T.C.
DÜZCE ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNİN VERİ BİLİMİ ANKETİ VERİ SETİ İLE KARŞILAŞTIRMALI ANALİZİ

ELVAN KÜBRA DOĞAN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DR. ÖĞR. ÜYESİ ARAFAT ŞENTÜRK

DÜZCE, 2021

T.C.
DÜZCE ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNİN VERİ BİLİMİ ANKETİ VERİ SETİ İLE KARŞILAŞTIRMALI ANALİZİ

Elvan Kübra DOĞAN tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Dr. Öğr. Üyesi Arafat ŞENTÜRK

Düzce Üniversitesi

Jüri Üyeleri

Dr. Öğr. Üyesi Arafat ŞENTÜRK
Düzce Üniversitesi _____

Doç. Dr. Devrim AKGÜN
Sakarya Üniversitesi _____

Dr. Öğr. Üyesi Enver KÜÇÜKKÜLAHLI
Düzce Üniversitesi _____

Tez Savunma Tarihi: 30.12.2021

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

30 Aralık 2021

Elvan Kübra DOĞAN

TEŞEKKÜR

Lisans dönemimden bu yana desteklerini esirgemeyip, vermiş olduğu fikirleri ile çalışmalarımda bana yol gösteren ve yakın ilgisini eksik etmeyen değerli hocam Dr. Öğr. Üyesi Arafat ŞENTÜRK'e, tüm hocalarıma ve dostlarıma en içten dileklerim ile teşekkür ederim.

Başaracağıma dair inançları ile bana benden çok güvenen, desteklerini sonuna kadar hissettiren ve başarılarımın asıl emektarları olan sevgili annem Gül KOK, babam Yasin KOK ve ablam Elif Başak'a sonsuz teşekkür ederim.

Hayallerime ve hedeflerime kavuşmak amacı ile attığım her adımda sevgisi ve anlayışı ile gücüme güç katan, gerek akademik gerekse özel hayatımda almış olduğum her kararımda beni sabırla destekleyen sevgili eşim, yol arkadaşım Mehmet DOĞAN'a sonsuz teşekkür ederim.

Hayatımın her anına şahitlik ettiği gibi, Yüksek Lisansa kabul edildiğimde de mutluluğuma şahitlik eden ancak tamamladığımı görmeye ömrü yetmeyen, üzerimde sonsuz emeği bulunan sevgili babaannem Nevim KOK'a emeklerinin karşılığını en iyi şekilde vermeye çalışacağıma dair söz veriyorum ve sonsuz minnet duyuyorum.

30 Aralık 2021 **Elvan Kübra Doğan**

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ vii

ÇİZELGE LİSTESİ ix

KISALTMALAR x

SİMGELER xi

ÖZET xii

ABSTRACT xiii

1. GİRİŞ 1

2. İLGİLİ ÇALIŞMALAR 5

3. MATERYAL VE YÖNTEM 11

3.1. VERİ MADENCİLİĞİ YÖNTEMLERİ 11

3.1.1. Sınıflandırma 12

3.1.2. Kümeleme 17

3.1.3. Birliktelik Kuralları 17

3.2. AÇIK KAYNAK KODLU VERİ MADENCİLİĞİ YAZILIMLARI 18

3.2.1. Weka (The Waikito Environment for Knowledge Analysis) 18

3.2.2. R Programlama 18

3.2.3. Orange 18

3.3. "VERİ BİLİMİ ANKETİ" İSİMLİ VERİ SETİNİN MODELLEMEYE HAZIR HALE GETİRİLMESİ 20

3.3.1. Kullanılan Veri Seti: "Veri Bilimi Anketi" 20

3.3.2. Problem Tanımlaması ve Veri Ön İşleme 24

3.3.3. Veri Setinin Yazılım Ortamına Aktarımı 31

4. MODEL OLUŞUMU VE DENEYSEL SONUÇLAR 37

4.1. EĞİTİM VE TEST VERİ SETİ ORANLAMASI 38

4.2. EĞİTİM, DOĞRULAMA VE TEST VERİ SETİ ORANLAMASI 38

4.3. ALGORİTMA/MODEL UYGULAMA DENEMELERİ 39

4.3.1. C4.5 Algoritmasının Weka'da Uygulanması 41

4.3.2. Rastgele Orman Algoritmasının Weka'da Uygulanması 45

4.3.3. K-En Yakın Komşu (KNN) Algoritmasının Weka'da Uygulanması 49

4.4. MODEL PERFORMANSININ DEĞERLENDİRİLMESİ 55

5. BULGULAR VE TARTIŞMA 58

6. SONUÇLAR VE ÖNERİLER 64

7. KAYNAKLAR 65

8. EKLER 70

8.1. EK 1:C4.5 ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (CROSS-VALİDATİON: FOLDS 10) 70

8.2. EK 2: C4.5 ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (PERCENTAGE SPLİT:%66)..... 71
8.3. EK 3: KNN ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (K=1, CROSS-VALİDATİON: FOLDS 10)..... 72
8.4. EK 4: KNN ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (K=1, PERCENTAGE SPLİT:%66) 73
8.5. EK 5: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=11, CROSS-VALİDATİON: FOLDS 10)..... 74
8.6. EK 6: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=11, PERCENTAGE SPLİT:%66) 75
8.7. EK 7: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=13, CROSS-VALİDATİON:FOLDS 10)..... 76
8.8. EK 8: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=13, PERCENTAGE SPLİT:%66) 77
8.9. EK 9: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=17, CROSS-VALİDATİON:FOLDS 10)..... 78
8.10. EK 10: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=17, PERCENTAGE SPLİT:%66) 79
8.11. EK 11: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=21, CROSS-VALİDATİON:FOLDS 10)..... 80
8.12. EK 12: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=21, PERCENTAGE SPLİT:%66) 81

ÖZGEÇMİŞ..... 82

ŞEKİL LİSTESİ

Sayfa No

Şekil 1.1. Veri, Enformasyon ve Bilgi Kavramlarının İfadesi. ... 1
Şekil 2.1. Veri Madenciliği Tarihsel Gelişim Süreci. ... 6
Şekil 2.2 1960'dan Günümüze Veri Madenciliği Kavramına Ulaşım. ... 7
Şekil 2.3 Veri Madenciliği Çalışma/Uygulama Alanları. ... 7
Şekil 3.1. Veri Madenciliği Aşamaları Genel Yapısının Akış Şeması. ... 11
Şekil 3.2. Veri Madenciliği Aşamaları Genel Yapısının Akış Şeması. ... 12
Şekil 3.3. Karar Ağacı Yapısı ve Ağacı Oluşturan Elemanların İfade Edilmesi. ... 13
Şekil 3.4.Entropi Tabanlı Algoritma Mantığının Akış Şemasında İfade Edilmesi. ... 14
Şekil 3.5.Yapay Sinir Ağı Modelinin Temel Yapısı. ... 16
Şekil 3.6.Uygulama Bölümü Çalışma Özeti. ... 20
Şekil 3.7.Veri Setinde Bir Kesit (.csv formatı). ... 21
Şekil 3.8. Veri Setinin Girdi ve Çıktı Niteliklerinin İfade Edilmesi. ... 25
Şekil 3.9. Veri Birleşimi ve Düzenleme Örneği. ... 27
Şekil 3.10. Veri Artırım Mantığının Veri Seti Üzerinden Örneklendirilmesi. ... 28
Şekil 3.11. WEKA Giriş Arayüzü. ... 31
Şekil 3.12. Veri Setinin Weka Ortamına Yüklenmesi. ... 32
Şekil 3.13. "Kullanılan Çalışma Aracı" Niteliğine Ait Selected Attribute Bölümü. ... 33
Şekil 3.14. "Tercih Edilen Dil" Niteliğine Ait Selected Attribute Bölümü. ... 34
Şekil 3.15. "Katılımcılar Tarafından Kullanılan Algoritma" Niteliğine Ait Selected Attribute Bölümü. ... 35
Şekil 3.16. "Çalışma Alanı" Niteliğine Ait Selected Attribute Bölümü. ... 36
Şekil 4.1. Eğitim ve Test Veri Seti Oranı. ... 38
Şekil 4.2. Eğitim, Doğrulama ve Test Veri Seti Oranlaması. ... 39
Şekil 4.3. Veri Setine Uygulanacak Olan Sınıflandırma Algoritmaları. ... 40
Şekil 4.4. Weka Ortamı -Classify Sekmesi. ... 40
Şekil 4.5. Weka Ortamında J48 Algoritması Yapılandırma Ekranı. ... 42
Şekil 4.6. Cross-validation:10 bölümlemeli, C4.5 Algoritması Weka Çıktısı. ... 44
Şekil 4.7. Percentage split:%66 bölümlemeli C4.5 Algoritması Weka Çıktısı. ... 45
Şekil 4.8.Rastgele Orman Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü. ... 46
Şekil 4.9.Cross-validation:10 bölümlemeli, Rastgele Orman Algoritması Weka Çıktısı. ... 48
Şekil 4.10.Percentage split:%66 bölümlemeli Rastgele Orman Algoritması Weka Çıktısı. ... 48
Şekil 4.11. KNN Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü (1). ... 49
Şekil 4.12. KNN Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü (2). ... 51
Şekil 4.13. KNN Algoritması Yapılandırma Seçenekleri Weka Arayüzü (3). ... 51
Şekil 4.14. Cross-validation:10 bölümlemeli, KNN Algoritması Weka Çıktısı. ... 52
Şekil 4.15. Percentage split:%66 bölümlemeli KNN Algoritması Weka Çıktısı. ... 53
Şekil 4.16. Weka Ortamında KNN Algoritması (Percentage split %66, k=21) Model Performansının Değerlendirilmesi. ... 57
Şekil 5.1. Orijinal ve İşlenmiş Veri Setlerinin Seçilen Algoritmaların Başarı Oran Grafikleri. ... 60
Şekil 5.2. İşlenmiş Veri Seti İçin Belirtilen Nitelikler Bazında Başarı Oranları (1). ... 61
Şekil 5.3. İşlenmiş Veri Seti İçin Belirtilen Nitelikler Bazında Başarı Oranları (2). ... 62
Şekil 5.4. Veri Madenciliği Aşamalarının Döngüselliği. ... 63
Şekil 8.1. C4.5 Algoritması Başarı Oranı Weka Özet Çıktısı (1). ... 70

Şekil 8.2. C4.5 Algoritması - Karışıklık Matrisi (1). 70
Şekil 8.3. C4.5 Algoritması Başarı Oranı Weka Özet Çıktısı (2). 71
Şekil 8.4. C4.5 Algoritması - Karışıklık Matrisi (2). 71
Şekil 8.5. KNN Algoritması Başarı Oranı Weka Özet Çıktısı (1). 72
Şekil 8.6. KNN Algoritması - Karışıklık Matrisi (1). 72
Şekil 8.7. KNN Algoritması Başarı Oranı Weka Özet Çıktısı (2). 73
Şekil 8.8. KNN Algoritması - Karışıklık Matrisi (2). 73
Şekil 8.9. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (3). 74
Şekil 8.10. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (4). 75
Şekil 8.11. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (5). 76
Şekil 8.12. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (6). 77
Şekil 8.13. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (7). 78
Şekil 8.14. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (8). 79
Şekil 8.15. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (9). 80
Şekil 8.16. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (10). 81

ÇİZELGE LİSTESİ

Sayfa No

Çizelge 3.1. Bazı Veri Madenciliği Ortamlarının Karşılaştırılması. 19
Çizelge 3.2. Veri Setinin “Kullanılan Çalışma Aracı” İsimli Niteliğine Ait Benzersiz Veri Çeşitleri. 22
Çizelge 3.3. Veri Setinin “Tercih Edilen Dil” İsimli Niteliğine Ait Benzersiz Veri Çeşitleri. 23
Çizelge 3.4. Veri Setinin “Katılımcılar Tarafından Kullanılan Algoritma” İsimli Niteliğine Ait Benzersiz Veri Çeşitleri. 23
Çizelge 3.5. Veri Setinin “Çalışma Alanı” İsimli Niteliğine Ait Benzersiz Veri Çeşitleri. 24
Çizelge 3.6. Veri Setinde "null" Değerine Sahip Örnek İki Veri Satırı. 26
Çizelge 3.7. Veri Çeşidi Birleşimi Örneği. 26
Çizelge 3.8. Orijinal Veri Setinden Bir Satır. 28
Çizelge 3.9. Çizelge 3.8’deki Veri Satırına Veri Artırım İşlemi Uygulaması Sonrası. 28
Çizelge 3.10. Çalışma Alanı "Akademik" Olan Veri Satırlarının, Katılımcılar Tarafından Kullanılan Algoritma Nitelik Değerlerinin Kullanım Sıklıkları. 30
Çizelge 3.11. Aykırı Değerlerin Düzenlenmesi Öncesinde Örnek Veri Satırları. 30
Çizelge 3.12. Aykırı Değerlerin Düzenlenmesi Sonrasında Örnek Veri Satırları. 30
Çizelge 4.1. KNN Algoritmasının Belirtilen Seçenekler Doğrultusunda Weka’da Elde Edilen Başarı Oranları (1). 53
Çizelge 4.2. KNN Algoritmasının Belirtilen Seçenekler Doğrultusunda Weka’da Elde Edilen Başarı Oranları (2). 54
Çizelge 4.3. Karışıklık Matris Yapısı. 55
Çizelge 5.1. Orijinal ve İşlenmiş Veri Seti Kullanıldığında, Seçili Algoritmaların Başarı Oranları ve Farkları. 59

KISALTMALAR

CART	Classification and Regression Trees (Sınıflandırma ve Regresyon Ağaçları)
CNNs	Convolutional Neural Networks (Evrişimli Sinir Ağları)
DN (TN)	Doğru Negatif
DNA	Deoksiribo Nükleik Asit
DP(TP)	Doğru Pozitif
DYS	Ders Yönetim Sistemi
İMKB	İstanbul Menkul Kıymetler Borsası
KNN	K-Nearst Neighbours (K-En Yakın Komşu)
LMT	Lojistik Model Ağacı
RNNs	Recurrent Neural Network (Tekrarlayan Sinir Ağları)
ROC	Receiving Operating Characteristics (Algılayıcı İşletim Özellikleri)
SDBK	Serum Demir Bağlama Kapasitesi
Weka	The Waikito Environment for Knowledge Analysis
YN(FN)	Yanlış Negatif
YP(FP)	Yanlış Pozitif
YSA	Yapay Sinir Ağları

SİMGELER

H(S)	$P_{olasılık}$ dağılımına sahip üreten S kaynağının entropisi
m_n	Kaynağın *n* adet üreteni
p_i	m_i üretileninin üretilme olasılığı
S	Veri Seti

ÖZET

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNİN VERİ BİLİMİ ANKETİ VERİ SETİ İLE KARŞILAŞTIRMALI ANALİZİ

Elvan Kübra DOĞAN
Düzce Üniversitesi
Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi
Danışman: Dr. Öğr. Üyesi Arafat ŞENTÜRK
Aralık 2021, 81 sayfa

Veri Madenciliği teknolojisi günden güne popülerliğini artıran bir teknoloji olmaktadır. Popülerliğinin artmasının en büyük sebeplerinden biri de çalışma alanı sınırının bulunmamasıdır. Teknik altyapı olarak bilişim sektörüne ait olan Veri Madenciliği teknolojisi, birçok sektöre kolaylık ve avantaj sağlamak için hizmet sunmaktadır. Çalışma kapsamında Veri Madenciliği teknolojisinde tercih edilen yazılım dili, kullanılan algoritma vb. kriterlerinin girdi olarak kabul edildiği, veri bilimcilerin tercih ettikleri bu teknik bilgilerden çıkarımla hangi sektörde çalıştıklarına dair çıktı bilgilerinin yer aldığı "Veri Bilimi Anketi" isimli veri seti kullanılmaktadır. Veri setinin Sınıflandırma Algoritmalarından C4.5 Algoritması, Rastgele Orman Algoritması ve K-En Yakın Komşu Algoritması ile modellenmesi sonucu başarı oranları değerlendirmelerine değinilmiştir. Modellerin başarı oranı kıyaslamaları gerçekleştirilirken, Sınıflandırma yöntemine ait olan algoritmalar hem orijinal hem de işlenmiş veri setini kullanmışlardır. Veri setleri bazında model başarı oranları değerlendirildiğinde, orijinal veri seti kullanılarak oluşturulan modellerin başarı oranları, veri ön işleme aşaması sonrasında oluşturulan işlenmiş veri seti kullanılarak da modellendiğinde başarı oranlarında %14-15 oranında artış olmaktadır. İşlenmiş veri seti, seçili sınıflandırma algoritmaları (C4.5, Rastgele Orman ve KNN) ve bu algoritmaların varsayılan algoritmik nitelikleri ile modellendiğinde elde edilen başarı oranları algoritmalar bazında kıyaslandığında sapma oranı çok düşük olmaktadır. Algoritmaların başarı oranları ön işleme öncesi kullanılan orijinal veri seti ve ön işleme sonrasında kullanılan işlenmiş veri seti ile algoritma bazında değerlendirildiğinde ise sapma değeri daha belirgin olmaktadır. Ayrıca KNN algoritmasına özgü olan "k" nitelik değerinin farklı değerler alması sonucu veya Eğitim-Test veri seti bölümleme seçenekleri gibi model başarı oranında sapma yaratacak durumlar için de başarı oranı değerleri gözlemlenmiştir. Ancak bahsedilen durumların model başarısına etkisi, ön işleme aşamasının model başarısına olan etkisi kadar belirgin değildir. Gerçekleştirilmiş olan bu kıyaslamalardan çıkarımla, başarılı modeller oluşturulabilmesi için Veri Madenciliği aşamalarının önem/etki seviyeleri değerlendirilmiş olup, Veri Madenciliği aşamaları "döngüsellik" ve "öznellik" kavramlarından faydalanılarak yorumlanmışlardır.

Anahtar sözcükler: C4.5 Algoritması, Döngüsellik, K-En Yakın Komşu, Öznellik, Rastgele Orman.

ABSTRACT

COMPARATIVE ANALYSIS OF DATA MINING CLASSIFICATION METHODS WITH DATA SCIENCE SURVEY DATA SET

Elvan Kübra DOĞAN
Düzce University
Graduate School of Natural and Applied Sciences, Department of Computer Engineering
Master's Thesis
Supervisor: Assist. Prof. Dr. Arafat ŞENTÜRK
December 2021, 81 pages

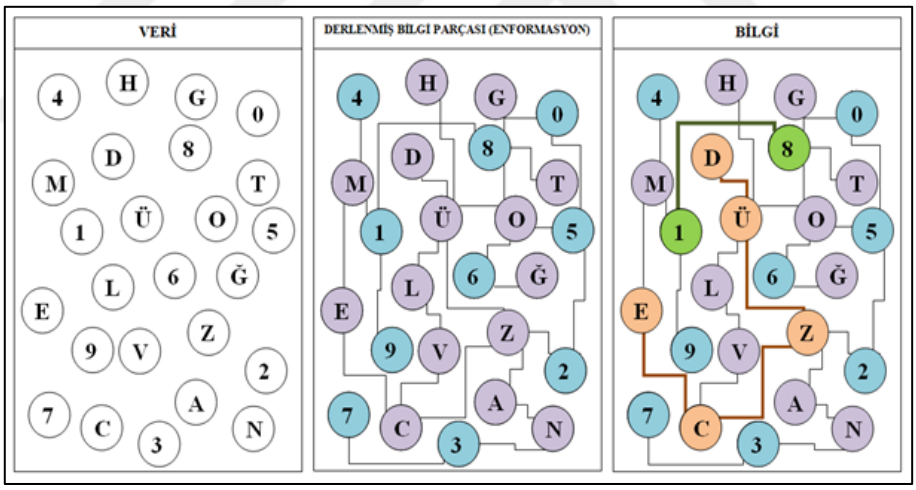
Data Mining technology is a technology that is increasing its popularity day by day. One of the biggest reasons for its increasing popularity is the absence of a workspace limit. Data Mining technology, which belongs to the IT sector as a technical infrastructure, provides services to many sectors to provide convenience and advantage. Within the scope of the study, the preferred software language in Data Mining technology, the algorithm used, etc. The data set named "Data Science Questionnaire" is used, in which the criteria of the data scientists are accepted as input, and the output information about which sector they work in with inferences from this technical information preferred by data scientists. As a result of modeling the data set with Classification Algorithms C4.5 Algorithm, Random Forest Algorithm and K-Nearest Neighbor Algorithm, success rates evaluations are mentioned. While comparing the success rate of the models, the algorithms belonging to the Classification method used both the original and the processed data set. When model success rates are evaluated on the basis of data sets, the success rates of models created using the original data set are increased by 14-15% when modeled using the processed data set created after the data preprocessing stage. When the processed data set is modeled with selected classification algorithms (C4.5, Random Forest and KNN) and the default algorithmic features of these algorithms, the deviation rate is very low when the success rates are compared on the basis of algorithms. When the success rates of the algorithms are evaluated on the basis of the original data set used before the preprocessing and the processed data set used after the preprocessing, the deviation value becomes more evident. In addition, the success rate values were observed for the situations that would cause deviations in the model success rate, such as the "k" attribute value, which is specific to the KNN algorithm, taking different values or the Training-Test data set partitioning options. However, the effect of the mentioned situations on model success is not as clear as the effect of the preprocessing stage on model success. By inferring from these comparisons, the importance/effect levels of Data Mining stages were evaluated in order to create successful models, and Data Mining stages were interpreted by using the concepts of "cyclicality" and "subjectivity".

Keywords: Circularity, C4.5 Algorithm, K-Nearest Neighbor, Random Forest, Subjectivity.

1. GİRİŞ

Bir araştırmada, tartışmada, akıl yürütmede sonuca ulaşabilmek için gereken bilgi ifadesi sözlükte "veri" kelimesi ile ifade edilmektedir. Hızla artan teknoloji ile doğru orantılı olarak veriler de artış göstermektedir. Ancak verilerin fazla olması demek, fazla bilgi sahibi olunması anlamına gelmemektedir [1]. Bir hedef erişimi doğrultusunda kullanıldığı takdirde veriler anlam kazanmaktadır.

Verilerin fazlalığı, birbirleriyle olan ilişkileri, birbirlerini tetiklemeleri gibi birçok ölçüt düşünüldüğünde durum Şekil 1.1'de ifade edilen "Harflerin anlamlı bir araya gelmesi gerekir." mantığındaki alfabe örneği kadar yalın olmayacaktır. Çünkü bu işlemler fazla miktardaki veriler için yapılacağından, istatistikten daha kapsamlı bir teknoloji olan veri analiz tekniği "Veri Madenciliği" kavramı devreye girmektedir [2].



Şekil 1.1. Veri, Enformasyon ve Bilgi Kavramlarının İfadesi.

Veri Madenciliği; istatistik ve matematiksel yöntemlerle birlikte, örüntü tanıma teknolojilerini de kullanarak depolanan veri yığınları arasında anlamlı yeni ilişki, örüntü ve eğilimlerin keşfedilmesi süreci olarak ifade edilmektedir. Bir başka tanımlama ile mevcut veriden anlamlı bilgileri, ilişki çıkarmada kullanılan tekniklere verilen genel isimdir [3].

Kullanılan teknikler;

- Makine Öğrenimi

- Yapay Zeka
- İstatistik
- Veri Tabanları
- Uzman Sistemler
- Veri Görselleştirme vb...

Kısaca Veri Madenciliği kavramı “Veriyi nitelikli bilgiye dönüştürme yolu” olarak tek cümlede belirtilebilir [3].

Bu tez çalışmasında Veri Madenciliği Sınıflandırma yöntemlerinden seçili (C4.5, Rastgele Orman ve K-En Yakın Komşu Algoritmaları) algoritmaların işleyişleri analiz edilmiş ve seçili veri seti için başarı oranları değerlendirilmiştir.

Yapılan literatür taraması kapsamında sınıflandırma yöntemi ya da karar ağacı algoritmaları gibi ortak noktalara sahip algoritmaların, belirlenen veri seti üzerine uygulanarak başarı oranlarının kıyaslamalarına birçok çalışma kapsamında rastlanmıştır. Örneğin, bakteri türlerinin farklı uzunluktaki DNA dizilimlerini ele alan bir çalışmada, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine Karar Ağacı Algoritmalarından J48, LMT (Lojistik Model Ağacı) ve Rastgele Orman algoritmaları uygulanarak sınıflandırma işlemi gerçekleştirilmiştir. Sınıflandırma sonucunda J48 Algoritmasının %93, LMT Algoritmasının %86 ve Rastgele Orman Algoritmasının %100 doğru sınıflandırma yaptığını gözlemlenmiştir [4]. Ders Yönetim Sistemi yazılımı (DYS) kullanan öğrencilerin, sistemle olan etkileşimlerine dayalı olarak farklı bir Veri Madenciliği problemi ele alınmıştır. Ele alınan çalışmada, Veri Madenciliği sınıflandırma algoritmaları ile öğrencilerin akademik başarılarının ne ölçüde tahmin edilebileceği ve onların akademik başarılarının tahmin edilmesinde algoritmalara hangi etkileşimlerin bilgi kazancı sağladığı ele alınmıştır. Araştırma verileri, Trakya Üniversitesi Eğitim Fakültesi’nde 2015-2016 eğitim öğretim yılında öğrenim görmekte olan 70 öğrencinin Bulut ders DYS yazılımını kullanma sürecindeki etkileşimlerinden elde edilmiştir. Veri Madenciliği sınıflandırma algoritmaları olarak Naive-Bayes, C4.5 ve K-En Yakın Komşu Algoritmaları kullanılmıştır. Çalışma sonucunda elde edilen doğruluk oranı değerlerinin %55,7 ile %64,3 arasında değişiklik gösterdiği görülmekte olup, en yüksek başarı oranının K-En Yakın Komşu algoritması k=3 parametresi ile elde edildiği gözlemlenmiştir [5]. Son

olarak da 699 hastadan toplanılmış görüntülerin sayısallaştırılması ile elde edilen 10 adet öznitelik içeren Wisconsin Üniversitesi meme kanseri veri setinin kullanılması ile gerçekleştirilen bir çalışma incelenmiştir. Veri seti K-En Yakın Komşu ve Karar Ağacı Algoritmalarının çeşitli parametreleri değiştirilerek sınıflandırılmak istenmiştir. Eğitim ve test verileri çapraz doğrulama ile karşılaştırılmış ve en yüksek sınıflandırma başarısı %97,30 ile K-En Yakın Komşu Algoritması ile elde edilmiştir [6].

İncelenen örnek çalışmalardan çıkarımla, farklı türdeki algoritmaların aynı veri seti üzerine uygulanıp başarı oranlarının kıyaslandığı görülmüştür. Bunun yanı sıra daha önce üzerinde çalışılmış olan veri setine (Meme kanseri verilerinin bulunduğu veri seti vb.) farklı veri ön işleme işlemlerinin gerçekleştirilip daha fazla doğruluk oranı elde edilmesi amaçlanan çalışmalar da incelenmiştir.

Bu tez çalışmasının diğer çalışmalardan farkı, çalışma kapsamında bahsedilen Veri Madenciliği ifadesinin modelleme değerlendirmelerinin "döngüsellik" ve "özgünlük" kavramları ile gerçekleştirilecek olmasıdır. Veri madenciliği çalışmalarında, çalışmayı oluşturan araştırmacının görüş ve önerileri doğrultusunda tekniklere yön verildiğinden, çalışmaların birebir denk gelmesi çok düşük bir ihtimale sahiptir. Veri seti seçimi, problemin yorumlanması, girdi/çıktı nitelikleri, veri seti bölümlemesi, model seçimi... gibi birçok faktör veri bilimcinin kararına bırakıldığından, belirli formül ve net yargı ile karar verilmediğinden ötürü her çalışma özgünlüğünü/farkını kendi içerisinde barındırmış olmaktadır. Bu çalışma kapsamında da "Veri Bilimi Anketi" isimli veri seti kullanılmaktadır. Veri seti dağılımı 3 girdi, 1 çıktı değeri elde edilecek şekilde bölümlenmiştir. Veri setinin oluşması için bilgileri alınan kullanıcıların tercih ettikleri çalışma aracı niteliği "Kullanılan Çalışma Aracı", algoritma tercihleri "Katılımcılar Tarafından Kullanılan Algoritma", yazılım dili tercihleri "Tercih Edilen Dil" başlıkları altında ele alınmaktadır. Bahsedilen bu sütunlar girdi değeri olarak kullanılıp, işverenin hangi sektöre/alana yönelik bir çalışma yaptığı "Çalışma Alanı" isimli sütunda çıktı değeri olarak bulunmaktadır. Bu veri seti üzerinde sınıflandırma yöntemini esas alan algoritmalardan C4.5, Rastgele Orman ve K-En Yakın Komşu Algoritmaları kullanılarak Çalışma Alanı (sınıf) değerinin tahmin edilmesi denemeleri gerçekleştirilmiştir.

Günümüzde bilişim sektörü alt yapısına sahip olan "Veri Madenciliği" kavramı birçok sektöre destek sağlamaktadır. Kullanılan veri setinde de sınıf değeri ile vurgulanmak istenilen durum bunu ifade etmektedir. Veri ön işlemenin önemi, sınıflandırma

probleminin farklı sınıflandırma algoritmaları tarafından işlenmesi, veri seti bölümlenmesi gibi birçok faktörün analizi, bahsedilen "Çalışma Alanı" çıktısının elde edilmesindeki başarı oranına etkileri bu tez çalışması kapsamında aktarılmaktadır.

Giriş bölümünde, bu tez çalışması kapsamında anlatılan konu başlıkları ve uygulama bölümüne değinilerek çalışmaya ön fikir oluşturulması sağlanmıştır.

İlgili Çalışmalar bölümünde, gerçekleştirilen literatür taraması sonrasında elde edilen bazı Veri Madenciliği tanımlamalarına yer verilmiş ve literatürdeki uygulama alanlarına değinilmiştir.

Materyal ve Yöntem bölümünde, Veri Madenciliği yöntem ve tekniklerinin tanımlamalarına ve çalışma prensiplerine değinilmiştir. Veri Madenciliği uygulama alanları olarak tercih edilen açık kaynak kodlu yazılımların tanıtım ve kıyaslamalarına yer verilmiştir. Uygulamada kullanılacak olan "Veri Bilimi Anketi" isimli veri seti tanıtımının yanı sıra, veri ön işleme aşamaları da açıklanmıştır. Ön işleme aşamaları gerçekleştirilmiş veri setinin, tercih edilen açık kaynak kodlu yazılım ortamı olan Weka'ya yüklenmesi ve veri analizlerinin gerçekleştirilmesine değinilmiştir.

Deneysel Sonuçlar bölümünde, modelleme ve test aşamasında kullanılmak üzere bölümlenecek veri seti oranlamalarının gerçekleştirilme tekniklerinden bahsedilmiştir. Sınıflandırma yöntemine ait tercih edilen C4.5, Rastgele Orman ve K-En Yakın Komşu algoritmalarının, Weka yazılım ortamında "Veri Bilimi Anketi" isimli veri setinin kullanılması ile modelleme denemelerine değinilmiştir. Aynı veri seti ile farklı algoritmalar kullanılarak, aynı algoritma için de orijinal ve işlenmiş (farklı) veri seti ile oluşturulan modeller gibi farklı durumların model başarılarının kıyaslamalarına ve başarı oranları hakkındaki yorumlamalara yer verilmiştir.

Bulgular ve Tartışma bölümünde, algoritmada kullanılan nitelikler, algoritma işleyişleri, veri kullanımları gibi birçok özellik hakkında model başarı oranları referans alınarak yorumlamalar yapılmıştır.

2. İLGİLİ ÇALIŞMALAR

Veri Madenciliği kavramının ortaya çıkışı ile ilgili zaman "günümüz" olarak belirtilebilir iken, Veri Madenciliği teknolojisinin temellerine dayanılmak istenirse tarihte çok eskilere gidilmesi gerekmektedir. Bilgisayarlarda veri depolanabilmesi, Veri Madenciliği teknolojisinin temeli olarak kabul edilecek olursa tarihte 1950'lere kadar inilebilmektedir. Hala gelişime, değişime açık bir kavram olan Veri Madenciliği kavramı ve teknikleri ile ilgili gün geçtikçe farklı tanımlamalar ve uygulamalar eklenmeye devam etmektedir [7].

