ÖZNİTELİK SEÇME VE MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE EĞİTMEN PERFORMANSININ TAHMİN EDİLMESİ

Yüksek Lisans Tezi

Fatih ÇİFÇİ

Eskişehir, 2018

ÖZNİTELİK SEÇME VE MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE EĞİTMEN PERFORMANSININ TAHMİN EDİLMESİ

Fatih ÇİFÇİ

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Cihan KALELİ

Eskişehir
Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Mart, 2018

JÜRİ VE ENSTİTÜ ONAYI

Fatih ÇİFÇİ'nin "Öznitelik Seçme ve Makine Öğrenmesi Yöntemleriyle Eğitmen Performansının Tahmin Edilmesi" başlıklı tezi 09/03/2018 tarihinde aşağıdaki Jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim- Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi kabul edilmiştir.

	Ünvanı - Adı Soyadı	**İmza**
Üye (Tez Danışmanı) :	**Doç. Dr. Cihan KALELİ**	**.....**
Üye :	**Doç. Dr. Serkan GÜNAL**	**.....**
Üye :	**Yrd. Doç. Dr. Mehmet KOÇ**	**.....**

.....

Enstitü Müdürü

ÖZET

NİTELİK SEÇME İLE BİRLEŞTİRİLMİŞ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE EĞİTMEN PERFORMANSININ TAHMİN EDİLMESİ

Fatih ÇİFÇİ

Bilgisayar Mühendisliği Anabilim Dalı

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Mart, 2018

Danışman: Doç. Dr. Cihan KALELİ

Günümüzde hayatın her sektöründe işlenen veri miktarının parabolik olarak artması, veri madenciliğin gitgide daha popüler hale gelmesine yol açmış ve yüksek miktarda verinin artan bir karmaşıklıkta işlenmesi ihtiyacı doğmuştur. Finanstan, sağlığa, savunmadan eğitime onlarca sektörün sorunlarını çözmek adına gün geçtikçe farklı yöntemler geliştirilmekte, sosyal, ekonomik, bilimsel birçok problemin çözümü adına veri madenciliğine başvurulmaktadır. Eğitilen ve eğiten sayısının gün geçtikçe arttığı eğitim sektöründe ise, sistemin başarısının geliştirilebilmesi için, gerek eğitilen gerekse eğitimcilerinin performanslarının takip edilmesi ve kıymetlendirilmesi ihtiyacı, Eğitimsel Veri Madenciliği (EVM) kavramını doğurmuştur. Bu alanda yapılan çalışmalar genel olarak, öğrenci performansı konularına yoğunlaştığından, eğitmen performansı konusunda daha çok çalışmaya ihtiyaç duyulmaktadır. Bu çalışmada Gazi Üniversitesi öğrencilerinin eğitmenleri hakkındaki değerlendirmelerini içeren bir veri seti üzerinde çalışılmış, çeşitli öznitelik indirgeme algoritmaları ve farklı makine öğrenme algoritmalarıyla eğitmenlerin performansları tahmin edilmiştir. Öznitelik indirgeme algoritmaları arasında en iyi sonucu Genetik Algoritma vermiş ve bu sayede daha az öznitelik kullanarak Tahmin Doğruluğu Performansı (TDP) arttırılmıştır. Kullanılan sınıflandırma algoritmaları arasında ise en doğru tahmin oranına Derin Öğrenme algoritması ulaşmıştır. Bu çalışmayı diğerlerinden farklı kılan özelliği ise, öznitelik indirgeme ve makine öğrenmesinin farklı kombinasyonlarını, işlem maliyetleriyle kıyaslayarak ortaya koymasıdır.

Anahtar Kelimeler: Öznitelik İndirgeme, Eğitimsel Veri Madenciliği, Derin Öğrenme, Karar Ağacı, K-En Yakın Komşuluk.

ABSTRACT

PREDICTING INSTRUCTOR PERFORMANCE BY FEATURE SELECTION AND MACHINE LEARNING METHODS

Fatih ÇİFÇİ
Department of Computer Engineering
Anadolu University, Graduate School of Sciences, March, 2018
Supervisor: Assoc. Prof. Dr. Cihan KALELİ

Today, parabolically increasing amount of data at all parts of life, make data mining more popular and high amount of data in increasing complexity demanded to acquist. Different methods developed day by day, for solving problems at many sectors like finance, health, defence, education etc., applied to data mining for many social, economical, scientifical problems. In education area, which both number of instructor and students always increase, for enhancing system performance, it is needed to observe and evaluate performance of students and instructors and this situation caused to born a new concept: Educational Data Mining (EDM).

Researches on this area generally focused on student performance. So, it is need to do more researches about instructor performance. A likert type questionnaire dataset which is about opinions of the Gazi University's student regarding their instructor's teaching performance is used in this research and different feature reduction, machine learning algorithms are used for evaluating the data set and performances of instructors.

Among attribute reduction algorithms that we used, Genetic Algortihm gave the best result. So, prediction performance is being increased via using less number of feature. Deep Learning algorithm gave the best performance among the classification algorithms we used.

The distinctiveness of this research is, applying different combinations of feature selection and maching learning with comparing costs.

Keywords: Attribute Reduction, Feature Selection Algorithms, Educational Data Mining, Deep Learning, Decision Tree, K-NN.

TEŞEKKÜR

Bu çalışmanın hazırlanmasında her türlü desteği sağlayan başta değerli hocam Doç. Dr. Cihan KALELİ olmak üzere, Anadolu Üniversitesi Bilgisayar Mühendisliği Bölümünde üzerimde çok değerli katkıları olan tüm hocalarıma ve kıymetli hocam Prof. Dr. Yaşar HOŞCAN'a ufuk açıcı destekleri ve eğitimleri için teşekkür ederim.

Kendisi başlı başına bir çalışma olan Hayat'ın her safhasında bana destek olan değerli eşim Özlem'e teşekkür ederim.

Fatih ÇİFÇİ

Mart, 2018

09/03/2018

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmanın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilemeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı" ile tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

Fatih ÇİFÇİ

İÇİNDEKİLER

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

KISALTMALAR DİZİNİ

ÖS : Öznitelik seçme
Bİ : Boyutsal İndirgeme
FSK : F-Skor Kriterlendirme
OBA : Ortak Bilgi Algoritması
YSA : Yapay Sinir Ağları
BB : Bulut Bilişim
İYS : İleri Yönlü Seçim
GYE : Geri Yönlü Eliminasyon
KA : Karar Ağacı
BK : Bilgi Kazancı
ST : Spekülatif Tur
MMAP : Maksimum Mutlak Azalma Parametresi
K-EYK : K-En Yakın Komşuluk algoritması
GA : Genetik Algoritma
ÇLR : Çokdeğişkenli Lineer Regresyon
SRA : Sınıflandırma ve Regresyon Ağaçları
EDA : Entropiye Dayalı Algoritmalar
OEAF : Olasılıksal Eğim Azalma Fonksiyonu
AÖA : Adaptif Öğrenme Oranı
SGÇO : Sınıf geri çağırma oranları
TDP : Tahmin Doğruluğu Performansı
ZM : Zıtlık Matrisi

