu veri setlerinin birlikte ele alındığı durumda ortalama *doğruluk* ve ortalama *f1* değeri açısından en başarılı sonuç A2, A3, A4, A9, A10 ve A12 özelliklerinin seçildiği durumda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8568, ortalama *kesinlik* değeri

0.8860 ve ortalama *f1* değeri 0.8594'tür. Bu koşullar altında toplam 1236 örnekten 1059 tanesi doğru kümelenirken 177 tanesi yanlış kümeler içerisinde yer almıştır. Bu 177 örnekten 50 tanesinin Tablo 4.90'da da gösterildiği gibi olması gereken kümelere ait üyelik dereceleri bulundukları kümelerin üyelik derecelerine çok yakındır.

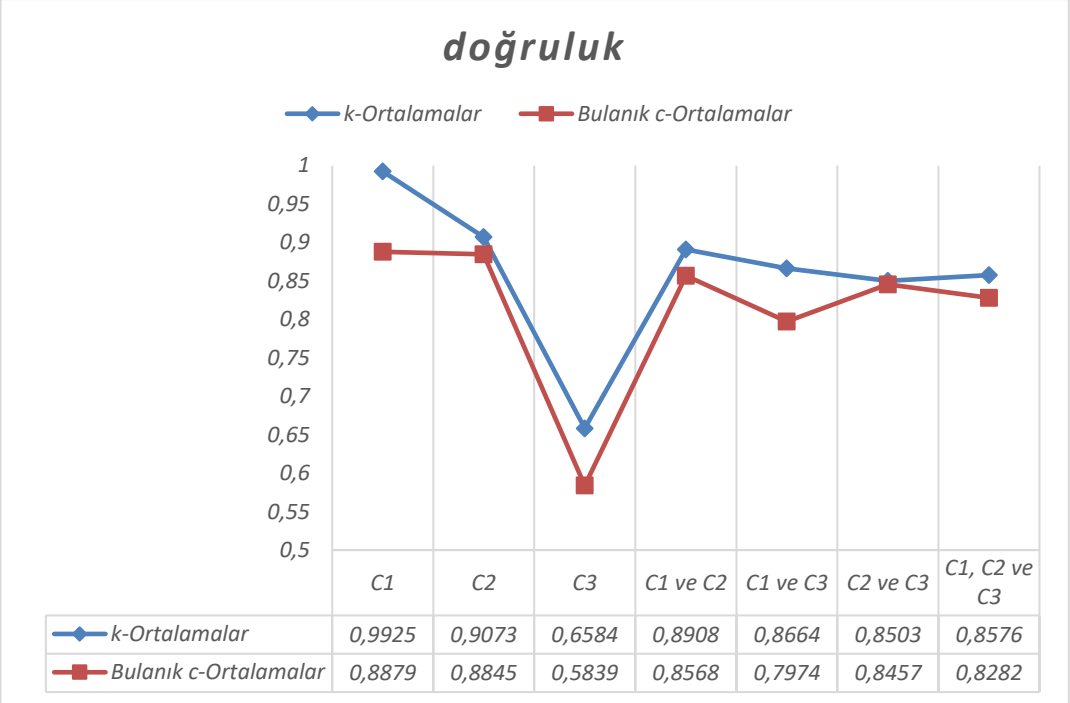
C1, C2 ve C3 veri setlerinin üçünün birlikte kullanılmasıyla tüm ölçütler için elde edilen en başarılı sonuçlar özellik olarak A2, A3, A4, A9 ve A10'un seçildiği durumda elde edilmiştir. Bu durumda elde edilen ortalama *doğruluk* değeri 0.8282, ortalama *kesinlik* değeri 0.8614 ve ortalama *f1* değeri 0.8321'dir. Bu koşullar altında toplam 1397 örnekten 1157 tanesi doğru kümelenirken 240 tanesi yanlış kümelenmiştir. Bu 240 tane örnekten de 110 tanesi Tablo 4.101'de de görüldüğü gibi yanlış kümelenmesine rağmen bulunduğu kümeyle olması gereken kümenin üyelik dereceleri birbirlerin çok yakın olan örneklerdir.

### 4.3.3 Kümeleme Sonuçları

Çalışma kapsamında önerilen özelliklerin kümeleme çalışmalarındaki başarımını test etmek için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar algoritmaları kullanılarak Bölüm 4.3.1 ve Bölüm 4.3.2'de yer alan uygulamalar gerçekleştirilmiştir.

Genel olarak bakıldığı zaman Uygulama 3'te geleneksel kümeleme yöntemlerinden birisi olan *k*-Ortalamalar yönteminin kullanılmasıyla elde edilen kümeleme sonuçları Uygulama 4'te Bulanık *c*-Ortalamalar kullanılarak elde edilen sonuçlardan daha başarılı sonuçlar vermiştir. Ancak Bulanık *c*-Ortalamalar yönteminin, *k*-Ortalamalar yöntemine göre faydası örneklere her bir küme için üyelik derecesi atamasıdır. Buna bağlı olarak üyelik derecelerine bakıldığı zaman yanlış kümelenen örneklerden birçoğunun aslında olması gereken küme üyelik derecelerinin bulunduğu küme üyelik derecesine çok yakın olduğu görülmektedir. Bu yöntemin başka bir olumlu tarafı da tüm kümelere yakın olan örneklerin tespit edilebilmesini sağlamasıdır. Bunun sağladığı yarar da buradaki kümeler siyasi partileri temsil ettiği için bu örnekler aslında diğerlerine göre siyasi görüşleri daha kolay değiştirilebilecek örnekler olarak tanımlanabilir.

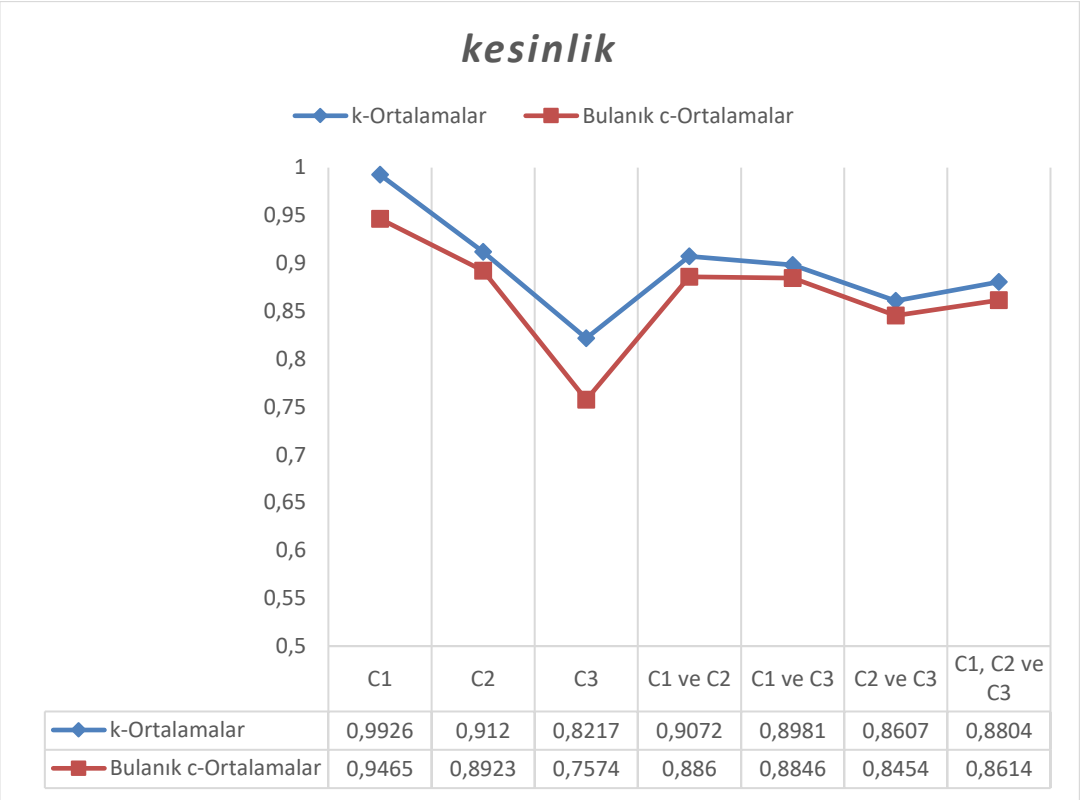
Şekil 4.19'da Uygulama 3 ve Uygulama 4'te kullanılan *k*-Ortalamalar ve Bulanık *c*-Ortalamalar algoritmalarının ortalama *doğruluk* değerine göre karşılaştırıldığı grafik görülmektedir.