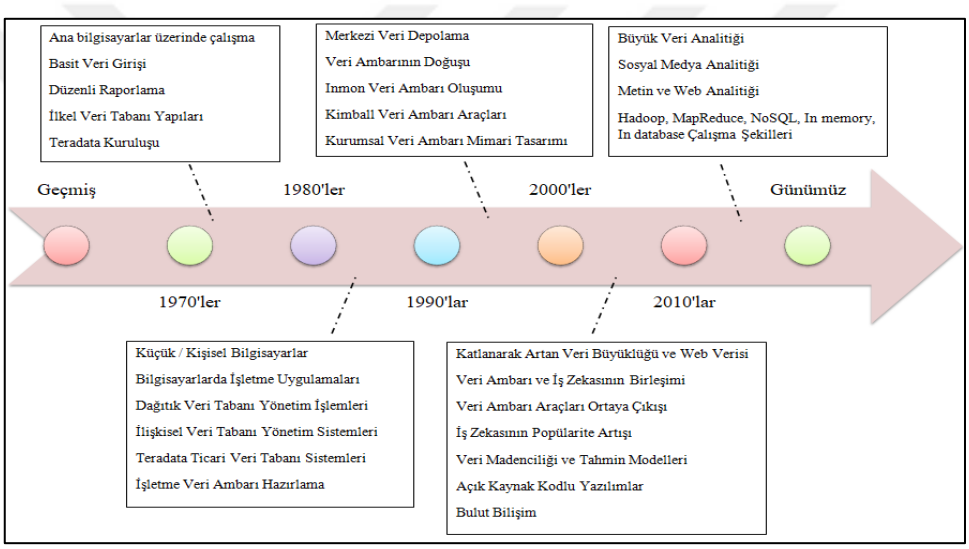
"Veri Madenciliği nedir?" sorusunun cevabı aranacak olursa, literatürde farklı bilim insanlarının, farklı yorumlamalar ile bu kavramı ele aldığı görülmektedir. Farklı görüşler doğrultusunda Veri Madenciliği kavramı ile ilgili yapılan tanımlamaların bir kısmı aşağıda belirtildiği şekildedir.

Veri Madenciliği;

- "Daha önceden bilinmeyen örüntüleri ve anlaşılabilir bilgileri geniş veri tabanlarından seçmeyi, keşfetmeyi ve bu bilgileri modellemeyi içerir." [8].
- "Veri tabanı, yapay zekâ ve istatistik dünyasını birbirine bağlayan, son zamanlarda çıkmış bir alandır." [9].
- "Bilgi keşfi sürecinde veriler arasındaki daha önceden bilinmeyen ilişki ve örüntüleri ayırt etmek için kullanılan yöntemler bütünüdür." [10].
- "Büyük veri tabanlarından daha önceden bilinmeyen, kavranabilir, işlenebilir verileri ayıklama ve bu verileri önemli kararlar vermek için kullanma sürecidir." [10].
- "Veri içerisindeki avantajlı örüntüleri açığa çıkarma sürecidir." [10].
- "Tek başına ham verinin sunamadığı bilgiyi ortaya çıkaran veri analizi sürecini karşılayan kavramdır." [11][12].

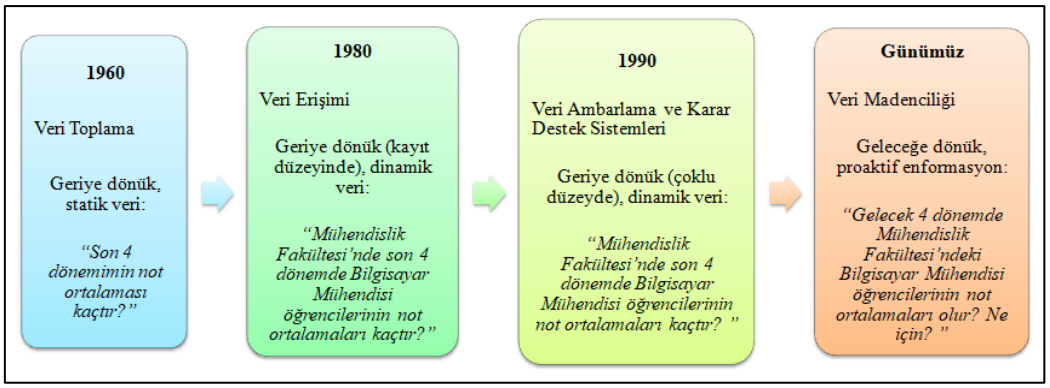
- “Veriden örüntü elde etmek için belirli algoritmaların uygulamasıdır.” bir başka ifade ile “Veri içerisindeki potansiyel olarak kullanışlı ve yararlı, anlaşılabilir örüntüleri belirleme sürecidir.” [10][13].
- 2009 yılında “Organizasyonların veri ambarlarındaki çok önemli bilgilere yoğunlaşmasına yardımcı olan, çok büyük veri tabanlarında saklı, akıllı bilgiyi ortaya çıkaran yeni bir teknolojidir.” [14].

Veri Madenciliği kavramının temelinden başlanarak, 1970’lerde ele alınan ilkel veri tabanı yapılarından, 2010’larda belirtilen büyük veri analitiğine kadar uzanan Veri Madenciliğinin tarihsel gelişim aşamaları Şekil 2.1’de belirtildiği gibi ele alınmaktadır [15].



Şekil 2.1. Veri Madenciliği Tarihsel Gelişim Süreci.

Zaman çizelgesinde gelişim sürecine bağlı kalınarak ele alınan kavramlar, örnekler verilerek somutlaştırmak istenmektedir. Veri Madenciliği kavramına ulaşıncaya kadar, kullanıcılara sunulan cevapların gelişim aşaması Şekil 2.2’de belirtildiği gibi ifade edilebilmektedir [16].

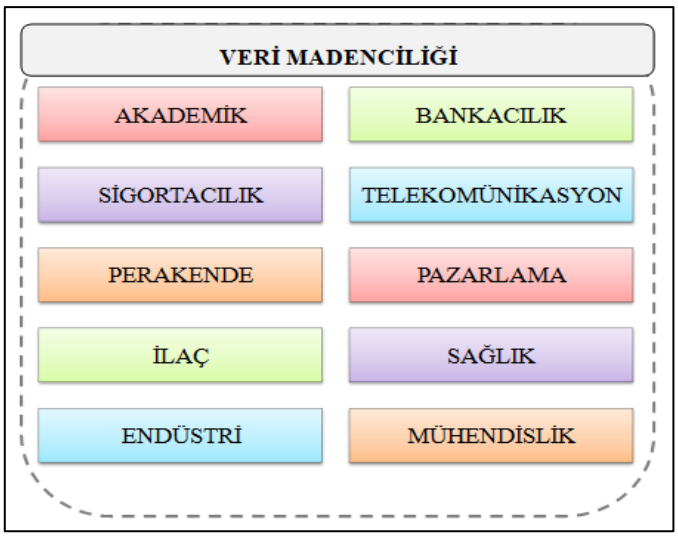


Şekil 2.2 1960'dan Günümüze Veri Madenciliği Kavramına Ulaşım.

Günümüz teknolojisinde veri kavramı ile güç, doğru orantılı olarak artmaktadır. Fazla veri, doğru kullanıldığı takdirde fazla gücü işaret etmektedir. Çünkü büyük hacimde veri bulunması Veri Madenciliği teknolojisinin daha verimli kullanılmasına olanak sağlamaktadır [17].

Büyük hacimde veri kullanımı, çalışma alanı genişliği gibi fonksiyonel özelliklere sahip olan Veri Madenciliği teknolojisi, karar verme süreçleri için kullanılacak en uygun yöntem olarak popülerliğini korumaya devam etmektedir [18].

Bahsedilen fonksiyonelliğin çalışma alanı genişliğine değinilecek olunursa Şekil 2.3'te belirtildiği gibi akademik, bankacılık, sigortacılık, telekomünikasyon, perakendecilik, pazarlama, ilaç ve sağlık, endüstri, bilim ve mühendislik gibi birçok alan Veri Madenciliği teknolojisinin kullanımı için örnek olarak gösterilebilmektedir [19].



Şekil 2.3 Veri Madenciliği Çalışma/Uygulama Alanları.

Veri Madenciliği çalışma alanı örneklerinin aynı zamanda bu tez çalışmasında kullanılan veri setinin sınıf niteliği değerlerini oluşturduğu belirtilmişti. Uygulama bölümünde veri seti özellikleri ve modelleme yöntemine detaylı bir şekilde yer verilecektir. Ancak, sınıf niteliğinin "Çalışma Alanı" çıktısı vermesinden dolayı kısaca bahsedilmek istenirse, veri setinde bulunan "Kullanılan Çalışma Aracı", "Tercih Edilen Dil" ve "Tercih Edilen Algoritma" nitelikleri ile sınıflandırma yöntemine ait seçili algoritmalar kullanılarak modellenmesi sonucu "Çalışma Alanı" sınıf niteliği elde edilmektedir. Bu sayede Veri Madenciliği teknolojisinin birçok alana hizmet verdiği, bahsedilen veri seti ve gerçekleştirilecek uygulama incelenerek de gözlemlenmiş olmaktadır.

Veri Madenciliği teknolojisinin yaygınlığından, fonksiyonelliğinden sıkça bahsedilmiştir. Bahsedilen bu özellikleri daha somut bir şekilde ifade edebilmek adına Veri Madenciliği kavramına dair gerçekleştirilen çalışmalar incelenmiş olup, aşağıda kısaca aktarılmıştır.

Akademik/Eğitim alanında; Web Tabanlı Uzaktan Eğitim Sisteminin başarısı, kullanılan öğretim modelinin doğru olarak seçimine bağlıdır [20]. 2009 yılında doğru model seçimi ile ilgili bir çalışma yapılmıştır. Bu çalışmada web tabanlı uzaktan eğitim için video ve animasyon destekli öğretim modeline uygun iki farklı öğretim materyalleri bulunmaktadır. Bu materyaller öğrenci başarısı üzerindeki etkileri incelenmek üzere hazırlanmıştır. Çalışma sonucunda video destekli öğretim materyallerinin animasyon destekli öğretim materyallerine göre öğrenci başarısını daha olumlu etkilediği görülmüştür.

Akademik/Eğitim alanında gerçekleştirilen bir başka çalışma da Hacettepe Üniversitesi Beytepe Yerleşkesi yemekhanelerine ait 2008 yılı günlük yemek listeleri ve satış kayıtlarını içeren veriler üzerine yapılan çıkarımlar hakkındadır [21]. Bahsedilen veriler iki farklı Veri Madenciliği yöntemi olan karar ağaçları ve birliktelik kuralları ile analiz edilmişlerdir. Tüketim miktarlarını etkileyen kalıplar ve önemli diğer faktörler Veri Madenciliği teknikleri ile belirlenmeye çalışılmıştır. Karar ağacı modelleri ile %80 oranında tahmini tüketim başarısı elde edilmiştir.

Bankacılık ve Sigortacılık alanında; gerçekleştirilen çalışmada İstanbul Menkul Kıymetler Borsası (İMKB) 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerini içeren veri

setinden yararlanılmıştır [22]. Veri Madenciliği tekniklerinden biri olan karar ağacı tekniği modellemesiyle, seçilen finansal göstergelere göre sanayi ve hizmet sektörlerinde faaliyet gösteren firmaları ayıran en önemli değişkenler saptanmıştır.

Telekomünikasyon alanında; Veri Madenciliğinin yaygın kullanıldığı uygulama alanlarından biri de müşteriler ile ilgili yapılan analizlerdir [23]. Ayrılma eğilimine sahip müşterilerin tahmin edilmesi (churn), şirketlerin bu potansiyeldeki müşterilere özel pazarlama kampanyaları oluşturmasına olanak sağlamaktadır. 2010 yılında yapılan bir çalışmada Türkiye'de telekomünikasyon sektöründe faaliyet gösteren büyük bir firmanın "churn" stratejisini geliştirmesi ele alınmıştır. Lojistik Regresyon Analizi ve sınıflandırma tekniklerinden Karar Ağaçları kullanılarak uygulama tamamlanmıştır.

Ticari alanda; 2007 yılında işletmelerin kalite iyileştirmeleri kategorisinde gerçekleştirilen çalışma, Veri Madenciliği tekniğinin kullanılmasına örnek olarak gösterilebilmektedir [24]. Çalışmada sürücü koltuğu kalitesi için müşteri memnuniyeti verileri analiz edilmiştir. Memnuniyeti etkileyen en önemli değişkenlerin belirlenmesi için karar ağacı yaklaşımları uygulanmıştır. Uygulama sonuçları, farklı bir çalışmada lojistik regresyon analizinden elde edilen sonuçlar ile kıyaslanmıştır.

Tıp/Sağlık alanında; 2008 yılında gerçekleştirilen bir çalışmada, kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde, hekime yardımcı olacak bir karar destek sistemi oluşturulmuştur [25]. Çalışmada örüntü tanıma sistemi esas alınmış olup, sistemin işleyişi Veri Madenciliği tekniklerinden karar ağaçları ile sağlanmıştır. Biyokimya parametrelerinden demir eksikliği anemisi hastalığı için temel belirleyiciler olan Serum demiri, Serum demir bağlama kapasitesi (SDBK) ve Ferritin enzimleri kullanılmış olup, çıkış/sınıf değeri olarak da Anemi(+) ve Anemi(-) değerlendirmelerinde bulunulmuştur. Tasarlanan sistemde 96 hasta verisi değerlendirilmiştir. Karar destek sisteminin sonuçları, doktorun verdiği kararlarla tamamen örtüştüğü sonucuna varılmıştır.

2010 yılında gerçekleştirilen bir diğer Veri Madenciliği tıp alanına ait olan çalışma da, kanser çeşitlerinden biri olan ve kadınlar arasında en sık görülen meme kanseri hastalığının teşhisine yarar sağlayan uygulamadır [26]. Xcyt örüntü tanıma programı yardımı ile dokular hakkında genel veriler elde edilerek, bu verilerin WEKA programında modellenmesi ile meme kanseri hücrelerinin tahmin ve teşhis işlemi gerçekleştirilmiştir.

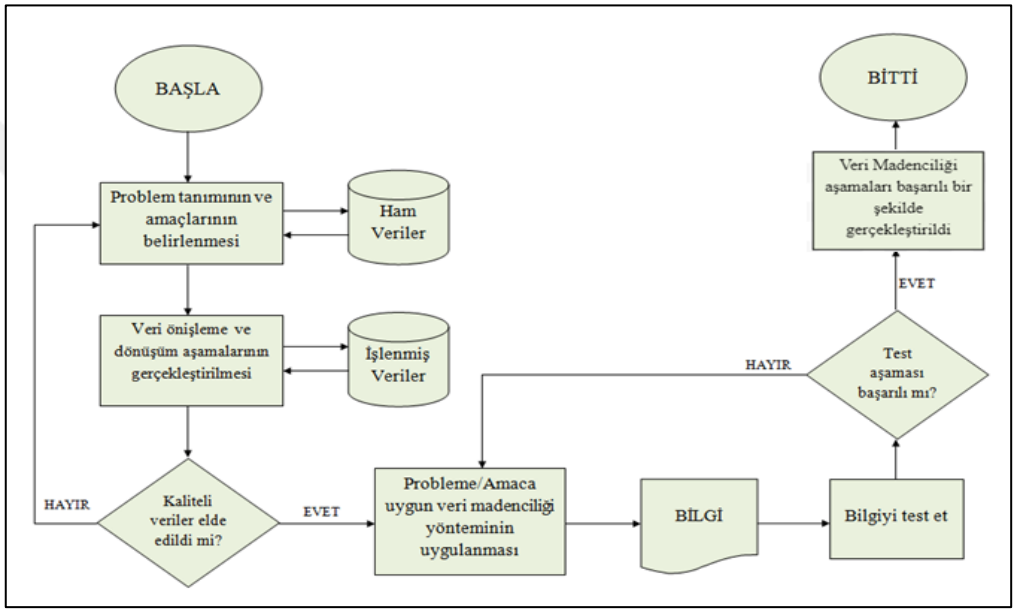
Bu tez çalışmasında Sınıflandırma yöntemleri ele alınacağından dolayı, yukarıda da Veri Madenciliği ile ilgili verilen örnek çalışmaların bu yöntemler ile oluşturuluyor olması tercih edilmiştir. Sınıflandırma yönteminin yaygın kullanım alanları, örneklerden de hareket ile genellenmek istenirse [27];

- Belirli bir sınıfın, olası üyesi olacak elemanların belirlenmesi
- Çeşitli vakaların, çeşitli kategorilere ayrılması
- Parametrik modellemede kullanılmak üzere çok sayıdaki değişkenden en önemlilerinin seçilmesi
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi

gibi ifadeler uygun olmaktadır.

3. MATERYAL VE YÖNTEM

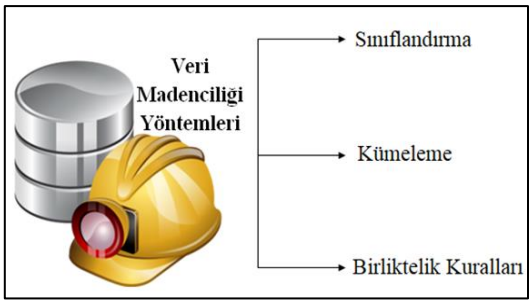
Veri Madenciliği kavramının kullanım şekli, kullanım nedeni, kullanım amacı vb. nitelikleri tek bir çatı altında ele alınabiliyor olsa da belirtilen bu niteliklere ulaşılabilmesi için farklı yöntemlerden faydalanılmaktadır. Veri Madenciliği kavramının genel işleyiş yapısı Şekil 3.1'de belirtilen akış şemasında ifade edilmektedir.



Şekil 3.1. Veri Madenciliği Aşamaları Genel Yapısının Akış Şeması.

3.1. VERİ MADENCİLİĞİ YÖNTEMLERİ

İstatistik, makine öğrenmesi, yapay zekâ ve veri tabanı yönetim sistemleri gibi farklı teknolojilerin harmanlanması olan Veri Madenciliği teknolojisinin temel amacı kullanılan verilerden anlamlı bir çıktı/bilgi edinmektir [28]. Bu amaç için Veri Madenciliği teknolojisinin farklı uygulama yöntemleri kullanılmaktadır. Kullanılan bu yöntemler Şekil 3.2'de belirtildiği gibi Sınıflandırma, Kümeleme ve Birliktelik Kuralları olarak ele alınmaktadır. Ele alınan bu yöntemlerin de kendi içlerinde özelleştikleri farklı algoritmik modeller bulunmaktadır.



Şekil 3.2. Veri Madenciliği Aşamaları Genel Yapısının Akış Şeması.

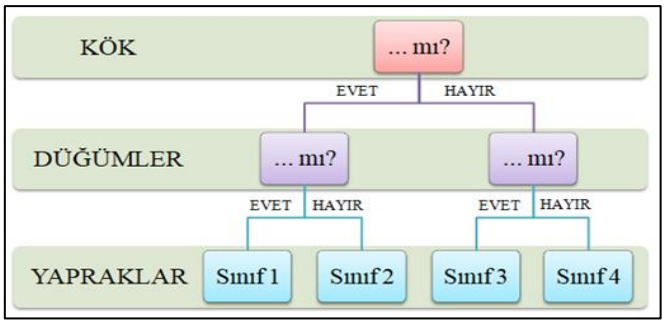
3.1.1. Sınıflandırma

Sınıflandırma yöntemi, tahmin edici ya da bir diğer ifadeyle kestirimci bir yöntemdir [29]. Örneklem içerisinde bulunan niteliklerden faydalanılarak eğitim seti verileri ile oluşturulan model kapsamında belirlenen çıktı/sınıf değişkeninin elde edilmesi işlemi Veri Madenciliği sınıflandırma yöntemini ifade etmektedir.

Kategorik konularda düzenin sağlanabilmesi, verilerin kolay kullanılabilirliği için sınıflandırma yöntemi tercih edilmektedir. Kredi başvuru sonuçları, hastalık teşhisleri, müşteri bilgileri doğrultusunda reklam gösterimi, öğrencilerin başarı/başarısızlık durumları gibi bankacılık, sağlık, pazarlama, eğitim, hukuk, sosyal medya vb. birçok sektör ve durum sınıflandırma yönteminin uygulama alanları arasındadır [30][31][32]. İlgili Çalışmalar bölümünde ifade edilen örnekler de bahsedilen bu yargı için geçerli olmaktadır.

Sınıflandırma yöntemi mantığını benimseyen birçok teknik ve bu tekniği esas alan birçok algoritma geliştirilmiştir. Karar Ağaçları, Yapay Sinir Ağları, Örnek/Bellek Tabanlı Algoritmalar, Genetik Algoritmalar, Naive-Bayes ve Lojistik Regresyon sınıflandırma mantığını benimseyen tekniklerdir [18].

Karar Ağaçları (Decision Trees): Karar ağacı yapısı, kuralları çok basit bir şekilde sunmaktadır. Karar ağacı dallanma yapısı ile niteliğin doğruluğu sorgulanıp yola o şekilde devam edilmektedir. Bu sorgulama tekniği ile ağaç, tüm kayıtları kullanacak şekilde büyümesini tamamlamaktadır [33]. Şekil 3.3'te karar ağacının genel yapısı ifade edilmiştir.



Şekil 3.3. Karar Ağacı Yapısı ve Ağacı Oluşturan Elemanların İfade Edilmesi.

Bir sistemdeki rastgelelik, düzensizlik "entropi" olarak tanımlanır. Karar ağacının, başarılı/verimli olması esasları arasında düğümlerin sıralaması yer almaktadır. Bu sıralamanın yapılmasında entropi kavramı etkisinde olan algoritmalara ID3 ve C4.5 Algoritmaları örnek gösterilebilmektedir [34]. Şekil 3.4'te entropi tabanlı algoritmaların mantığı, genel yapısı itibari ile akış şemasında ifade edilmektedir. Karar ağacı oluşumunda, dallanma aşamalarında entropi mantığını esas alan algoritmalar "Bilgi Kazanımı" kavramını da kullanmaktadırlar. Bilgi kazanımı, entropi ifadesinin tam tersi olarak formüllendirilmektedir [35].

Entropi ve Bilgi Kazanımı ifadeleri sırasıyla Formül (3.1) ve Formül (3.2)'de belirtilen şekilde hesaplanmaktadır.

Veri Seti: S

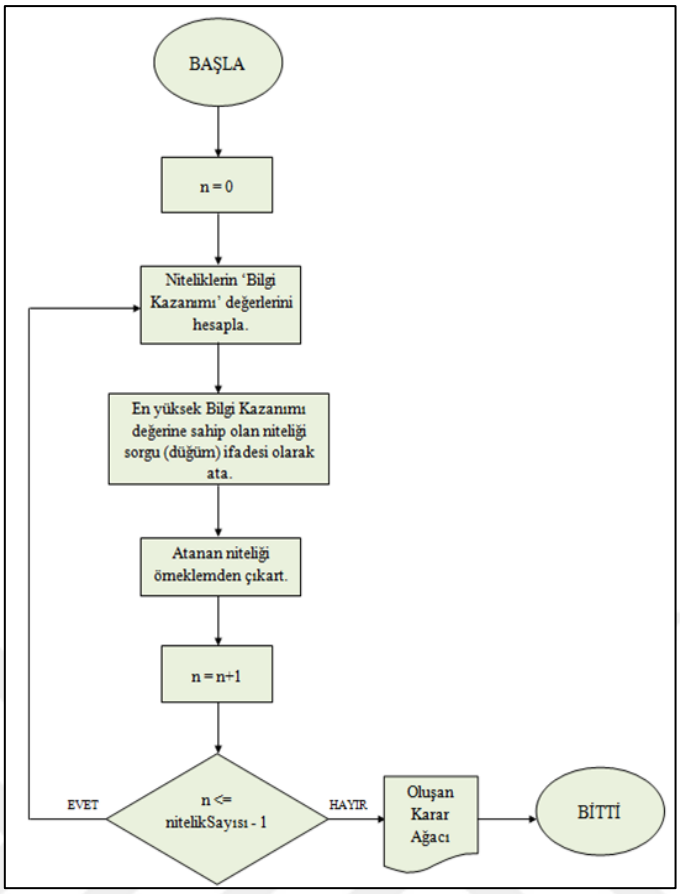
Kaynağın n adet üreteni: $\{m_1, m_2, \ldots, m_n\}$ (Üretilenler bağımsızdır.)

m_i üretilenin üretilme olasılığı: p_i

$P = \{p_1, p_2, \ldots, p_n\}$ olasılık dağılımına sahip üreten S kaynağının entropisi: H(S)

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (3.1)$$

$$\text{Bilgi Kazanımı}(S, \textit{öznitelik}) = H(S) - H(\textit{öznitelik}) \quad (3.2)$$



Şekil 3.4. Entropi Tabanlı Algoritma Mantığının Akış Şemasında İfade Edilmesi.

C4.5 ve ID3 Algoritmaları; Entropi tabanlı algoritmalar olarak ifade edilmiş olsalar dahi, işleyişlerinde bazı farklılıklar bulunmaktadır [36].

- ID3 Algoritmasında entropi veya bilgi kazanımı değeri hesabı yapılır. Bu değerlere göre karar belirlenir. C4.5 Algoritmasında ise normalleştirme kavramı kullanılmakta olup, entropi değerleri birer oran olarak saklanır.
- C4.5 Algoritmasında, karar ağacı üzerindeki erişim sıklıklarına göre alt ağaçların (subtree) farklı seviyelere taşınması mümkün olmaktadır.
- C4.5 Algoritmasında, karar ağacının verimini arttırabilmek adına budama (prunning) işlemi yapılabilmektedir.

Belirtilen farklılıklar izinde C4.5 Algoritmasının çalışma prensibi;

- Her adımda özellikleri kontrol et,
- Her özelliğin normalize edilmiş bilgi kazanım değerini hesapla,
- En iyi bilgi kazanım değerini veren özellikleri karar noktaları olarak belirle,

- Karar düğümünün ardına alt liste oluşturarak alt karar ağacı inşa et,

şeklinde aşamalar halinde ifade edilebilir [36].

Karar ağacı oluşturma tekniğini içinde barındıran bir diğer yöntem olan Sınıflandırma ve Regresyon kavramı birçok kaynakta "CART (Classification and Regression Trees: Sınıflandırma ve Regresyon Ağaçları)" olarak da isimlendirilmektedir [37].

Ortak özellikler barındırıp, belli bir grup altında ele alınabilen veriler kategorik değişken olarak ifade edilmektedir. Alt ve üst sınırları olup, bu sınır aralığındaki değerlere sahip olan veriler ise sürekli değişken olarak ifade edilmektedir [38]. Problemlere, kategorik ve sürekli değişkenlerin kullanılıp karar ağacı oluşturulması ile çözüm sunan Veri Madenciliği yöntemi Sınıflandırma ve Regresyon yöntemi altında ele alınmaktadır. Veri setindeki değişkenler kategorik ise sınıflandırma yöntemi karar ağaçları (Classification Trees), kullanılan değişkenler sürekli ise regresyon yöntemi karar ağaçları (Regression Tree) kullanılmaktadır [39]. Gini Algoritması, Rastgele Orman, Regresyon Ağaçları ve Twoing Algoritması sınıflandırma ve regresyon yöntemi altında ele alınan algoritmalara verilebilecek örneklerdir.

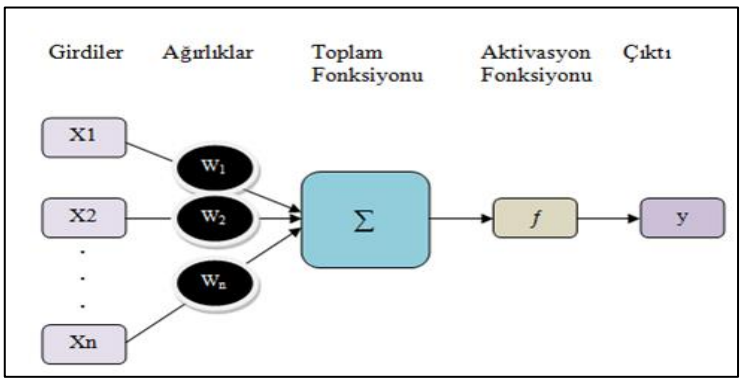
Gini Algoritması; Sorgu düğümlerinin hangi sıra ile dallanacağı işlemi için gini indeks değeri hesaplaması ile karar verilir. Gini indeks değeri aynı olan sorgulamalar için gini algoritması mantığı gereği ikili dallanma sağlanabileceğinden, sorgular arasında ilişkisellik sağlanarak düğüm sayısı iki olarak düzenlenmektedir [40].

Rastgele Orman (Random Forest); Rastgele Orman Algoritması, bir karar ağacına güvenmek yerine, her ağaçtan tahminlerinin alınıp oy çoğunluğundan hareketle sonuç olabilecek çıktının tahmin edilmesi şeklinde ifade edilebilmektedir [41].

Twoing Algoritması; Sınıflandırma ve Regresyon çatısı altında olduğundan CART özelliklerini barındırmaktadır. CART kapsamında farklı bir algoritma olarak belirtilmesinin sebebi düğüm noktalarındaki ayrımların oluşmasında kullanılmış olan teknik farklılığıdır. Eğitim kümesi her adım için iki kısma ayrılarak, grupların sağ ve sol kısımları için nitelik değerlerinin tekrar sayıları dikkate alınır. Uygunluk değeri en yüksek olan değer bölünme noktası olarak seçilmiş olmaktadır [42].

Yapay Sinir Ağları (YSA) (Artificial Neural Networks): Yapay sinir ağı modellemesi girdi, ağırlık, toplam fonksiyonu, aktivasyon fonksiyonu ve çıktı kavramlarından meydana gelmektedir. Girdi kavramı ağdan öğrenme gerçekleştirilecek örnekler, ağırlık kavramı gelen örnek bilginin önemi olarak ifade edilebilmektedir. Toplam fonksiyonu

ile hücreye gelen net hesaplamalar yapılırken, aktivasyon fonksiyonu net girdinin işlenip çıktı kavramını üretmesi işlemi olarak ifade edilebilmektedir [43].



Şekil 3.5.Yapay Sinir Ağı Modelinin Temel Yapısı.

Örnek/Bellek Tabanlı Algoritmalar: Örnek tabanlı öğrenmede eğitim örnekleri birebir saklanmaktadır. Bilinmeyen bir test örneğinin eğitim setinin hangi üyesine en yakın olduğuna karar vermek için uzaklık fonksiyonu kullanılmaktadır. En yakın eğitim örneği bulunduktan sonra, test örneği için sınıf tahmin edilmektedir [44]. K-En Yakın Komşu Algoritması (KNN), örnek/bellek tabanlı algoritmalar kategorisi için verilebilecek en uygun örnek olmaktadır.

K-En Yakın Komşu (K-Nearst Neighbours - KNN) Algoritması; K-en yakın komşu algoritması çoğunlukla tüm özellikler sürekli değer (lineer) olduğunda tercih edilmektedir. Ancak algoritmanın kategorik özelliklerle de başa çıkması için üzerinde değişiklik yapılabilmektedir [45]. Algoritmanın isminde de geçen k değeri, komşu olan kayıtların sayısını temsil etmektedir. KNN algoritması, verilen n adet prototip örüntüye ve bu örüntülerin doğru sınıflandırılmasına göre, sınıfı bilinmeyen bir örüntüyü en yakın komşu gruba atamaktadır [46].

Genetik Algoritmalar (Genetic Algorithms): Genetik algoritmaların yapısı evrimsel sürece benzetildiğinden dolayı bu isme sahip olan modelleme yöntemidir. Örnekleme bu algoritma yöntemi uygulandığında, örnek girdi çözüm değil alternatif çözümleri barındıran çözüm kümesi elde etmiş olur. Girdilere çözüm kümesi sunan bu yöntemin işleyişi popülasyon oluşturma, seçim, çaprazlama ve mutasyon gibi evrimsel aşamalara benzetilmektedir [47].

Naive-Bayes: Naive Bayes sınıflandırıcısı, Bayes teoreminin bağımsızlaştırılarak basitleştirilmiş hali olarak ifade edilebilmektedir. Bayes teoremine göre A ve B şeklinde

iki olay ele alınmış olsun. B olayının olması durumu ile birlikte A olayının da olması ihtimalinin, yalnızca A olayının olması durumu olasılığı ile çarpılıp, yalnızca B olayının olması durumuna oranlanması değeri elde edilir [48].