1. GİRİŞ

Eğitim, her toplumda olmazsa olmaz, kaybedilemeyecek ve ihmal edilemeyecek en temel olgu ve kurallardan biri olarak görülmüş, yazının icadından günümüze gelene dek yaşanan binlere yıllık süreçte milyonlarca format değişikliğine uğramış, bütün bu süreçlere rağmen milletleri yaşatan temel dinamik olma özelliğini ise hiç kaybetmemiştir. Tüm dünyada eğitim sistemlerinin giderek büyümesi, eğitim kalitesinin toplumu her yönden etkilemesi ve nitelikli toplum yetiştirmenin ancak nitelikli eğitimle mümkün olabileceği gerçeği, başta öğrenci ve eğitmen olmak üzere sistemin tüm bileşenleri tarafından kabul edilmektedir. Nitelikli eğitime duyulan ihtiyaç çok eski dönemlerde fark edilmesi karşın, sürekli nüfus artışı, artan karmaşıklık, sürekli değişen sosyo-psikolojik şartlar eğitim sistemin iyileştirilmesini gün geçtikçe daha zor hale getirmiş ve artık günümüze gelindiğinde ise, tüm alanlarda olduğu bu alanda da iyileşme sağlanabilmesi adına teknolojik yeniliklerden faydalanılmıştır. Eğitim sistemlerini iyileştirmede yaşanan sorunların çözümü adına son 50 yıldır kullanılan basit anket ve örneklemler ise, günümüzde artık etkin bir kıymetlendirmeden uzak kalmıştır. Eğitim eko sistemin yüksek hızda büyümesi, yüksek hız ve tutarlılıkta çözüm ihtiyacı doğurmuş ve bu durum özellikle yükseköğrenimdeki eğitim yöneticileri için önemli bir problem haline gelmiştir [1].

Son zamanlarda büyük önem kazanmaya başlayan veri madenciliği, büyük verileri işleyerek anlamlandırmada büyük başarı elde ederek; başta finans, sağlık, haberleşme ve eğitim olmak birçok sektörde büyük problemleri çözmekte etkin tahminlerde bulunmaktadır [2]. Veri madenciliği uygulamalarının eğitim alanında uygulanmasına yönelik son dönemde yapılan çalışmalar, bu alandaki geniş boşluğu doldurma adına büyük katkı sağlamış, Eğitim Veri Madenciliği (EVM) kavramının doğmasına yol açmıştır. Bu çalışmalar ile eğitim sistemlerinin etki, kalite ve verimliliğini arttırmak hedeflenmiştir [3]. Ayrıca EVM, başta öğrenciler olmak üzere eğitmenler, dersler, eğitim ortamı, eğitim yöneticileri hakkında olmak üzere birçok ilave aydınlatıcı fikir oluşturulabilmesine olanak sağlamıştır [4]. EVM üzerine yapılan araştırmalar genel olarak öğrencilerin başarıları üzerine odaklanmış, eğitim sistemi başarısını tespitte tek kaynağın öğrenciler olduğu vurgulanmıştır. Bu tespitler yapılırken de öğrencilerden alınan dönütlerin yeterince objektif olamayacağı kabul edilmiş, ancak eğitimle ilgili büyük miktarda ve çeşitteki verilerin yeterli miktarda ancak öğrencilerden

alınabildiği belirtilmiştir [5-7]. Bu nedenle eğitim sisteminin başarısı ile öğrencilerin başarısı arasında doğrusal bir ilişki kurularak, eğitim sistemi başarısı, öğrenci başarısı üzerinden değerlendirmiştir [8].

Bu tez çalışmasında, EVM alanında bir üzerinde sınırlı sayıda araştırma [9-17] yapılmış bir konudan, eğitmen performansının değerlendirilmesinden bahsedeceğiz. Aslında, iyi bir eğitimcinin her seviyeden öğrenciyi belirli ölçeklerde mutlaka başarıya ulaştırabileceği gerçeği eğitim uzmanları tarafından genel bir kabul olarak görülmekte ve birçok öğrenci tarafından, öğretilen dersten çok öğreten kişinin (eğitmenin) başarıyı etkilediğini bildirmesi [18], eğitim sistemlerinde eğitmenlerin ne denli büyük bir öneme sahip olduğunu göstermektedir.

Bu çalışmada, UC (University of California) Irvine Machine Learning Repository'de bulunan, Gazi Üniversitesinde yapılmış, öğrencilerin eğitmenlerinin eğitme performanslarını 33 öznitelik üzerinden değerlendirdiği Likert ölçekli bir veri seti kullanılmıştır. Bu veri seti üzerinde öznitelik indirgemesi ve çeşitli makine öğrenmesi algoritmaları kullanılarak eğitmen performans tahmini yapılmış ve başarılı sonuçlar elde edilmiş, EVM alanında gelecekte de yapılabilecek çalışmalara ışık tutmak amaçlanmıştır.

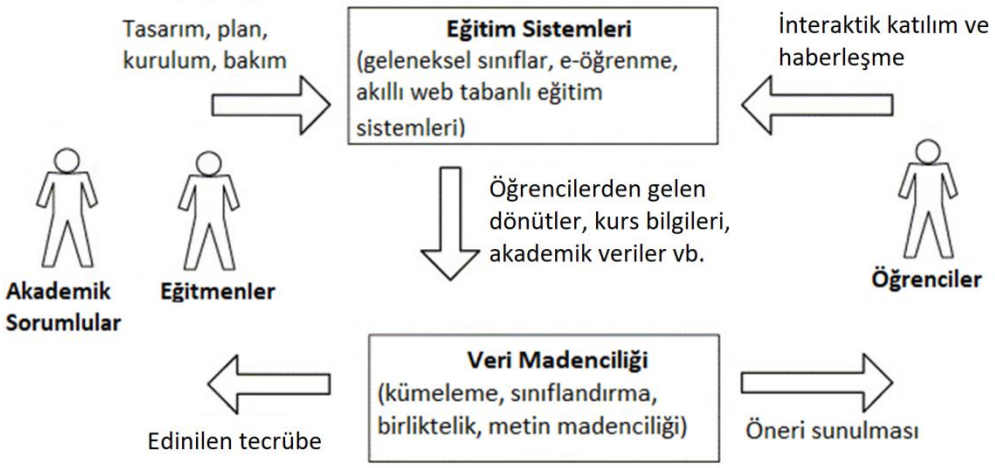
Bu tezdeki bölümler şu şekilde devam etmektedir: İkinci bölümde EVM alanında eğitmen performansı tahminine dair daha önce yapılmış olan çalışmalar incelenmiş, üçüncü ve dördüncü bölümlerde sırasıyla çalışmada kullanılan öznitelik seçme ve makine öğrenmesi yöntemlerinde bahsedilmiştir. Beşinci bölümde ise öznitelik seçme ve makine öğrenme yöntemlerini birleştirdiğimiz hibrid modelden bahsedilmiş, altıncı bölümde yaptığımız modellemelerin deneysel sonuçları gösterilmiş, yedinci ve son bölümde ise, tezimizdeki bulguların genel değerlendirmesi yapılarak gelecekteki çalışmalar için önemli çıkarımlar yapılmıştır.

2. İLGİLİ ÇALIŞMALAR

Daha önce yapılan çalışmalara ilişkin envanter taraması yapılırken birinci aşama olarak, öznitelik seçme ve çeşitli makine öğrenmesini yöntemlerini birleştirerek tahmin yapan çeşitli konulardaki çalışmalar incelenmiş ve bunların farklılıkları tespit edilmiştir. İkinci aşama olarak EVM alanında öznitelik seçme ile birleştirilmiş makine öğrenmesi bazlı tahmin çalışmaları incelenmiş ve bunların genel olarak öğrenci performansı üzerine yoğunlaştığı, ancak az sayıdaki çalışmanın eğitmen performansı üzerinde anılan hibrid yaklaşımı kullandığı görülmüştür.