| | C1 | C2 | C3 | C1 ve C2 | C1 ve C3 | C2 ve C3 | C1, C2 ve C3 |
|---|---|---|---|---|---|---|---|
| k-Ortalamalar | 0,9925 | 0,9073 | 0,6584 | 0,8908 | 0,8664 | 0,8503 | 0,8576 |
| Bulanık c-Ortalamalar | 0,8879 | 0,8845 | 0,5839 | 0,8568 | 0,7974 | 0,8457 | 0,8282 |

Şekil 4.19: C1, C2 ve C3 veri setlerinin farklı kombinasyonları için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle elde edilen kümelemelerin karşılaştırmalı ortalama *doğruluk* değerleri

Şekil 4.19'da *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle farklı veri setleri üzerinde elde edilen ortalama *doğruluk* değerlerinin karşılaştırması görülmektedir. *k*-Ortalamalar algoritmasıyla veri seti olarak C2 ve C3'ün birlikte seçildiği ve C1, C2, C3'ün birlikte seçildiği durumlar dışında Bulanık *c*-Ortalamalar'a göre daha yüksek ortalama *doğruluk* değerleri elde edilmiştir. Her iki algoritma için de ortalama *doğruluk* açısından en başarılı sonuçlar C1 veri setiyle elde edilirken en başarısız sonuçlar C3 veri setiyle elde edilmiştir. Bunun nedeni C1'in siyasi partilerin resmi Twitter hesaplarının arkadaş listesindeki kullanıcılardan yani siyasetçiler, milletvekilleri, bakanlar gibi siyasi partinin resmi Twitter hesabıyla çok benzer yapıya sahip kullanıcı hesaplarına ait verilerden oluşmasıdır. Bunun yanında en başarısız sonuçların elde edildiği C3 veri seti ise rastgele seçilen kullanıcı verilerinden oluşmaktadır.

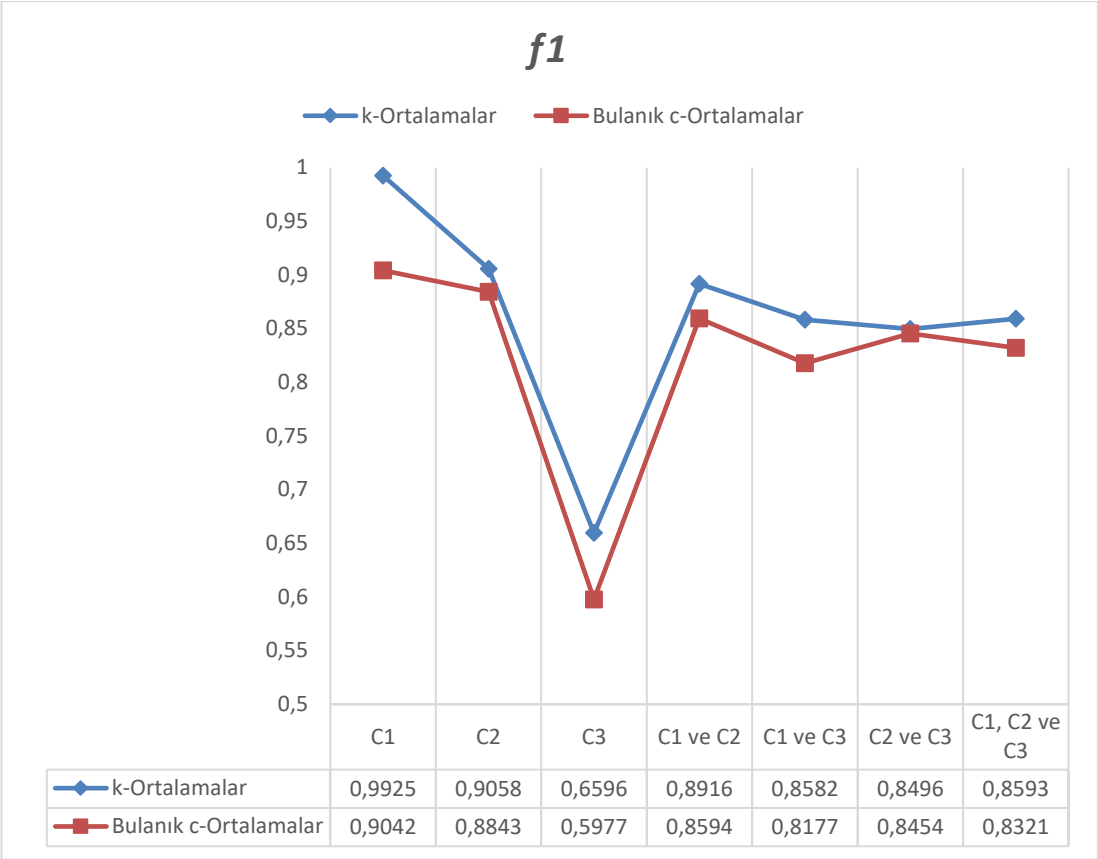



| | C1 | C2 | C3 | C1 ve C2 | C1 ve C3 | C2 ve C3 | C1, C2 ve C3 |
|---|---|---|---|---|---|---|---|
| k-Ortalamalar | 0,9926 | 0,912 | 0,8217 | 0,9072 | 0,8981 | 0,8607 | 0,8804 |
| Bulanık c-Ortalamalar | 0,9465 | 0,8923 | 0,7574 | 0,886 | 0,8846 | 0,8454 | 0,8614 |

Şekil 4.20: C1, C2 ve C3 veri setlerinin farklı kombinasyonları için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle elde edilen kümelemelerin karşılaştırmalı ortalama *kesinlik* değerleri

*k*-Ortalamalar ve Bulanık *c*-Ortalamalar algoritmalarının farklı veri setlerine göre gösterdikleri performasın ortalama *kesinlik* değeri açısından karşılaştırılması Şekil 4.19'da görülmektedir. Her iki algoritma içinde en yüksek ortalama *kesinlik* değerleri C1'in veri seti olarak kullanılması durumunda elde edilmiştir. Bu durumda *k*-Ortalamalar ile elde edilen ortalama *kesinlik* değeri 0.9926 iken Bulanık *c*-Ortalamalar ile elde edilen değer 0.9465'tir.