Açıklanan bu olasılık ifadesinin daha iyi kavranabilmesi için örnek durum üzerinden somutlaştırılma gerçekleştirilecek olunursa; iş çıkışı arkadaşlarla kahve içmeye gitmek ya da gitmemek çıktısını etkileyen tek faktör işin yoğunluğu değildir. Hava durumu, farklı bir plana davet edilme vb. birçok durum etkisi altında kalınarak kahve içmeye gitmek ya da gitmemek kararına varılmış olur.

Lojistik Regresyon: Sınıflandırma yönteminin bir başka algoritmik çeşidi olan lojistik regresyon yöntemi, tahmin edilmek istenilen değişken tipinin kategorik olduğu durumlarda tercih edilmektedir. Örneğin, örneklemin sınıf değerlerinin medeni durum (evli/bekar), cinsiyet (kadın/erkek) vb. olduğu durumlar lojistik regresyon kullanılması için en ideal örneklem kümeleridir.

3.1.2. Kümeleme

Küme kavramı anlam olarak, ortak özellikleri bulunan örnekleri bir arada barındıran örneklemdir. Kümeleme yöntemi de küme kavramının anlamına uygun olacak şekilde örneklemlerin ayrılmasını/bölünmesini sağlamaktadır. Sınıf değeri oluşturacak kadar keskin bilgi bulundurmayan örneklemler için tercih edilen bu Veri Madenciliği yöntemi, örneklem niteliklerinden faydalanılarak benzerliklere göre kümelenmesi işlemi olarak ifade edilebilmektedir [49][50][51].

3.1.3. Birliktelik Kuralları

Eş zamanlı-bağlantılı işlenen durumların ilişkiselliklerini belirlemek için kullanılan Veri Madenciliği yöntemidir. Birliktelik kurallarının bu tanımlamasına genellikle pazar/market alışverişi örnek olarak verilmektedir [2].

Birliktelik Kuralları yönteminin bir başka kullanım şekli, ardışık zamanlı işlenen durumların ilişkiselliklerinin belirlenmesi için kullanılmasıdır. Ardışık zamanlı olarak gerçekleşen birliktelik kuralları yöntemine de,koşullu dersler durumu ile ifade edilmek istenilen “A dersini başarılı bir şekilde geçme kuralı ile B dersinin seçilebilmesi işlemi” örnek olarak verilebilmektedir.

3.2. AÇIK KAYNAK KODLU VERİ MADENCİLİĞİ YAZILIMLARI

Veri Madenciliği kavramı karar verme aşamasında çıktı üretirken yaptığı hesaplamalarda istatistik biliminden faydalanmaktadır [52]. İstatistik biliminden en önemli farkı, yapılan bu hesaplamaların yoğunluğu, kullanılan veri sayısı ve niteliklerin fazlalığı ile bir hedef çıktı elde edilmesidir. Bu sebeple Veri Madenciliği yöntemlerinin kullanılması için yarar sağlayacak uygulamaların gerçekleştirileceği yazılım ortamları bulunmaktadır. Weka, R, Orange ve sıralanabilecek daha birçok Veri Madenciliği yazılım ortamları, uygulamaların gerçekleştirildiği ortamlara örnek verilebilmektedir.

3.2.1. Weka (The Waikito Environment for Knowledge Analysis)

Yazılımda kullanılan öğrenme algoritmaları birçok dile uyarlanabilir ve farklı formatlarda kullanılabilmektedir [53]. Weka ortamında hazır halde bulunan algoritmalar sayesinde, kullanılan veri seti direkt istenilen algoritma çeşidi ile modellenmektedir. Tercih edildiği takdirde, kullanıcı tarafından Java kodu ile de algoritma oluşumu sağlanabilmektedir. Weka'ya entegre edilme işlemi için ".arff" ve ".csv" formatlarındaki dosyalar uygunluk sağlamaktadır.

3.2.2. R Programlama

Auckland Üniversitesi İstatistik Bölümündeki bilim insanı Robert Gentleman ve RossIhaka tarafından istatistiksel hesaplamalar ve veri analizleri için geliştirilen bir yazılımdır [54]. İstatistiksel hesaplama ve veri analizlerinin yanı sıra aynı zamanda bir programlama dilidir.

3.2.3. Orange

Slovenya Ljubljana Üniversitesi Bilgisayar ve Enformatik Bilimleri bölümü yapay zeka ekibi tarafından C++ programlama dili ile geliştirilmiş bir yazılımdır. Arayüz tasarımı ve grafik oluşum ortamı için Python programlama dilinden yararlanılmıştır [55]. Orange ortamına, yalnızca metin dosyaları entegre edilebilmektedir.

Bahsedilen bu üç Veri Madenciliği yazılım ortamlarının teknik özellikleri Çizelge 3.1'de kıyaslanmaktadır [56].

Çizelge 3.1. Bazı Veri Madenciliği Ortamlarının Karşılaştırılması.

Yazılım Ortamı / **Özellik**	**Weka**	**R**	**Orange**
Logo			
Veri Madenciliği Algoritmaları	Var (En fazla)	Var (En az)	Var
İstatistiksel Hesaplama	Var (Zayıf)	Var (Çok güçlü)	Var (Zayıf)
Veri Analizi	Var	Var	Var
Ön işleme	Var	Var	Var
Sınıflandırma	Var	Var	Var
Kümeleme	Var	Var	Var
Birliktelik Kuralları Çıkarımı	Var	Var (Paket olarak)	Var
Nitelik Seçimi	Var	Var	Var
Komut Satırı Arayüz Bağlantısı	Var (Çok iyi)	Var (Çok zayıf)	Var
Kullanım/Öğrenim Kolaylığı	Evet (En kolay)	Evet	Evet
Eklenebilir Paketler	Evet	Evet	Evet
Veri Alma/Verme	Var	Var (Çok kolay)	Var
Desteklenen Dosya Formatları	.arff .csv	.r .txt .ods .csv .xml	.tab .basket .names .data .txt .xls .arff ve .csv sadece okunabilir.
Veritabanları ile Çalışabilme	Var	Var	Var
Excel Dosyaları İle Çalışabilme	Evet (Kötü Bağlantı)	Evet	Hayır
Bellek Kullanımı	Arttırılabilir / Ayarlanabilir	Limitli	Limitli
Yazıldığı Dil	Java	C C++ R Fortran	Python C++
Kurulum Şartı	Var	Var	Var
Gerekli Minimum İşletim Sistemi	MS Windows, GNU/Linux, Mac OsX	MS Windows, GNU/Linux, Unix, Mac OsX	MS Windows, GNU/Linux, Mac OsX

3.3. "VERİ BİLİMİ ANKETİ" İSİMLİ VERİ SETİNİN MODELLEMEYE HAZIR HALE GETİRİLMESİ

Veri Madenciliği kavramı ve bu kavrama ait yöntemlerden Sınıflandırma yöntemi algoritma çeşitlerine ve kullanımına teorik olarak değinilmiştir. Kavramların ve yöntemlerin teorik bilgilerinin yanı sıra uygulamalı olarak da aktarılması için, seçilen veri seti üzerinde, seçilen sınıflandırma yöntemi algoritmaları ile denemeler gerçekleştirilmiştir. Uygulama bölümünde aktarılacak çalışma, Şekil 3.6'da ifade edilmektedir.



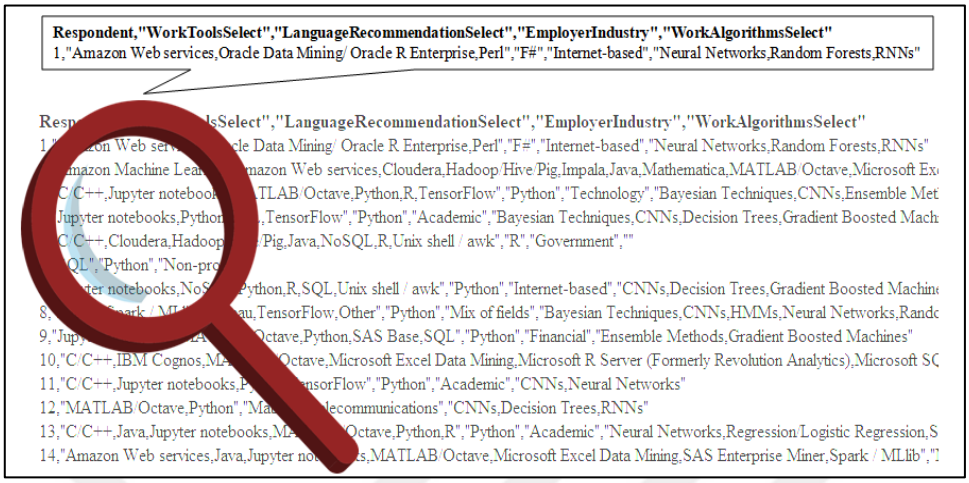
Şekil 3.6.Uygulama Bölümü Çalışma Özeti.

3.3.1. Kullanılan Veri Seti: "Veri Bilimi Anketi"

Ele alınan "Veri Bilimi Anketi" [19] isimli veri setinde, veri bilimi üzerine çalışmalar yapan 10.000'den fazla bilim insanının vermiş oldukları cevaplar doğrultusunda tercih ettikleri bilimsel yöntemler, algoritmalar, sistemler ve çalışmakta oldukları sektörler gibi bilgiler yer almaktadır.

İşlenmemiş, ham halde bulunan veri seti incelenecek olursa; veri seti 10153 satır ve 5 nitelikten oluşmaktadır. Bu 5 nitelik "Numara (Respondent)", "Kullanılan Çalışma Aracı (Work Tools Select)", "Tercih Edilen Dil (Language Recommendation Select)", "Çalışma Alanı (Employer Industry)" ve "Katılımcılar Tarafından Kullanılan Algoritma (Work Algorithms Select)" başlıkları altında ele alınmışlardır. Herhangi bir satır ele alınıp, o satıra ait nitelik verileri incelenecek olur ise, bir nitelik alanında birden fazla

veri değerinin olduğu gözlemlenmektedir. Niteliklerin veri türleri karakter dizisi formatındadır. Şekil 3.7'de veri setinin bir bölümü görsel olarak ifade edilmektedir.



Şekil 3.7.Veri Setinde Bir Kesit (.csv formatı).

- **Numara/ID (Respondent):**

Veri Madenciliği yöntemleri kullanılarak karar verme, bilgi üretme aşamasında bir yarar sağlamayan bu nitelik, yalnızca veri setinde bulunan diğer nitelikleri numaralandırmak üzere kullanılan benzersiz (unique) değerler içermektedir. Karar vermede katkısı olmadığından, model oluşumunda kullanılacak işlenmiş veri setinde de bu niteliğe yer verilmemiştir.

- **Kullanılan Çalışma Aracı (Work Tools Select):**

Anketi cevaplamış olan veri bilimcilerin, kullandıkları çalışma araçları ile ilgili bilgilerin yer aldığı sütundur. Bir veri bilimcinin birden fazla çalışma aracı kullanma durumu olabileceğinden dolayı, tek satırda birden fazla çalışma aracı verileri bulunabilmektedir. Kullanılan Çalışma Aracı niteliği olarak belirtilen sütunun %22'si veri değeri içermemekte olup, boş bulunmaktadır. Veri setinde, Kullanılan Çalışma Aracı niteliğine ait benzersiz veri değerleri Çizelge 3.2'de belirtilen şekilde bulunmaktadır.

Çizelge 3.2. Veri Setinin “Kullanılan Çalışma Aracı” İsimli Niteliğine Ait Benzersiz Veri Çeşitleri.

Kullanılan Çalışma Aracı Niteliği Veri Çeşitleri	
Amazon Machine Learning	Microsoft SQL Server Data Mining
Amazon Web Services	Minitab
Angoss	NoSQL
C/C++	Oracle Data Mining/Oracle R Enterprise
Cloudera	Orange
Data Robot	Perl
Diğer	Python
Flume	QlikView
Google Cloud Compute	R
Hadoop/Hive/Pig	RapidMiner (commercial version)
IBM Cognos	RapidMiner (free version)
IBM SPSS Modeler	Salfrod Systems CART /MARS/ TreeNet/ RF/ SPM
IBM SPSS Statistics	SAP Business Objects Predictive Analytics
IBM Watson / Waton Analytics	SAS Base
Impala	SAS Enterprise Miner
Java	SAS JMP
Julia	Spark / MLlib
Jupyter notebooks	SQL
KNIME (commercial version)	Stan
KNIME (free version)	Statistica (Quest/Dell-formerly Statsoft)
Mathematica	Tableau
MATLAB/Octave	TensorFlow
Microsoft Azure Machine Learning	TIBCO Spotfire
Microsoft Excel Data Mining	Unix shell / awk
Microsoft R Server (Formerly Revolution Analytics)	

- **Tercih Edilen Dil (Language Recommendation Select):**

Anketi cevaplamış olan veri bilimcilerin tercih ettikleri yazılım dili ile ilgili bilgilerin yer aldığı sütundur. Kullanıcılar birden fazla yazılım dili tercih edebildiklerinden dolayı tek satırda birden fazla veri çeşidi yer alabilmektedir. Tercih Edilen Dil niteliğinin %36’lık bölümünde veri değeri bulunmamakta olup, hücreler boş bulunmaktadır. Veri

setinde Tercih Edilen Dil niteliğine ait benzersiz veri değerleri Çizelge 3.3'te belirtilen şekilde kullanılmaktadır.

Çizelge 3.3. Veri Setinin "Tercih Edilen Dil" İsimli Niteliğine Ait Benzersiz Veri Çeşitleri.

Tercih Edilen Dil Niteliği Veri Çeşitleri		
F#	C/C++/C#	Diğer
Haskell	Java	Julia
Matlab	Python	R
SAS	Scala	SQL
Stata		

- **Katılımcılar Tarafından Kullanılan Algoritma (Work Algortihms Select):**

Anketi cevaplamış olan veri bilimcilerin kullandıkları algoritma bilgilerinin yer aldığı sütundur. Kullanıcılar, birden fazla algoritma kullanımı sağlayabildiklerinden dolayı, bu nitelik türüne ait tek satırda birden fazla veri değeri yer alabilmektedir. Katılımcılar tarafından kullanılan algoritma türü bilgilerinin yer aldığı bu sütunda %28 oranında boş hücre değeri bulunmaktadır. Veri setinde bu niteliğe dair kullanılan benzersiz veri değerleri, Çizelge 3.4'te belirtilen şekildedir.

Çizelge 3.4. Veri Setinin "Katılımcılar Tarafından Kullanılan Algoritma" İsimli Niteliğine Ait Benzersiz Veri Çeşitleri.

Katılımcılar Tarafından Kullanılan Algoritma Niteliği Veri Çeşitleri	
Bayesian Techniques (Bayesian Tekniği)	CNNs (Evrişimli Sinir Ağı)
Decision Trees (Karar Ağaçları)	Diğer
Ensemble Methods (Çoklu Öğrenme Algoritmaları Yöntemi)	Evolutionary Approaches (Evrimsel Yaklaşımlar)
GANs (Üretici Muhalif Ağı)	Gradient Boosted Machines (Gradyan Yükseltmeli Makine)
HMMs (Gizli Markov Modeli)	Markov Logic Networks(Markov Mantık Ağı)
Neural Networks(Yapay Sinir Ağları)	Random Forests (Rastgele Orman)
Regression/Logistic Regression (Regresyon/Lojistik Regresyon)	RNNs (Yinelenen Sinir Ağı)
SVMs (Destek Vektör Mekanizması)	

- **Çalışma Alanı (Employer Industry):**

Veri Madenciliği teknolojisinin birçok sektöre katkı sağladığından önceki bölümlerde bahsedilmiş idi. Veri setinde Çalışma Alanı olarak belirtilen bu nitelik de, "Bahsedilen veri bilimi teknikleri kullanılarak hangi sektöre hizmet sunulmuştur?" bilgisini barındırmaktadır. Çalışma Alanı niteliğinin veri seti içerisinde %12'lik bölümü boş hücrelerden oluşmaktadır. Çalışma Alanı niteliğinde kullanılan benzersiz veri değerleri Çizelge 3.5'te belirtilen şekilde veri setinde bulunmaktadır.

Çizelge 3.5. Veri Setinin "Çalışma Alanı" İsimli Niteliğine Ait Benzersiz Veri Çeşitleri.

Çalışma Alanı Niteliği Veri Çeşitleri		
Akademik	Alanların karışımı	Askeri/Güvenlik
Devlet	Diğer	Eczacılığa dair
Finansal	İmalat	İnternet tabanlı
Kâr amacı gütmeyen	Konaklama/Eğlence/Spor Dalları	Müşteri İlişkileri Yönetimi/Pazarlama
Perakende	Sigorta	Teknoloji
Telekomünikasyon		

3.3.2. Problem Tanımlaması ve Veri Ön İşleme

Veri ön işleme işlemi, algoritma seçimi, algoritma başarı hesabı gibi Veri Madenciliğinin bütününü oluşturan işlem adımlarının etkin/verimli bir biçimde yürütülebilmesi için en önemli ve ilk aşama, problemin iyi bir şekilde anlaşılabilmesidir. Yukarıda belirtilen veri seti ile ne yapılmak isteniyor, nasıl yapılmak isteniyor, hangi nitelikler girdi/çıktı değeri gibi yargılar netleştirildiğinde diğer işlem adımları birbirini daha kolay bir biçimde takip etmiş olmaktadır.

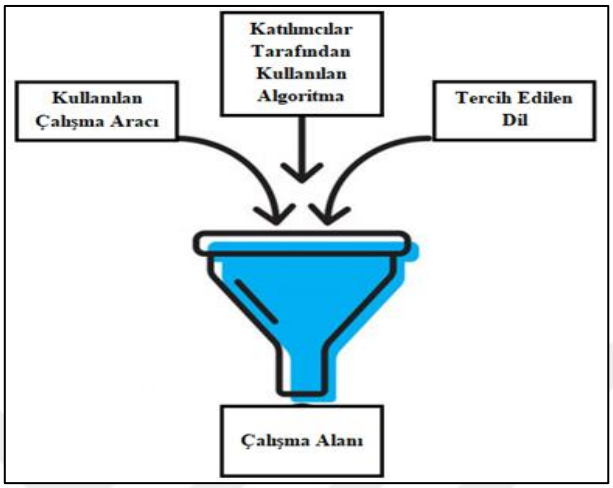
"Veri Bilimi Anketi" isimli veri setindeki ID, Kullanılan Çalışma Aracı, Tercih Edilen Dil, Katılımcılar Tarafından Kullanılan Algoritma nitelikleri girdi olarak kabul edilip, Çalışma Alanı niteliği çıktısının tahmini üzerine bir Veri Madenciliği çalışması planlanması üzerine Veri Hazırlık aşamasına geçilmiştir.

- **Gereksiz Veri Alanları Tespiti**

Veri seti niteliklerinin tanımlamaları yapılırken bahsedilen ID niteliği benzersiz (unique) değerler içerdiğinden dolayı, veriler arası bağlantı sağlanmasında ya da çıktı

değerinin belirlenmesinde bir katkısı olmamaktadır. Bu yüzden veri seti girdi nitelikleri arasından, ID sütununun kaldırılmasının uygun olduğu kanısına varılmıştır.

Gereksiz alan olarak belirlenen ID nitelik değerlerinin kaldırılması sonrasında çalışma kapsamında ele alınan problemin girdi ve çıktı nitelikleri Şekil 3.8'de ifade edilmiştir.



Şekil 3.8. Veri Setinin Girdi ve Çıktı Niteliklerinin İfade Edilmesi.

Veri seti analizi, problem tanımlama ve anlaşılma işlemleri gerçekleştirildikten sonraki aşama, anlaşılan bu probleme oluşturulması gereken çözümdür. Kullanılan Çalışma Aracı, Katılımcılar Tarafından Kullanılan Algoritma ve Tercih Edilen Dil niteliklerinin veri değerleri kullanılarak hangi sınıfa (Çalışma Alanı) ait olduğu bilgisinin elde edilmesi hedeflenen problem olduğundan, bu çalışma bir sınıflandırma problemi kategorisi altında değerlendirilmektedir.

Ele alınan probleme ait verilerin, seçilen sınıflandırma algoritmalarında daha verimli bir şekilde kullanılıp, üretilecek çıktı/sonuç değerinin daha başarılı bir şekilde elde edilebilmesi için veri ön işleme adımlarının gerçekleştirilmesi gerekmektedir.

- **Eksik Verilerin Temizlenmesi**

Veri setinde bulunan eksik veriler normalizasyon vb. gibi işlemler ile tamamlanabilir. Ancak "Veri Bilimi Anketi" isimli veri setindeki veri sayısının yeterli seviyede olduğu düşünüldüğünden, eksik veri niteliğine sahip satırın silinmesi ile veri temizleme aşaması tamamlanmıştır.

Veri seti, satır bazında incelendiğinde tek hücresinde dahi "null" değeri barındıran veri satırı veri setinden çıkartılarak eksik verilerin temizlenmesi işlemi gerçekleştirilmiştir.

Çizelge 3.6’da, veri setinde eksik veri hücresine sahip 2 örnek veri satırı ifade edilmiştir. Çizelge 3.6’da belirtilen örnek satırlar gibi,veri setinde “null” değere sahip tüm veri satırları veri setinden çıkartılarak eksik verilerin temizlenmesi aşaması gerçekleştirilmiştir.

Çizelge 3.6. Veri Setinde "null" Değerine Sahip Örnek İki Veri Satırı.

Kullanılan Çalışma Aracı	**Tercih Edilen Dil**	**Katılımcılar Tarafından Kullanılan Algoritma**	**Çalışma Alanı**
Java	null	Random Forest	Akademik
null	null	null	null

- **Nitelik Düzenlemesi/Veri Çeşidi Birleşimi**

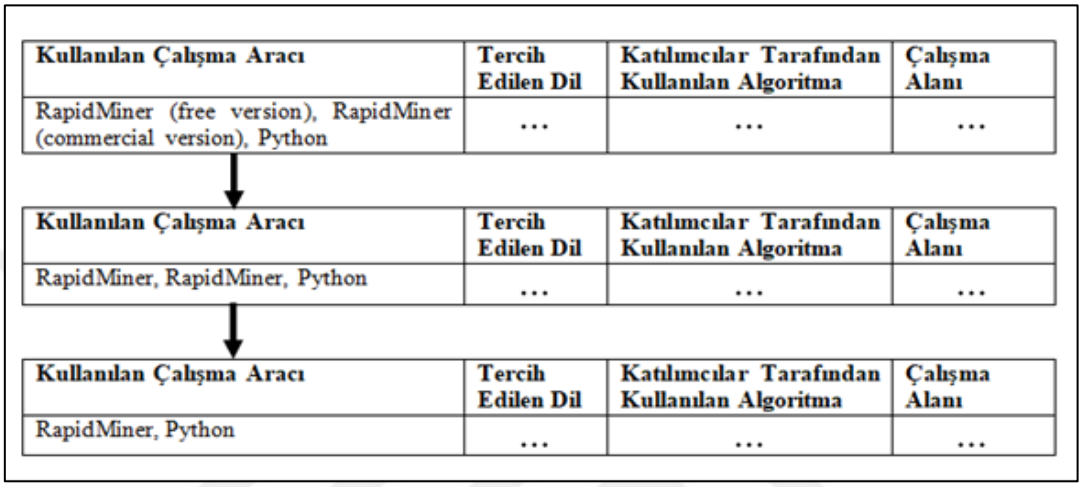
“Veri Bilimi Anketi” isimli veri setinde kullanılan niteliklerin benzersiz veri çeşitleri, her bir nitelik için oluşturulan başlık altında ifade edilmişti. Nitelikler temel alınarak değerlendirildiği takdirde, bazı veri çeşitlerinin aynı/benzer kavramı ifade etmesine rağmen farklı şekilde ifade edildiği gözlemlenmiştir. Bu veriler farklı ifade edildiklerinden dolayı algoritma modeli tarafından kullanıldığında ayrı/farklı veri değeri olarak ele alınmış olacaktır. Ancak aynı/benzer durumu ifade eden veri çeşitleri belirlenip düzenleme gerçekleştirildiği takdirde, algoritma için daha verimli veri nitelik değerleri kazanılmış olacaktır.

Örneğin; Kullanılan Çalışma Aracı sütununun veri çeşitleri incelendiğinde RapidMiner (free version) ve RapidMiner (commercial version) olmak üzere iki farklı veri çeşidi içerdiği gözlemlenmiştir. Versiyon farkı gibi, sınıf bilgisine ulaşmada iki farklı veri değeri olarak algılanıp, sapma oluşturabilecek veriler için birleşim işlemi uygulanmıştır. Çizelge 3.7’de veri çeşidi birleşim işleminin ifade edilebilmesi için iki farklı durum örnek olarak sunulmuştur.

Çizelge 3.7. Veri Çeşidi Birleşimi Örneği.

Kullanılan Çalışma Aracı Niteliği	
Veri Çeşidi	**Veri Çeşidi Birleşimi**
RapidMiner (free version)	RapidMiner
RapidMiner (commercial version)	
KNIME (free version)	KNIME

Veri çeşidi birleşimi aşamasından sonra, veri setinin satır bazında nitelikleri incelenmiştir. Aynı satır için nitelik bazında tekrarlı verilerin olduğu gözlemlenmiştir. Her satır farklı veri bilimciyi temsil ettiğinden, tekrarlı verilerin olması demek bahsedilen nitelik değerine veri bilimcinin tekrar aynı cevabı vermiş olması demektir. Bu yüzden satırlardaki nitelik değerlerinin tekrarlı durumları kaldırılmıştır. Bahsedilen durum veri setinden örnek bir satır ile Şekil 3.9'da aşamalı olarak ifade edilmiştir.

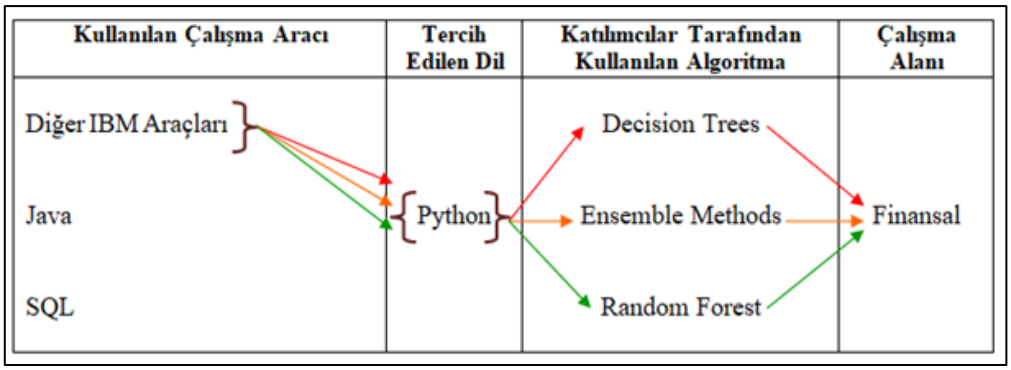


Şekil 3.9. Veri Birleşimi ve Düzenleme Örneği.

- **Veri Artırımı**

Veri setinde bulunan değerlere normalizasyon işlemi uygulanmayıp temizleme işlemi uygulanmasının bir diğer sebebi de, veri setine artırım uygulanacak olmasıdır. Veri setinde tek satırda aynı nitelik için farklı veri değerleri bulunmaktadır. Bu durumun da veriler arasında bazı çelişki/karışıklığa yol açacağı düşünüldüğünden dolayı tek hücrede tek veri olacak şekilde, veri tutarlılığını da artırmaya yönelik bir uygulama gerçekleştirilmiştir. Bu sayede sınıf bilgisi (Çalışma Alanı) çıkarımı yapılırken, verilerden daha fazla yarar sağlanabileceği düşünülmektedir.

Tek hücre tek veri mantığı ile anlatılmak istenilen durum Şekil 3.10'da görsel olarak ifade edilmektedir.



Şekil 3.10. Veri Artırım Mantığının Veri Seti Üzerinden Örneklendirilmesi.

Veri seti düzenleme öncesi Çizelge 3.8'de belirtilen şekilde,veri setinin bazı satırlarına ait bazı hücreleri içerisinde birden fazla veri değeri bulunmaktadır.

Çizelge 3.8. Orijinal Veri Setinden Bir Satır.

Kullanılan Çalışma Aracı	**Tercih Edilen Dil**	**Katılımcılar Tarafından Kullanılan Algoritma**	**Çalışma Alanı**
Diğer IBM Araçları, Java, SQL	Python	Decision Trees, Ensemble Methods, Random Forest	Finansal

Açıklanmış olan veri artırım işlemi Çizelge 3.8'de örnek gösterilen satıra uygulandığı takdirde Çizelge 3.9'da ifade edilen şekle dönüşmüş olmaktadır.

Çizelge 3.9. Çizelge 3.8'deki Veri Satırına Veri Artırım İşlemi Uygulaması Sonrası.

Kullanılan Çalışma Aracı	**Tercih Edilen Dil**	**Katılımcılar Tarafından Kullanılan Algoritma**	**Çalışma Alanı**
Diğer IBM Araçları	Python	Decision Trees	Finansal
Diğer IBM Araçları	Python	Ensemble Methods	Finansal
Diğer IBM Araçları	Python	Random Forest	Finansal
Java	Python	Decision Trees	Finansal
Java	Python	Ensemble Methods	Finansal
Java	Python	Random Forest	Finansal
SQL	Python	Decision Trees	Finansal
SQL	Python	Ensemble Methods	Finansal
SQL	Python	Random Forest	Finansal
Python	Python	Decision Trees	Finansal
Python	Python	Ensemble Methods	Finansal
Python	Python	Random Forest	Finansal

Çizelge 3.8'de örnek bir satır ele alınarak, Çizelge 3.9'daki şekle dönüştürülmesini sağlayan veri artırım işlemi tüm veri seti için gerçekleştirilmiştir.

- **Aykırı Değerlerin Düzenlenmesi**

Veri setine uygulanacak olan son veri ön işleme aşaması, aykırı değerlerin tespit edilmesidir. Nitelikler sayısal değerler ile ifade edildiğinde, belirlenen aralığın dışında kalan değerler ya da belirli bir formül uygulandığında çıkan değerin altında/üstünde kalan değerler aykırı değerler olarak belirlenip, gerekli düzenlemeler gerçekleştirilebilmektedir. Kullanılan veri setindeki nitelik değerleri metinsel ifadeler olduğundan dolayı aykırı değerlerin belirlenmesi işleminde, sınıf değeri temel alınarak verilerin kullanım sıklıkları incelenmiştir. Sınıf bazında nitelik sıklıklarının ortalaması alınarak, ortalama altında kalan değerler niteliklerin veri çeşidine göre belirlenen gruplara dahil edilmişlerdir.

Ortalama altında kalan ve diğer veri çeşitleri ile ortak noktaları bulunmayan değerler, genel yargılar olarak ifade edilen "Diğer", "Alanlar Karması" gibi veri değerlerine dahil edilmişlerdir.

Ortalama altında kalan ve diğer veri çeşitleri ile ortak noktaları bulunan değerler de tek nitelik çatısı altında toplanarak hem sayısını hem de söz hakkını arttırmış olmaktadırlar. Örneğin; Katılımcılar Tarafından Kullanılan Algoritma niteliğinde "Yapay Sinir Ağları", "Tekrarlayan Sinir Ağları (RNNs)" ve "Evrişimli Sinir Ağları (CNNs)" bulunan veri çeşitleri tek çatı altında toplanarak "YSA/RNNs/CNNs" şeklinde bir grubu oluşturmuşlardır. Bu şekilde ortak noktaları bulunan azınlık veri gruplarında düzenlemeler yapılarak fazla söz sahibi ve modele yön oluşturucu etkiye sahip olmaları sağlanmaya çalışılmıştır.

Uygulamada kullanılan veri setinin "Akademik" değere sahip Çalışma Alanı (sınıf) niteliğinin, veri çeşitleri ve kullanım sıklıkları Çizelge 3.10'da ifade edilmektedir. Sınıf ve nitelik değerleri bazında sıklıkları hesaplanan veri çeşitlerinin, genel sıklık ortalamaları da hesaplanmış olup, ortalamanın altında ya da üstünde kalma durumlarına göre gerekli düzenlemeler gerçekleştirilmiştir.