2.1. EVM Alanında Yapılmış Genel Çalışmalar:

Eğitim sistemleri ekosistemi, içinde eğitim ortamı, eğitim materyalleri, eğitmenler, eğitim yöneticileri, eğitim kıymetlendirme sistemleri olmak üzere onlarca bileşeni içeren ve karmaşıklığı giderek artan sistemlerdir. Şekil 2.1 de görülen [19], eğitim sisteminin genel içindeki bileşenleri arasındaki etkileşim yoğunluğu, EVM alanındaki artan ihtiyacı gözler önüne sermektedir. EVM alanında bugüne de yapılan çalışmalar da bu ihtiyacın belirli oranda giderilmesi ve sistemin iyileştirilmesini hedeflemiştir.



Şekil 2.1 *Eğitim Sistemi bileşenlerinin etkileşimi*

1995-2005 yılları arasındaki veri madenciliğinin eğitim alanındaki uygulamalarına yönelik bir inceleme Romero ve Ventura [19] tarafından yapılmış,

geleneksel eğitim sistemlerine veri madenciliği uygulanmasından, özel web tabanlı kurslara birçok konu incelenmiş, bunların ortak noktasının ise veri ön işleme işlemleri sonrasında veri madenciliği tekniklerinin uygulandığı vurgulanmıştır.

Minaei ve Bidgolim [20] öğrencilerin yılsonu notlarını tahmin edebilmek için ilk kez Genetik Algoritma (GA) tabanlı sınıflandırma yapmışlar, web tabanlı bir eğitim sisteminin kayıtları üzerindeki öznitelik vektörlerini GA ile optimize ederek Tahmin Doğruluğu Performansını(TDP) arttırmış ve birleşik sınıflandırma yöntemleri kullanmanın performansı arttırdığını tespit etmişlerdir.

Öğrencilerin akademik yıl içindeki performanslarını 3 kademeli olarak derecelendirilerek önceden tahmin etmeyi amaçlayan Superby ve ark. [21] tarafından yapılan çalışmada, Karar Ağacı (KA) ve Yapay Sinir Ağları (YSA), Diskriminat Analizi (DA) olmak üzere çeşitli veri madenciliği yöntemleri kullanılarak sınıflandırma yapılmış, bunlar arasından KA yöntemi ile en başarılı şekilde öğrencilerin dereceleri tahmin edilmiştir.

Cortez ve Silva [22] tarafından yapılan çalışmada Portekiz'in Alentejo bölgesindeki ortaokul öğrencilerinin matematik ve Portekizcelerinin neden zayıf oldukları konusuna yoğunlaşılmış, kullanılan dört veri seti için KA ve YSA'nın her ikisinin de %72 lik bir tahmin değeri yakaladığı belirtilmiştir.

Ionian Üniversitesi Enformatik Yüksek Lisans öğrencilerinin yılsonu notlarının en verimli şekilde tahmin edildiği Koutina ve Kermanidis [23] tarafından yapılan çalışmada ise, az sayıda katılımcı verileri kullanılarak yapılan tahmin işlemlerinde Naïve Bayesian ve K-En Yakın Komşuluk algoritmalarının en başarılı sonuçları verdiği vurgulanmıştır.

Natek ve Zwilling'in [24] çalışmasında Microsoft Excel Table Tools with Data Mining add-in ve MS SQL Server yazılımları ile Weka yazılımları kullanılarak küçük bir veri seti üzerinde veri madenciliği teknikleri uygulanmış, bu iki veri madenciliği yazılımının öğrenci başarısını tahmin etmedeki performansları karşılaştırılmıştır.

Veri madenciliği kullanılarak öğrenci başarısını tespit etme alanına yeni bir bakış açısı getiren Sorour ve ark. [25], Latent Dirichlet Allocation (LDA) ve Destek Vektör Makineleri (DVM) algoritmaları ile öğrencilerin her ders sonrası aldıkları eğitime dair çıktılar değerlendirilerek öğrencilerin dönem sonu notları tahmin edilmiş, elde edilen sonuçların Latent Semantik Analiz (LSA)'ya göre daha başarılı olduğu belirtilmiştir.

Öğrencilerden gelen dönütleri tekil yöntemlerle kıymetlendirmenin yeterli olmayacağının belirtildiği Hajizadeh ve Ahmadzadeh'in [26] çalışmasında Sınıflandırma ve Birliktelik Kuralları Çıkarımı (BKÇ) yöntemleri kullanılarak öğrencilerin bazı dersleri tekrar almamalarındaki kök faktörler araştırılmıştır.

YSA kullanılarak bazı özniteliklere göre öğrencilerin hangi dersi kaç kez tekrar ettiği tahmin edilmeye çalışıldığı Oyedotun ve ark. [27] tarafından yapılan çalışmada, öğrenme ortamı, öğrenci-eğitmen arasındaki ilişkiler tahmin edilmeye çalışılmış, bunun için de Radial Tabanlı Fonksiyon Ağları (RTFA) kullanılmıştır.

Zimmerman ve ark. [28], Bilgisayar Bilimleri Yüksek Lisans ve Lisans öğrencilerine ait çok boyutlu veriler üzerinde değişken seçimi ile birleştirilmiş Regresyon modelleri kullanmış, veri setindeki değişkinlerin %54'ünün lisans öğrencisi başarısı üzerinde önemli etkisinin olduğu tespit edilmiş, elde edilen verilen eğitim yöneticilerine yol göstereceği belirtilmiştir.

Başka bir çalışmada, mühendislik öğrencilerinin Üretim Süreçleri Kursundaki başarılarına ilişkin tahminler üretmek üzere, Kentli ve Şahin [29] tarafından Destek Vektör Makineleri (DVM) ve YSA birlikte kullanılarak elde edilen bulgular, Çokdeğişkenli Lineer Regresyon (ÇLR) algoritması ile yapılan tahminler ile kıyaslanmış, en başarılı sonuçları SVM'nin verdiği bildirilmiştir.

EVM'de Birliktelik Kuralları ile çıkarım yapma işlemine dair bir çalışma, Kumar ve Chadha [33] tarafından yapılmıştır. Bu araştırmada, veri analizinde veri ön işlemenin ne denli büyük bir öneme sahip olduğu ve TDP'nın nasıl büyük oranda etkilediğinden bahsedilmiş, Apriori algoritması kullanılarak başarılı sonuçlar elde edildiği anlatılmıştır.

Microsoft Azure bulut platformunda büyük veriler üzerinde Punlumjeak ve ark. [34] tarafından yapılan çalışma, Bulut Bilişim (BB) ve EVM alanlarını kesişten bir çalışma olması bakımından ilgi çekicidir. Tayland Rajamangala Üniversitesi öğrencilerine ait veri setleri kullanılan çalışmada, büyük veri üzerinde sınıflandırmaya geçmeden önce öznitelik indirgeme yapılarak en etkili özniteliklerin seçilmesi ve buna bağlı olarak tahmin performansının iyileştirilmesi amaçlanmıştır. Çalışma sonucunda Ortak Bilgi Algoritması (OBA)'nın öznitelik seçmede kullanılması ve YSA sınıflandırıcısının kullanılmasının %91 lik bir tahmin performansı doğruluğu sağladığı vurgulanmıştır.