Yöntem açısından bakıldığı zaman da tüm veri setlerinde *k*-Ortalamalar algoritmasıyla, Bulanık *c*-Ortalamalar'a göre daha yüksek ortalama *kesinlik* değeri elde edilmiştir.




| | C1 | C2 | C3 | C1 ve C2 | C1 ve C3 | C2 ve C3 | C1, C2 ve C3 |
|---|---|---|---|---|---|---|---|
| k-Ortalamalar | 0,9925 | 0,9058 | 0,6596 | 0,8916 | 0,8582 | 0,8496 | 0,8593 |
| Bulanık c-Ortalamalar | 0,9042 | 0,8843 | 0,5977 | 0,8594 | 0,8177 | 0,8454 | 0,8321 |

Şekil 4.21: C1, C2 ve C3 veri setlerinin farklı kombinasyonları için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle elde edilen kümelemelerin karşılaştırmalı ortalama *kesinlik* değerleri

Ortalama *f1* değerleri açısından her iki algoritmanın karşılaştırma sonuçları Şekil 4.21'de görülmektedir. Diğer ölçütlerde olduğu gibi ortalama *f1* içinde en başarılı sonuç C1 veri setiyle elde edilmiştir.

Yöntem açısından bakıldığı zaman veri setinin C2 ve C3'ün birlikte kullanılması durumu dışında *k*-Ortalamalar ile daha başarılı sonuçlar elde edilmiştir.

Sonuç olarak tüm bu testler göz önünde bulundurulduğunda yalnızca kümeleme başarımı dikkate alındığında C2 ve C3'ün birlikte veri seti olarak kullanıldığı durumda ortalama *doğruluk* ve ortalama *f1* için Bulanık *c*-Ortalamalar algoritmasının, diğer tüm durumlarda *k*-Ortalamalar algoritmasının kullanılması önerilmektedir.

Bulanık *c*-Ortalamalar algoritması genel olarak *k*-Ortalamalar'a göre daha başarısız sonuçlar elde etse de bu yöntem kümeleme sonucunda her bir örneğin her bir kümeye ait üyelik derecelerini vermektedir. Bu sayede de yanlış kümelenen örneklerin

hangi kümelere de yakın olduğu ya da tüm kümelere yakın olan örneklerin tespit edilmesi gibi işlemler kolaylıkla gerçekleştirilebilmektedir.

# 5. SONUÇ VE ÖNERİLER

Bu çalışmada, temeli sosyal ağ analizi kavramlarına dayanan, sınıflandırma ve kümeleme gibi veri madenciliği uygulamalarında kullanılabilecek bir dizi özellikler önerilmiştir. Bu özelliklerle amaçlanan, sosyal medya kullanıcılarının benzerliklerinin nicel bir şekilde ifade edilebileceği yeni metrikler tanımlayabilmektir. Tablo 4.2'de ayrıntılı şekilde açıklanan bu özelliklerin temeli, bir hedef düğümü oluşturan kullanıcıyla benzerliği hesaplanacak kullanıcının arkadaş listelerinin, takipçi listelerinin ve hatta ikinci derece arkadaş ve takipçi listelerinin kesişimlerine dayanmaktadır.

Önerilen özelliklerin başarımını test etmek için Twitter üzerinde kullanıcıların siyasi eğilimlerine yönelik analizler gerçekleştirilmiştir. *k*-NN ve CART algoritmaları kullanılarak yapılan sınıflandırma çalışmaları Bölüm 4.2'de, *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleri kullanılarak yapılan kümelendirme çalışmaları da Bölüm 4.3'te ayrıntılı olarak açıklanmıştır. Ayrıca *k*-NN algoritmasında en uygun *k* değerinin belirlenebilmesi için tüm veri seti ve özellik kombinasyonları *k*=[1, 10] aralığında bir Greedy Araması gerçekleştirilmiş ve en yüksek başarım değerlerinin elde edildiği *k* değerleri tespit edilmiştir. CART algoritmasının kullanıldığı analizlerde ise karar ağacının dallarının ayrımı sırasında maliyet fonksiyonu olarak gini kullanılırken, ağacın çok fazla büyümesini ve dallanmasını önlemek için maksimum derinlik değeri 6 olarak ayarlanmıştır.

Veri setleri arasında yer alan C1 siyasi partilerin arkadaş listesinden elde edilen Twitter kullanıcılarının, C2 siyasi partilerin takipçi listesinden elde edilen Twitter kullanıcılarının, C3 ise Pamukkale Üniversitesi'nin resmi Twitter hesabının takipçi listesinden rastgele seçilen kullanıcıların önerilen 13 özellik değerlerini içermektedir. Sınıflandırma uygulamalarında C1 ve C2 veri setlerinin farklı kombinasyonları eğitim veri seti olarak kullanılırken, C3 eğitim veri seti olarak kullanılmıştır. Kümeleme uygulamalarında ise her üç veri seti farklı kombinasyonlarla kullanılarak analizler gerçekleştirilmiştir.

Sınıflandırma için *k*-NN ve karar ağacı yöntemlerinin Tablo 4.1'de yer alan C1 ve C2 veri setlerinin farklı kombinasyonlarının eğitim, C3'ün ise test veri seti olarak kullanıldığı analizler yapılmıştır. Bu analizler sonucunda *k*-NN için en başarılı