Çizelge 3.10. Çalışma Alanı "Akademik" Olan Veri Satırlarının, Katılımcılar Tarafından Kullanılan Algoritma Nitelik Değerlerinin Kullanım Sıklıkları.

Çalışma Alanı = "Akademik"	
"Katılımcılar Tarafından Kullanılan Algoritma" Niteliği	**Kullanım Sıklığı (Satır Adedi)**
Bayesian Techniques	621
Decision Trees	715
Random Forest	700
Regression/Logistic Regression	920
CNNs	389
Gradient Boosted Machines	254
Neural Networks	659
SVMs	555
HMMs	142
Ensemble Methods	465
Evolutionary Approaches	205
Diğer	125
RNNs	265
GANs	63
Markov Logic Networks	82
***ORTALAMA* =**	**410,667**

Aykırı değerlerin düzenlenmesi işlemine, veri setinden birkaç veri satırı ile örnek gösterilmek istenerek düzenleme öncesi durumu Çizelge 3.11'de, sonrası durumu Çizelge 3.12'de ifade edilmiştir.

Çizelge 3.11. Aykırı Değerlerin Düzenlenmesi Öncesinde Örnek Veri Satırları.

Kullanılan Çalışma Aracı	**Tercih Edilen Dil**	**Katılımcılar Tarafından Kullanılan Algoritma**	**Çalışma Alanı**
Python	R	Yapay Sinir Ağları	Akademik
Excel	R	Tekrarlayan Sinir Ağları (RNNs)	Akademik
Orange	Other	Evrişimli Sinir Ağları (CNNs)	Akademik

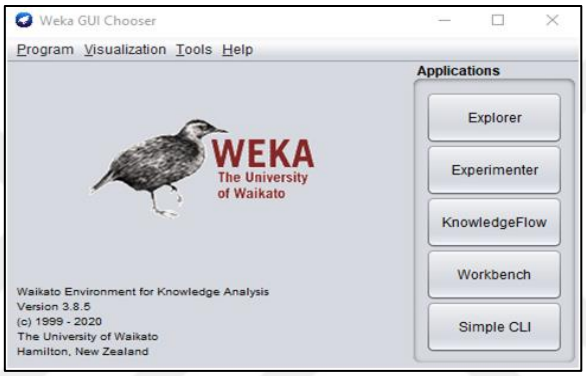
Çizelge 3.12. Aykırı Değerlerin Düzenlenmesi Sonrasında Örnek Veri Satırları.

Kullanılan Çalışma Aracı	**Tercih Edilen Dil**	**Katılımcılar Tarafından Kullanılan Algoritma**	**Çalışma Alanı**
Python	R	YSA/RNNs/CNNs	Akademik
Excel	R	YSA/RNNs/CNNs	Akademik
Orange	Other	YSA/RNNs/CNNs	Akademik

3.3.3. Veri Setinin Yazılım Ortamına Aktarımı

Veri Madenciliği teknolojisi, Sınıflandırma yöntemi adı altında farklı algoritmaların bulunduğundan Giriş bölümünde detaylı olarak bahsedilmiştir. Çalışma kapsamında sınıflandırma algoritmalarından, modellemesinde farklı kavramları (entropi, mesafe vb.) esas alan algoritmalar kullanılarak, aynı problem/veri seti üzerindeki değerlendirmeleri ele alınmaktadır.

Seçilen algoritmaların kıyaslama işleminin WEKA [57] ortamında gerçekleştirilmesi tercih edilmiştir. Ön işleme ve modelleme özelliklerini barındıran WEKA programının giriş ekranı Şekil 3.11'de belirtildiği gibidir.



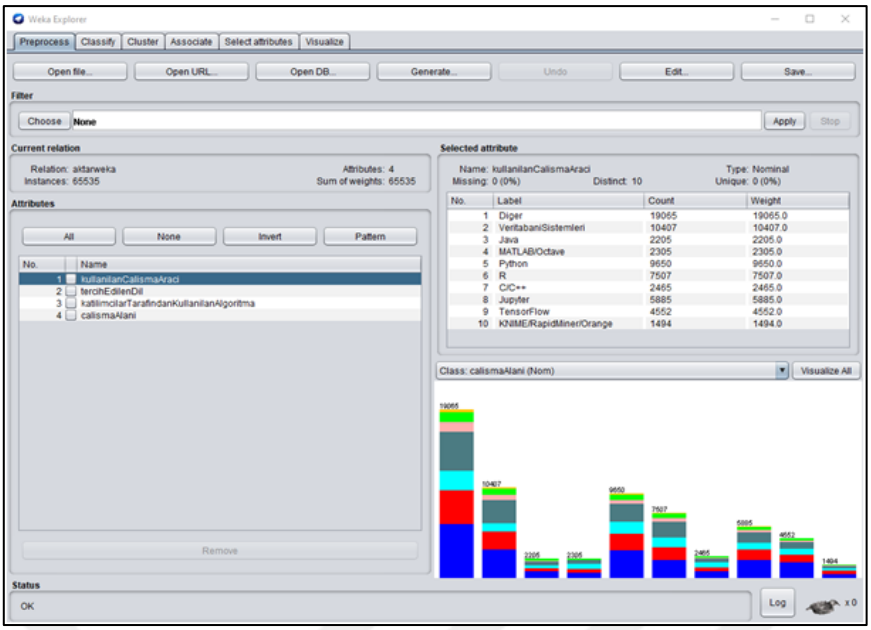
Şekil 3.11. WEKA Giriş Arayüzü.

WEKA ortamı giriş ekranında bulunan seçeneklerin görevlerinden aşağıda kısaca bahsedilmektedir.

- Experimenter: Büyük ölçekli uygulamalar için performans karşılaştırılması yapılmasını sağlamaktadır.
- KnowledgeFlow: Grafik arayüzü işlemleri gerçekleştirilmektedir.
- Workbench: Grafiksel kullanıcı arayüzlerinin birleşim ortamıdır.
- Simple CLI: Yorum satırı arayüzüdür.

Veri seti yükleme/aktarım işlemi için WEKA giriş ekranından Explorer seçeneği seçilerek, karşılaşılan arayüzde "Preprocess" sekmesi seçimi gerçekleştirilmektedir.

Ön işleme adımları tamamlanan güncel (işlenmiş) veri seti 3 girdi, 1 çıktı (sınıf) niteliği olmak üzere 4 sütun, 65535 satır içermektedir. Niteliklerin durumları analiz edilmek üzere veri seti Weka ortamına yüklenip incelemeleri gerçekleştirilmiştir. Şekil 3.12'de veri seti incelemelerinin gerçekleştirileceği Weka arayüz ekranı ifade edilmektedir.

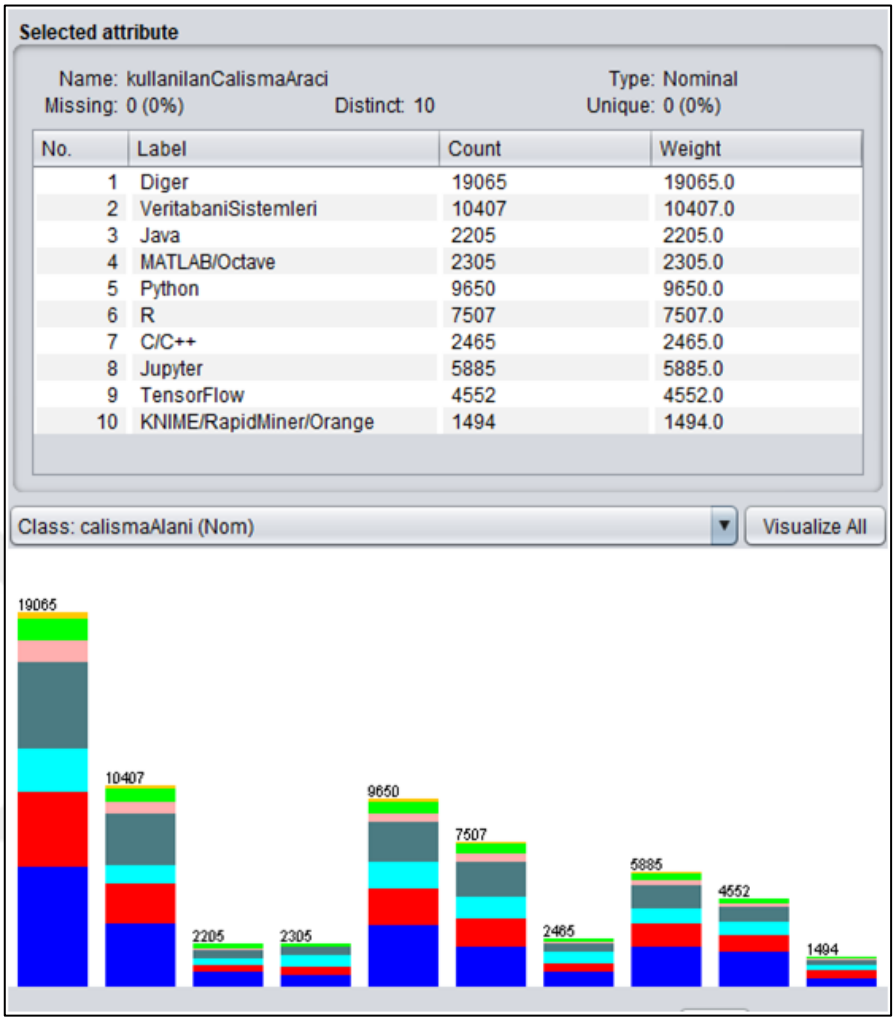


Şekil 3.12. Veri Setinin Weka Ortamına Yüklenmesi.

Weka, “Selected attribute” bölümünde niteliklerin veri değerlerinin sayısal ve görsel dağılımları yer almaktadır. Bölümün üst kısmında niteliğe ait veri çeşitlerinin sınıf değerleri bazında adet ve ağırlık bilgileri yer almakta olup, alt kısmında yer alan grafikte de veri çeşitlerinin ait oldukları sınıflara ait dağılımlar renk bazında kategorize edilmiştir.

(Renk=Sınıf, Lacivert=Teknoloji/İletişim, Kırmızı=Alanlar Karması/Diğer, Turkuaz=Akademik, Gri=Ticari/İmalat/Finansal, Açık pembe=Sağlık/Sigorta, Yeşil=Devlet/Askeri/Güvenlik, Sarı=Konaklama/Eğlence/Spor)

Nitelik bazında Selected attribute bölümü ekran görüntüleri aşağıda sırası ile yer almaktadır.

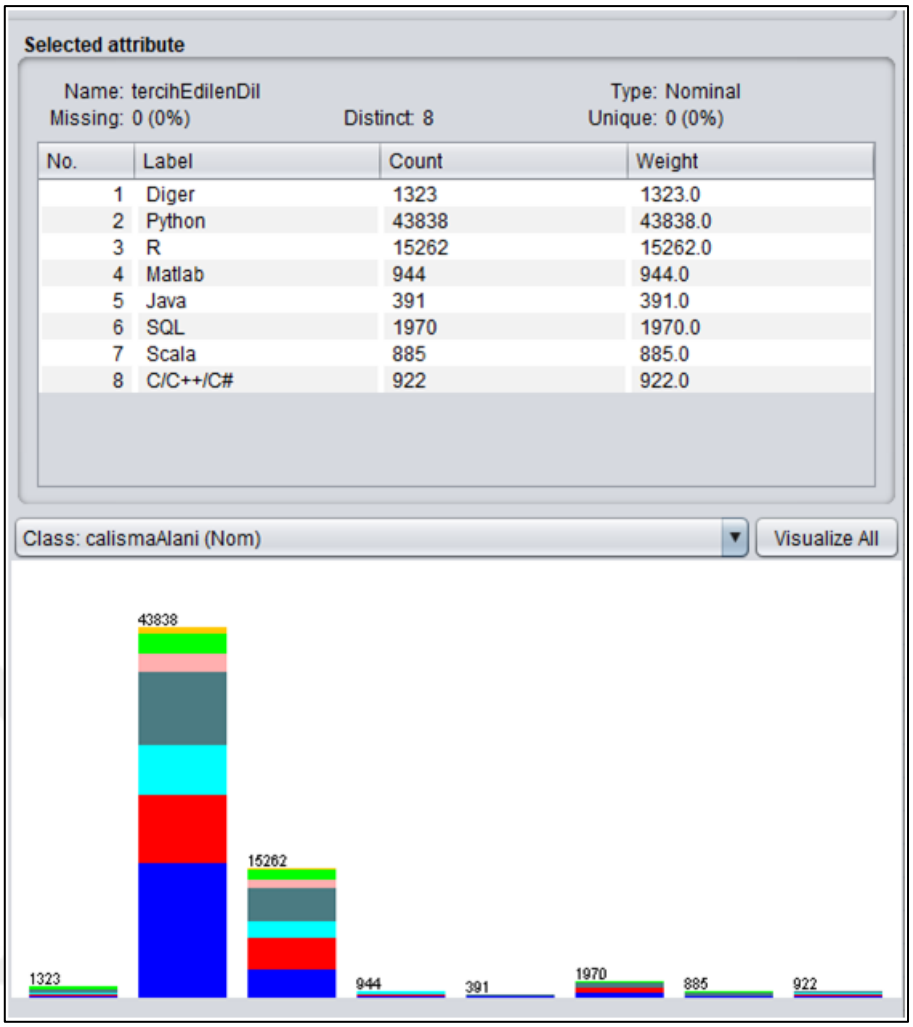


Şekil 3.13. "Kullanılan Çalışma Aracı" Niteliğine Ait Selected Attribute Bölümü.

Kullanılan Çalışma Aracı Niteliği; Selected attribute bölümü incelendiğinde ön işleme aşamaları sonrası Kullanılan Çalışma Aracı niteliği için 10 farklı veri değerinin bulunduğu ve bu veri değerlerinin sıklık adetleri yer almaktadır.

Sıklık adetlerinde de görüleceği üzere grafiklerde ilk göze çarpan "Diger" değerine sahip olan veri çeşidinin diğer verilere oranla daha fazla olması durumudur.

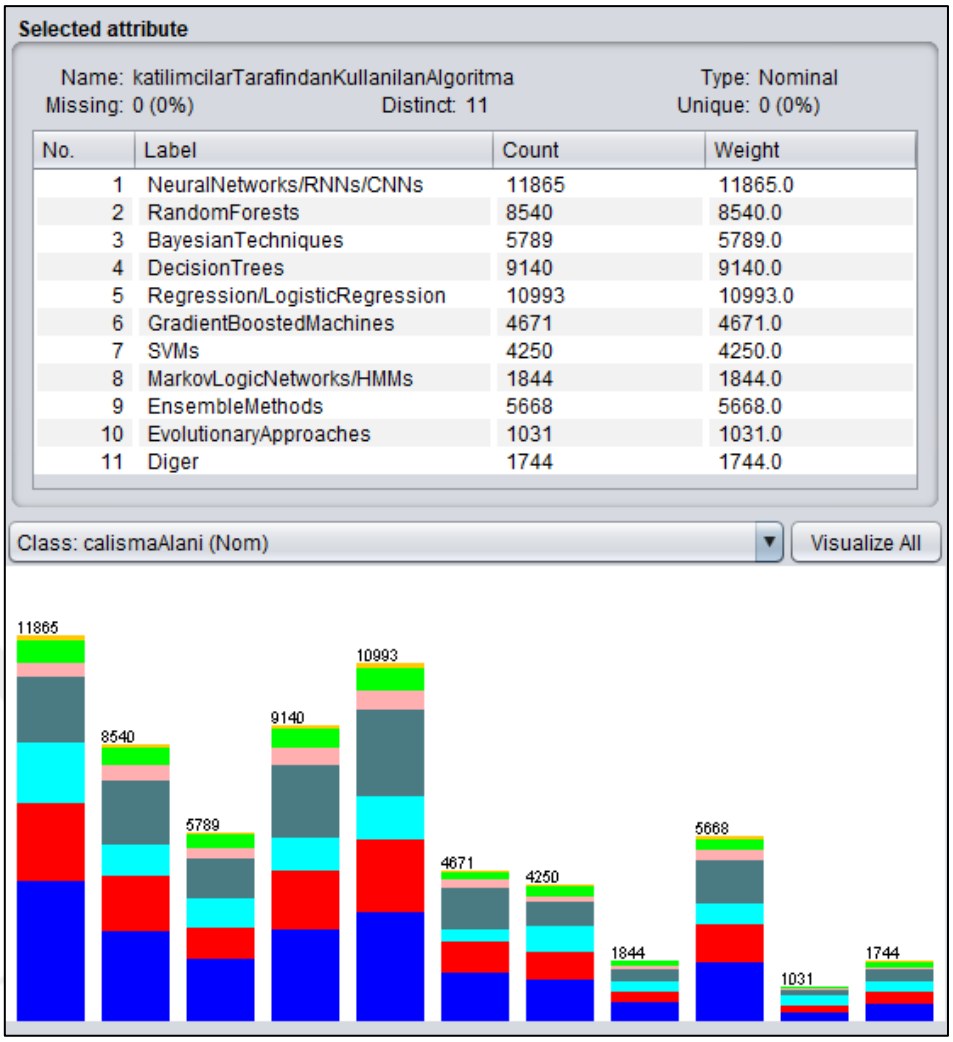
Tüm veri çeşitlerine ait detaylı bilgi edinilmek istenir ise tabloda da grafikte de bu bilgilere yer verilmiştir.



Şekil 3.14. "Tercih Edilen Dil" Niteliğine Ait Selected Attribute Bölümü.

Tercih Edilen Dil Niteliği; Selected attribute bölümü incelendiğinde ön işleme aşamaları sonrası Tercih Edilen Dil niteliği için 8 farklı veri değerinin bulunduğu ve bu veri değerlerinin sıklık adetleri ile ilgili bilgilerin yer aldığı görülmektedir.

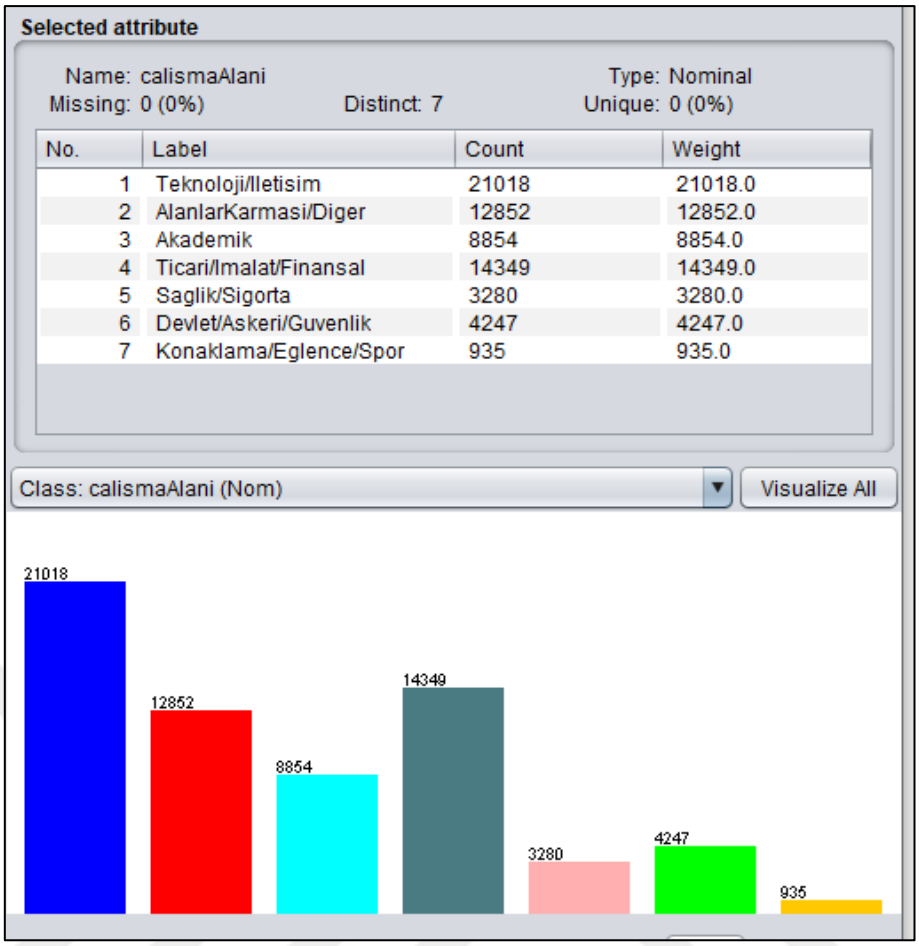
Alt kısımda yer alan grafikte de Çalışma Alanı niteliği baz alınarak nitelik değerleri incelendiğinde, Tercih Edilen Dil kategorisinde fazlaca "Python" veri değerinin kullanıldığı açıkça görülmektedir.



Şekil 3.15. "Katılımcılar Tarafından Kullanılan Algoritma" Niteliğine Ait Selected Attribute Bölümü.

Katılımcılar Tarafından Kullanılan Algoritma Niteliği; Selected attribute bölümü incelendiğinde ön işleme aşamaları sonrası Katılımcılar Tarafından Kullanılan Algoritma niteliği için 11 farklı veri değerinin bulunduğu ve bu veri değerlerinin sıklık adetleri ile ilgili bilgilerin yer aldığı görülmektedir.

Alt kısımda yer alan grafikte de veri çeşitlerinin ait oldukları sınıflara ait dağılımlar renk bazında kategorize edilmiştir. Sınıf değeri geneli incelendiğinde, Katılımcılar Tarafından Kullanılan Algoritma niteliği dağılımlarının açıkça gözükecek bir fazlalık farkının olmadığı çıkarımına varılmıştır.



Şekil 3.16. "Çalışma Alanı" Niteliğine Ait Selected Attribute Bölümü.

Çalışma Alanı (Sınıf) Niteliği; Ön işleme aşamaları sonrasında Çalışma Alanı niteliği 7 farklı sınıf değerini belirtmektedir. Bu değerlerin sıklık dağılımları ile ilgili bilgiler Şekil 3.16'da Selected Attribute bölümünde belirtilmektedir.

Şekil 3.16'nın alt kısmında yer alan grafikte de sınıf veri türlerinin sayısal dağılımları renkler bazında grafiksel olarak ifade edilmiştir. Grafikte "Konaklama/Eğlence/Spor" veri grubu azınlık olarak dikkat çekmekte olup, diğer veri grupları için dikkat çekici bir azınlık/çoğunluk durumu gözlenmemektedir.

Weka ortamına veri seti aktarımı gerçekleştiği takdirde yukarıda ifade edildiği gibi veriler ve dağılımları ile ilgili birçok farklı yorumlama/çıkarım ile veri analizleri gerçekleştirilebilmektedir.

4. MODEL OLUŞUMU VE DENEYSEL SONUÇLAR

Veri Madenciliği algoritmalarının "Başarılı" seviyeye ulaşabilmesinin en önemli sebebi, algoritmada kullanılan verinin (girdinin) kendisidir. Veri seçimi, işleme, veri seti dağılım oranları gibi birçok etmen başarı durumunu birincil olarak, doğrudan veri bağlantısı ile etkilemektedir. Uygulamada kullanılan Veri Madenciliği yöntemi, model seçimi gibi algoritma başarı oranını etkileyen diğer etmenler ise başarı oranını etkileyen ikincil öneme sahip durumlardır. Kullanılan veri ne kadar verimli, etkileyici, başarılı bir şekilde işlendiyse, kullanılan Veri Madenciliği yöntemi de başarısını o oranla ön plana çıkarmış olacaktır.

Aynı Veri Madenciliği yöntemi ve algoritma nitelikleri kullanılmış olsa da verilerin daha verimsiz işlenip uygulamaya geçmesi sonrasında elde edilen başarı oranı düşük olacaktır. Verinin önem derecesi belirtilen bu durum ile de ifade edilmeye çalışılmıştır.

Başarı oranı değeri, kullanılan modelin doğru çıktı oluşturup oluşturmaması ile ilgili kavramlardan yararlanılarak gerçekleştirilen hesaplamalar sonucu elde edilmektedir. Yapılan hesaplamalarda bahsedilen Doğru Pozitif (True Positive – TP), Doğru Negatif (True Negative - TN), Yanlış Pozitif (False Positive – FP) ve Yanlış Negatif (False Negative – FN) değerleri modelin test aşamasında kullanılan, veri setindeki sınıfların doğruluk ve hata oranlamaları yardımı ile hesaplanan değerlerdir. Bu değerler model oluşturulup, başarısı sorgulanırken somut yorumlamalar yapılabilmesi açısından hesaplanmaktadırlar.

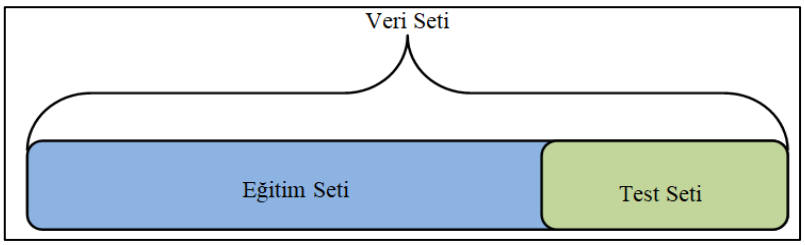
Başarı kelimesine sahip olunabilmesi için bahsedilen somut hesaplamaları, veri ve model harici etkileyen bir diğer husus da veri seti dağılım oranlamasıdır. Veri Madenciliği teknolojisinin genel değerlendirilmesi göz önünde bulundurulduğunda veri seçimi, model seçimi gibi birçok aşamada kullanılan kavramların belirlenmesinde net bir kural bulunmaması, veri seti dağılımı için de geçerlidir. Deneme yanılma yöntemi ile karar verilen kavram seçimlerinin, başarı oranı maksimum seviyeye ulaştığında en iyi değerlere erişmiş olduğu düşünülmektedir.

Yapılan literatür taramaları sonrasında 2 farklı veri seti dağılım yönteminin sıklıkla kullanıldığı gözlemlenmiştir. Bahsedilen bu veri seti dağılım yöntemlerinden ilkinde veri seti, eğitim ve test şeklinde iki farklı veri grubuna bölünmektedir. Bahsedilen ikinci dağılım yönteminde ise, eğitim ve test veri grubu haricinde bir de doğrulama setini oluşturan farklı bir veri grubu bulunmaktadır [58].

- Eğitim Seti; modelin eğitilmesi için kullanılan veri topluluğudur.
- Doğrulama Seti; eğitim setinin kullanılmasıyla oluşturulan modelin performansını değerlendirmek amacı ile kullanılan veri grubudur. Model kıyaslaması, modele uygun parametrelerin belirlenmesi gibi eğitim aşamasının test aşamasına geçmesini başarı ile sağlamak amacıyla köprü görevi gören aşamada kullanılmaktadır.
- Test Seti; eğitim setinde bulunan veriler yardımı ile oluşturulmuş olan modelin başarısını belirlemek amacı ile modelin test edilmesi sırasında kullanılan veri topluluğudur. Kullanım yeri ve şekli olarak doğrulama seti ile aynı hedefe sahiptir.

4.1. EĞİTİM VE TEST VERİ SETİ ORANLAMASI

Eğitim ve Test Seti olarak iki bölüme ayrılmak istenen veri seti oranlamasının yaklaşık değerleri genel olarak%70'e %30 ya da %60'a %40 şeklinde tercih edilmektedir. Veri setinin eğitim/test bölümlemesi Şekil 4.1'de de ifade edildiği oranda dağılım sağlanmaktadır [59].

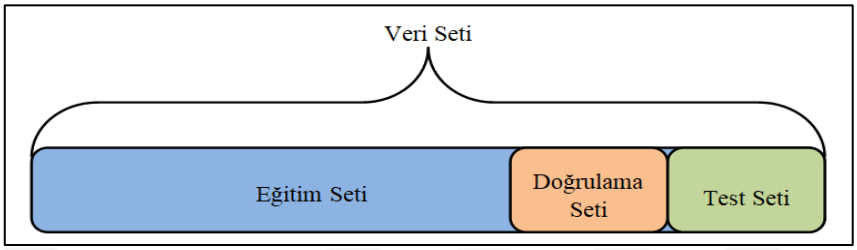


Şekil 4.1. Eğitim ve Test Veri Seti Oranı.

4.2. EĞİTİM, DOĞRULAMA VE TEST VERİ SETİ ORANLAMASI

Eğitim seti tanımlaması göz önünde bulundurulacak olursa, genel olarak veri setinin yaklaşık %60'lık bölümü bu oranlama yönteminde eğitim seti miktarı olarak tercih

edilmektedir. Model doğruluğunun test aşamasına geçmesinden önce bir teyidinin yapıldığı aşamada kullanılan Doğrulama Seti de veri setinin genel olarak yaklaşık %20'lik bölümünü kullanmaktadır. Son aşama olarak belirtilen test aşamasında kullanılan Test Seti de veri setinin kalan kısmı olan yaklaşık %20'lik bölümünü kullanarak, test aşaması gerçekleştirilmiş olmaktadır. Belirtilen bu oranlamalar Şekil 4.2'de görsel olarak da ifade edilmektedir [59].

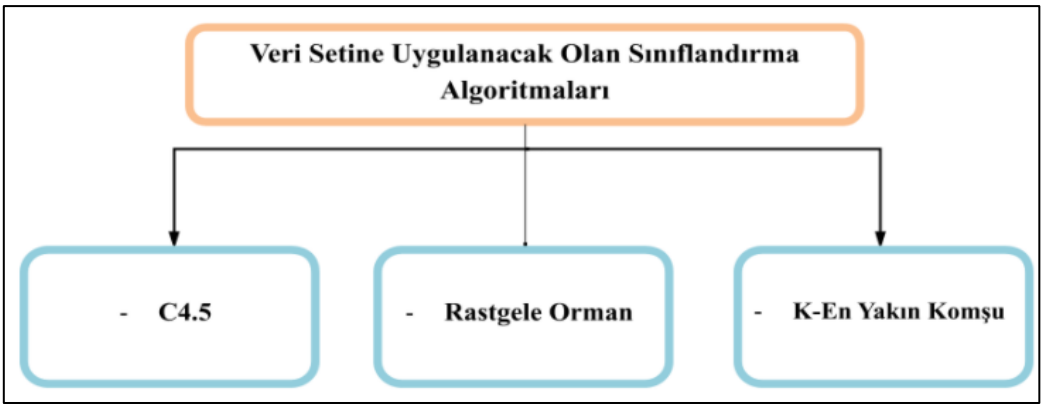


Şekil 4.2. Eğitim, Doğrulama ve Test Veri Seti Oranlaması.

İlk yöntem olarak açıklanmış olan Eğitim ve Test Veri Seti Oranlaması ve ikinci yöntem olarak açıklanmış olan Eğitim, Doğrulama ve Test Veri Seti Oranlaması yöntemlerinin tercih edilmesi, kullanılan veri setinin boyutuna bağlı olarak değişmektedir. Veri setinin boyutu yeterince fazla ise üç bölmeye ayrılıp Eğitim, Doğrulama ve Test Veri Seti Oranlaması dağılım yönteminin kullanımı tercih edilebilmektedir. Ancak yeteri kadar veri seti boyutuna sahip olunmadığı durumlarda veriyi fazla bölümlere ayırmak model başarısı açısından uygun olmayacağından Eğitim ve Test Veri Seti Oranlaması dağılım yönteminin kullanılmasının daha uygun olduğu düşünülmektedir.