EVM alanında son 10 yılda (2007-2017) yapılmış çalışmaların incelenerek kategorize edildiği bir araştırma, Bakhshinategh ve ark. [35] tarafından yapılmış, envanterdeki çalışmalar genel olarak 9 kategoride toplanmıştır: 1-sınıflandırma ve regresyon, 2-kümeleme, 3- birliktelik kuralları, 4- aykırı değer tespiti, 5- sosyal ağ analizi, 6- metin madenciliği, 7- modeller ile çıkarım, 8- ardışık desen madenciliği 9- görselleştirme teknikleri. Çalışmada, öğrenci modelleme sürecine de değinilmiş, son olarak, yapılan çalışmalar hedef kitleleri bakımından ele alınmıştır.

2.2 Eğitmen Performansına Yönelik Yapılmış Çalışmalar

Karar Ağacı (KA) algoritmaları, Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) ve Diskriminant Analizi olmak üzere dört farklı sınıflandırma tekniği kullanılan Ağaoğlu'nun [30] çalışmasında, Marmara Üniversitesi öğrencilerinin eğitmenlerinin eğitim performanslarını 26 adet öznitelik üzerinden değerlendirdikleri bir veri seti kullanılmıştır. Anılan makine öğrenmesi yöntemlerinin sınıflandırma performansları kıyaslanmış, en başarılı sonucu doğruluk, hassasiyet ve özgünlük bakımından C5.0 KA algoritmasının verdiği vurgulanmıştır. Bu çalışmada ayrıca öğrenci verileri üzerinden eğitmen başarısını tahmin etmenin "öğrencilerin derse olan ilgisine" bağlı olduğu çıkarımı yapılmış, elde edilen veriler ile daha gelişmiş veri madenciliği bileşenlerinin oluşturulabileceği bildirilmiştir.

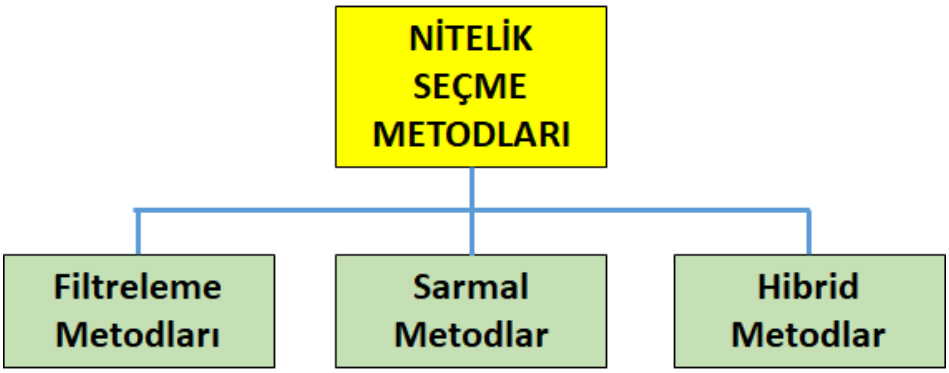
Ahmed ve ark. [31] tarafından yapılan çalışmada, J48 KA, Çok Katmanlı Algılayıcı (ÇKA), Naïve Bayes, ve Ardışık Minimal Optimizasyon (AMO) veri sınıflandırıcılar kullanılarak öğrenci başarısına etki eden faktörler ve eğitmen başarısı tahmini yapılmıştır. Çalışmada, WEKA veri madenciliği yazılımına ait OneR algoritması ile öznitelik seçimi yapılmış ve sınıflandırma işlemi öncesi en anlamlı öznitelikler seçilmiştir. Öznitelik seçme işleminin tahmin performansını tüm sınıflandırma işlemlerinde arttırdığı ve en başarılı tahmini J48 KA algoritmasının verdiği ifade edilmiştir.

Hindistan'da bir üniversitenin mühendislik fakültesinde okuyan öğrencilerin eğitmenleri hakkındaki görüşlerinin temel alındığı Pal ve Pal [32] tarafından yapılan çalışmada Naive Bayes, ID3, Sınıflandırma ve Regresyon Ağaçları (SRA) sınıflandırma algoritmaları kullanılmış, kullanılan algoritmalar içinde en iyi tahmin performansını %80 oranında bir doğrulukla Naïve Bayes sınıflandırıcının verdiği belirtilmiştir.

Sanjay ve Keshav [36] tarafından yapılan çalışmada, eğitmen performansını tahmin etmek için, Naïve Bayes sınıflandırma, C4.5 KA ve YSA kullanılmıştır. 10 öznitelikten oluşan bir anket veri setinin kullanıldığı çalışmada, kullanılan 3 algoritma içinde en iyi TDP'nı C4.5 KA algoritmasının verdiği ifade edilmiştir.

3. KULLANILAN NİTELİK SEÇME YÖNTEMLERİ

Öznitelik seçme (ÖS), temel olarak mevcut veri seti ile ulaşılmak istenen tahmin performansına en yakın değere, -hatta daha iyi tahmin değerlerine- daha az nitelik kullanarak ulaşmayı hedefleyen bir işlemler dizisidir. Daha az nitelik kullanarak daha iyi sonuçlar elde etmek, sonuçların kıymetlendirmesinde işlem maliyetini büyük oranda aşağı çekmekte, sürecin daha hızlı ve başarılı yürütülmesini sağlamaktadır. ÖS işleminin kendisine benzer gibi görünen Boyutsal İndirgeme (Bİ)'den temel farkı ise, Bİ'de birleşik yeni nitelikler oluşturulabilirken, ÖS'de mevcut niteliklerin yapıları korunmaktadır [37].



Şekil 3.1 *Öznitelik Seçme metotları ([38]'den esinlenilmiştir)*

ÖS metotları üç sınıfa ayrılır (Şekil 3.1):

3.1 Filtreleme Metotları: Filtreleme metotları, veri setinde bulunan her özniteliğe bir ağırlık atamak için istatistiksel yöntemler kullanır. Ağırlıklandırma sonrasında, seçilen eşik değerine göre genel sonuca etkisinin önemsiz olduğu değerlendirilen öznitelikler atılarak yeni bir veri seti kümesi elde edilir. Bu işlemler yapılırken özniteliklerin her birisi birbirinden bağımsız olarak ele alınır. Hesaplama bakımından etkin olduğu için özellikle yüksek boyutlu problemlerde tercih edilir. Popüler filtreleme metotları; F-Skor Kriterlendirme (FSK), OBA, Bilgi Kazancı (BK) ve Korelasyon olarak sıralanabilir [39-42]. Çalışmada BK yöntemini kullanacağız.

3.1.1 Bilgi Kazancı (BK):

Bilgi kazancı, entropinin tersi olarak ifade edilebilen ve 0 ile bir arasında değerler alabilen bir çıkarımdır. BK, işlenen veri setindeki her bir niteliğin, yapılan her bir sınıflandırmaya yatkınlığının incelendiği bir ölçüdür. Bir öznitelik, yapılan tüm sınıflandırmalar içinde ne kadar çok sınıftan ne kadar çeşitli değer alıyorsa; BK 1 değerine o kadar yakın bir değer alır. Eğer, bir öznitelik örneklemdeki tüm sınıflara ait farklı değerler taşıyorsa BK=1 ve Entropi=0 olur. Başka bir deyişle, bir öznitelik, sınıflardan genel olarak ne kadar uzaksa, BK o kadar düşük olur. Bir niteliğe ait BK üç aşamada bulunur.