sınıflandırma 0.77 *doğruluk* ile elde edilirken, karar ağacı için 0.75 *doğruluk* elde edilmiştir. Benzer şekilde kümeleme için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemlerinin kullanıldığı, *k* ve *c* değerlerinin 3 olarak kabul edildiği analizler yapılmıştır. Analizler sonucunda elde edilen en başarılı *doğruluk* değerleri *k*-Ortalamalar için 0.99 iken Bulanık *c*-Ortalamalar için 0.88'dir. Ancak bu değerler veri seti olarak C1'in kullanıldığı durumda elde edilen değerlerdir. C1 veri seti siyasi partilerin arkadaş listesinde yer alan kullanıcılara ait verilerden oluşmaktadır, dolayısıyla siyasi partilerle benzerlikleri oldukça yüksektir. Yalnızca C3 yani rastgele toplanan verilerden oluşan veri setinin kullanıldığı durumda elde edilen *doğruluk* değerleri de *k*-Ortalamalar için 0.65 iken Bulanık *c*-Ortalamalar için 0.58'dir. Yapılan analizler sonucunda her bir yöntem için en yüksek başarım değerleri farklı veri setleriyle farklı özellikler kullanılarak elde edilmiştir. Bu farklı veri seti ve farklı özellik kombinasyonlarının her bir yöntem ile test edilmesi sonucunda elde edilen *doğruluk* değerleri Tablo 5.1'de gösterilmiştir. Örneğin; *k*-NN ile C1 veri setinin eğitim veri seti, C3'ün test veri seti olarak kullanıldığı durumda en yüksek *doğruluk* değeri A3 ve A10 özelliklerinin seçilmesi durumunda *k*=4 için 0.7516 olarak hesaplanmıştır. Yine benzer şekilde A3 ve A10 özelliklerinin seçildiği durumlarda CART algoritması için eğitim veri setinin C1, test veri setinin C3 olması durumunda 0.4472, *k*-Ortalamalar için veri setinin C1 olduğu durumda 0.9570, aynı durumda Bulanık *c*-Ortalamalar için 0.5963 *doğruluk* değerleri elde edilmiştir. Ayrıca bu tablonun YKÖ sütunu Bulanık *c*-Ortalamalar yöntemiyle yanlış kümelenmesine rağmen ait olduğu küme üyelik derecesiyle olması gereken küme üyelik derecesi arasındaki fark 0.05'ten az olan küme sayısının yanlış kümelenen örnek sayısına oranını göstermektedir. CART algoritması için eğitim veri setinin C1, test veri setinin C3 seçilmesi durumunda elde edilen en yüksek *doğruluk* değeri Tablo 5.1'de de görüldüğü gibi 0.7578'dir. Bu değer özellik olarak A2, A5 ve A13'ün seçilmesi durumunda elde edilmiştir, aynı koşullarda *k*-NN algoritmasıyla elde edilen en yüksek *doğruluk* değeri *k*=3 için 0.5652'dir. Kümeleme yöntemleri için aynı özelliklerle C1'in veri seti olarak kullanılması durumunda elde edilen en yüksek *doğruluk* değerleri, *k*-Ortalamalar ile 0.8579, Bulanık *c*-Ortalamalar ile 0.5364 olarak hesaplanmıştır.

Sınıflandırma ve kümeleme sonuçlarına ayrı ayrı bakıldığı zaman, C2'nin eğitim veri seti, C3'ün test veri seti olarak seçildiği durumda A1, A4, A7, A8, A10 ve A11 özelliklerinin birlikte kullanılması durumunda *k*-NN ile *k*=1 için 0.6894, CART

ile 0.7267 *doğruluk*, C1 ile C2'nin birlikte eğitim veri seti, C3'ün test veri seti olarak seçildiği durumda A2, A4, A10 ve A11 özelliklerinin birlikte kullanılması durumunda $k$=2 için $k$-NN ile 0.6522, CART ile 0.7081 *doğruluk* elde edilmiştir.

Tablo 5.1: En yüksek *doğruluk* değerlerinin elde edildiği durumlarda tüm yöntemlerin başarım değerleri

| **koleksiyon** | **özellik** | ***k*-NN** | | **CART** | ***k*-Ortalamalar** | **Bulanık *c*-Ortalamalar** | |
|---|---|---|---|---|---|---|---|
| | | ***k*** | ***doğruluk*** | ***doğruluk*** | ***doğruluk*** | ***doğruluk*** | **YKÖ[1]** |
| C1 | A3 ve A10 | 4 | 0.7516 | 0.4472 | 0.9570 | 0.5963 | 5/216 |
| | A2, A5 ve A13 | 3 | 0.5652 | 0.7578 | 0.8579 | 0.5364 | 3/248 |
| | A2, A3, A4, A8, A9 ve A12 | 4 | 0.5590 | 0.3416 | 0.9925 | 0.8673 | 17/71 |
| | A1, A2, A3, A4, A6, A7, A9, A10 ve A12 | 1 | 0.5590 | 0.5839 | 0.8729 | 0.8879 | 3/60 |
| C2 | A9 | 6 | 0.7391 | 0.5963 | 0.3923 | 0.3723 | 8/274 |
| | A1, A4, A7, A8, A10 ve A11 | 1 | 0.6832 | 0.7267 | 0.9073 | 0.8845 | 29/81 |
| | A2, A4 ve A7 | 1 | 0.6894 | 0.6400 | 0.9073 | 0.8845 | 29/81 |
| | A4 ve A7 | 2 | 0.6025 | 0.5963 | 0.9073 | 0.8845 | 29/81 |
| C3 | A7 | | | | 0.6584 | 0.4783 | 35/84 |
| | A2, A6, A7, A9, A11 ve A12 | | | | 0.6584 | 0.5839 | 18/67 |
| C1 ve C2 | A9 ve A11 | 10 | 0.7702 | 0.5714 | 0.6513 | 0.4903 | 15/630 |
| | A2, A4, A10 ve A11 | 2 | 0.6522 | 0.7081 | 0.8778 | 0.8439 | 76/193 |
| | A4 ve A7 | 6 | 0.6149 | 0.5031 | 0.8908 | 0.8204 | 74/222 |
| | A2, A3, A4, A9, A10 ve A12 | 1 | 0.5342 | 0.6646 | 0.8738 | 0.8568 | 50/177 |

[1] YKÖ: Yanlış kümelenmesine rağmen bulunduğu küme ile olması gereken kümenin üyelik dereceleri arasındaki fark 0.05'ten az olan örnek sayısının toplam kümelenen örnek sayısına oranı

Tablo 5.1 (devam): En yüksek *doğruluk* değerlerinin elde edildiği durumlarda tüm yöntemlerin başarım değerleri

| **koleksiyon** | **özellik** | ***k*-NN** | | **CART** | ***k*-Ortalamalar** | **Bulanık *c*-Ortalamalar** | |
|---|---|---|---|---|---|---|---|
| | | ***k*** | ***doğruluk*** | ***doğruluk*** | ***doğruluk*** | ***doğruluk*** | **YKÖ** |
| C1 ve C3 | A2, A3, A8, A9, A12 ve A13 | | | | 0.8664 | 0.7040 | 15/206 |
| | A3, A4 ve A7 | | | | 0.8549 | 0.7974 | 1/141 |
| C2 ve C3 | A4 ve A7 | | | | 0.8503 | 0.8399 | 24/138 |
| | A1, A2, A3, A4, A5, A6 A7 ve A11 | | | | 0.8503 | 0.8457 | 13/133 |
| C1, C2 ve C3 | A4, A6 ve A7 | | | | 0.8503 | 0.7946 | 65/287 |
| | A2, A3, A4, A9 ve A10 | | | | 0.8261 | 0.8282 | 110/240 |

Kümeleme için yapılan analizlere bakıldığı zaman, rastgele seçilen verilerden oluşan C3'ün veri seti olarak seçildiği durumda özellik olarak A7'nin kullanılması durumunda *k*-Ortalamalar ile 0.6584, Bulanık *c*-Ortalamalar ile 0.4783, aynı veri setiyle özellik olarak A2, A6, A7, A9, A11 ve A12'nin kullanılması durumunda *k*-Ortalamalar ile yine 0.6584 *doğruluk* elde edilirken, Bulanık *c*-Ortalamalar ile 0.5839 *doğruluk* elde edilmiştir.