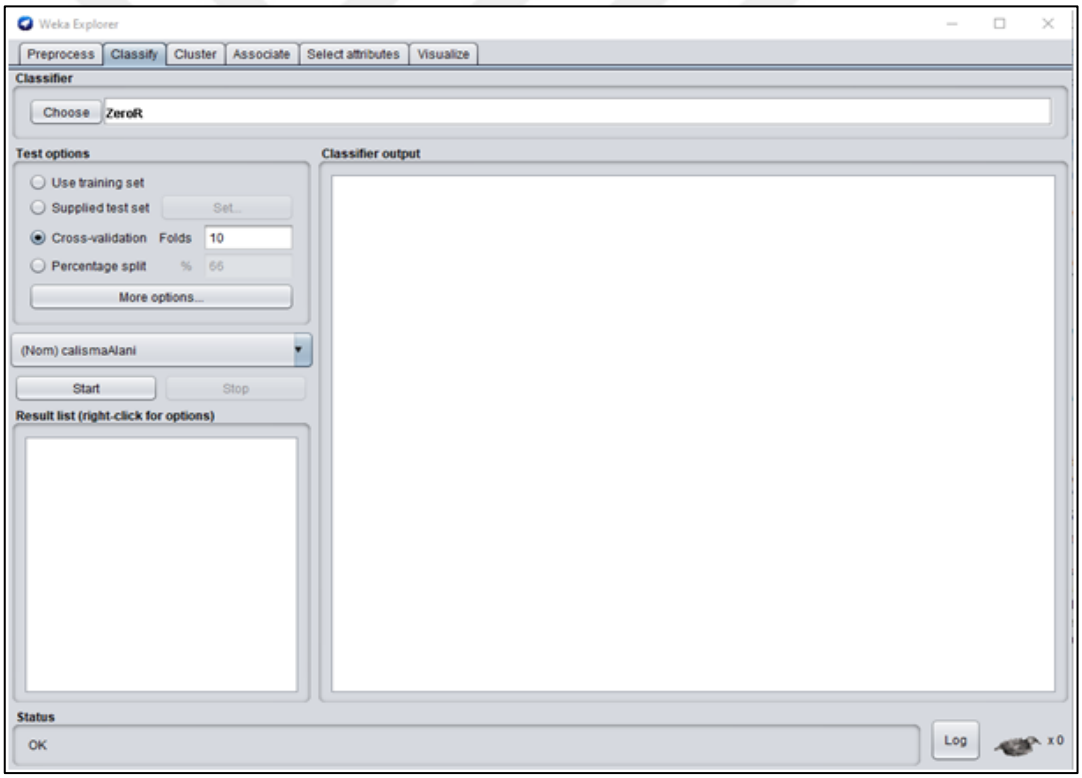
4.3. ALGORİTMA/MODEL UYGULAMA DENEMELERİ

Veri Madenciliği Sınıflandırma yönteminin ifade edildiği bölümde algoritma çeşitlerine ve işleyişlerine yer verilmişti. Uygulama kapsamında bu algoritmalar arasından ortak yönü Sınıflandırma yöntemine ait olan, farklı yönü ise model oluşum teknikleri olan 3 farklı algoritma çeşidi ile modelleme işlemleri gerçekleştirilmiştir. "Veri Bilimi Anketi" isimli veri setine uygulanacak olan sınıflandırma algoritmaları Şekil 4.3'te de ifade edildiği gibi, modelleme tekniğinde entropi kavramını esas alan C4.5 Algoritması, birçok karar ağacı içerisinden seçilen karar ağacını esas alan Rastgele Orman Algoritması ve modellemesinde mesafe kavramını esas alan K-En Yakın Komşu Algoritması olmaktadır.



Şekil 4.3. Veri Setine Uygulanacak Olan Sınıflandırma Algoritmaları.

Weka'ya yüklenilen verilerin analizi gerçekleştirildikten sonra Şekil 4.4'te görülen arayüzde Classify sekmesine geçilerek eğitim/test veri oranlaması Test options bölümünden, algoritma seçim aşamaları da Choose bölümünden gerçekleştirilmektedir.



Şekil 4.4. Weka Ortamı -Classify Sekmesi.

Test options bölümünde farklı seçenekler ile veri seti bölümlemesi gerçekleştirilmektedir.

Usetraining set seçeneği ile yüklenilen veri setinin tamamı model oluşturmak için eğitim seti olarak kullanılmaktadır. Model oluşumu gerçekleştirildikten sonra aynı şekilde veri setinin tamamı test seti olarak analiz gerçekleştirilmektedir.

Supplied test set seçeneği ile aynı problem için farklı dosyalarda veri setleri düzenlenerek bir dosya eğitim seti, diğer dosya test seti olarak seçilerek veri seti bölümlemesi dosya bazında gerçekleştirilmiş olmaktadır.

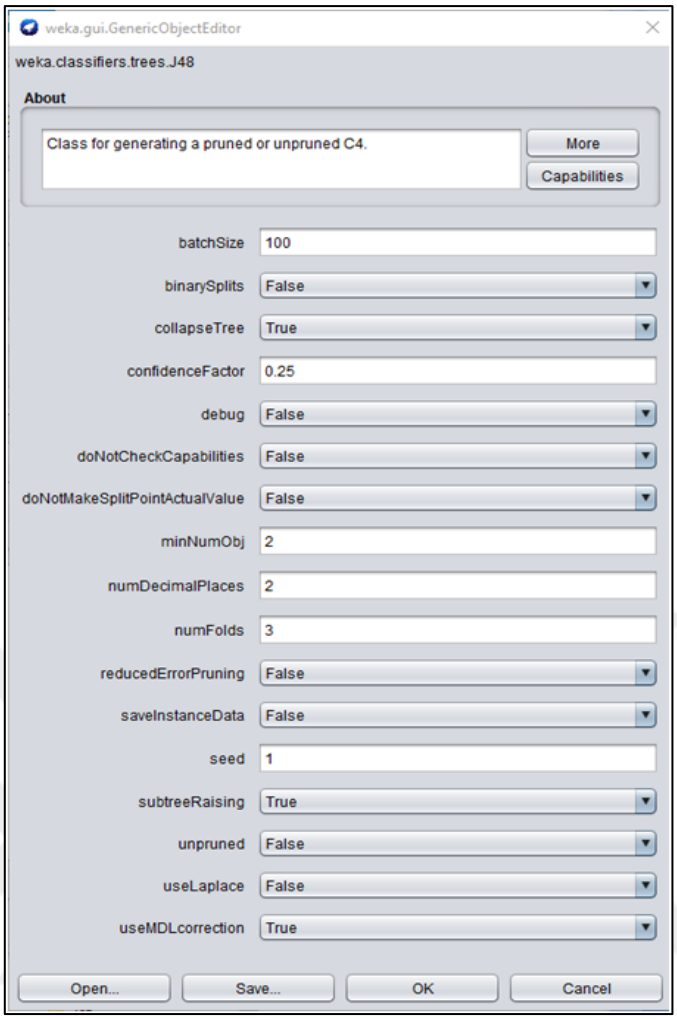
Cross-validation seçeneği ile girilen değer bazında veri seti bölümlemesi gerçekleştirilerek bir bölüm test verileri için kullanılıp kalan bölümler için modelleme gerçekleştirilmesi mantığına dayanmaktır. Varsayılan değer olarak 10 değeri kullanılır. Veri dosyası 10 parçaya bölünür. 9 parça model geliştirilmesinde eğitim seti olarak kullanılırken, 1 parça analiz işlemini gerçekleştirmek üzere tekrarlanarak test seti olarak kullanılmaktadır. Analiz işlemi sonucu tek bir başarı değeri ortaya çıkmaktadır. Varsayılan değer değiştirilerek bölümleme, işlem tekrar sayısı ve başarı oranı farklılaştırılabilmektedir.

Percentage split seçeneği ile girilen değer baz alınarak eğitim seti oranlaması gerçekleştirilmektedir. Varsayılan olarak "66" değeri ile gerçekleştirilen işlemler için veri setinin %66'sı eğitim seti, %34'ü test seti olarak kullanılmaktadır. Farklı yüzdelik değerleri ile farklı başarı oranları elde edilebilmektedir.

Eğitim/test veri seti oranlaması belirlenip, kullanılmak istenilen algoritma modeli de seçildikten sonra "Start" komutu ile işleme başlanan algoritma çıktıları Şekil 4.4'teki Weka arayüzünde görünen Classifier output bölümünde ifade edilmektedir.

4.3.1. C4.5 Algoritmasının Weka'da Uygulanması

Bazı algoritmalar Weka ortamında farklı isimlendirmeler ile yer almaktadır. C4.5 algoritması da farklı isimlendirilen algoritmalardan biridir. Classify sekmesi, algoritma seçim bölümünde ağaçlar başlığı altında "J48" olarak ifade edilen algoritma türü C4.5 Algoritmasını temsil etmektedir. Seçimi yapılan J48 algoritmasının yapılandırılması işlemi için algoritma ismi tıklanarak Şekil 4.5'te belirtilen seçenek ekranına erişilmiş olunmaktadır.



Şekil 4.5. Weka Ortamında J48 Algoritması Yapılandırma Ekranı.

Seçilen algoritma için yapılandırma yapılmaksızın modellenmesi sonucu, algoritmanın kullanım seçeneklerine ait değerler Şekil 4.5'te belirtildiği gibi olmaktadır. Yapılandırma seçeneklerinin Weka'da belirtilen açıklamaları;

- batchSize: Toplu tahmin gerçekleştiriliyor ise işlenmesi tercih edilen örnek sayısı. Daha fazla veya daha az örnek sağlanabilir. Ancak bu durum, uygulamalara tercih edilen bir toplu iş boyutunu belirleme şansı verir.

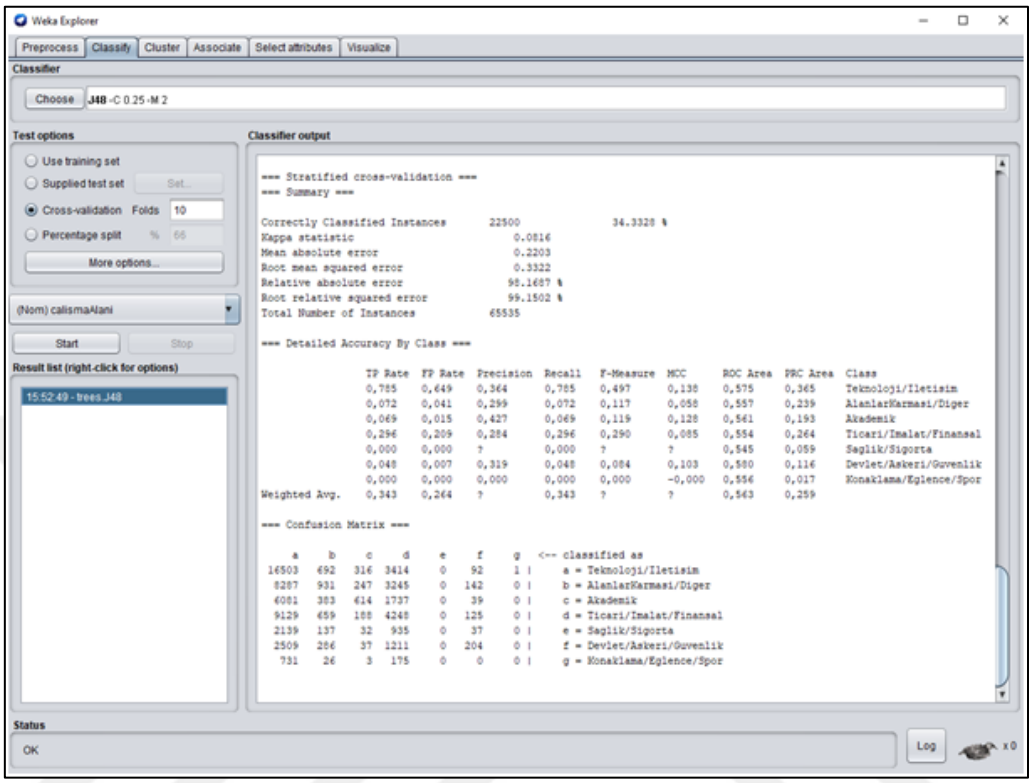
- binarySplits: Ağaçları oluştururken nominal özniteliklerde ikili bölmelerin kullanılıp kullanılmayacağı.

- collapseTree: Eğitim hatasını azaltmayan parçaların çıkarılıp çıkarılmadığı.

- confidenceFactor: Budama için kullanılan güven faktörü (daha küçük değerler daha fazla budamaya neden olur).

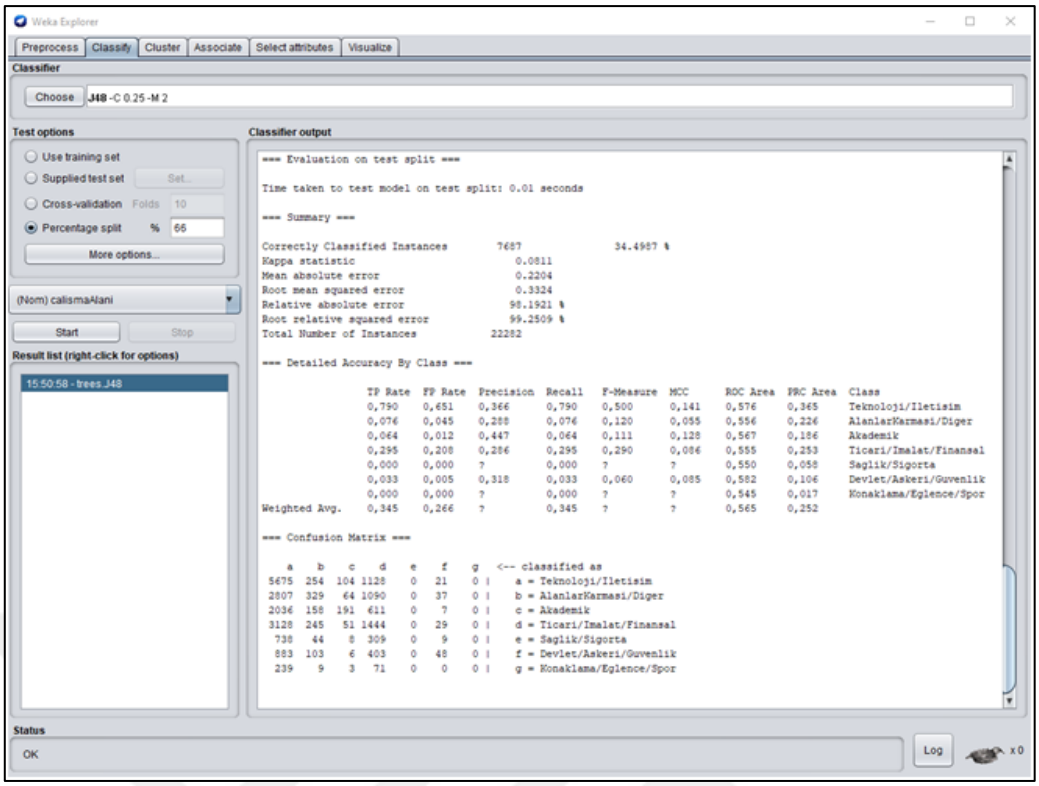
- debug: “true” olarak ayarlanırsa, sınıflandırıcı konsola ek bilgi verebilir.
- doNotCheckCapabilities: Ayarlanırsa, sınıflandırıcı oluşturulmadan önce sınıflandırıcı yetenekleri kontrol edilmez (Çalışma zamanını azaltmak için dikkatli kullanılmalı).
- doNotMakeSplitPointActualValue: Doğruysa, ayırma noktası gerçek bir veri değerine yeniden yerleştirilmez. Bu, sayısal niteliklere sahip büyük veri kümeleri için önemli hızlanmalar sağlayabilir.
- minNumObj: Yaprak başına minimum örnek sayısı.
- numDecimalPlaces: Modeldeki sayıların çıktısı için kullanılacak ondalık basamak sayısı.
- numFolds: Azaltılmış hata ayıklama için kullanılan veri miktarını belirler. Bir kat budama için, geri kalanı ağacı büyütmek için kullanılır.
- reducedErrorPruning: C4.5 budaması yerine azaltılmış hata budamasının kullanılıp kullanılmadığı.
- saveInstanceData: Görselleştirme için eğitim verilerinin kaydedilip kaydedilmeyeceği.
- seed: Hataları azaltılmış budama kullanıldığında verileri rastgele sıralamak için kullanılan tohum.
- subtreeRaising: Budama sırasında alt ağaç yükseltme işleminin dikkate alınıp alınmayacağı.
- unpruned: Budama yapılıp yapılmadığı.
- useLaplace: Yapraklardaki sayıların Laplace'a göre düzeltilip düzeltilmediği.
- useMDLcorrection: Sayısal öznitelikler üzerinde bölmeler bulunurken MDL (Minimum Description Length - Minimum Tanım Uzunluğu) düzeltmesinin kullanılıp kullanılmadığı.

Varsayılan yapılandırma seçenekleri kullanılarak Cross-validation seçeneğinde 10 bölümleme, Percentage split seçeneğinde %66'lık eğitim seti seçimiyle C4.5 algoritmasının analizi gerçekleştirilmektedir.



Şekil 4.6. Cross-validation:10 bölümlemeli, C4.5 Algoritması Weka Çıktısı.

Varsayılan yapılandırma seçenekleri kullanılarak Cross-validation seçeneği ile 10 bölümleme gerçekleştirilen veri setinin, C4.5 Algoritması doğruluk oranı %34,3328 şeklinde elde edilmektedir.

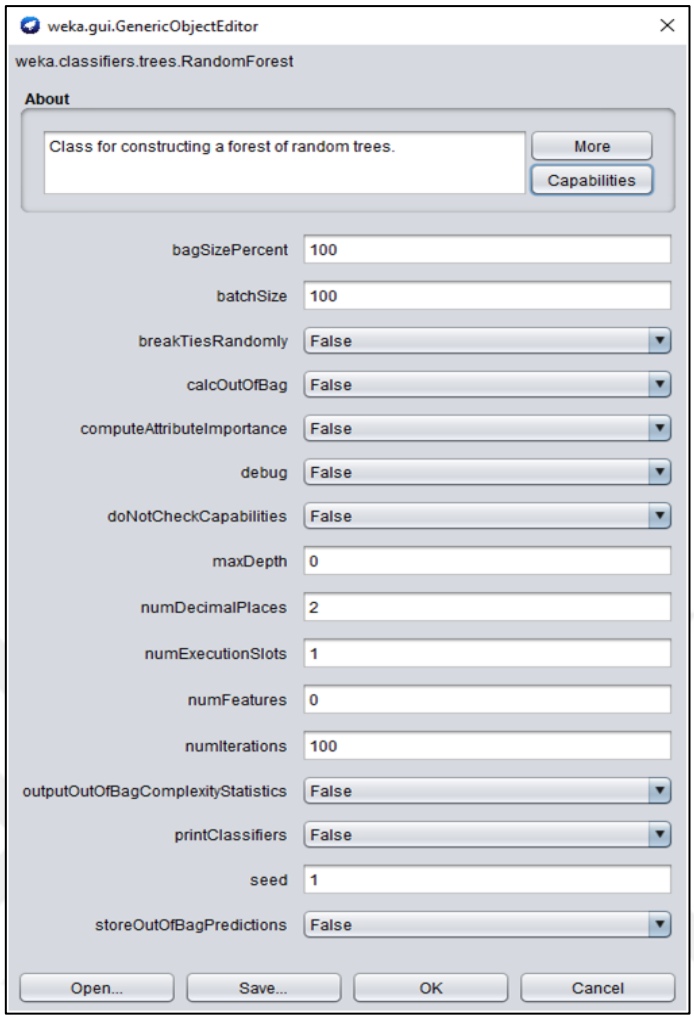


Şekil 4.7. Percentage split:%66 bölümlemeli C4.5 Algoritması Weka Çıktısı.

Varsayılan yapılandırma seçenekleri kullanılarak Percentage split seçeneği ile %66 bölümleme gerçekleştirilen veri setinin, C4.5 Algoritması doğruluk oranı %34,4987 şeklinde elde edilmektedir.

4.3.2. Rastgele Orman Algoritmasının Weka'da Uygulanması

Rastgele ağaçlardan meydana gelen orman modelinin kullanılması mantığına dayanan Rastgele Orman sınıflandırma algoritması için Random Forest algoritma seçimi gerçekleştirilir. Algoritma yapılandırılması için Weka'nın varsayılan seçenek değerleri Şekil 4.8'de belirtildiği gibi olup, seçeneklerin açıklamalarına yer verilmiştir.

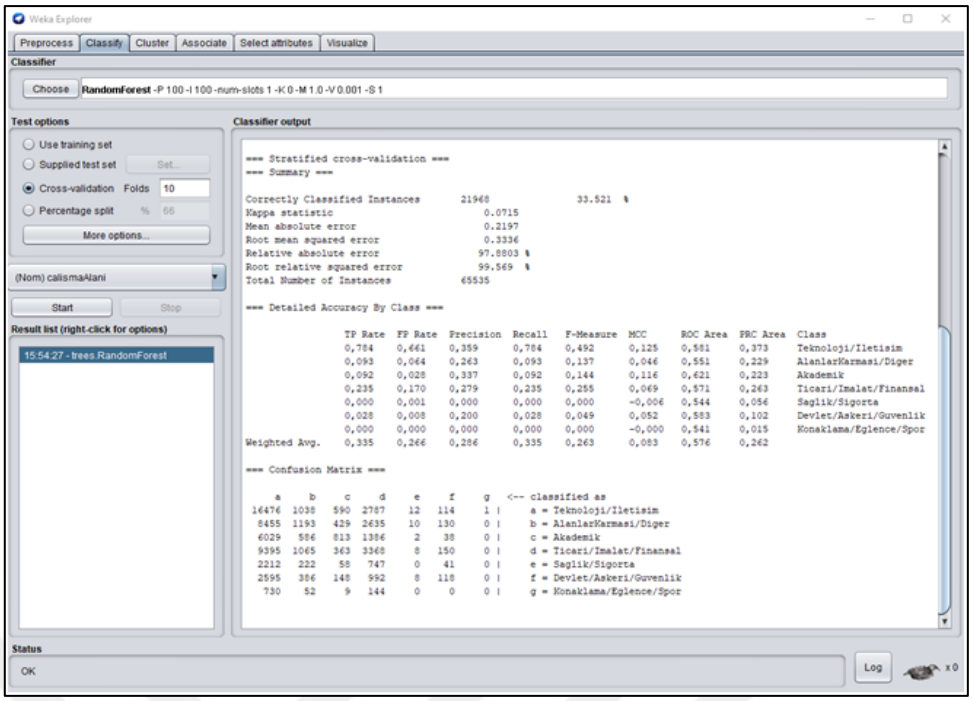


Şekil 4.8.Rastgele Orman Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü.

- bagSizePercent: Eğitim seti boyutunun yüzdesi olarak her torba boyutu.
- batchSize: Toplu tahmin gerçekleştiriliyor ise işlenmesi tercih edilen örnek sayısı. Daha fazla veya daha az örnek sağlanabilir. Ancak bu durum, uygulamalara tercih edilen bir toplu iş boyutunu belirleme şansı verir.
- breakTiesRandomly: Birkaç özellik eşit derecede iyi göründüğünde bağları rastgele koparın.
- calcOutOfBag: Torba dışı hatasının hesaplanıp hesaplanmadığı.
- computeAttributeImportance: Ortalama kirlilik azalması yoluyla öznitelik önemini hesaplayın.
- debug:“true” olarak ayarlanırsa, sınıflandırıcı konsola ek bilgi verebilir.

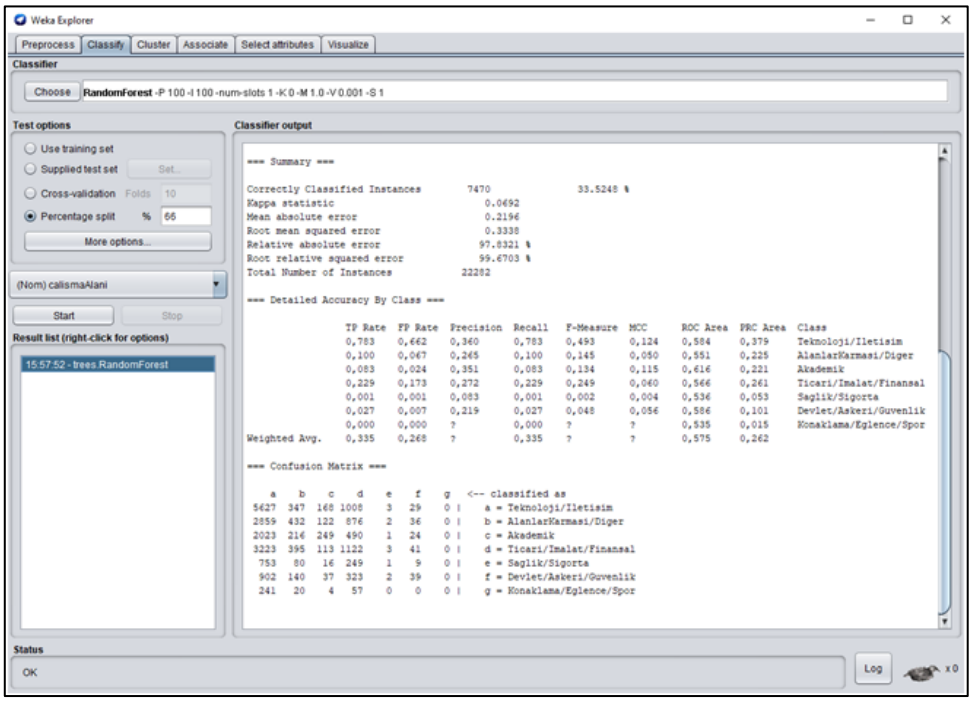
- doNotCheckCapabilities: Ayarlanırsa, sınıflandırıcı oluşturulmadan önce sınıflandırıcı yetenekleri kontrol edilmez (Çalışma zamanını azaltmak için dikkatli kullanılmalı).
- maxDepth: Ağacın maksimum derinliği, sınırsız için 0.
- numDecimalPlaces: Modeldeki sayıların çıktısı için kullanılacak ondalık basamak sayısı.
- numExecutionSlots: Topluluğu oluşturmak için kullanılacak yürütme yuvalarının (iş parçacığı) sayısı.
- numFeatures: Rastgele seçilen niteliklerin sayısını ayarlar. 0 ise int(log_2(#predictors) + 1) kullanılır.
- numIterations: Rastgele ormandaki ağaç sayısı.
- outputOutOfBagComplexityStatistics: Torba dışı değerlendirme yapıldığında karmaşıklığa dayalı istatistiklerin çıktısının verilip verilmeyeceği.
- printClassifiers: Çıktıdaki bireysel sınıflandırıcıları yazdırın.
- seed: Kullanılacak rastgele sayı çekirdeği.
- storeOutOfBagPredictions: Torba dışı tahminlerin saklanıp saklanmayacağı.

Varsayılan yapılandırma seçenekleri kullanılarak Cross-validation seçeneği ile 10 bölümleme gerçekleştirilen veri setinin Rastgele Orman Algoritması doğruluk oranı %33.521 şeklinde elde edilmektedir.



Şekil 4.9.Cross-validation:10 bölümlemeli, Rastgele Orman Algoritması Weka Çıktısı.

Varsayılan yapılandırma seçenekleri kullanılarak Percentage split seçeneği ile %66 bölümleme gerçekleştirilen veri setinin Rastgele Orman Algoritması doğruluk oranı %33.5248 şeklinde elde edilmektedir.



Şekil 4.10.Percentage split:%66 bölümlemeli Rastgele Orman Algoritması Weka Çıktısı.

4.3.3. K-En Yakın Komşu (KNN) Algoritmasının Weka'da Uygulanması

Weka ortamında farklı bir isimlendirme ile anılan bir diğer algoritma modeli de K –En Yakın Komşu algoritmasıdır. Algoritma modelleri seçim bölümünde KNN algoritması "IBk" olarak ifade edilmektedir. Algoritmanın çalışma prensibinde, eğitim veri setindeki en benzer "k" adet örneğin modu kullanılmaktadır. k olarak belirtilen komşu sayısı manüel olarak belirlenebilmektedir.

Algoritmanın düzenlenebilir özellikleri için Weka'da algoritma ismine tıklandığında Şekil 4.11'de ifade edilen seçenek ekranı belirmektedir.



Şekil 4.11. KNN Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü (1).

- KNN: Kullanılacak komşu sayısı.
- batchSize: Toplu tahmin gerçekleştiriliyor ise işlenmesi tercih edilen örnek sayısı. Daha fazla veya daha az örnek sağlanabilir. Ancak bu durum, uygulamalara tercih edilen bir toplu iş boyutunu belirleme şansı verir.
- crossValidate: 1 ile KNN parametresi olarak belirtilen değer arasındaki en iyi k değerini seçmek için çapraz doğrulamanın kullanılıp kullanılmayacağı.

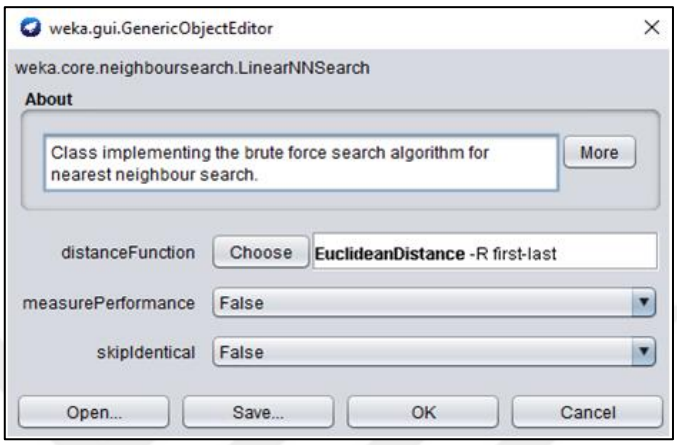
- debug:“true” olarak ayarlanırsa, sınıflandırıcı konsola ek bilgi verebilir.
- distanceWeighting: Kullanılan mesafe ağırlıklandırma yöntemini alır.
- doNotCheckCapabilities: Ayarlanırsa, sınıflandırıcı oluşturulmadan önce sınıflandırıcı yetenekleri kontrol edilmez (Çalışma zamanını azaltmak için dikkatli kullanılmalı).
- meanSquared: Regresyon problemleri için çapraz doğrulama yapılırken ortalama mutlak hata yerine ortalama karesi alınmış hatanın kullanılıp kullanılmadığı.
- nearestNeighbourSearchAlgorithm: Kullanılacak en yakın komşu arama algoritması (Varsayılan: weka.core.neighboursearch.LinearNNSearch).
- numDecimalPlaces: Modeldeki sayıların çıktısı için kullanılacak ondalık basamak sayısı.
- windowSize: Eğitim havuzunda izin verilen maksimum örnek sayısını alır. Bu değerin üzerine yeni örneklerin eklenmesi, eski örneklerin kaldırılmasına neden olur. 0 değeri, eğitim vakalarının sayısında sınır olmadığını belirtir.

Açıklanan bu parametreler arasında k değerinin değişimi başarı niteliğini tetikleyen önemli etkenlerden biridir. K parametresine 1 değerinin atanması ile modele en çok benzeyen tek eğitim örneği kullanılarak model analiz işlemi gerçekleştirilmektedir. K parametresi değeri veri kümesi boyutu ile doğru orantılı olarak artış gösterebilmektedir. Yapılan literatür taramalarında uygulama örnekleri incelendiğinde, k değeri için genellikle 3, 7, 11, 21 gibi veri seti boyutuna bağlı olarak değişen ortak değerlerin kullanıldığı gözlemlenmiştir.

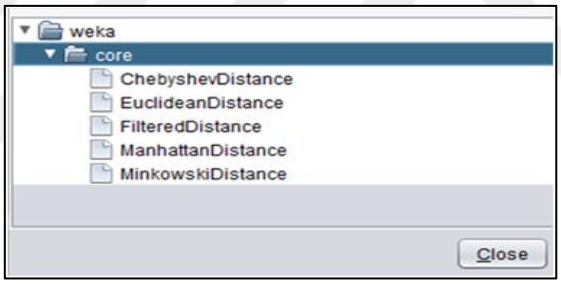
Sunulan parametre seçenekleri arasında “crossValidate” parametresinin “true” değeri ile algoritma için çapraz doğrulanma seçeneği aktifleşmiş olmaktadır. Bu sayede k parametre değeri için en iyi değer otomatik olarak belirlenebilmektedir.

Model başarı değerini etkileyen bir diğer önemli faktör ise KNN algoritmasının temeline dayanan “mesafe” kavramıdır. Mesafe ölçümü için kullanılabilen farklı mesafe formülleri bulunmaktadır. Bu mesafe formüllerinin seçim kısmı da “nearestNeighbourSearchAlgorithm” parametresinin seçenekleri ile gerçekleşmektedir.