Öncelikle veri setinin genel bilgi değeri bulunur (Eşitlik 3.1):

$$\boldsymbol{Bilgi(D) = -\sum_{i=1}^{m} p_i \,.\, log_2(p_i)} \qquad (\mathbf{3.1})$$

D: Bilgisi hesaplanmak istenen veri setinin satır sayısı

$\boldsymbol{p_i}$: Ele alınan özniteliğe ait bir değerin, tüm veri seti içinde bulunma olasılığı

$\boldsymbol{i}$ **:** Bir özniteliğin alabileceği değer adedi

$\boldsymbol{Bilgi\ (D)}$ **:** Bir özniteliğin aldığı öznitelik değerlerinin dağılımına bağlı olarak, veri setinde bulunan bilginin değeri

2. aşamada bir niteliğin tüm veri setinde bulunan bilgisi bulunur (Eşitlik 3.2):

$$\boldsymbol{Bilgi_A\,(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Bilgi} \qquad (\mathbf{3.2})$$

(D_j) : İşleme alınan örneklemin (parça veri kümesinin) satır sayısı

$Bilgi\ (D_j)$: Satır satısı (D_j) olan örneklemin bilgi değeri

$Bilgi_A\ (D)$: A özniteliğinin tüm veri setinde bulunan bilgisi

Sonuç olarak A niteliği için bilgi kazancı değeri Eşitlik 3.3'te belirtilen BK formülü ile bulunur [43].

$$\boldsymbol{Bilgi\ Kazancı\ (A) = Bilgi\ (D) - Bilgi_A\ (D)} \qquad \mathbf{(3.3)}$$

3.2 Sarmal Metotlar:

Bu metotlar bir öğrenme algoritması içerir. Öznitelik alt kümelerinin seçimi bir arama problemi olarak ele alınır ve seçimler diğer kombinasyonlarla karşılaştırılır. Seçilen özniteliklerin başarısı, bir öğrenme algoritmasında denenmektedir. Sarmal metotlar daha iyi doğruluk üretmelerine karşın, daha yüksek işlem maliyetine sahiptir. Bu metotlara örnek olarak İleri Yönlü Seçim (İYS) ve Geri Yönlü Eliminasyon (GYE) gösterilebilir. Bu çalışmada kullanacağımız GYE'e değinelim.

3.2.1 Geri Yönlü Eliminasyon (GYE):

GYE işleminde amaç, sonuç olarak performansa yüksek oranda etki eden öznitelikler kümesi elde etmektir. Bunun için öncelikle Maksimum Mutlak Azalma Parametresi (MMAP) belirlenir. Bu değer iterasyonlarda karşılaştırma işlemleri için kullanılacak eşik değeridir. GYE işlemine örnek veri setinin tamamı ile başlanır ve her turda öznitelikler azaltılarak devam edilir. Her öznitelik kaldırma işleminde yerel kıymetlendirme öğeleri ile performans değerlendirmesi yapılır ve elde edilen değer MMAP değeri ile kıyaslanır. GYE'nin amacı performansa etkisi büyük (atıldığında büyük performans azalmasına yol açacak) verileri korumak olduğundan, iterasyona (istenmeyen öznitelikleri atmaya) devam etmek için;

1) Elde edilen performans değerinin MMAP değerinden küçük olması

2) Atıldığında performans artışına neden olan bir niteliğin tespiti durumlarının oluşması gerekir. MMAP değerine eşit ya da büyük bir durum tespit edildiğinde iterasyon durdurulur.

Durdurma işlemin kesin sonuç verdiğinden emin olmak için turlara devam edilebilir. Bu durumda atılacak turlara Spekülatif Tur (ST) denir. Belirlenen ST sayısı kadar tur atıldıktan sonra ilk karşılaşılan durma şartında durulur, durma şartı oluşmamışsa karşılaşılana dek iterasyona devam edilir.

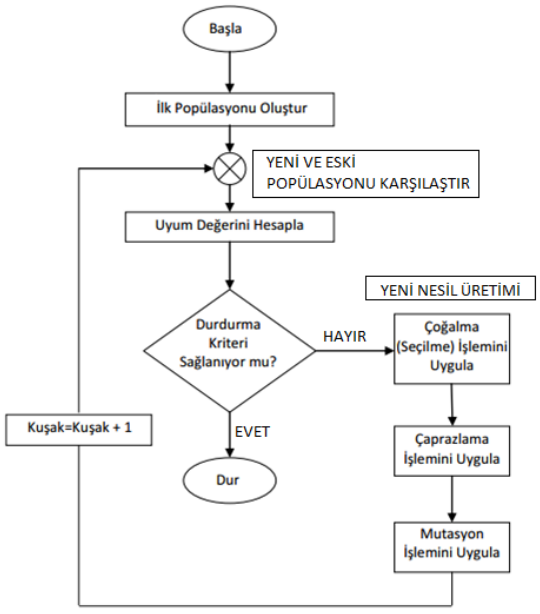
3.3 Hibrid Metodlar:

Filtreleme metotları ve sarmal metotların iyi yönlerini kendinde birleştiren bir yaklaşımdır. Bu metotlar, tahmin doğruluğuna en yüksek katkıyı veren özniteliklerin belirlenmesi amaçlar. Hibrid metotların kullanımıyla hesaplama maliyeti azalır, TDP artar. Genetik Algoritma en bilinen hibrid algoritmadır [38].

3.3.1 Genetik Algoritma:

Genetik algoritma 1970 lerin başlarında ortaya çıkmış, canlılardaki genetik çoğalma ve değişimi temel alan matematiksel ve algoritmik bir yapıdır. İlk kez John Holland [44] tarafından yapılan çalışmada canlılardaki gen aktarımı, çeşitli nedenlerle oluşan kalıtsal değişimler, yeni türlerin ortaya çıkması, seçili genlerin aktarımıyla yeni özelliklerin ortaya çıkması, kaliteli gen seçimi ile üstün bireyler elde edilmesi vb. konulardan ilham alınarak bütün bu süreçlerden *çoklu öğrenme* yapabilen *Genetik Algoritmalar* adlı yeni bir yöntem geliştirilmiştir.

Genetik Algoritmaların (GA) temel amacı, herhangi bir problemin çözümünde, o problem için uygulanabilecek tüm çözümlerin arasından en iyi çözümü seçmektir. GA bunu yaparken gen biliminde olduğu gibi kaliteli nesiller üretmeyi hedefler. Popülasyondan (işleme alınan veri kümesinden) üretilecek her yeni nesilde kötü çözümlerin gitgide yok olması, iyi çözümlerin ise daha iyi çözümler elde etmek üzere yeni nesillerin oluşturulmasında kullanılması beklenir. GA'nın genel işleyişi Şekil 3.2 de görülmektedir:



Şekil 3.2 *Genetik Algoritma Akış Diyagramı ([56]'dan esinlenilmiştir.)*

GA'lar, tüm veri setini taramak yerine bir alt çözüm uzayını tarayarak çözüm elde etmeyi amaçlarlar. Bu yöntem işlem maliyetini azaltır, hızlı ve etkin bir çözüm sağlar. Diğer bir avantajı ise ele aldığı uzayda paralel işlemler yaptığından yerel çözümlere takılmaz. GA'lar genel olarak sonuca etki eden öznitelik sayısının fazla

olduğu durumlarda kullanılırlar ve ne kadar iyi sonuçlar vereceklerini önceden tahmin edebilmek güçtür. GA'lardaki temel kavramları ele alırsak:

Gen: genetik bilgi taşıyan anlamlı en küçük parçacıklara gen adı verilir. Genler, kromozomların yapıtaşıdır. Genler, harflerden oluşabileceği gibi ikili bitlerden (0 ve/veya 1) de oluşabilir.