C1, C2 ve C3 veri setlerinin farklı kombinasyonlarının önerilen özellikler kullanılarak *k*-NN, CART, *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle elde edilen ortalama *doğruluk* değerleri Tablo 5.2'de gösterilmiştir. Tabloda sınıflandırma algoritmalarında yalnız C1, yalnız C2 ve C1 ile C2 birlikte eğitim veri seti olarak kullanılırken, C3 test veri seti olarak kullanılmış ve her bir eğitim veri seti kombinasyonu için en yüksek ortalama *doğruluk* değerleri, bu değerlerin elde edildiği özellikler ve *k*-NN için en uygun *k* değeri listelenmiştir. Benzer şekilde C1, C2 ve C3 veri setlerinin 7 farklı kombinasyonu için *k*-Ortalamalar ve Bulanık *c*-Ortalamalar yöntemleriyle elde edilen en yüksek ortalama *doğruluk* değerine sahip kümeleme sonuçları ve bu sonuçların elde edildiği özellikler de bu tabloda listelenmiştir. Tablo 5.2'de ortalama *doğruluk* için listelenen değerler sırasıyla Tablo 5.3'te ortalama *kesinlik*, Tablo 5.4'te ise ortalama *f1* ölçütü için listelenmiştir. Bu tablolar sayesinde seçilen yöntem ve veri setine göre en uygun özellik ve parametreler, seçilen veri setine göre en uygun yöntem ve özellikler açısından çıkarımlar yapılabilmektedir. Örneğin sınıflandırma için en yüksek ortalama *doğruluk* ve ortalama *f1* *k*-NN yöntemiyle *k*=10 için eğitim veri seti olarak C1 ile C2'nin birlikte kullanıldığı, özellik olarak da A2, A4, A10 ve A11'in seçildiği durumda, en yüksek ortalama *kesinlik* bir karar ağacı yöntemi olan CART algoritmasıyla veri seti olarak C1'in kullanıldığı, özellik olarak da A1 ve A7'nin seçildiği durumda elde edilmiştir. Kümeleme uygulamalarında tüm başarım ölçütleri açısından en yüksek değerler C1 veri seti üzerinde A2, A3, A4, A8, A9 ve A12 özelliklerinin kullanılmasıyla *k*-Ortalamalar yöntemiyle elde edilmiştir. C1 veri seti içerisinde yer alan örnekler siyasi partilerin arkadaş listelerinde yer alan kullanıcılara ait verilerden oluşmaktadır ki bu kişiler de genellikle o partinin temsilcileri, siyasetçileri veya o parti ile öne çıkan ve gündeme gelen isimlerdir. Bundan dolayı bu kişilerin arkadaş listesinde bulundukları parti ile benzerlikleri oldukça yüksektir. Yine benzer şekilde en yüksek başarım değerlerinin elde edildiği özelliklere bakıldığı zaman, A2 ve A3'ün siyasi partinin arkadaş listesiyle, seçilen

örneğin arkadaş listesinin ne kadar kesiştiğini gösteren bir benzerlik ölçütü olduğu, A4'ün o örnek kullanıcının o siyasi partiyi takip edip etmediğini gösteren bir değişken olduğu, A8 ve A9, örnek kullanıcının arkadaş listesinin, siyasi partinin arkadaş listesindeki her bir kullanıcının arkadaş listesiyle ne kadar kesiştiğini gösteren bir benzerlik ölçütü olduğu ve A12'nin ise siyasi partinin takipçi listesindeki her bir kullanıcının arkadaş listesiyle, o örnek kullanıcının arkadaş listesinin ne kadar kesiştiğini gösteren bir benzerlik ölçütü olduğu görülmektedir. Daha önce de bahsedildiği üzere C1 veri seti siyasi partilerle oldukça benzer yapıya sahip kullanıcılardan oluşmaktadır ve bundan dolayı bu 6 özellik değerinin oldukça yüksek çıkması da beklenen bir sonuçtur. Rastgele seçilmiş örneklerden oluşan C3 veri seti üzerinde ortalama *doğruluk* için en yüksek değer A7 özelliği kullanılarak, ortalama *kesinlik* için A2, A5, A6, A8, A11 ve A13 özellikleri kullanılarak, *f1* için yine A7 özelliği kullanılarak *k*-Ortalamalar yöntemiyle elde edilmiştir. Bulanık *c*-Ortalamalar yönteminde, *k*-Ortalamalar yöntemine göre daha düşük başarım elde edilse de bu yöntemin avantajı örneklerin küme üyelik derecelerini gösteriyor olmasıdır. Bu sayede yanlış kümelenen örneklerin üyelik derecelerine bakılarak çeşitli yorum ve tahminlemeler yapılabilmektedir.

Tablo 5.2: Seçilen veri seti ve yönteme bağlı olarak elde edilen *doğruluk* değerleri

| **koleksiyonlar** | **Sınıflandırma** | | | | | **kümeleme** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***k*-NN** | | | **CART** | | ***k*-Ortalamalar** | | **Bulanık *c*-Ortalamalar** | | |
| | ***doğ.*** | **özellik** | ***k*** | ***doğ.*** | **özellik** | ***doğ.*** | **özellik** | ***doğ.*** | **özellik** | **YKÖ** |
| **C1** | 0.7516 | A3, A10 | 4 | 0,7578 | A2, A5, A13 | 0,9925 | A2, A3, A4, A8, A9, A12 | 0,8879 | A1, A2, A3, A4, A6, A7, A9, A10, A12 | 3/60 |
| **C2** | 0,7391 | A9 | 6 | 0,7267 | A1, A4, A7, A8, A10, A11 | 0,9073 | A2, A4, A7 | 0,8845 | A4, A7 | 29/81 |
| **C3** | - | - | - | - | - | 0,6584 | A7 | 0,5839 | A2, A6, A7, A9, A11, A12 | 18/67 |
| **C1 ve C2** | 0,7702 | A9, A11 | 10 | 0,7081 | A2, A4, A10, A11 | 0,8908 | A4, A7 | 0,8568 | A2, A3, A4, A9, A10, A12 | 50/177 |
| **C1 ve C3** | - | - | - | - | - | 0,8664 | A2, A3, A8, A9, A12, A13 | 0,7974 | A3, A4, A7 | 1/141 |
| **C2 ve C3** | - | - | - | - | - | 0,8503 | A4, A7 | 0,8457 | A1, A2, A3, A4, A5, A6, A7, A11 | 13/133 |
| **C1, C2 ve C3** | - | - | - | - | - | 0,8576 | A4, A6, A7 | 0,8282 | A2, A3, A4, A9, A10 | 110/240 |

Tablo 5.3: Seçilen veri seti ve yönteme bağlı olarak elde edilen *kesinlik* değerleri

| **koleksiyonlar** | **sınıflandırma** | | | | | **kümeleme** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***k*-NN** | | | **CART** | | ***k*-Ortalamalar** | | **Bulanık *c*-Ortalamalar** | | |
| | ***kesinlik*** | **özellik** | ***k*** | ***kesinlik*** | **özellik** | ***kesinlik*** | **özellik** | ***kesinlik*** | **özellik** | **YKÖ** |
| **C1** | 0,8700 | A2, A3 | 1 | 0,8750 | A1, A7 | 0,9926 | A2, A3, A4, A8, A9, A12 | 0,9465 | A2, A3, A4, A5, A9, A10 | 21/71 |
| **C2** | 0,7740 | A4, A9, A12, A13 | 1 | 0,7993 | A2, A4, A5, A6, A8, A9, A10 | 0,9120 | A2, A4, A7 | 0,8923 | A3, A4, A7, A10 | 29/81 |
| **C3** | - | - | - | - | - | 0,8217 | A2, A5, A6, A8, A11, A13 | 0,7574 | A4, A7, A10, A12 | 1/81 |
| **C1 ve C2** | 0,8070 | A4 | 1 | 0,7775 | A4, A6, A7, A8, A10 | 0,9072 | A2, A4, A9, A10 | 0,8860 | A2, A3, A4, A9, A10, A12 | 50/177 |
| **C1 ve C3** | - | - | - | - | - | 0,8891 | A4, A5, A12 | 0,8846 | A8 | 31/281 |
| **C2 ve C3** | - | - | - | - | - | 0,8607 | A2, A4, A10, A12, A13 | 0,8454 | A1, A2, A3, A4, A5, A6, A7, A11 | 13/133 |
| **C1, C2 ve C3** | - | - | - | - | - | 0,8804 | A2, A4, A6, A8, A9, A10 | 0,8614 | A2, A3, A4, A9, A10 | 110/240 |