Varsayılan seçeneği “LinearNNSearch” olan ölçüm formüllerinde değişiklik sağlanabilmesi için Şekil 4.11’de belirtilen “nearestNeighbourSearchAlgorithm” parametre sekmesi yardımı ile Şekil 4.12’deki Weka ekranı görüntülenmektedir. Görüntülenen ekranda “distanceFunction” değeri seçildiğinde Şekil 4.13’te belirtilen Öklid, Manhattan… gibi mesafe formülleri kullanılabilmektedir.



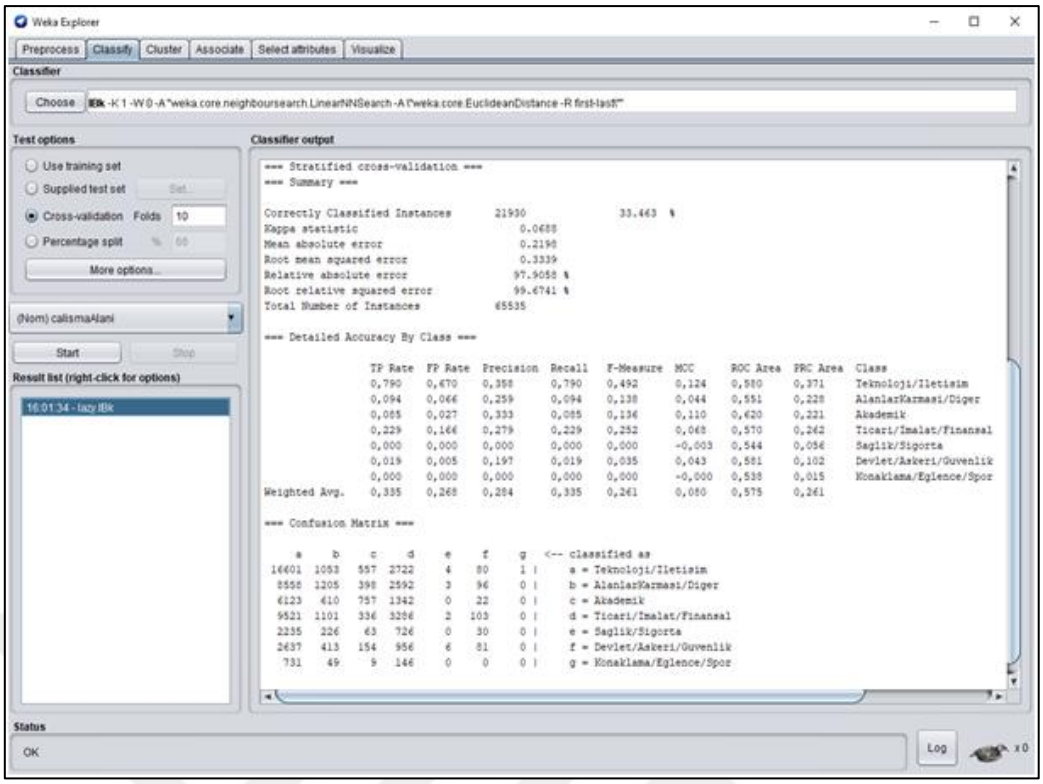
Şekil 4.12. KNN Algoritmasının Yapılandırma Seçenekleri Weka Arayüzü (2).



Şekil 4.13. KNN Algoritması Yapılandırma Seçenekleri Weka Arayüzü (3).

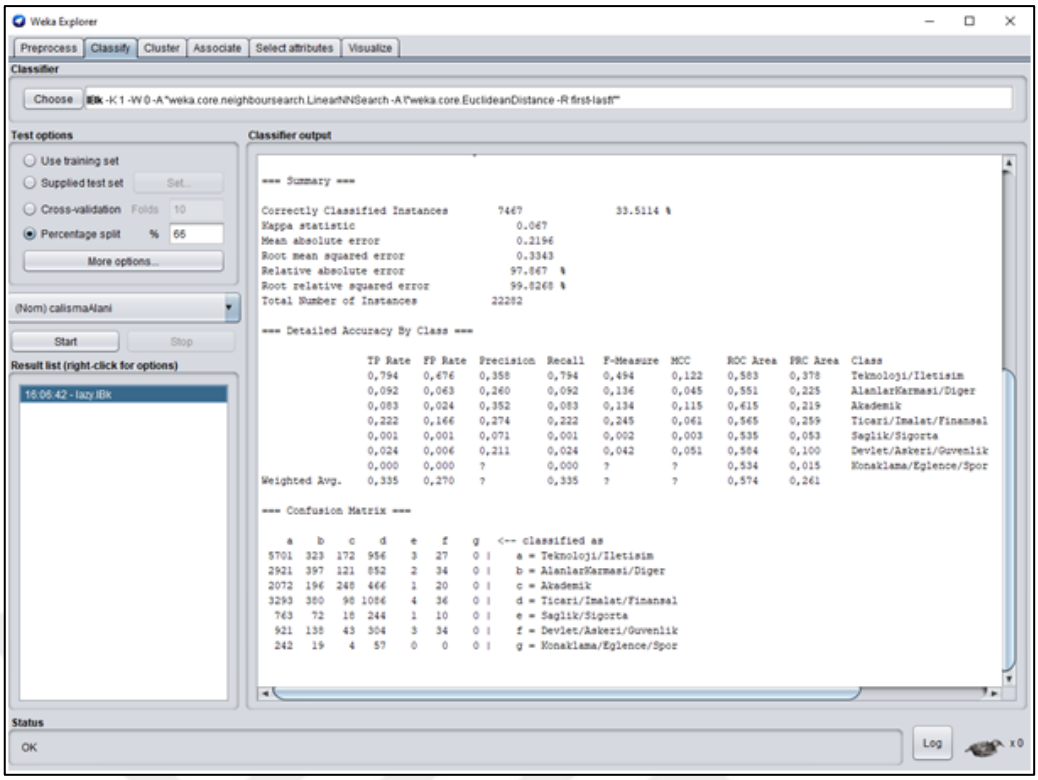
En iyi k değeri, en uygun mesafe ölçüm formülü seçimi gibi kesin yargılar bulunmadığından, algoritma başarısının en yüksek elde edilebilmesi için bu parametreleri deneme yanılma yöntemi ile belirlemek en uygun yöntem olmuş olmaktadır.

KNN algoritmasının uygulanmasında, Çizelge 4.1’de bulunan tüm durumların Weka ekran görüntüleri ekte yer almaktadır. Varsayılan seçenek değerleri ile oluşturulan modellerin başarı oranları ise Şekil 4.14 ve Şekil 4.15’te ifade edilmektedir.



Şekil 4.14. Cross-validation:10 bölümlemeli, KNN Algoritması Weka Çıktısı.

Varsayılan yapılandırma seçenekleri kullanılarak Cross-validation seçeneği ile 10 bölümleme gerçekleştirilen veri setinin KNN Algoritması doğruluk oranı %33,463 şeklinde elde edilmektedir.



Şekil 4.15. Percentage split:%66 bölümlemeli KNN Algoritması Weka Çıktısı.

Varsayılan yapılandırma seçenekleri kullanılarak Percentage split seçeneği ile %66 bölümleme gerçekleştirilen veri setinin KNN Algoritması doğruluk oranı %33,5114 şeklinde elde edilmektedir.

Çizelge 4.1. KNN Algoritmasının Belirtilen Seçenekler Doğrultusunda Weka'da Elde Edilen Başarı Oranları (1).

Başarı Oranı **Algoritma Çeşidi**	**Cross-validation: Folds 10**	**Percentage split: %66**
C4.5	%34,3328	%34,4987
Rastgele Orman	%33,521	%33,5248
KNN	%33,463	%33,5114

Aynı veri seti 3 farklı algoritma tarafından modellendiğinde farklı sonuçlar verdiği gözlemlenmiştir. Bu durum algoritmaların işleyiş, mantık farklılıklarından dolayı farklı başarı oranları sunabilmesi ile açıklanabilmektedir. Aynı veri setinin aynı algoritma ile modellenip, analizi gerçekleştirildiği takdirde de farklı sonuçlar verdiği Weka ortamında denemeler yapılarak Çizelge 4.2'de ifade edilmektedir.

C4.5 Algoritması: Uygulama bölümünde belirtilen Şekil 4.5'teki seçeneklerin deneme yanılma yöntemi ile değiştirilmesi, başarı oranını arttırıp/azaltan etkiye sahiptir.

Örneğin “unpruned” seçeneğinin true/false durumları ile algoritmada budama yapılıp yapılmaması veya “minNumObj” seçeneğine atanan değer ile yaprakların temsil ettiği örnek sayısı gibi seçenekler düzenlenerek algoritmanın başarı oranı arttırılabilmektedir.

Rastgele Orman Algoritması: Şekil 4.8’de belirtilen seçenekler yardımı ile algoritma işleyiş mantığına aykırı olmadan, algoritmayı doğrudan etkileyen değerler üzerinde düzenlemeler gerçekleştirilebilmektedir. Bu düzenlemeler algoritma başarı seviyesini arttırma amacına yönelik olmaktadır. Örneğin algoritma seçenekleri bölümünde “numIterations” şeklinde bahsedilen kavram Rastgele Orman Algoritmasında kullanılacak ağaç sayısını temsil ettiğinden, bu algoritmaya özgü bir başarı oranı etkileme niteliğidir.

KNN Algoritması: KNN algoritması mantığına özgü kavramlardan en önemlileri de “KNN” ve “nearestNeighbourSearchAlgorithm” seçeneklerinin düzenlenmesidir. Mesafe mantığına dayanan KNN Algoritması Şekil 4.11’deki “KNN” seçeneği ile kullanılacak komşu sayısı belirlenirken, “nearestNeighbourSearchAlgorithm” seçeneği ile de mesafe ölçüm formülünün (Öklid, Manhattan…) seçilebilmesi ile başarı oranına değişim sağlayabilecek kendine özgü özellikler barındırmaktadır.

Örneğin; varsayılan Öklid mesafe formülü ile “KNN” niteliği ele alınarak gerçekleştirilen ufak denemeler ile bile model başarı oranlarında farklılıklar olduğu Çizelge 4.2’te gözlemlenebilmektedir.Matematiksel olarak Çizelge 4.2’de ifade edilen farklı k değerlerine ait hesaplanan algoritma başarı oranı durumlarının, çalışma içerisinde Weka sonuç ekran görüntüleri yer almayan durumlar Ekler bölümünde ifade edilmektedir.

Çizelge 4.2. KNN Algoritmasının Belirtilen Seçenekler Doğrultusunda Weka’da Elde Edilen Başarı Oranları (2).

Başarı Oranı **k Değeri**	**Cross-validation: Folds 10**	**Percentage split: %66**
k = 1	%33,463	%33,5114
k = 11	%33,5775	%33,6191
k = 13	%33,5821	%33,6146
k = 17	%33,5332	%33,6505
k = 21	%33,5073	%33,6909

4.4. MODEL PERFORMANSININ DEĞERLENDİRİLMESİ

Veri Madenciliği teknolojisi altında birçok modelleme yöntemi ve bu yöntemler kapsamında birçok algoritma bulunmaktadır. Bu yöntem ve algoritmaların performans, doğruluk, hassasiyet vb. kavramları göz önünde bulundurularak tercih edilebilirlik anlamında kesin bir yargılamada bulunmak söz konusu değildir. Her algoritma hitap ettiği veri seti ile uyumundan, o veri setine ait ihtiyaçları karşılıyor olmasından (veri türleri, veri seti boyutu vb.) gibi birçok özellikten kaynaklı olarak tercih edilmektedir. Bu yüzden en ideal model seçimi problem bazında olmaktadır. Zaten en ideal model diye bir model belirlenmiş olsa diğer modellerin kullanım amacı kalmayacaktır. İdeal olarak karar verilmek istenen model seçimi için veri seti ihtiyaçlarını karşılayan yöntemler araştırılıp, en başarılı algoritma bulunana kadar deneme yanılma yöntemi ile seçim gerçekleştirilmiş olmaktadır.

Model başarısının değerlendirilmesi, modelin idealliğinin belirlenmesi için bazı ölçütler bulunmaktadır. Bu ölçütlerin hesaplanabilmesi için model oluşum ve çıktılarına dayanan Çizelge 4.3'teki gibi bir matris oluşturulmaktadır. Matristeki bu bilgiler doğrultusunda ölçütlerin hesaplamaları yapılmaktadır. Bu matris bazı kaynaklarda "karışıklık matrisi" olarak da adlandırılmaktadır.

Çizelge 4.3. Karışıklık Matris Yapısı.

Tahmin Sınıfı / **Gerçek Sınıf**	**Pozitif**	**Negatif**
Pozitif	Doğru Pozitif (DP)(TP)	Yanlış Negatif (YN)(FN)
Negatif	Yanlış Pozitif (YP)(FP)	Doğru Negatif (DN)(TN)

Doğruluk Oranı;

Sınıflayıcı modelin doğru tahminlerinin ölçütüdür. Formül (4.1)'de belirtilen şekilde hesaplanmaktadır.

$$Doğruluk\ Oranı = \frac{TP+TN}{TP+FN+FP+TN} \quad (4.1)$$

Hata Oranı;

Sınıflayıcı modelin yanlış tahminlerinin ölçütüdür. Formül (4.2)'de belirtilen şekilde hesaplanmaktadır.

$$Hata\ Oranı = \frac{FP+FN}{TP+FN+FP+TN} \quad (4.2)$$

Duyarlılık;

Pozitif olarak doğru bir şekilde sınıflandırılmış örneklerin, gerçek sınıf değeri pozitif olan tüm örneklere oranlanması ile Formül (4.3)'te belirtilen şekilde hesaplanmaktadır.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (4.3)$$

Hassasiyet/Kesinlik;

Sınıflandırılan örneklerin doğru olarak tahminliğinin bulunmasının ölçütüdür. Formül (4.4)'te belirtilen şekilde hesaplanmaktadır.

$$Hassasiyet = \frac{TP}{TP+FP} \quad (4.4)$$

F- Ölçütü;

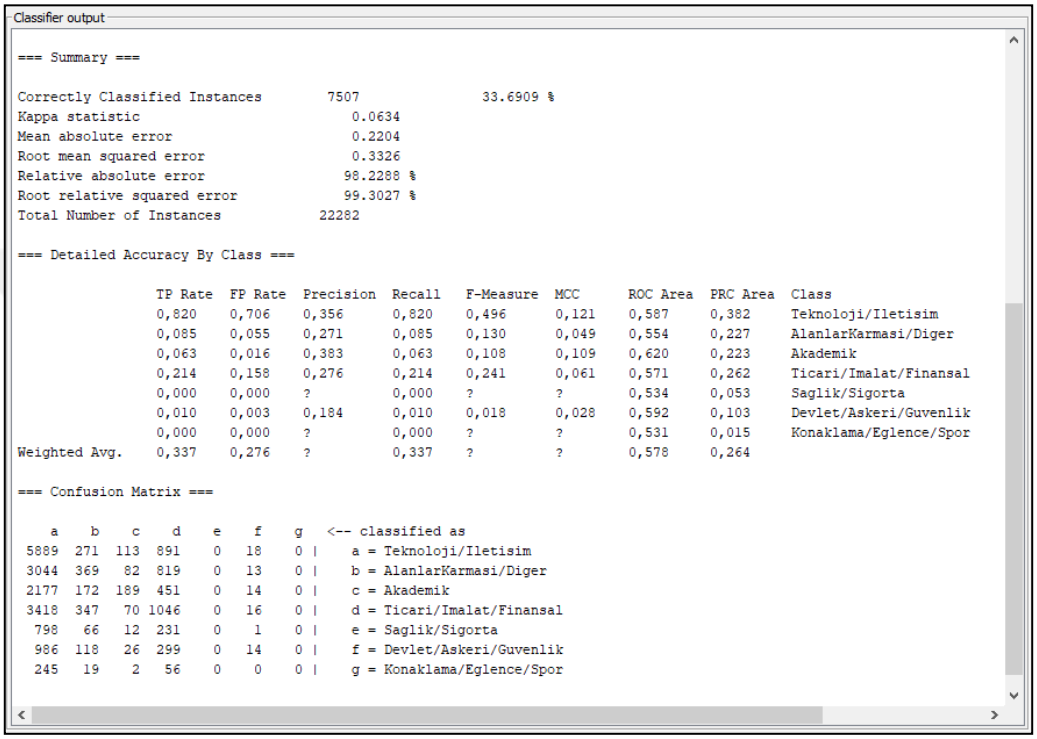
Duyarlılık ve hassasiyet ölçütlerinden daha detaylı anlamlar çıkarılması için beraber kullanımının yararlı olduğu hesaplanmış olup F-ölçütü kavramı adı altında değerlendirilmektedir. Formül (4.5)'te belirtilen şekilde duyarlılık ve hassasiyet ölçütlerinin harmonik ortalaması ile hesaplanmaktadır.

$$F\text{- }Ölçütü = \frac{2\ X(Duyarlılık)X\ (Hassasiyet)}{(Duyarlılık)+\ (Hassasyet)} \quad (4.5)$$

ROC (Receiving Operating Characteristics – Algılayıcı İşletim Özellikleri) Eğrisi;

Model performans başarısının görsel olarak belirtilebilmesi için kullanılmaktadır. Roc eğrisinde yatay eksen (x ekseni) yanlış pozitif oranlamasını, düşey eksen (y ekseni) doğru pozitif oranlamasını temsil etmektedir. Roc eğrisinin altında kalan alanın büyük olması, sınıflandırma modelinin başarısının fazla olması anlamına gelmektedir [39].

Bahsedilen model performansının değerlendirilmesine yardımcı olan kavramlar somut bir örnek ile ifade edilmek istenerek, Çizelge 4.2'den yüksek başarı oranını sunan model ele alınmıştır. Bu model Percentage split %66 ve k=21 seçeneklerinin kullanılması ile elde edilen KNN algoritmasıdır. Weka ortamında %33,6909 başarı oranına sahip algoritmanın model performansı değerlendirme seçenekleri ile birlikte ifade edilmesi Şekil 4.16'da belirtilmektedir.



Şekil 4.16. Weka Ortamında KNN Algoritması (Percentage split %66, k=21) Model Performansının Değerlendirilmesi.

5. BULGULAR VE TARTIŞMA

Başarı oranının yanı sıra, formüllerinden bahsedilen doğruluk oranı, hata oranı, duyarlılık, hassasiyet/kesinlik, F-ölçütü, ROC alanı gibi değerlendirme kavramları algoritma analizleri gerçekleştirildikten sonra Weka ortamı, Classify sekmesi içeriğinde, Classifier output bölümünde sonuçları ile yer almaktadırlar. Verimlilik ya da başarı kıyaslaması yapılmak istenildiği takdirde bu kavram değerleri kullanılarak daha somut yorumlamalar yapılabilmektedir.

Algoritmaların başarı oranı kavramlarının sonuçları esas alınarak değerlendirmeler yapılmak istenmekte olup, bu oranların farklılıklarının sebepleri ve farklılık oranları farklı durumlar öncülüğünde değerlendirilmek istenmektedir. Literatür taraması çalışmasında da bolca rastlanılan algoritma çeşidi, veri madenciliği aşamaları, ele alınan algoritmaya özgü kavram güncellemeleri gibi birçok faktörün model başarı oranında yarattığı sapma değerleri ele alınarak faktörlerin önem durumları ve etkileri değerlendirilmişlerdir. Örneğin C4.5, Rastgele Orman, KNN ve Destek Vektör Makinesi (SVM) algoritmaları, aynı veri setinin ön işleme aşaması öncesi ve sonrası halinin kullanılması ile modellenmişlerdir [60]. Elde edilen model başarı oranlarının algoritma çeşitleri baz alınarak hem algoritma çeşitlerinin model başarıları hem de orijinal ve işlenmiş veri setleri ile oluşan model başarı oranlarının sapmaları kıyaslanmıştır. Algoritma çeşitliliğinden dolayı oluşan model başarı oranı sapmaları, orijinal ve işlenmiş veri seti kullanımından kaynaklı oluşan sapma değerinden daha düşük değerdedir.

Ele alınan bir diğer konu, KNN algoritmasının “k” değerinin farklı değerler alması ile elde edilen model başarı oranlarının farklılaşma seviyeleridir. Bu tez çalışması kapsamında da tespit edildiği gibi k değeri değişimi başarı oranında düşük sapmalar yaratmaktadır. Literatürde karşılaşılan bir çalışmada Rastgele Orman, KNN, Naive Bayes ve C4.5 algoritmalarının arasında aynı veri seti ile elde edilen başarı oranı durumları hesaplanmıştır [61]. Elde edilen sonuçlar yine bu çalışma ile fikir birliği sağlar niteliğinde olup, modellenen algoritmaların Sınıflandırma (aynı) yöntemine ait olmalarından kaynaklı olarak model başarı oranlarında aşırı sapma yaratmadığı

gözlemlenmiştir. Aynı şekilde farklı bir çalışmada sınıflandırma alt yapılı algoritmalardan Naive Bayes, C4.5 ve Rastgele Orman sınıflandırıcıları kıyaslamış, model başarı oranları arasında aynı teknikli algoritmalar olmalarından kaynaklı olarak çok ufak sapmalara sahip olduğu sonucuna varılmıştır [62].

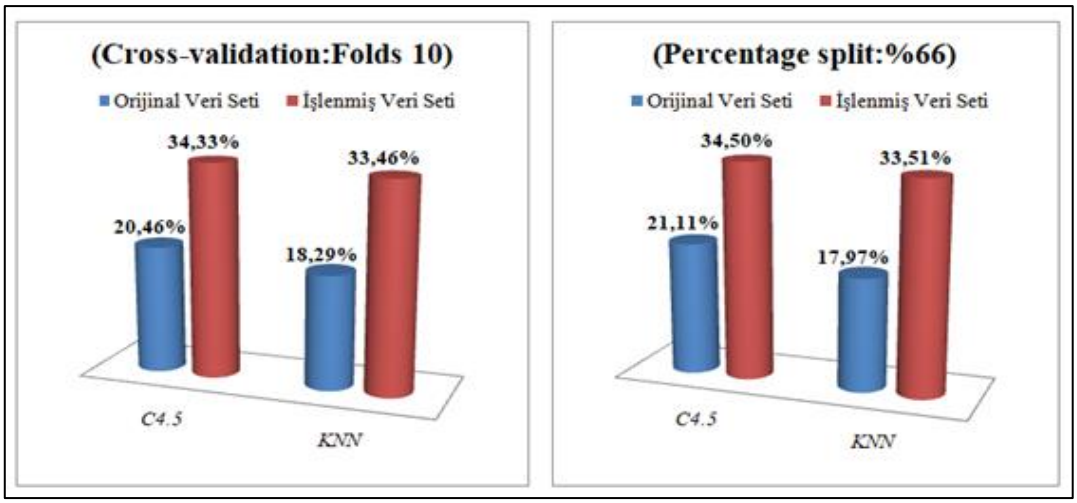
Bu çalışma kapsamında da Veri Madenciliğinin döngüsünde bulunan algoritma seçimi, ön işleme aşamaları, algoritmalara özgü kavram belirlemeleri gibi durumların model başarı oranına etkisinin gözlemlenebilmesi için literatür araştırmasında karşılaşılan örneklerden de hareketle model başarı oranı çıktıları yorumlanmıştır.

İlk durum Veri Madenciliği kavramı aşamalarının önemini vurgulayabilmek adına veri ön işleme adımı gerçekleştirilmemiş olan orijinal veri seti üzerinde, işlenmiş verilere uygulanan sınıflandırma algoritmalarından olan C4.5 ve KNN algoritma türlerinin modellenmelerinin gerçekleştirilmesidir. Başarı oranları Çizelge 5.1'de ifade edilmektedir. Matematiksel olarak ifade edilen ancak çalışma içerisinde yer almayan tüm başarı oranı hesaplama durumlarının Weka ekran görüntüleri Ekler bölümünde yer almaktadır.

Çizelge 5.1. Orijinal ve İşlenmiş Veri Seti Kullanıldığında, Seçili Algoritmaların Başarı Oranları ve Farkları.

Başarı Oranı / Algoritma Çeşidi	**Orijinal Veri Seti**	**İşlenmiş Veri Seti**	**Farkı**
C4.5 (Cross-validation: Folds 10)	%20,464	%34,3328	%13,8688
C4.5 (Percentage split: %66)	%21,1095	%34,4987	%13,3892
KNN (Cross-validation: Folds 10)	%18,2941	%33,463	%15,1689
KNN (Percentage split: %66)	%17,9676	%33,5114	%15,5438

Çizelge 5.1 incelendiğinde aynı algoritma çeşidi ve bu algoritmaların aynı seçenek değerleri, orijinal veri seti ve işlenmiş veri seti kullanılması ile modellenmeleri sonucunda elde edilen başarı oran değerleri, veri ön işleme aşaması sayesinde ortalama olarak %14-%15 oranında artış gösterdiği gözlenmektedir. Çizelge 5.1'de belirtilen artış değerlerinin grafiksel olarak ifade edilişi Şekil 5.1'de belirtilmektedir.

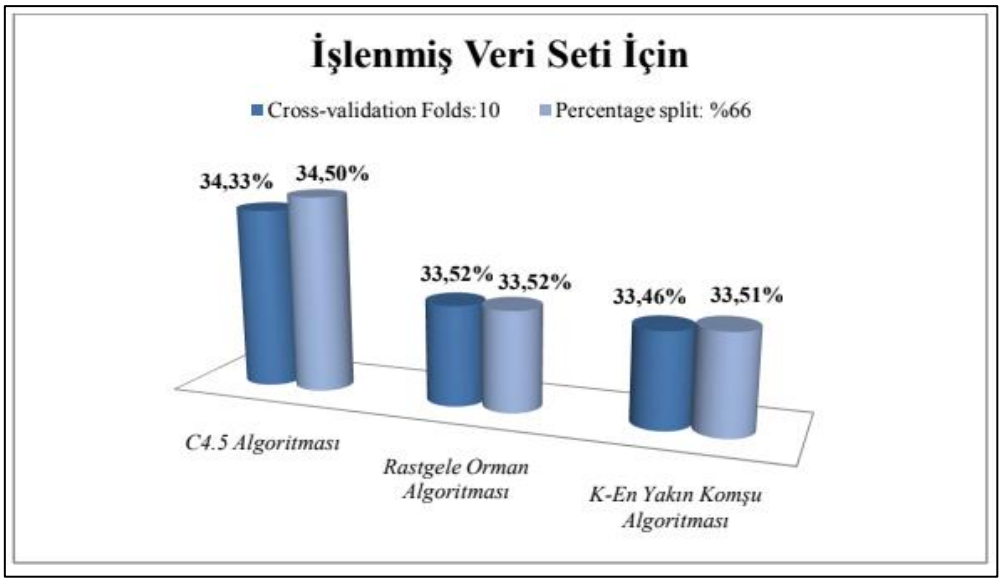


Şekil 5.1. Orijinal ve İşlenmiş Veri Setlerinin Seçilen Algoritmaların Başarı Oran Grafikleri.

Diğer bir durum orijinal veri seti kullanımı ile Veri Madenciliği aşamalarının öneminin anlaşılmasının yanı sıra, karşılaştırma değerlerine aynı veri seti kullanımı –işlenmiş veri seti- ile devam edilecek olmasıdır. Bu durum da gelişme bölümünde uygulamaları gerçekleştirilmiş olan iki farklı bakış açısının modellenmesidir.

Birincisi aynı veri seti için farklı sınıflandırma algoritmaları kullanılarak, aynı varsayılan seçenek değerleri ile modellenmeleri sonucunda elde edilen başarı oranı kıyaslamalarının gerçekleştirilmesi işlemidir. Sınıflandırma yöntemi adı altında ele alınan algoritmaların hepsi sınıflandırma mantığına dayalı çıktı değeri oluşturmaktadır. Farklı olan durum, sınıflandırma mantığının model oluşturma aşamasında kullanmış olduğu kavramlar ve teknikler olmaktadır.

Veri Madenciliği Yöntemleri bölümünde modelleme aşamaları da ifade edilen C4.5, Rastgele Orman ve KNN algoritmaları kapsamında işlenmiş veri seti kullanılarak bu üç algoritmanın varsayılan seçenek değerleri ve eğitim/test veri seti oranlaması ile elde edilen model başarıları Şekil 5.2'de grafiksel olarak ifade edilmektedir.



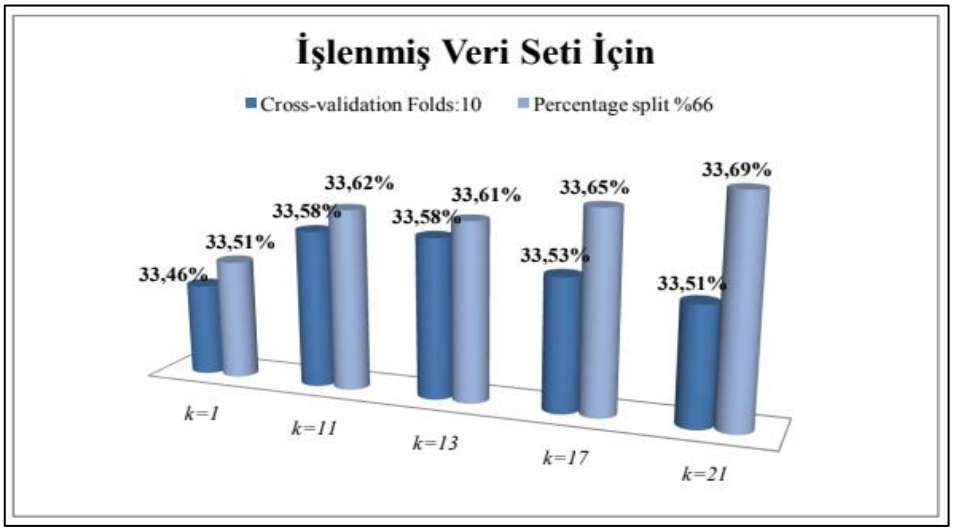
Şekil 5.2. İşlenmiş Veri Seti İçin Belirtilen Nitelikler Bazında Başarı Oranları (1).

Aynı veri seti için kullanılan algoritma türlerinin başarı oranları analiz edildiğinde iki bakış açısı ile değerlendirme sağlanabilmektedir.

- Sınıflandırma mantığı ile oluşturulmuş olsa dahi bu üç algoritmanın da model oluşturma yöntemi, model oluştururken kullanmış oldukları kavramlar farklı olduğundan, aynı veri seti üzerine uygulanmış olsalar da farklı başarı oranları sunmaktadırlar. Ele alınan problem sınıflandırma problemi olsa da algoritmaların belirtilen işleyiş farklılıklarından dolayı hangisinin kullanımının uygun olduğuna, deneme yanılma yöntemi ile karar verilmesi düşüncesi Şekil 5.2'deki grafikler ile de desteklenmektedir. Grafiklerde gözlenen başarı oranı sapma değerleri doğrultusunda kullanılması gereken algoritma seçimi gerçekleştirilmiş olmaktadır.
- Veri Madenciliği yöntemi, modellenen veri seti, algoritma türü gibi önemli faktörler aynı olmuş olsa bile, varsayılan değerleri ile kullanılan veri seti bölümleme seçeneklerinin farklı olmaları başarı oranını etkilemekte olduğu uygulamalı bir şekilde gözlemlenerek başarı oranları grafikler üzerinde ifade edilmektedir.

Gelişme bölümünde uygulaması gerçekleştirilmiş olan son durum ise, işlenmiş veri setinin KNN Algoritması ile modellenmiş olmasıdır. KNN algoritmasının en önemli kavramlarından biri olan k değeri farklılaşması ile gerçekleştirilen modellemelerin

başarı oranları ele alınmış idi. Bu değerler yorumlamaya destek olunabilmesi açısından grafikler ile de Şekil 5.3'te ifade edilmektedir.