Kromozom: birden çok genin birleşmesinden oluşan ve çözümü hedeflenen probleme ilişkin tüm bilgileri içeren yapılardır. Kromozomlar, gen biliminde popülasyondaki bireylere karşılık gelmektedir. Amaç, her genetik çoğalmada daha iyi kromozomlar (çözümler) elde etmektir.

Popülasyon: kromozomlardan (bireylerden) oluşan topluluk, başka bir deyişle çözümü hedeflene probleme ait kullanılabilecek alternatif çözüm yöntemleri kümesidir. GA'larda, popülasyondaki birey sayısı sabit tutularak, zayıf bireylerin (kötü çözüm üreten etkenlerin) yok olarak, daha iyi bireylerin yaşamını sürdürmesi amaçlanır. Birey sayısı, araştırmacı tarafından problemin niteliğine göre önceden belirlenir.

Kodlama: Kromozomların gösteriliş biçimidir. Sıfır ve birlerden oluşan *ikili kodlama* (1001110, 10111), çeşitli rakamlardan oluşan (254896, 1547964) *permütasyon kodlama,* çeşitli değerler içerebilen (2.3, AERFT, ileri, geri) *değer kodlama,* birbirini izleyen ve gelişen çeşitli öznitelikteki değerler için *ağaç kodlama,* kodlama türleri olarak sıralanabilir.

3.3.1.1 Seçilim:

Çaprazlama ve *mutasyon* işlemi kullanılarak, mevcut popülasyondan yeni bireyler oluşturulması işlemine *seçilim* denir. Seçilim işlemi, kaliteli bireylerin yaşamasına izin verilerek, bunlardan yeni bireyler elde etmeyi amaçlar. Seçilim işleminde, *uygunluk değeri* yüksek olan birey ile yola devam edilmesi esas alınır. *Rulet, turnuva* ve *sıralı seçilim*, başlıca seçilim çeşitleridir.

Rulet seçimi: Bir bireyin rulet seçimindeki seçilme olasığı kendi uygunluk değerinin, tüm popülasyonun toplam uygunluk değerine bölünmesiyle bulunur.

Sıralı seçim: En kötü uygunluğa sahip kromozoma 1 değeri vererek başlamak üzere, uygunluk değeri arttıkça kromozomlara verilen değer de arttırılarak tüm örneklem uzayındaki bireylere değerler atanır.

Turnuva Seçimi: topluluk içerisinden rastgele seçilen **k** adet birey içinden, en yüksek uygunluk değerine sahip olan bireyin seçilmesi işlemidir.

3.3.1.2 Çaprazlama

Çaprazlama işlemi, aynen canlılarda olduğu gibi, anne ve babada bulunan genlerden karma yeni bir kromozom dizisine sahip çocukların oluşturulmasıdır. Bu işlemde amaç, uygunluk değeri yüksek olan ata kromozomları kullanılarak, daha yüksek uygunluk düzeyine sahip çocuklar (yeni kromozomlar) üretmektir. Tüm çaprazlama işlemleri sonunda, çocukların kromozom sayıları her bir atasında bulunan kromozom sayılarına eşit olur.

Tek Noktalı Çaprazlama: Yeni çocukları oluşturmak üzere kromozomları kullanılacak her iki atanında kromozomları iki parçaya ayrılır. Her bir çocuk, kendi kromozomlarının bir parçasını bir atasından, diğer parçasını ise diğer atasından alır. Ataların tüm kromozomları çocuklar tarafından kullanılmış olur. 1. çocuk, genlerin bir kısmını 1. atasından alırken, aynı hizaya karşılık gelen 2. atanın genleri 2. çocuk tarafından alınır. Bir çocuğun kromozomu oluşurken, her bir atadan birer kez olmak üzere sadece toplam iki kez gen bloğu alınır.

Çift Noktalı Çaprazlama: Atalara ait kromozomlardaki karşılıklı genlerin çocuklar tarafından paylaşımında, çocuklar tarafından karşılıklı gen sırası takip edilir. Örneğin, 1. çocuk, 1. atasının kromozomundan 1. genini alıyorsa, 2. çocuk da 2. atadan 1.geni alır. Sonra, 1. çocuk 2.atadan kromozom almaya geçince, 2. çocuk da 1. atadan gen alır. Bu değişim, kromozomdaki gen sayısı tamamlanana dek sürer.

Sıralı Çaprazlama: Bu çaprazlama işleminde, bir çocuk üretmek için bir atanın kromozomu temel alınıp, 2. atadan bu ataya gen kümesi transfer edilir. 2. atadaki gen kümesi geldiği yerdeki kromozom diziliminde hangi konumda bulunuyorsa, hedefte de aynı konuma yerleştirilir. Yerleştirme öncesinde, kaynaktan gelen genler hedefte de bulunduğundan, yeni gelen genlerin hedefte dağınık halde bulunduğu tüm noktalar işaretlenir ve gen transferi ile yeri işgal edilmiş olan genler (eski konumlarından konuşlandıkları yerdeki soldan sağa dizilişlerine göre) kromozomun en solundan en sağına doğru işaretli yerlere yerleştirilirler. Bir başka deyişle, hedefe giden gen kümesi, hedefteki kendi türdeş genlerini imha etmiş, gittiği bölgedeki genleri de, türdeşlerinden boşalan bölgelere sürgün ettirmiştir.

3.3.1.2 Mutasyon

Mutasyon, çaprazlama ile yeni nesil kromozomlar elde edilmeden, mevcut krozomlarda bulunan gen ya da gen parçalarının değiştirilerek yeni kromozomlar elde edilmesidir. Mutasyon kullanılarak (yerel maksimum) çalışmanın kısır döngüde kalması engellenir ve problemin çözümünü bulmada yön değişikliği sağlanabilir. Rastgele seçilen iki karakterin yer değiştirdiği mutasyon türüne *pozisyona göre mutasyon*, ikinci seçilen karakterin, birinci karakter yerine konmasıyla değişim sağlayan mutasyona da *sıraya göre mutasyon* denir.

3.3.1.3 Genetik Algoritma Performans Faktörleri

Kromozom sayısının azalması, üretilecek yeni nesillerde kalitenin azalmasına, kromozom sayısının artması ise toplam işlem maliyetinin artmasına yol açar. Yeni nesiller üretildikçe kromozomların gitgide birbirine benzemeye başladığı durumlarda ise, müdahale gerekir ve yüksek bir mutasyon oranı verilerek etkin bir çözüm yaklaşılmaya çalışılır.