Tablo 5.4: Seçilen veri seti ve yönteme bağlı olarak elde edilen *f1* ölçütü değerleri

| **koleksiyonlar** | **sınıflandırma** | | | | | **kümeleme** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***k*-NN** | | | **CART** | | ***k*-Ortalamalar** | | **Bulanık *c*-Ortalamalar** | | |
| | ***f1*** | **özellik** | ***k*** | ***f1*** | **özellik** | ***f1*** | **özellik** | ***f1*** | **özellik** | **YKÖ** |
| **C1** | 0,7454 | A10, A13 | 9 | 0,7597 | A2, A6, A7, A13 | 0,9925 | A2, A3, A4, A8, A9, A12 | 0,9042 | A1, A2, A3, A4, A6, A7, A9, A10, A12 | 3/60 |
| **C2** | 0,7442 | A9, A11 | 10 | 0,7263 | A1, A4, A7, A8, A10, A11 | 0,9058 | A2, A4, A7 | 0,8843 | A4, A7 | 29/81 |
| **C3** | - | - | - | - | - | 0,6596 | A7 | 0,5977 | A2, A6, A7, A9, A11, A12 | 18/67 |
| **C1 ve C2** | 0,7744 | A9, A11 | 10 | 0,7106 | A2, A4, A10, A11 | 0,8916 | A4, A7 | 0,8594 | A2, A3, A4, A9, A10, A12 | 50/177 |
| **C1 ve C3** | - | - | - | - | - | 0,8582 | A2, A3, A8, A9, A12, A13 | 0,8177 | A3, A4, A7 | 1/141 |
| **C2 ve C3** | - | - | - | - | - | 0,8496 | A4, A7 | 0,8454 | A1, A2, A3, A4, A5, A6, A7, A11 | 13/113 |
| **C1, C2 ve C3** | - | - | - | - | - | 0,8593 | A4, A6, A7 | 0,8321 | A2, A3, A4, A9, A10 | 110/240 |

Bu çalışmada Twitter'ın API ve zaman kısıtlarından dolayı ancak belirli sayıda örnek toplanarak bunlar üzerinde analizler gerçekleştirilmiştir. Gelecekte yapılacak analiz çalışmaları için her bir partinin resmi Twitter hesabının arkadaş ve takipçi listesinde yer alan tüm kullanıcılara ait bilgiler çekilerek oluşturulan ego ağlarının kullanılmasının başarıyı arttıracağı düşünülmektedir. Bu durum aynı zamanda yarı denetimli veya denetimli yöntemlerde eğitim veri setinde yer alan örnek sayısını ve çeşitliliği arttıracağından, denetimsiz yöntemlerde ise küme ve grupların merkezlerinin daha doğru belirlenmesini sağlayacağından daha başarılı sonuçlar elde edilebilir. Ayrıca kullanıcıların düğümleri, kullanıcılar arasındaki ilişkilerin de kenarları oluşturduğu bu sosyal ağda tüm ilişkiler var veya yok şeklinde eşit olarak kabul edilmiştir. Aslında gerçek hayatta insanların birbirleriyle olan yakınlıkları, samimiyetleri ve ilişki durumları farklılık göstermektedir. Bundan dolayı sosyal ağ analizinde yer alan yakınlık, arasındalık, merkezilik, yoğunluk ve yarıçap gibi kavramlardan yararlanılarak ilişkilerin sabit ya da değişken katsayılarla ağırlıklandırılması veya bulanıklaştırılması durumunda elde edilecek tahminleme sonuçlarının daha başarılı olacağı öngörülmektedir.

Önerilen bu özellikler, burada olduğu gibi sadece siyasi parti eğilimlerinin ya da gruplanmalarının tahmin edilmesi için değil aynı zamanda kullanıcıların favori spor kulüplerinin tahmin edilmesi, arkadaş gruplarının belirlenmesi, arkadaş ve içerik öneri sistemlerinin geliştirilmesi, reklam ve pazarlama için hedef kitlenin belirlenmesi gibi çalışmalara da uygulanabilir. Bu özellikler sadece Twitter için değil Facebook, Instagram vb. farklı sosyal medyalara da uygulanabilir.

Gelecekteki çalışmalarımızda bahsedilen bu iyileştirmelerin yanında Bölüm 4.2 ve Bölüm 4.3'te yer alan yöntemlerden farklı yöntemler kullanılarak ve bu çalışmada yer alan 3 siyasi partinin dışındaki diğer partileri de dahil ederek yapılan analizlerin genelleştirilmesi hedeflenmektedir.

# 6. KAYNAKLAR


Akarsu, C. ve Diri, B., "Turkish TV rating prediction with Twitter", *24th Signal Processing and Communication Application Conference(SIU)*, Zonguldak Turkey, 345-348, doi: 10.1109/SIU.2016.7495748, (2016).

Al, U., Sezen, U. ve Soydal, İ. "Hacettepe Üniversitesi bilimsel yayınlarının sosyal ağ analizi yöntemiyle değerlendirilmesi", *Hacettepe Üniversitesi Edebiyat Fakültesi Dergisi(JFL)*, 29(1), 53-71, (2012).

Alam, F., Stepanov, E. and Riccardi, G., "Personality traits recognition on social network - Facebook", *Computational Personality Recognition*, (2013).

Albornoz, M. R. G., "Voting intention analysis of Twitter users", Master Thesis, *Federico Maria Technical University*, Chile, (2013).

Anaconda, "What is Anaconda[online]", (07.06.2018), Web adresi: https://www.anaconda.com/what-is-anaconda/, (2018).

Ateş, Ü., "Inference of personality using social media profiles", Master Thesis, *Informatics Institute of Middle East Technical University*, Ankara, (2014)

Baykara, M. ve Gürtürk, U., "Sosyal medya paylaşımlarının duygu analizi yöntemiyle sınıflandırılması", *2017 International Conference on Computer Science and Engineering (UBMK)*, IEEE, Antalya Türkiye, 911-916, doi: 10.1109/UBMK.2017.8093536, (2017).

Bernard, J., *Python Recipes Handbook*, Berkeley: Apress, (2016).

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer. P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, Arnaud, Holt and B., Varoquaux, G. "API design for machine learning software: experiences from the scikit-learn project", *Europan Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECMLPKDD'13)*, Prague Czech Republic, (2013).