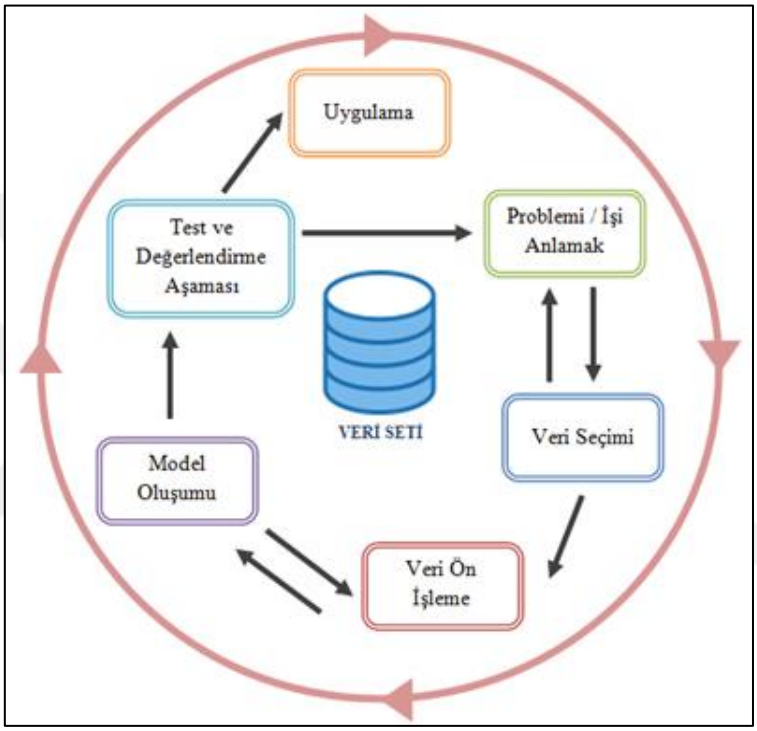
Şekil 5.3. İşlenmiş Veri Seti İçin Belirtilen Nitelikler Bazında Başarı Oranları (2).

Bu uygulama bazında işlenen veri setine KNN algoritmasının uygulanması ile başarı oranları elde edilmiştir. KNN algoritmasına özgü kavramlardan olan k-komşu sayısı kavramı farklılaşmasının model başarı oranına olan etkileri uygulamalı bir şekilde gözlemlenmiştir. En iyi k değeri bulma yöntemi elbette yoktur. Deneme yanılma yöntemi ile model başarısı artırımı o probleme en uygun k değerinin hangisi olduğunu belirlemekteki en garanti yöntemdir.

Aktarılan tüm çizelge ve grafiklerin ortak özelliğinin Veri Madenciliği aşamalarının kullanılması sonucu model başarı oranını arttırmaya yönelik olduğu gözlemlenebilmektedir. Model başarı oranı artırımı için belirli bir yöntem, formül olmaması ile birlikte başarı oranının kısıtı da bulunmamaktadır. Veri Madenciliği aşamaları her işlem adımlarının kendi içerisinde güncellenebilir olması ile veya en iyi seviyeye ulaşabilmesi ile bir bütünlük sağlayarak, model başarı oranını arttırmaya yönelik düzenlenmektedir.

Model başarısının yalnızca tek bir etkene bağlı olmadığı uygulama bölümünde gerçekleştirilen örnekler ve sonuç bölümünde ele alınan model başarılarının yorumlamaları ile somutlaştırılabilmektedir. Yalnızca veri seti farklılığı, yöntem-model-algoritma farklılığı, model oluşumunda kullanılan kavram farklılıkları vb. gibi birçok etmen başlı başına veya bütünsel olarak model başarısını etkileyebilmektedir.

Model başarısının tek bir etkene bağlı olmamasının yanı sıra model başarısını arttırmadaki en önemli bir diğer etken gelişimdir. Gelişim-güncelleme kavramları sonu bulunabilen kavramlar olmadığından her geçen gün algoritma başarısının, gerek ön işleme adımlarında yapılan bir güncelleme ile, gerek kullanılan algoritma özelliklerinde bulunan bir kavramın düzenlenmesi ile, gerek veri seti oranlamasının değiştirilmesi ile artırımının sağlanması mümkün olmaktadır. Şekil 5.4'te döngüsellik takibi ile model başarısının her daim artarak uygulama doğruluğuna katkı sağlaması ifade edilmektedir.



Şekil 5.4. Veri Madenciliği Aşamalarının Döngüselliği.

6. SONUÇLAR VE ÖNERİLER

Model başarı oranını etkileyen durumların ve ele alınan bu durum değerlerinin en iyi seviyeye ulaştırılması için gerçekleştirilen döngünün belirli kurallarının olmaması Veri Madenciliği problemlerinin "öznellik" yönünün vurgulanmasını desteklemektedir. Nitelik seçimleri, veri işlemeleri, yöntem belirleme vb. gibi durumlar için elbette literatür kısmında bahsedildiği üzere belirli teknikler bulunmaktadır. Ancak bu tez çalışması kapsamında da ön plana çıkarılmak istenilen döngüsellik ve öznellik kavramları ile bahsedilmek istenilen durum, belirli teknikler içerisinden seçim yapılırken herhangi bir kurala tabii tutulmadan, geliştiricinin yorum ve tecrübeleri doğrultusunda gerçekleştirilen seçimler ile model başarısı elde edilmesidir.

Örneğin; "X problemi için Y yöntemi uygundur. Y yöntemi var ise eğitim/test veri seti oranlaması a/b olmalıdır. Y yöntemi algoritmalarından Z algoritmasının kullanması üzerine, algoritma seçeneklerinin belirli kurallar doğrultusunda ayarlanması gerekmektedir" şeklinde nesnel bir yargı kullanılması mümkün değildir. Gerek ele alınan problem, gerek problemi ele alan kullanıcı gibi her türlü görüş farklılıklarına açık olunan Veri Madenciliği aşamaları verimli bir şekilde bütünleştiği takdirde döngü sürekli olarak başarı oranı arttırılabilir şekilde güncellenmektedir.

"Yerinde duran, geriye gidiyor demektir." demiş Mustafa Kemal Atatürk. Bu söz her algoritmanın gelişim ve başarısı için sloganı olmalıdır. Döngüsellik, algoritmalara gelişim ve güncelleme imkânı sunarak başarı oranı artırımı sağlayabiliyor iken; yerinde duran statik algoritmalar model başarısına katkı sağlayamamak ile birlikte gelişen/güncellenen modellemeler arasında günden güne başarı sırası gerilemektedir. Verimli ve başarılı sonuçlar sunan modeller elde edilebilmesi için, Veri Madenciliği aşamalarında seçim gerektiren her türlü durumlar ve teknikler için döngü mantığı en iyi seçenekler tercih edilecek şekilde güncellenmelidir. Bu çalışmadan yapılan çıkarım ve sonrasında gerçekleştirilebilecek çalışmalar, model başarısını maksimum seviyede tutabilmek için gerçekleştirilmesi gereken güncelleme seçenekleri ve bu seçeneklerin belirlenme durumları olarak özet bir şekilde ifade edilebilir.

7. KAYNAKLAR

[1] A. Kalikov, "Veri madenciliği ve bir e-ticaret uygulaması," Yüksek Lisans Tezi, Elektronik - Bilgisayar Eğitimi, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, Türkiye, 2006.

[2] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği,"*İstanbul Üniversitesi İşletme Fakültesi Dergisi*, c. 29, ss. 1–22, 2000.

[3] P. Giudici, *Applied data mining statistical methods for business and industry*, 2. baskı, Hoboken, New Jersey, ABD: John Wiley & Sons, 2009.

[4] B. Daş ve İ. Türkoğlu, "DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması," *Eleco 2014 Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu*, Bursa, Türkiye, 2014, ss. 381-383.

[5] H. Güldal ve Y. Çakıcı, "Ders yönetim sistemi yazılımı kullanıcı etkileşimlerinin sınıflandırma algoritmaları ile analizi," *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, c. 21, sayı. 4, ss. 1355–1367, 2017.

[6] M. A. Pala, M. E. Çimen, Ö. F. Boyraz, M. Z. Yıldız ve A. F. Boz, "Meme kanserinin teşhis edilmesinde karar ağacı ve knn algoritmalarının karşılaştırmalı başarım analizi," *7. Uluslararası Mühendislik ve Bilimde Yenilikçi Teknolojiler Sempozyumu*, Şanlıurfa, Türkiye, 2019, ss. 544–552.

[7] Anonim, (2021, 13 Şubat). [Online]. Erişim: https://www.bilisimle.com/gecmisten-gunumuze-veri-depolama-birimleri-tarihi/

[8] M. J. Show, C. Subramaniam, G. W. Tan, ve M. E. Welge, "Knowledge management and data mining for marketing," *Decision Support Systems*, c. 31, sayı. 1, ss. 127–137, 2001.

[9] Y. Lindell ve B. Pinkas, "Privacy preserving data mining," *Journal of Cryptology*, c. 15, sayı 3, ss. 177–206, 2002.

[10] F. Ö. Koçoğlu, "Veri madenciliğinde veri ayrıklaştırma yöntemlerinin karşılaştırılması ve bir uygulama," Yüksek Lisans Tezi, Enformatik, Fen Bilimleri Enstitüsü, İstanbul Üniversitesi, İstanbul, Türkiye, 2012.

[11] P. Jacobs, "Data mining: what general manager sneed to know," *Harvard Management Update*, c. 4, sayı 8, 1999.

[12] M. M. Uluyardımcı ve M. Zontul, "Veri madenciliği yöntemleri ile uçuş biletleme analizi," *Aurum Mühendislik Sistemleri ve Mimarlık Dergisi*, c. 3, sayı 2, ss. 153–168, 2019.

[13] U. Fayyad, G. Piatetsky-Shapiro, ve P. Smyth, "From Data mining to knowledge discovery in databases," *AI Magazine*, c. 17, sayı 3, ss. 37–54, 1996.

[14] S. Çankırı, E. Kartal, K. Yıldırım, ve S. Gülseçen, "Organizasyonlarda bilgi yönetimi sürecinde veri madenciliği yaklaşımı," *Bilgi Çağında Varoluş: "Fırsatlar ve Tehditler" Sempozyumu*, İstanbul, Türkiye, 2009, ss. 148–167.

[15] R. Sharda, D. Delen ve E. Turban, *Business intelligence : a managerial perspective on analytics*, 3.baskı, Hoboken, New Jersey, ABD: Prerice Hall, 2013.

[16] W. A. Aldana, "Data mining industry: emerging trends and new opportunities," Yüksek Lisans Tezi, Elektrik Mühendisliği ve Bilgisayar Bilimi, Massachusetts Teknoloji Enstitüsü, Cambridge, Massachusetts, ABD, 2020.

[17] S. Savaş, N. Topaloğlu, ve M. Yılmaz, "Veri madenciliği ve Türkiye'deki uygulama örnekleri," *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, c. 11, sayı 21, ss. 1–23, 2012.

[18] A. Çalış, S. Kayapınar, ve T. Çetinyokuş, "Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama," *Gazi Üniversitesi - Endüstri Mühendisliği Dergisi*, c. 25, sayı 3–4, ss. 2–19, 2014.

[19] A. A. Awan. (2020). *Data science survey on Kaggle*. [Online] Erişim: https://www.kaggle.com/kingabzpro/datascience-survey-on-kaggle

[20] S. Savaş ve N. Arıcı, "Web tabanlı uzaktan eğitimde iki farklı öğretim modelinin öğrenci başarısı üzerindeki etkilerinin incelenmesi," *5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09),* Karabük, Türkiye, 2009.

[21] A. S. Bozkır ve E. A. Sezer, "Usage of data mining techniques in discovering the food consumption patterns of students and employees of university," *Balkan-Kafkas ve Türk Devletleri Uluslararası Mühendislik Sempozyumu,* Isparta, Türkiye, 2009, ss. 104–109.

[22] A. S. Albayrak ve Ş. K. Yılmaz, "Veri madenciliği karar ağacı algoritmaları ve imkb verileri üzerine bir uygulama," *Süleyman Demirel Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi*, c. 14, sayı 1, ss. 31–52, 2009.

[23] U. Tuğba ve Ş. Gürsoy, "Customer churn analysis in telecommunication sector," *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, c. 39, sayı 1, ss. 35–49, 2010.

[24] F. Güntürkün, "A comprehensive review of data mining applications in quality improvement and a case study," Yüksek Lisans Tezi, İstatistik Bölümü, Fen Bilimleri Enstitüsü, Orta Doğu Teknik Üniversitesi, Ankara, Türkiye, 2007.

[25] Ş. Doğan ve İ. Türkoğlu, "Iron-deficiency anemia detection from hematology parameters by using decision trees," *International Journal of Science&Technology*, c. 3, sayı 1, ss. 85–92, 2008.

[26] M. Danacı, M. Çelik, ve A. E. Akkaya, "Veri madenciliği yöntemleri kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi," *Akıllı Sistemlerde Yenilikler ve Uygulama Sempozyumu,* Kayseri, Türkiye, 2010, ss. 9–12.

[27] M. VanDiepen ve P. H. Franses, "Evaluating chi-squared automatic interaction detection," *Information Systems*, c. 31, sayı 8, ss. 814–831, 2006.

[28] C. Güngör ve S. Şekeroğlu, "Hizmet sektöründe bir veri madenciliği uygulaması," Yüksek Lisans Tezi, Mühendislik Yönetimi, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, Türkiye, 2010.

[29] J. Han, M. Kamber, ve J. Pei, *Data Mining: Concepts and Techniques*, 3. baskı, Burlington, Massachusetts, ABD: Morgan Kaufmann Publishers, 2012.

[30] M. E. Taşçı ve R. Şamlı, "Veri madenciliği ile kalp hastalığı teşhisi," *Avrupa Bilim ve Teknoloji Dergisi*, Özel Sayı, ss. 88–95, 2020.

[31] M. Çelik, “Veri madenciliğinde kullanılan sınıflandırma yöntemleri ve bir uygulama,” Yüksek Lisans Tezi, Ekonometri, Sosyal Bilimler Enstitüsü, İstanbul Üniversitesi, İstanbul, Türkiye, 2009.

[32] P. Cihan ve O. Kalıpsız, “Öğrenci proje anketlerini sınıflandırmada en başarılı algoritmanın belirlenmesi,” *TBV - Türkiye Bilişim Vakfı Bilgisayar Bilimleri Ve Mühendisliği Dergisi*, c. 8, sayı 1, 2015.

[33] E. Küçüksille, “Veri madenciliği süreci kullanılarak portföy performansının değerlendirilmesi ve imkb hisse senetleri piyasasında bir uygulama,” Doktora Tezi, İşletme, Sosyal Bilimler Enstitüsü, Süleyman Demirel Üniversitesi, Isparta, Türkiye, 2009.

[34] E. Uzun. [Online]. Erişim: https://erdincuzun.com/makine_ogrenmesi/decision-tree-karar-agaci-id3-algoritmasi-classification-siniflama/

[35] Y. Özkan ve Ç. S. Erol, *Biyoenformatik DNA Mikrodizi Veri Madenciliği*, 2. baskı, İstanbul, Türkiye: Papatya Yayıncılık, 2015.

[36] Ş. E. Şeker. (2012). [Online]. Erişim: https://bilgisayarkavramlari.com/2012/11/13/c4-5-agaci-c4-5-tree/

[37] F. Bozan. (Ocak, 2010). [Online]. Erişim: http://www.farukbozan.com/2010/01/cartclassification-and-regression-tree/

[38] S. Şensoy, “Kategorik değişkenler arası ilişki katsayılarının simülasyon yoluyla karşılaştırılması,” Yüksek Lisans Tezi, Zootekni, Fen Bilimleri Enstitüsü, Ordu Üniversitesi, Ordu, Türkiye, 2020.

[39] E. Deconinck, T. Hancock, D. Coomans, D. Massart, ve Y. VanderHeyden, “Classification of drugs in absorption classes using the classification and regression trees (cart) methodology,” *Journal of Pharmaceutical and Biomedical Analysis*, c. 39, ss. 91–103, 2005.

[40] L. Breiman, J. Friedman, ve C. J. Stone, "Classification and Regression Trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, ss. 14-23, 2011.

[41] E. Korkem, “Mikroarray gen ekspresyon veri setlerinde random forest ve naive bayes sınıflandırma yöntemleri yaklaşımı,” Yüksek Lisans Tezi, Biyoistatistik, Sağlık Bilimleri Enstitüsü, Hacettepe Üniversitesi, Ankara, Türkiye, 2013.

[42] D. Şengür ve A. Tekin, “Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini,” *Bilişim Teknolojileri Dergisi*, c. 6, sayı 3, ss. 7–16, 2014.

[43] B. Ataseven, “Yapay sinir ağları ile öngörü modellemesi,” *Öneri Dergisi*, c. 10, sayı 39, ss. 101-115, 2013.

[44] S. Chakrabarti, E. Cox, E. Frank, R. Güting, J. Han, X. Jiang, M. Kamber, S. Lightstone, T.Nadeau, R. E. Neapolitan, D.Pyle, M. Refaat, M. Schneider, T. Teorey ve I. Witten, *Data Mining Know It All*, 1. baskı, Burlington, Massachusetts, ABD: Morgan Kaufmann Publishers, 2008.

[45] M. Bramer, *Principles of data mining*, Berlin, Almanya: Springer Science+Business Media, 2020.

[46] S. Atadan ve T. Tosun, “Veri madenciliği teknikleriyle kredi kartlarında müşteri kaybetme analizi,” Yüksek Lisans Tezi, Bilgisayar Mühendisliği, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, Türkiye, 2006.

[47] A. Gülten ve Ş. Doğan, "Genetik algoritmalar yönteminin biyomedikal verileri üzerindeki uygulamaları," *Fırat Üniversitesi Doğu Araştırmaları Dergisi*, c. 7, sayı 1, ss. 12–16, 2008.

[48] R. Solmaz, M. Günay, ve A. Alkan, "Fonksiyonel tiroit hastalığı tanısında naive bayes sınıflandırıcının kullanılması," *Akademik Bilişim'14 - XVI. Akademik Bilişim Konferansı Bildirileri,* Mersin Üniversitesi, Mersin, Türkiye, 2014, ss. 891-897.

[49] E.Şirin. (Temmuz, 2017). [Online]. Erişim: https://www.veribilimiokulu.com/kumeleme-notlari-1-giris/

[50] T. Altıntaş, "Veri madenciliği metotlarından olan kümeleme algoritmalarının uygulamalı etkinlik analizi," Yüksek Lisans Tezi, Endüstri Mühendisliği, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, Türkiye, 2006.

[51] A. Baltacı, "Nitel araştırmalarda örnekleme yöntemleri ve örnek hacmi sorunsalı üzerine kavramsal bir inceleme," *Bitlis Eren Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, c. 7, sayı 1, ss. 231–274, 2018.

[52] A. S. Koyuncugil ve N. Özgülbaş, "Veri madenciliği: tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları," *Bilişim Teknolojileri Dergisi*, c. 2, sayı 2, ss. 21-32, 2009.

[53] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, ve I. H. Witten, "The weka data mining software," *ACM SIGKDD Explorations Newsletter*, c. 11, sayı 1, ss. 10–18, 2009.

[54] A. Tekerek, "Veri madenciliği süreçleri ve açık kaynak kodlu veri madenciliği araçları," *Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri 2,* Malatya, Türkiye, 2011, ss. 2-4.

[55] T. Bilgin, "Veri akışı diyagramları tabanlı veri madenciliği araçları ve yazılım geliştirme ortamları," *Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri,* Harran Üniversitesi, Şanlıurfa, Türkiye, 2009, ss. 807-814.

[56] M. Kaya ve S. A. Özel, "Açık kaynak kodlu veri madenciliği yazılımlarının karşılaştırılması," *Akademik Bilişim'14 - XVI. Akademik Bilişim Konferansı Bildirileri,* Mersin, Türkiye, 2014, ss. 47-53.

[57] Anonim, (2014, 15 Aralık) [Online]. Erişim: https://www.cs.waikato.ac.nz/ml/weka/index.html

[58] E. Uzun. (2021). [Online]. Erişim: https://erdincuzun.com/makine_ogrenmesi/makine-ogrenmesinden-elde-edilen-modellerin-degerlendirmesi/

[59] G. Bland. (Ekim 2020). [Online]. Erişim: https://algotrading101.com/learn/train-test-split/

[60] J. P. Kandhasamy ve S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science,* c. 47, ss. 45-51, 2015.

[61] M. Alehegn, R. R. Joshi ve P. Mulay, "Diabetes analysis and prediction using random forest, knn, naive bayes and j48: an ensemble approach," *International Journal Of Scientific&Technology Research*, c. 8, sayı 9, ss. 1346-1354, 2019.

[62] N. Ashraf, W. Ahmad ve R. Ashraf, “A comparative study of data mining algorithms for high detection rate in intrusion detection system,” *Annals of Emerging Technologies in Computing (AETiC)*, c. 2, sayı 2, ss. 49-57, 2018.

8. EKLER

8.1. EK 1:C4.5 ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (CROSS-VALİDATİON: FOLDS 10)

Orijinal veri setinin C4.5 Algoritması ile modellenmesi sonucu %20,464 başarı oranı elde edilmiştir. Şekil 8.1'de orijinal veri setinin C4.5 Algoritması, "Cross-validation: Folds 10" seçeneği ile başarı oranı hesabının özet çıktısı, Şekil 8.2'de ise bu hesaba ait Karışıklık Matrisi'nin çıktısı ifade edilmektedir.

Classifier output

```
=== Summary ===

Correctly Classified Instances        1226               20.464  %
Kappa statistic                          0.0413
Mean absolute error                      0.1094
Root mean squared error                  0.2502
Relative absolute error                 98.0537 %
Root relative squared error            105.9455 %
Total Number of Instances             5991
```

Şekil 8.1. C4.5 Algoritması Başarı Oranı Weka Özet Çıktısı (1).

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   <-- classified as
  25  14 322  35   8   0   8   0   0   3   1   1   3   2   1 |   a = Internet-based
  20  20 432  47  16   0   7   1   1   7   2   5   0   6   0   2 |   b = MixOfFields
  43  39 886 104  31   6  18   1   3  11   5   3   1   3   1   3 |   c = Technology
  31  40 599 236  24   6  15   2   8   5   8   7   1   5   1   3 |   d = Academic
  17  16 459  63  25   4  13   1   0   8   4  10   0   7   2   2 |   e = Financial
   2   2 168  11   6   0   4   0   1   0   0   2   0   1   1   0 |   f = Telecommunications
  22  18 407  55  23   3  14   4   0   6   1   2   1   4   3   1 |   g = Other
   2   2  75   6   3   0   5   0   0   0   2   2   0   1   0   0 |   h = Pharmaceutical
   1   2  53  20   0   1   1   0   0   0   1   1   0   0   0   2 |   i = Military/Security
   8  10 172  14  13   1   6   0   0   0   1   4   0   1   0   0 |   j = CRM/Marketing
   6   4 146  25   8   0   3   2   0   3   2   4   1   1   1   1 |   k = Manufacturing
   6   8 213  37  15   2   2   1   1   7   3   9   1   6   2   0 |   l = Government
   4   2  56   8   1   0   2   0   0   1   2   1   0   0   1   0 |   m = Hospitality/Entertainment/Sports
   5   7 140  16  11   0   3   0   0   3   1   2   0   7   0   0 |   n = Insurance
   9   5 117  17   5   0   5   2   0   1   1   2   1   1   0   0 |   o = Retail
   3   6  53  11   6   0   4   0   1   1   0   1   0   0   0   2 |   p = Non-profit
```

Şekil 8.2. C4.5 Algoritması - Karışıklık Matrisi (1).

8.2. EK 2: C4.5 ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (PERCENTAGE SPLİT:%66)

Orijinal veri setinin C4.5 Algoritması ile modellenmesi sonucu %21,1095 başarı oranı elde edilmiştir. Şekil 8.3'te orijinal veri setinin C4.5 Algoritması, "Percentage split:%66" seçeneği ile başarı oranı hesabının özet çıktısı, Şekil 8.4'te ise bu hesaba ait Karışıklık Matrisi'nin çıktısı ifade edilmektedir.

Classifier output

```
=== Summary ===

Correctly Classified Instances         430               21.1095 %
Kappa statistic                          0.0511
Mean absolute error                      0.109
Root mean squared error                  0.2486
Relative absolute error                 97.6735 %
Root relative squared error            105.1924 %
Total Number of Instances              2037
```

Şekil 8.3. C4.5 Algoritması Başarı Oranı Weka Özet Çıktısı (2).

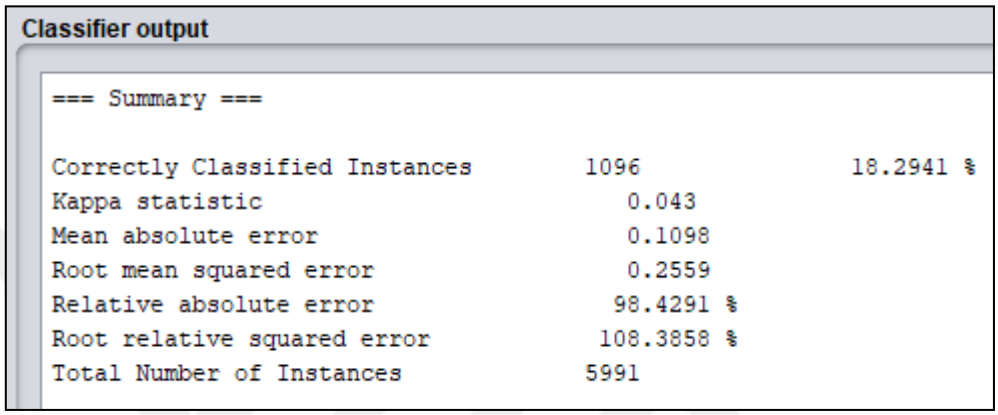
```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   <-- classified as
   6   3 126  10   2   0   5   0   0   2   0   0   1   1   0   0 |   a = Internet-based
   8  11 155  13   6   0   1   1   0   2   0   1   0   2   0   1 |   b = MixOfFields
  15  13 306  35   8   1   0   0   1   2   1   2   0   0   0   1 |   c = Technology
  11  13 208  80   7   2   5   1   1   2   2   3   1   1   0   2 |   d = Academic
   6   3 138  22  12   1   5   0   0   4   2   2   0   1   1   0 |   e = Financial
   0   0  55   2   1   0   1   0   0   0   0   0   0   0   0   1 |   f = Telecommunications
   7   3 139  26   5   0   7   1   0   1   4   1   1   1   3   1 |   g = Other
   1   0  20   2   1   0   1   0   0   1   0   1   0   0   1   0 |   h = Pharmaceutical
   1   0  21   8   0   0   1   0   0   0   0   0   0   0   0   0 |   i = Military/Security
   2   5  71   4   4   0   2   0   0   1   1   2   0   0   0   0 |   j = CRM/Marketing
   2   0  56  10   2   0   1   0   0   1   1   1   1   0   0   0 |   k = Manufacturing
   2   1  58  16   4   1   1   0   1   0   1   2   1   1   0   0 |   l = Government
   2   1  15   1   3   0   1   0   0   0   0   0   0   0   1   0 |   m = Hospitality/Entertainment/Sports
   3   0  46   4   2   0   3   0   0   4   1   1   0   3   0   0 |   n = Insurance
   2   1  49   4   4   0   1   1   0   0   1   1   0   0   0   0 |   o = Retail
   3   0  18   3   2   0   1   0   0   0   0   1   0   0   0   1 |   p = Non-profit
```

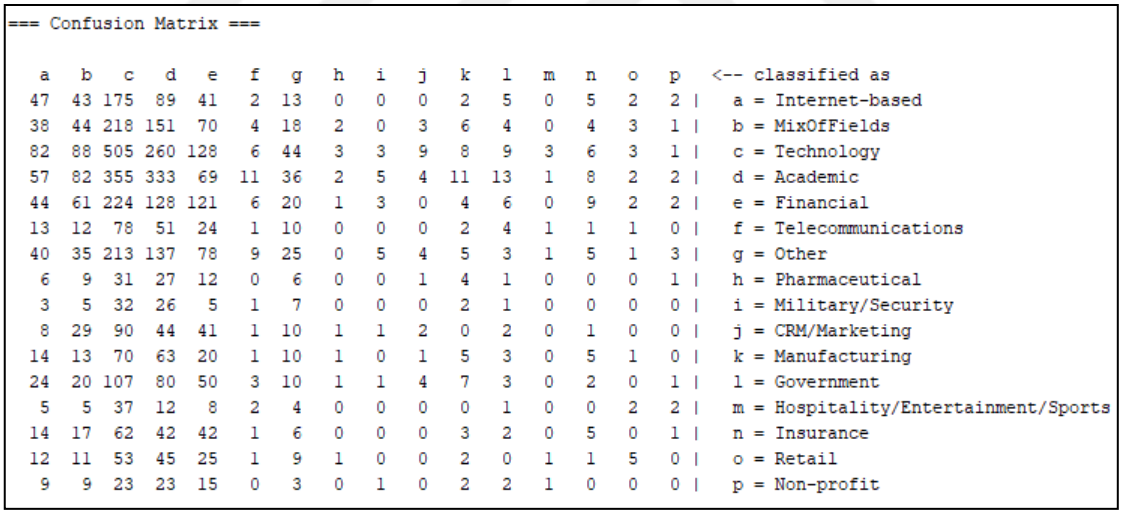
Şekil 8.4. C4.5 Algoritması - Karışıklık Matrisi (2).

8.3. EK 3: KNN ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (K=1, CROSS-VALİDATİON: FOLDS 10)

Orijinal veri setinin KNN Algoritması ile modellenmesi sonucu %18,2941 başarı oranı elde edilmiştir. Şekil 8.5'te orijinal veri setinin KNN Algoritması ile k=1 ve "Cross-validation: Folds 10" seçenekleri ile başarı oranı hesabının özet çıktısı, Şekil 8.6'da ise bu hesaba ait Karışıklık Matrisi'nin çıktısı ifade edilmektedir.



Şekil 8.5. KNN Algoritması Başarı Oranı Weka Özet Çıktısı (1).



Şekil 8.6. KNN Algoritması - Karışıklık Matrisi (1).

8.4. EK 4: KNN ALGORİTMASI KULLANILARAK ORJİNAL VERİ SETİNİN MODELLENMESİ (K=1, PERCENTAGE SPLİT:%66)

Orijinal veri setinin KNN Algoritması ile modellenmesi sonucu %17,9676 başarı oranı elde edilmiştir. Şekil 8.7'de orijinal veri setinin KNN Algoritması, k=1 ve "Percentage split:%66" seçenekleri ile başarı oranı hesabının özet çıktısı, Şekil 8.8'de ise bu hesaba ait Karışıklık Matrisi'nin çıktısı ifade edilmektedir.

Classifier output