Bazı çalışmalarda istenen hedefe ulaşabilmek adına çok noktalı çaprazlamaya gereksinim duyulmuş, her çaprazlama sonucunda elde edilen bireyin ise eşzamanlı kullanılmaması istenen kalitede nesiller elde etmede önemli faydalar sağlamıştır. Kullanılacak olan parametrelerin, ele alınan probleme uygun nitelikte (logaritmik/doğrusal) hazırlanması ve tüm sürecin uygun şekilde kıymetlendirilmesi, tüm işlem sürecinin performansına yansıyan önemli işlemlerdir.

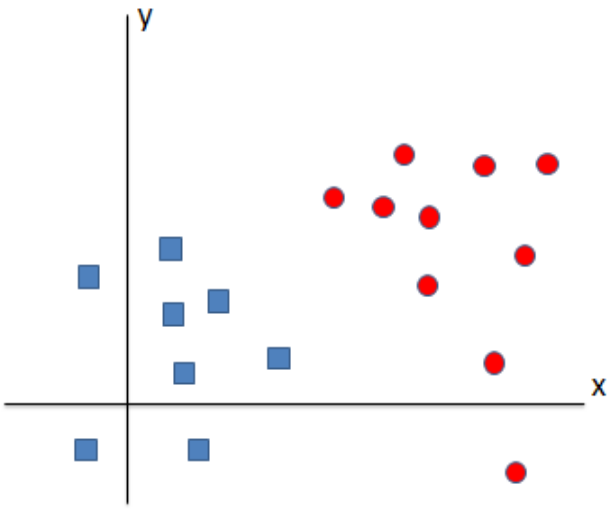
4. KULLANILAN MAKİNE ÖĞRENMESİ METODLARI

Çalışmada, veri setine Bilgi Kazancı, Geri Yönlü Eliminasyon ve Genetik Algoritmayı ayrı ayrı uygulayarak en efektif öznitelik seçme işlemini tamamladıktan sonra bir sonraki aşama olarak, indirgenmiş veri setinde çeşitli makine öğrenmesi yöntemleri kullanarak tahmin yapma işlemine geçilmiştir.

Sırasıyla tahmin performansları kıyaslanacak makine öğrenme yöntemleri, K-En Yakın Komşuluk, Karar Ağacı, Naive Bayes ve Derin Öğrenme'dir.

4.1. K-En Yakın Komşuluk

K-En Yakın Komşuluk (K-EYK) algoritması, önceden sınıflandırma yapılmış bir veri kümesinde, yeni bir elemanı sınıflandırmada kullanılır. Algoritmanın temelinde, sınıflandırılmak istenen öğenin, en yakın mesafede olduğu K adet komşusunun tespiti vardır. Bu K adet en yakın koşunun çoğunluğu hangi sınıfa mensup ise, işleme alınan eleman da o sınıfa dâhil edilir [45].

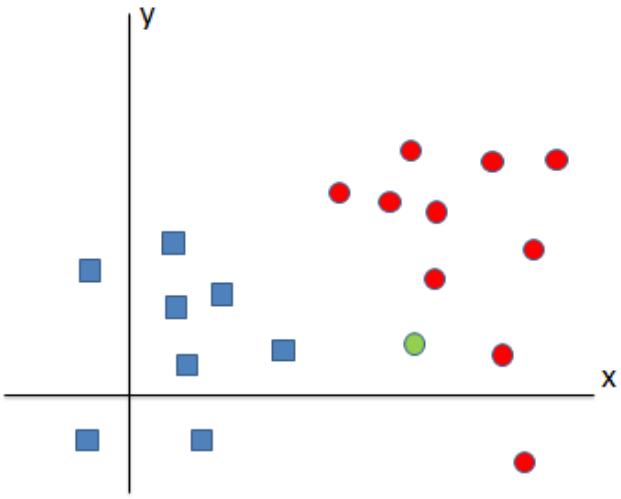


Şekil 4.1 *Önceden sınıflandırılmış Veri Kümesi ([57]'den esinlenilmiştir.)*

Örneğin, Şekil 4.1 de gösterilen ve Doğrusal Ayrıştırma yöntemiyle birbirinden ayrıştırılmış iki boyutlu düzlemde bulunan (sınıflara bölünmüş) veri kümesine Şekil 4.2 de gösterildiği şekilde sınıflandırılmak üzere yeni bir eleman dâhil edilsin. K=3 alındığında, bu elemana en yakın uzaklığa sahip 3 elemanın tespit edilmesi gerekmektedir. Bu uzaklık hesabında Öklid Mesafesi kullanılır (Eşitlik 4.1).

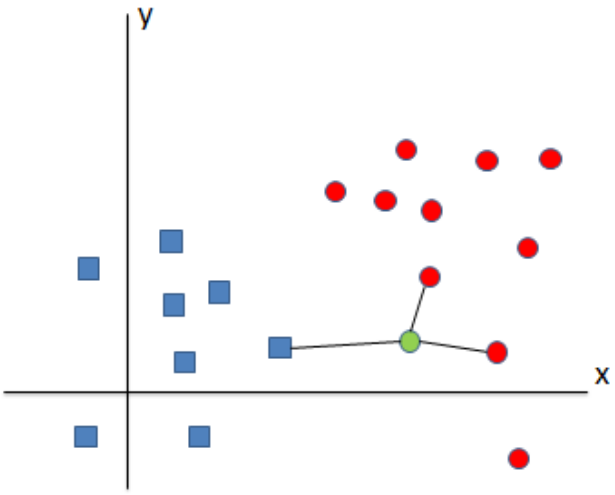
$$\boldsymbol{d(A,B) = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}} \qquad \textbf{(4.1)}$$

A, B: aralarındaki mesafe hesaplanacak olan 2 boyutlu koordinatsal konuma sahip noktalar

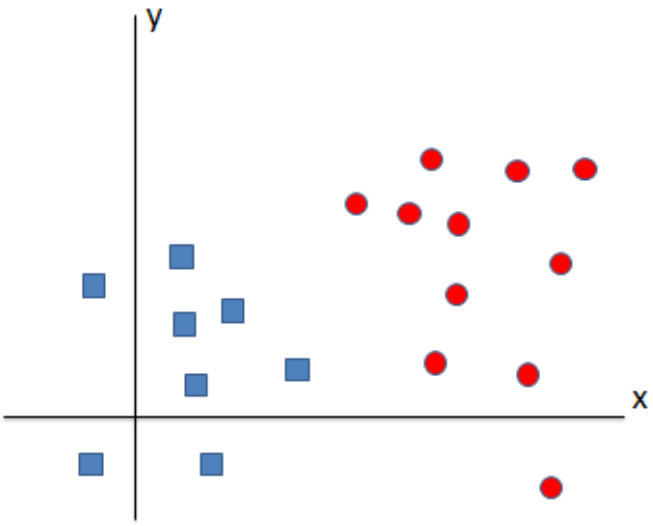


Şekil 4.2 *Sınıflandırılacak yeni elemanın yerleştirilmesi ([57]'den esinlenilmiştir.)*

Yeni elemanın dâhil olacağı sınıfın bulunması için Öklid Uzaklığı kullanılarak en yakın 3 komşusu bulunur **(Şekil 4.3)** ve bu 3 koşudan ikisi (çoğunluk) *yuvarlak* sınıfa mensup olduğundan yeni eleman da *yuvarlak* sınıfa dâhil edilir. **(Şekil 4.4)**