Burger, J. D., Henderson, J., Kim, G. and Zarrella, G., "Discriminating gender on Twitter", *Proceedings of Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh Scotland UK, 1301-1309, (2011).

Caruana, R. and Niculescu-Mizil, A., Data mining in metric space: an empirical analysis of supervised learning performance criteria“, *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, New York USA, 69-78, doi: 10.1145/1014052.1014063, (2004).

Choi, S., Park, J. and Woo, H., "Using social media data to explore communication processes within South Korean online innovation communities", *Scientometrics*, 90(1), 43-56, doi: 10.1007/s11192-011-0514-7, (2012).

Choudhury, M. D., Diakopoulos, N. and Naaman, M., "Unfolding the event landscape on Twitter: classification and exploration of user categories", *Proceedings of the 2012 ACM Conference on Computer Supported Cooperative Work(CSCW-2012)*, ACM, Seattle Washington, 241-244, doi: 10.1145/2145204.2145242, (2012).

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A. And Menczer, F., "Predicting the political alignment of Twitter users", *IEEE Third International Conference Social Computing*, IEEE, Boston MA USA, 192-199, doi: 10.1109/PASSAT/SocialCom.2011.34, (2011).

Coşkun, C. ve Baykal, A., "Veri madenciliği sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması", *Akademik Bilişim 2011*, Malatya Türkiye, (2011).

Çoban, Ö. ve Özyer, G., "Türkçe Twitter mesajları için LDA ile duygu sınıflandırması", *24th Signal Processing and Communication Application Conference (SIU)*, IEEE, Zonguldak Türkiye, doi: 10.1109/SIU.2016.7495693, (2016).

Decision_Tree, "Decision Tree[online]", (23.05.2018), Web adresi: http://scikit-learn.org/stable/modules/tree.html, (2017).

Doran, P. R., Doran, C. and Mazur, A., "Social network analysis as a method for analyzing interaction in collaborative online learning enviroments", *Journal of Systemics*, Cybernetics and Informatics, 9(7), 10-16, (2011).

Drobnjak, A., "Fuzzy clustering in social networks", Master Thesis, *University of Freibourg*, Freibourg, (2012).

Dülger, Ü., "Stratejik veri yönetiminin yatırımlar üzerindeki etkileri", Yüksek Lisans Tezi, *İstanbul Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul, (2015).

Elmas, Ç., *Yapay Zeka Uygulamaları*, Ankara: Seçkin Yayıncılık, 203-249 (2016).

Ellson J., Gansner, E., Koutsofios, L., North Srephen, C. And Woodhull, G., "Graphiz-Open Source Graph Drawing Tools", (eds: Mutzel, P., Jünger, M. and Leipert, S.), *Graph Drawing GD 2001 Lecture Notes in Computer Science*, 2265, Berlin Heidelberg: Springer, 483-484, (2002).

Firan, C. S., Nejdl, W. and Paiu, R., "The benefit of using tag-based profiles", *Web Conference 2007(LA-WEB 2007)*, IEEE, Santiago Chile, 32-41, doi: 10.1109/LA-Web.2007.13, (2007).

Golbeck, J., Robles, C. and Turner, K., "Predicting personality with social media", *The 29th CHI Conference on Human Factors in Computing Systems(CHI2011)*, ACM, Vancouver British Columbia Canada, 253-262, doi: 10.1145/1979742.1979614, (2011).

Golbeck, J., Robles, C. and Turner, K., "Predicting personality from Twitter", *Third International Conference on Social Computing 2011 (SocialCom 2011)*, IEEE, Boston MA USA, 149-156, doi: 10.1109/PASSAT/SocialCom. 2011.33, (2011).

Goldberg, D., Nichols, D., Oki, B. M. and Terry, D., "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, 35(12), 61-70, doi: 10.1145/138859.138867, (1992)

Gossart, C. and Özman, M., "Co-authorship networks in social sciences: The case of Turkey", *Scientometrics*, 78(2), 323-345, doi: 10.1007/s11192-007-1963-x, (2009).

Gürsakal, N., Tüzüntürk, S. ve Sert, F., "Sosyal ağ verilerinin kuvvet yasası olasılık dağılımına uygunluk analizi: Twitter örneği", *15. Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu (15.EYİ)*, Isparta Türkiye, 501-523, (2014).

Han, J., Pei, J. and Kamber, M. *Data Mining: Concepts and Techniques Third Edition*, Waltham: Elsevier, (2011).

Hansen, D. L., "Exploring social media relationships", *On the Horizon*, 19(1), 43-51, doi: 10.1108/10748121111107726, (2011).

Hung, C., Huang, Y., Hsu, J. Y. and Wu, D. K., "Tag-based user profilling for social media recommendation", *Proceedings of the Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, the 23rd*

*AAAI Conference on Artifical Intelligence (AAAI-08)*, Chicago Illinois, 49-55, (2008).

Hunter J., "matplotlib: a 2D graphics environment", *Computing in Science & Engineering*, 9(3), 90-95, doi: 10.1109/MCSE.2007.55, (2007).

İşeri, İ., Atasoy, Ö. F. ve Alçiçek, H., "Türkiye'deki telekomünikasyon firmalarının sosyal medya verisi kullanılarak duygu sınıflandırması", 2017 International Conference on Computer Science and Engineering (UBMK), IEEE, Antalya Türkiye, 1015-1019, doi: 10.1109/UBMK.2017.8093419, (2017).

Jeya, T. and Bala, V. M., "Socirank: Identifying and ranking prevalent newstopics using social media factors", *International Journal of Advance Research in Science and Engineering*, 7(1), 13-18, (2018).

Kadushin, C., *Undestanding Social Networks: Theories, Concepts, and Findins*, New York: Oxford University Press, (2012).

Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S. and Bhattacharyya, D. K., "Big data analtyics in bioinformatics: a machine learning perspective", *Journal of Latex Class Files,* 13(9), 1-20, (2014).

KMeans, "*k*-means clustering[online]", (07.06.2018), Web adresi: https://www.mathworks.com/help/stats/kmeans.html, (2018).

Leydesdorff, L., "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals", *Journal of the American Society for Information Science and Technology*, 58(9), 1303-1319, doi: 10.1002/asi.20614, (2007).

Lima, A., C., E., S. and de Castro, L., N., "A multi-label, semi-supervised classification approach applied to personality prediction in social media", *Neural Networks*, 58, 122-130, doi: 10.1016/j.neunet.2014.05.020, (2014).

Liu, H. and Maes, P., "InterestMap: Harvesting social network profiles for recommendations", *Proceedings of the Beyond Personalization (IUI2005)*, San Diego California USA, 54-59, (2005).
Lupton, D., *Digital Sociology*, London: Routledge, (2014).

Marres, N., *Digital Sociology*, Cambridge: Polity Press, (2017).

Matlab, "The Language of Technical Computing[online]", (08..06.2018), Web adresi: https://www.mathworks.com/help/matlab/index.html, (2018).

McCarthy, J. F. and Lehnert, W. G., "Using decision trees for coreference resolution", *Proceedings of the Fourteenth International Joint Conference on Artifical Intelligence (IJCAI'95)*, Montreal Quebec Canada, 1060-1065, (1995).

Milgram, S., "The small world problem", *Psychology Today*, 1(1), 60-67, (1967).