```
=== Summary ===

Correctly Classified Instances         366               17.9676 %
Kappa statistic                          0.0412
Mean absolute error                      0.1096
Root mean squared error                  0.2564
Relative absolute error                 98.1978 %
Root relative squared error            108.5028 %
Total Number of Instances             2037
```

Şekil 8.7. KNN Algoritması Başarı Oranı Weka Özet Çıktısı (2).

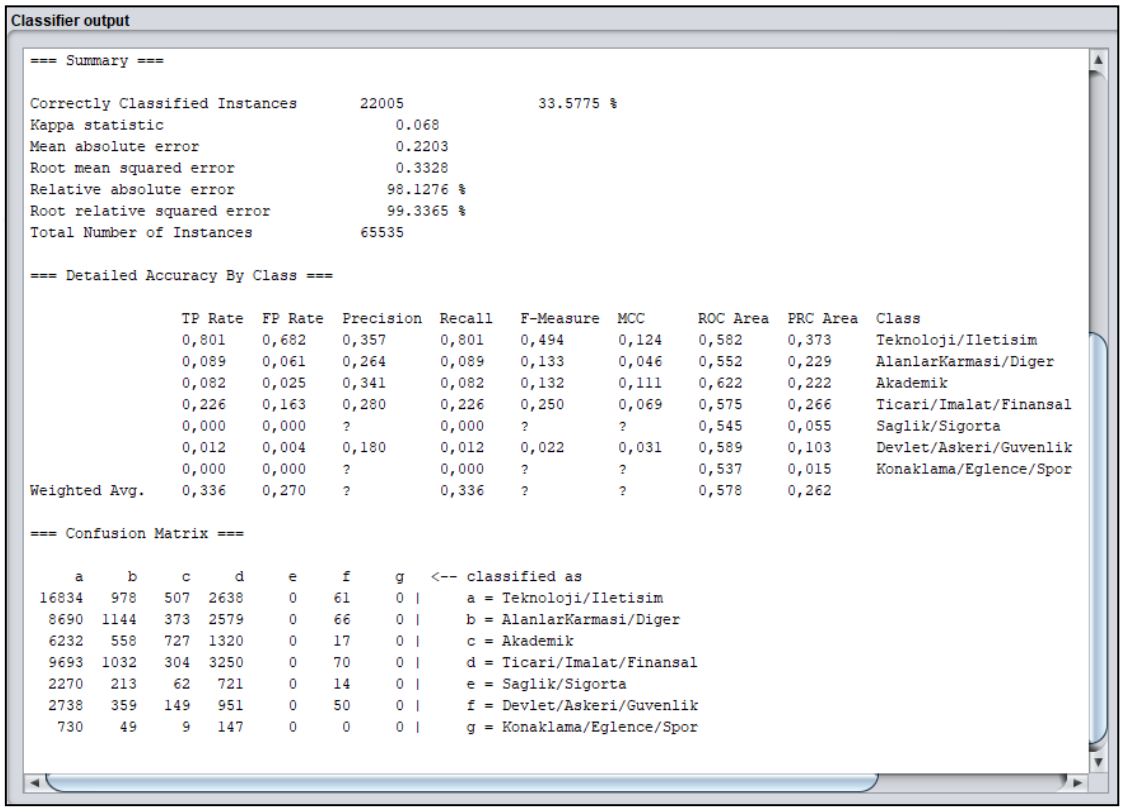
```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   <-- classified as
  17  14  63  39  14   0   5   0   0   0   1   0   0   1   1   1 |   a = Internet-based
  18  16  80  47  26   1   5   1   0   0   2   1   0   4   0   0 |   b = MixOfFields
  30  37 168  85  41   1  12   3   0   0   1   3   0   4   0   0 |   c = Technology
  23  31 118 108  33   1  13   0   2   1   2   4   0   2   1   0 |   d = Academic
  12  20  73  38  39   2   5   1   1   1   0   2   0   2   1   0 |   e = Financial
   2   4  24  16   9   0   2   0   0   0   0   1   0   1   1   0 |   f = Telecommunications
  17  15  72  45  26   6   8   0   1   1   5   0   0   2   1   1 |   g = Other
   3   3  12   7   1   0   1   0   0   0   0   1   0   0   0   0 |   h = Pharmaceutical
   1   2  15   9   2   0   0   0   0   2   0   0   0   0   0   0 |   i = Military/Security
   3  15  37  12  16   1   6   0   1   1   0   0   0   0   0   0 |   j = CRM/Marketing
   3   7  27  23   7   0   3   0   1   0   3   1   0   0   0   0 |   k = Manufacturing
   5   9  35  21  12   1   2   0   2   0   1   1   0   0   0   0 |   l = Government
   2   1  14   3   1   1   0   0   0   0   0   1   0   0   1   0 |   m = Hospitality/Entertainment/Sports
   6  10  18  15  10   1   2   0   0   0   1   1   0   3   0   0 |   n = Insurance
   7   9  13  17  12   0   1   1   0   0   2   0   0   0   2   0 |   o = Retail
   4   2   7  11   3   0   0   0   0   0   0   1   1   0   0   0 |   p = Non-profit
```

Şekil 8.8. KNN Algoritması - Karışıklık Matrisi (2).

8.5. EK 5: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=11, CROSS-VALİDATİON: FOLDS 10)

İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,5775 başarı oranı elde edilmiştir. Şekil 8.9'da işlenmiş veri setinin KNN Algoritması, k=11 ve "Cross-validation: Folds 10" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Classifier output

```
=== Summary ===

Correctly Classified Instances       22005               33.5775 %
Kappa statistic                          0.068
Mean absolute error                      0.2203
Root mean squared error                  0.3328
Relative absolute error                 98.1276 %
Root relative squared error             99.3365 %
Total Number of Instances            65535

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,801    0,682    0,357      0,801    0,494      0,124    0,582     0,373     Teknoloji/Iletisim
                 0,089    0,061    0,264      0,089    0,133      0,046    0,552     0,229     AlanlarKarmasi/Diger
                 0,082    0,025    0,341      0,082    0,132      0,111    0,622     0,222     Akademik
                 0,226    0,163    0,280      0,226    0,250      0,069    0,575     0,266     Ticari/Imalat/Finansal
                 0,000    0,000    ?          0,000    ?          ?        0,545     0,055     Saglik/Sigorta
                 0,012    0,004    0,180      0,012    0,022      0,031    0,589     0,103     Devlet/Askeri/Guvenlik
                 0,000    0,000    ?          0,000    ?          ?        0,537     0,015     Konaklama/Eglence/Spor
Weighted Avg.    0,336    0,270    ?          0,336    ?          ?        0,578     0,262

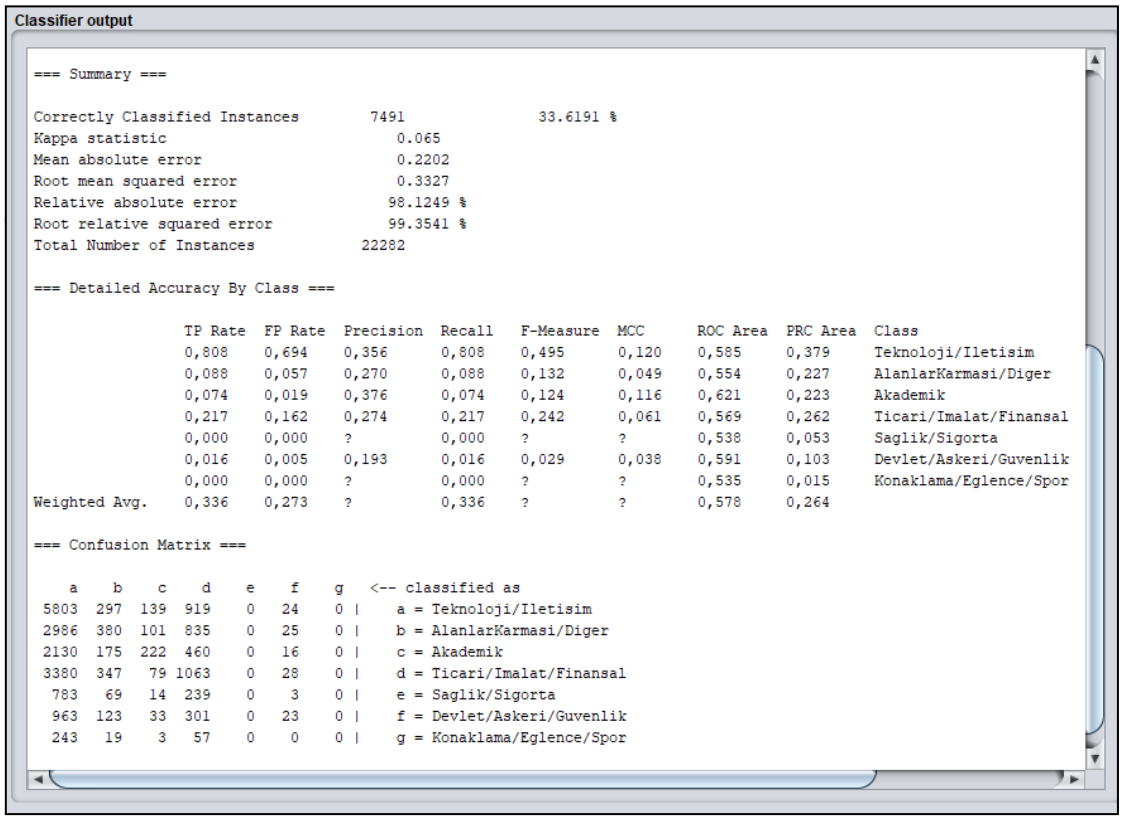
=== Confusion Matrix ===

     a     b     c     d     e     f     g   <-- classified as
 16834   978   507  2638     0    61     0 |     a = Teknoloji/Iletisim
  8690  1144   373  2579     0    66     0 |     b = AlanlarKarmasi/Diger
  6232   558   727  1320     0    17     0 |     c = Akademik
  9693  1032   304  3250     0    70     0 |     d = Ticari/Imalat/Finansal
  2270   213    62   721     0    14     0 |     e = Saglik/Sigorta
  2738   359   149   951     0    50     0 |     f = Devlet/Askeri/Guvenlik
   730    49     9   147     0     0     0 |     g = Konaklama/Eglence/Spor
```

Şekil 8.9. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (3).

8.6. EK 6: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=11, PERCENTAGE SPLİT:%66)

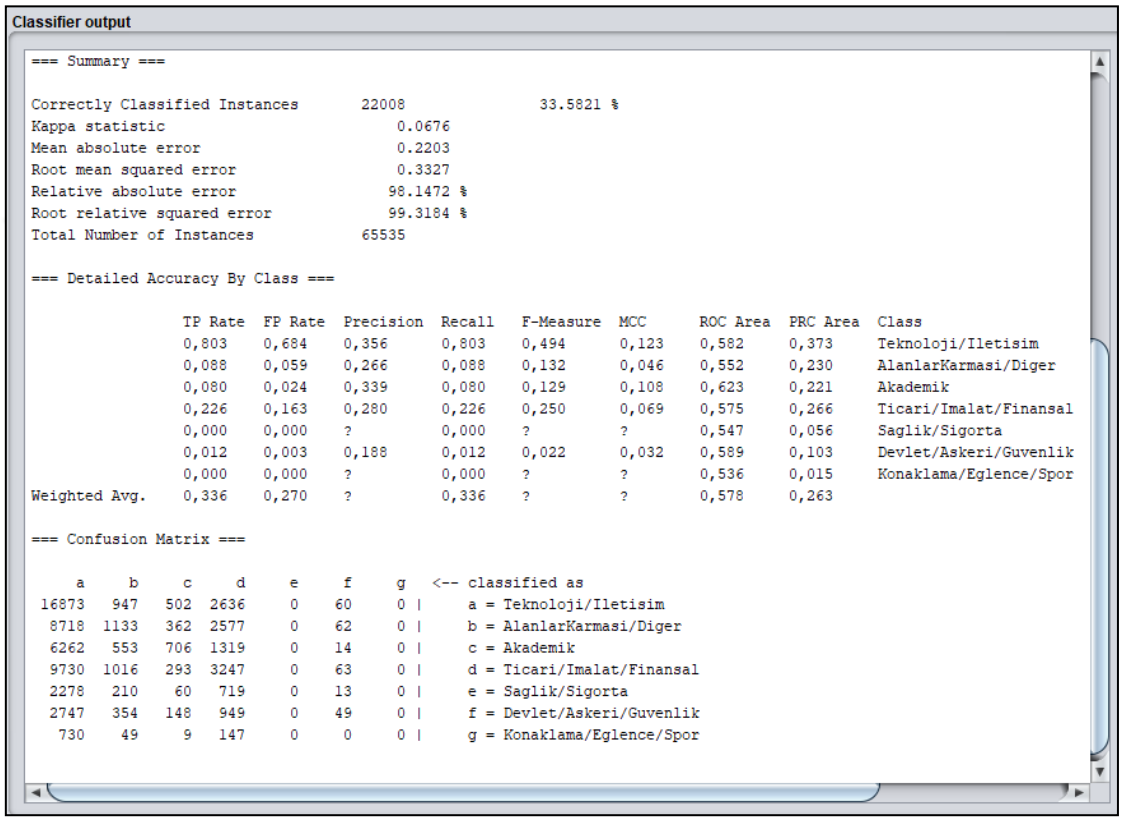
İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,6191 başarı oranı elde edilmiştir. Şekil 8.10'da işlenmiş veri setinin KNN Algoritması, k=11 ve "Percentage split:%66" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Şekil 8.10. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (4).

8.7. EK 7: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=13, CROSS-VALİDATİON:FOLDS 10)

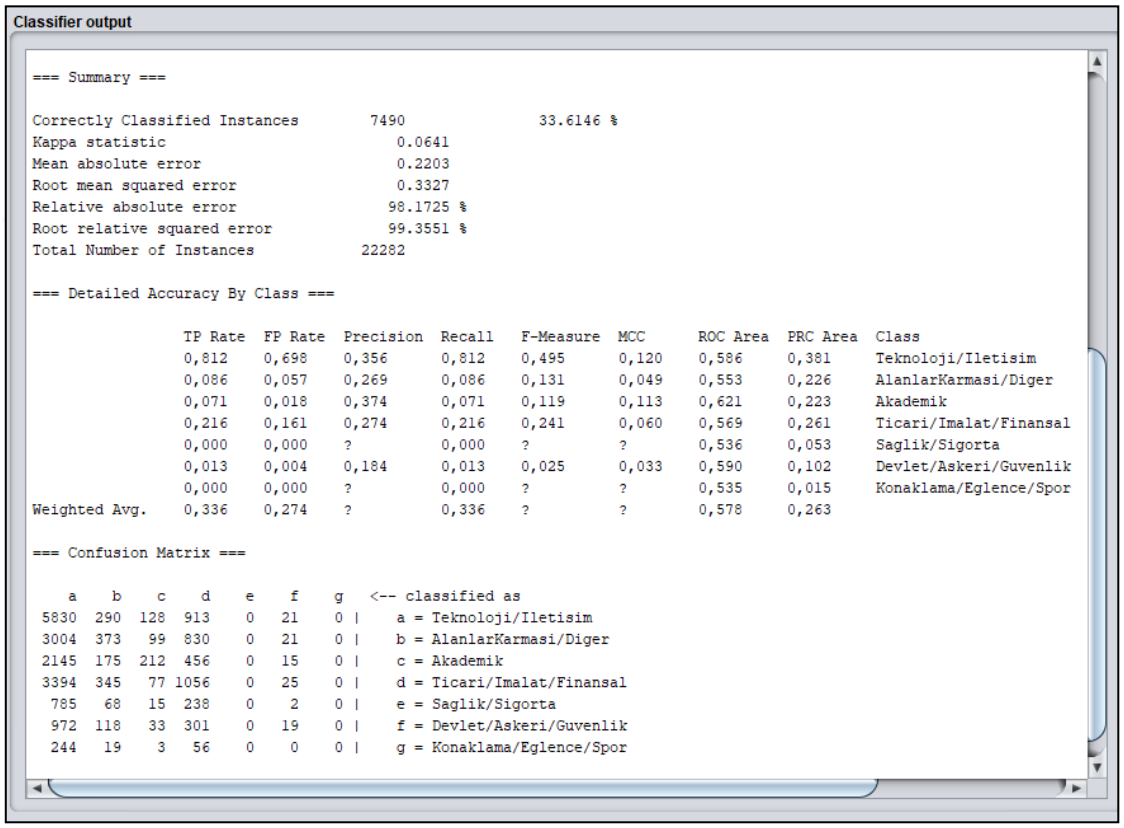
İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,5821 başarı oranı elde edilmiştir. Şekil 8.11'de işlenmiş veri setinin KNN Algoritması, k=13 ve "Cross-validation: Folds 10" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Şekil 8.11. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (5).

8.8. EK 8: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=13, PERCENTAGE SPLİT:%66)

İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,6146 başarı oranı elde edilmiştir. Şekil 8.12'de işlenmiş veri setinin KNN Algoritması, k=13 ve "Percentage split:%66" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Classifier output

```
=== Summary ===

Correctly Classified Instances        7490               33.6146 %
Kappa statistic                          0.0641
Mean absolute error                      0.2203
Root mean squared error                  0.3327
Relative absolute error                 98.1725 %
Root relative squared error             99.3551 %
Total Number of Instances            22282

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,812     0,698     0,356       0,812   0,495      0,120    0,586     0,381     Teknoloji/Iletisim
               0,086     0,057     0,269       0,086   0,131      0,049    0,553     0,226     AlanlarKarmasi/Diger
               0,071     0,018     0,374       0,071   0,119      0,113    0,621     0,223     Akademik
               0,216     0,161     0,274       0,216   0,241      0,060    0,569     0,261     Ticari/Imalat/Finansal
               0,000     0,000     ?           0,000   ?          ?        0,536     0,053     Saglik/Sigorta
               0,013     0,004     0,184       0,013   0,025      0,033    0,590     0,102     Devlet/Askeri/Guvenlik
               0,000     0,000     ?           0,000   ?          ?        0,535     0,015     Konaklama/Eglence/Spor
Weighted Avg.  0,336     0,274     ?           0,336   ?          ?        0,578     0,263

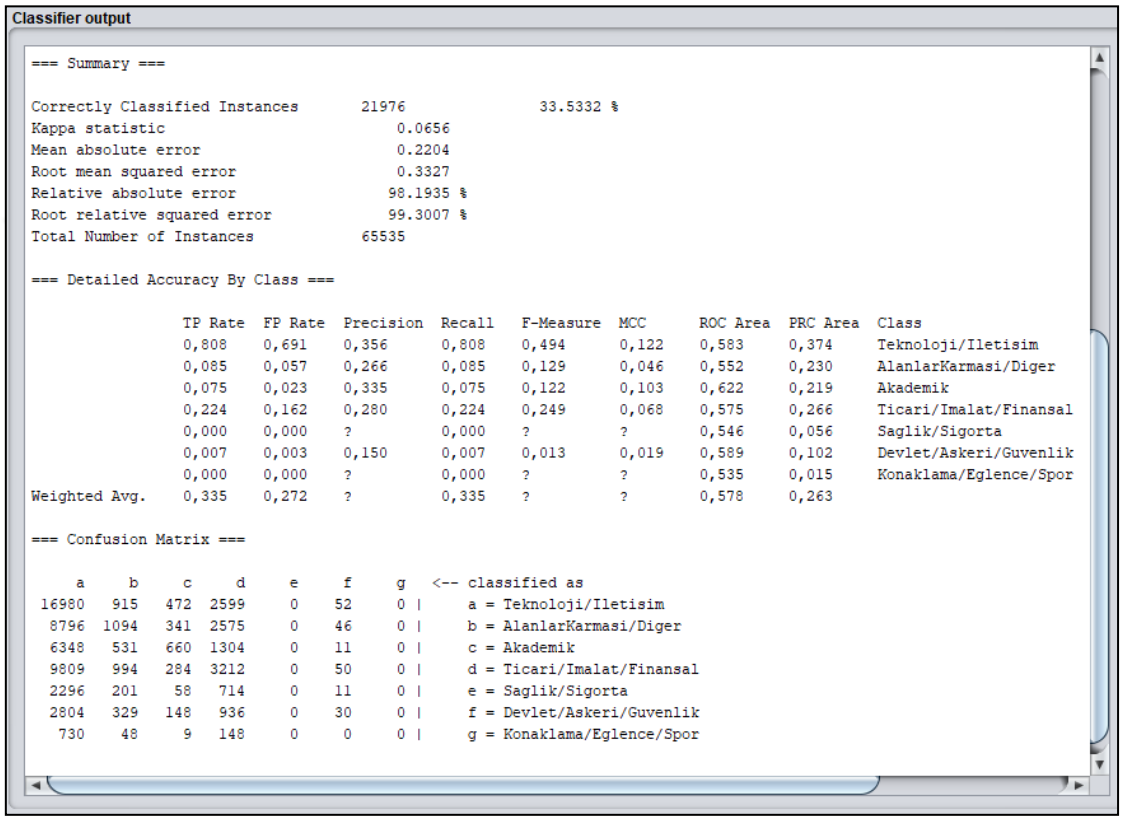
=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
 5830  290  128  913    0   21    0 |    a = Teknoloji/Iletisim
 3004  373   99  830    0   21    0 |    b = AlanlarKarmasi/Diger
 2145  175  212  456    0   15    0 |    c = Akademik
 3394  345   77 1056    0   25    0 |    d = Ticari/Imalat/Finansal
  785   68   15  238    0    2    0 |    e = Saglik/Sigorta
  972  118   33  301    0   19    0 |    f = Devlet/Askeri/Guvenlik
  244   19    3   56    0    0    0 |    g = Konaklama/Eglence/Spor
```

Şekil 8.12. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (6).

8.9. EK 9: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=17, CROSS-VALİDATİON:FOLDS 10)

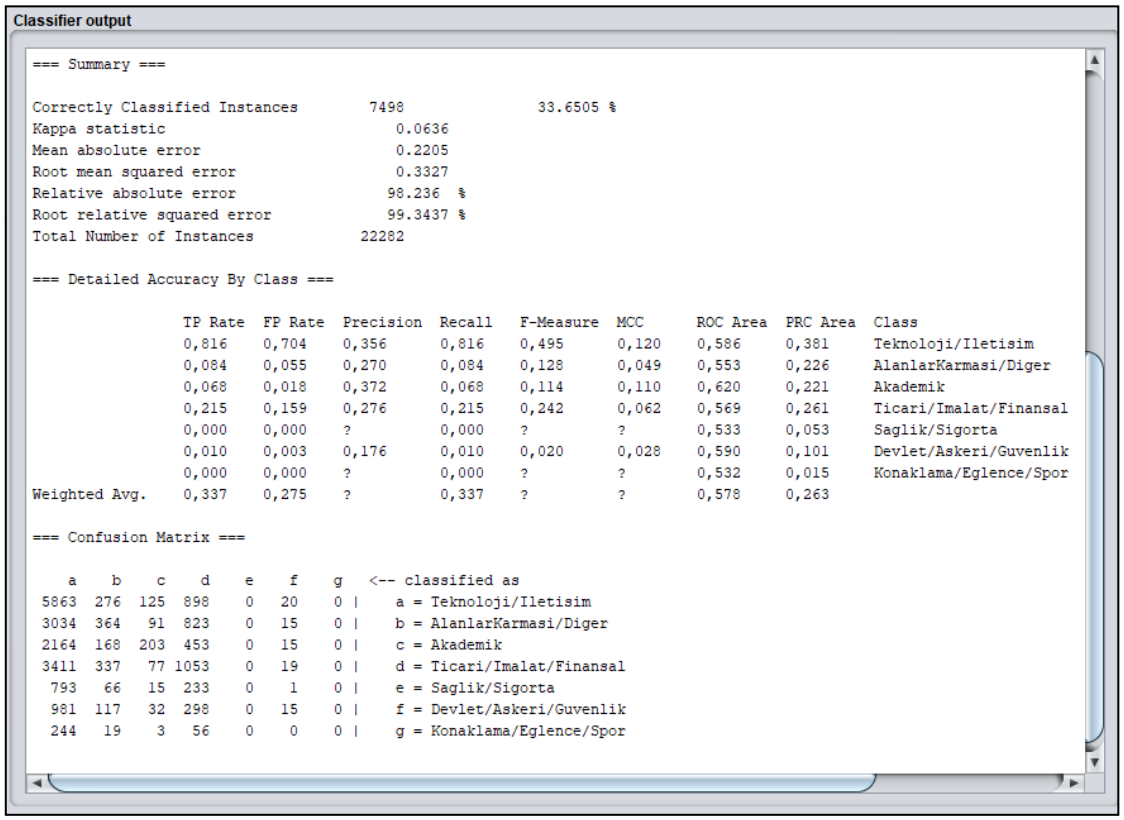
İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,5332 başarı oranı elde edilmiştir. Şekil 8.13'te işlenmiş veri setinin KNN Algoritması, k=17 ve "Cross-validation:Folds 10" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Şekil 8.13. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (7).

8.10. EK 10: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=17, PERCENTAGE SPLİT:%66)

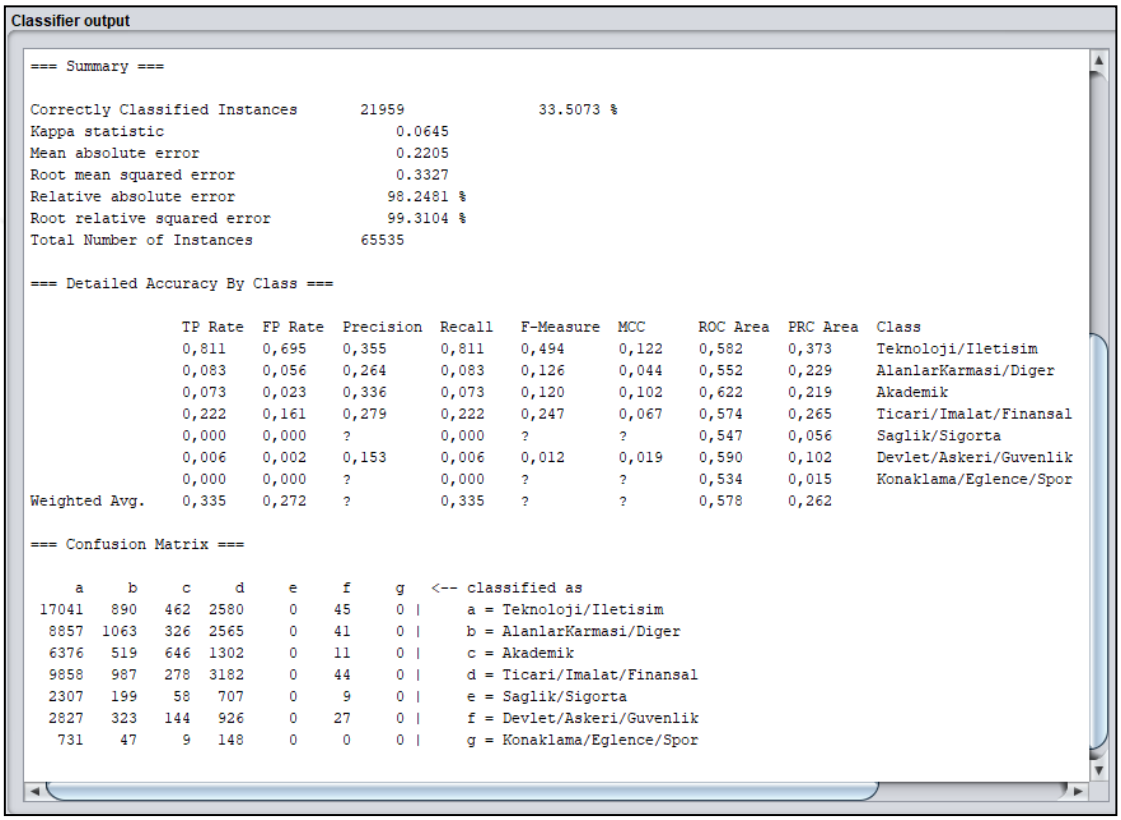
İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,6505 başarı oranı elde edilmiştir. Şekil 8.14'te işlenmiş veri setinin KNN Algoritması, k=17 ve "Percentage split:%66" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Şekil 8.14. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (8).

8.11. EK 11: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=21, CROSS-VALİDATİON:FOLDS 10)

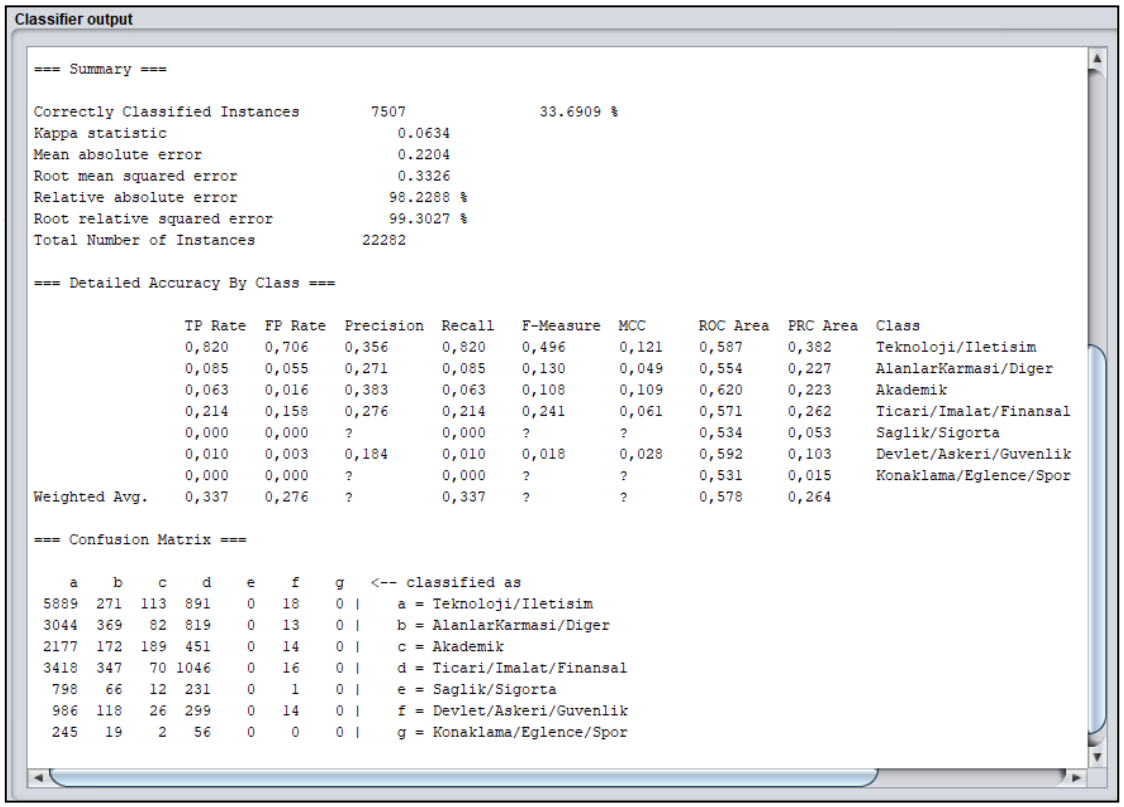
İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,5073 başarı oranı elde edilmiştir. Şekil 8.15'te işlenmiş veri setinin KNN Algoritması, k=21 ve "Cross-validation:Folds 10" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Şekil 8.15. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (9).

8.12. EK 12: KNN ALGORİTMASI KULLANILARAK İŞLENMİŞ VERİ SETİNİN MODELLENMESİ (K=21, PERCENTAGE SPLİT:%66)

İşlenmiş veri setinin KNN Algoritması ile modellenmesi sonucu %33,6909 başarı oranı elde edilmiştir. Şekil 8.16'da işlenmiş veri setinin KNN Algoritması, k=21 ve "Percentage split:%66" seçenekleri ile başarı oranı hesabının Weka ortamındaki sonuç çıktısı ifade edilmektedir.



Classifier output

```
=== Summary ===

Correctly Classified Instances        7507               33.6909 %
Kappa statistic                          0.0634
Mean absolute error                      0.2204
Root mean squared error                  0.3326
Relative absolute error                 98.2288 %
Root relative squared error             99.3027 %
Total Number of Instances            22282

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,820    0,706    0,356      0,820    0,496      0,121    0,587     0,382     Teknoloji/Iletisim
               0,085    0,055    0,271      0,085    0,130      0,049    0,554     0,227     AlanlarKarmasi/Diger
               0,063    0,016    0,383      0,063    0,108      0,109    0,620     0,223     Akademik
               0,214    0,158    0,276      0,214    0,241      0,061    0,571     0,262     Ticari/Imalat/Finansal
               0,000    0,000    ?          0,000    ?          ?        0,534     0,053     Saglik/Sigorta
               0,010    0,003    0,184      0,010    0,018      0,028    0,592     0,103     Devlet/Askeri/Guvenlik
               0,000    0,000    ?          0,000    ?          ?        0,531     0,015     Konaklama/Eglence/Spor
Weighted Avg.  0,337    0,276    ?          0,337    ?          ?        0,578     0,264

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
 5889  271  113  891    0   18    0 |    a = Teknoloji/Iletisim
 3044  369   82  819    0   13    0 |    b = AlanlarKarmasi/Diger
 2177  172  189  451    0   14    0 |    c = Akademik
 3418  347   70 1046    0   16    0 |    d = Ticari/Imalat/Finansal
  798   66   12  231    0    1    0 |    e = Saglik/Sigorta
  986  118   26  299    0   14    0 |    f = Devlet/Askeri/Guvenlik
  245   19    2   56    0    0    0 |    g = Konaklama/Eglence/Spor
```

Şekil 8.16. KNN Algoritması Başarı Oranı Weka Ekran Çıktısı (10).

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı :Elvan Kübra DOĞAN

Yabancı Dili :İngilizce

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Müh. (Tezli)	Düzce Üniversitesi	2021
Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2018
Lise		Hacı Fahri Zümbül Anadolu Lisesi	2014

YAYINLAR

Veri Madenciliği Yöntemleri İle İşveren Sektörünün Sınıflandırılması, *Uluslararası Araştırma-Geliştirme ve Tasarım Konferansı,* 15-18 Aralık 2021.