Mitchell, T., *Machine Learning*, New York: McGaw Hill Education, (1997).

Michlmayr, E. and Cayzer, S., "Learning user profiles from tagging data and leveraging them for personal(ized) information access", *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference(WWW2007)*, Banff Canada, 1-7, (2007).

Nanjekye, J., *Python 2 and 3 Compatiblity*, Berkeley: Apress, (2017).

Özkan, Y., *Veri Madenciliği Yöntemleri*, İstanbul: Papatya Yayıncılık Eğitim, (2016).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel V., Thirion, B., Grisel, O, Blondel, M., Prettenhofer, P., Weiss, R., Dubourg. V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, 12, 2825-2830, (2011).

Peersman, C., Daelemans, W. and Vaerenbergh, L. V, "Predicting age and gender in online social networks", *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents(SMUC'11)*, ACM, Glasgow Scotland UK, 37-44, doi: 10.1145/2065023.2065035, (2011).

Peng, S., Tseng, V. S., Liang, C. W. and Shan, M. K., "Emerging product topics prediction in social media without social structure information", *Proceedings of The Web Conference 2018 (WWW'18)*, Republic and Canton of Geneva Switzerland, 1661-1668, doi: 10.1145/3184558.3191625, (2018).

Pennacchiotti, M. and Popescu, A., M., "A machine learning approach to Twitter user classification", *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media(ICWSM-11)*, DBLP, Barcelona Catalonia Spain, 281-288, (2011).

Pratama B. Y. and Sarno R. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM", *2015 International Conference on Data and Software Engineering(ICoDSE)*, IEEE,Yogyakarta Indonesia, 170-174, doi: 10.1109/ICODSE.2015.7436992, (2015).

Prell, C., *Social Network Analysis: History Theory and Methodology*, London: SAGE, (2012).

PyMongo, "pymongo 3.6.1[online]", (06.06.2018), Web adresi: https://pypi.org/project/pymongo/, (2018).

Python, "What is Python? Executive Summary[online]", (08.06.2018), Web adresi: https://www.python.org/doc/essays/blurb/, (2018).

Qi, J., Yu, Y., Wang, L., Liu, J. ve Wang, Y., "An effective and efficient hierarchical k-means clustering algorithm", *International Journal of Distributed Sensor Networks*, 13(8), doi: https://doi.org/10.1177/1550147717728627, (2017).

Quinlan, J. R., "Induction of decision trees", *Machine Learning*, 1(1), doi: https://doi.org/10.1007/BF00116251, 81-106, (1986).

Rao, D., Yarowsky, D., Shreeevats, A. and Gupta, M., "Classifying latent user attributes in Twitter", *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents(SMUC'10)*, ACM, Toronto Ontario Canada, doi: 10.1145/1871985.1871993, 37-44, (2010).

Rao., D., Paul, M., Fink, C., Yarowsky, D., Oates, T. and Coppersmith, G., "Hierarchical Bayesian models for latent attribute detection in social media", *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media(ICWSM-11)*, Barcelona Spain, (2011).

Rigolin, V. H., "What is Twitter? How do I get started? Why should I become a user?", *Journal of the American Society of Echocardiography*, 31(3), A31-A32, doi: https://doi.org/10.1016/j.echo.2018.01.005, (2018).

Rosen, P and Kluemper, D., "The impact of the Big Five Personality Traits on the acceptance of social networking website", *Proceedings of Americas Conference on Information Systems(AMCIS 2008)*, Toronto Canada, 274-274, (2008).

Ross, T., J., *Fuzzy Logic with Engineering Applications Second Edition*, England: John Wiley & Sons Ltd., (2004).

Scott, J., *Social Network Analysis 4th Edition*, London: SAGE, (2017).

Sert, F., Tüzüntürk, S. ve Gürsakal, N., "NodeXL ile sosyal ağ analizi: #akademikzam örneği", *15. Uluslararası Ekonometri, Yöneylem Araştırmaları ve İstatistik Sempozyumu Bildiriler Kitabı*, Isparta, 464-482, (2014).

Stetco, A., Zeng, X., J. ve Keane, J., "Fuzzy C-means++: Fuzzy c-means with effective seeding initialization", *Expert Systems with Applications*, 45(1), 7541-7548, doi: https://doi.org/10.1016/j.eswa.2015.05.014, (2015).

Şavklı, N. E., *mongoDB*, İstanbul: Dikey Eksen 6-14, (2009).

Şeker, S. E., *İş Zekası ve Veri Madenciliği*, İstanbul: Cinius Yayınları, (2013).

Talebi, M. ve Köse, C., "Facebook yorumlarının analiziyle cinsiyet, yaş ve eğitim düzeti belirleme", *21st Signal Processing and Communications Applications Conference(SIU)*, IEEE, Haspolat Turkey, 1-4, doi: 10.1109/SIU.2013.6531599, (2013).

Taşçı, E. ve Onan, A., "k-En Yakın Komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi", *XVIII. Akademik Bilişim Konferansı(AB 2016)*, Aydın Türkiye, (2016).

Tuttle, H., "Facebook Scandal Raises Data Privacy Concerns", *Risk Management*, 65(5), 6-9, (2018).

Tweepy, "Introduction to tweepy, Twitter for Python[online]", (06.06.2018), Web adresi: https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/, (2013).

Van Der Walt, S., Colbert, S. C. And Varoquaux, G., "the numpy array: a structure for efficient numerical computation", *Computing in Science & Engineering*, 13(2), 22-30, doi: 10.1109/MCSE.2011.37, (2011).

Wani, I. A., *The Sociology: A Study of Society*, New Delhi: Education Publishing, (2014).

Zaki, M. J. and Jr., W. M., *Data Mining and Analysis Fundamental Concepts and Algorithms*, New York: Cambridge University Press, (2018).

Zeng, D., Chen, H., Lusch, R. and Li, S. H., “Social media analytics and intelligence”, *IEEE Intelligent Systems*, 25(6), 13-16, doi: 10.1109/MIS.2010.151, (2010).

# 7. ÖZGEÇMİŞ


Adı Soyadı : Emre ŞAHİN

Doğum Yeri ve Tarihi : 29.04.1991

Lisans Üniversite : Pamukkale Üniversitesi

Elektronik posta : sahinnemre@gmail.com

İletişim Adresi : Pamukkale Üniversitesi İktisadi ve İdari Bilimler Fakültesi Yönetim Bilişim Sistemleri Bölümü B Blok Kat: 2 No: B-02-005 Pamukkale Üniversitesi, Kınıklı Kampüsü Pamukkale/DENİZLİ


**Bildiri Listesi :**

• Akkaş, S. and Şahin, E., "Prediction of the possible number of cinemagoers of Turkish movies via neural networks by taking into account the number of people who watch the trailer", *3rd International Conference on Science, Ecology and Technology(ICONSETE'2017)*, Rome Italy, (2017).

• Yeşilyurt, M. E., Elbi, M. E., Emrouznejad, A., Koyuncuoğlu M. U., Şahin, E., Yeşilyurt, F. and Kızılkaya, A., "Computing single outputs for DEA", *15th International Conference on Data Envelopment Analysis(DEA2017)*, Prague Czech Republic, (2017).