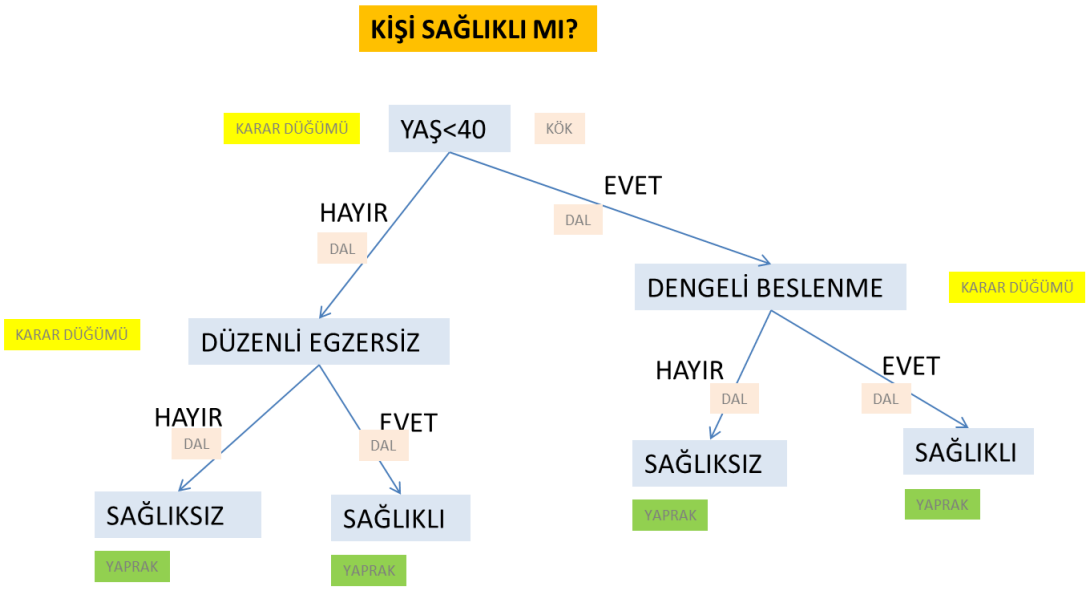
Şekil 4.3 *En yakın K adet komşunun belirlenmesi ([57]'den esinlenilmiştir.)*



Şekil 4.4 *Yeni elemanın sınıfına atanması ([57]'den esinlenilmiştir.)*

4.2. Karar Ağacı

Karar ağaçları (KA), verilerin bir ağaç yapısı şeklinde sürekli olarak kümelendiği ve bölündüğü, tümevarım mantığıyla hareket eden bir makine öğrenmesi yöntemidir. KA, ayrık parametrelerden oluşur ve gürültüye dayanıklı bir yapıya sahiptir. KA ana bileşenler olarak karar düğümleri ve yapraklardan oluşmaktadır. Karar düğümleri, veriler kümelenirken hangi yöne gideceklerini tayin ederken, yapraklar son noktadaki sonucu, bir başka deyişle öğenin dâhil edileceği sınıfı gösterir **(Şekil 4.5)**.



Şekil 4.5 *Karar Ağacı örneği ([58]'den esinlenilmiştir.)*

KA, sınıflandırma işlemlerinde yaygın olarak kullanılır. Başta müşteri kredi notunun hesaplanması olmak üzere, ses/karakter tanıma, kullanıcı davranışı analizi, hastalık teşhisi, web madenciliği vb. birçok alanda kullanılmaktadır ve diğer sınıflandırma metotlarına göre oldukça başarılıdır. KA, akış şemalarında olduğu gibi zincirleme bir yapıya sahip olup, dallar, yapraklar ve kökten oluşur. Dallar, başlangıç düğümünden (kök) karar (yapraklara) giden ve seçimlere göre şekillenen yolları ifade eder.

Sınıflandırma işlemi KA'da iki yöntemle yapılır. Bunlar Sınıflandırma ve Regresyon Ağaçları (SRA –CART) ve Entropiye Dayalı Algoritmalar (EDA) dır. EDA'ya örnek olarak ID3 ve C4.5 algoritmaları ve SRA için örnek olarak da Twoing ve Gini algoritmaları verilebilir.

ID3 algoritması metinsel kategorik (nominal) öznitelikleri Bölüm 3'de belirtilen *entropi* ve *bilgi kazancı* eşitsizliklerini kullanarak sınıflandırma yaparken, C4.5 algoritması buna ilave olarak budama işlemi ve nümerik verilerin sınıflandırmasını da yapabilmektedir. SRA algoritmaları ise her düğümden ikili dallanmalar yaparak tüm örneklemin sınıflandırmasını esas alan algoritmalardır.

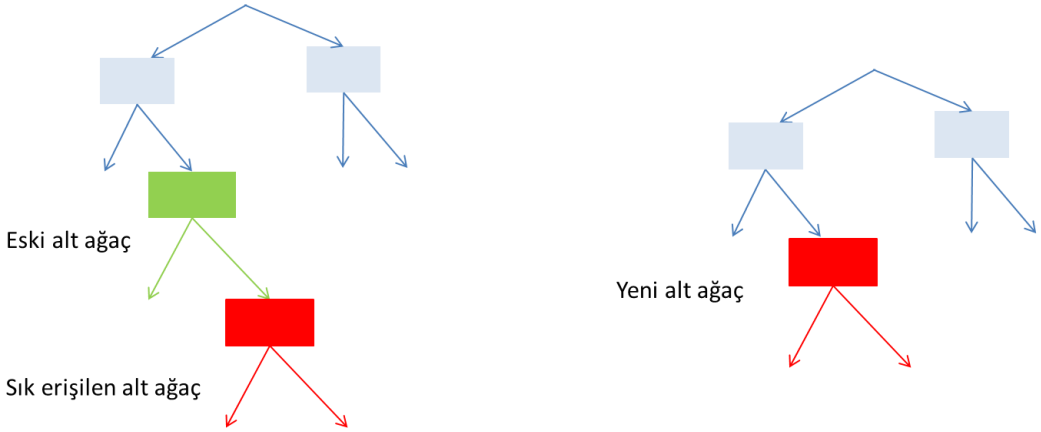
4.2.1. Karar Ağacı Oluşturma Metotları

4.2.1.1. Ağaç Oluşturulması

Karar ağaçları, ters duran bir ağaç şeklinde olduğundan, en tepede kök bulunur ve KA inşa işlemine kökten başlanır. Kök için öznitelik seçiminden sonra, seçilen özniteliklere göre yapılan yeni seçimlerin oluşturduğu dallar ve karşılaşılacak yeni seçimler (düğümler) in eklenmesi, istenen hedefe kadar devam eder. Bir düğüm oluşturulduktan sonra eldeki örneklerin tamamı aynı sınıfa ait durumdaysa o düğüm yaprak olarak sonlandırılır ve o sınıfa ait etiketi alır. Eğer örneklerin tamamı aynı sınıfa ait değilse dallanma devam eder ve düğümü sınıflara en iyi şekilde bölecek diğer niteliğe geçilir. Dallanma işlemi; örneklerin çoğunluğu aynı sınıfa ait olana, örnekleri tekrar bölebilecek başka öznitelik kalmayana ve kalan öznitelikleri temsil eden örnek kalmayana dek sürer. Dallanma işlemi, eğitim verilerinin en iyi şekilde sınıflandırılacağı duruma gelene dek KA dallarının derinleşmesiyle devam eder. Böylece, zincirleme alternatif birçok seçimle elde edilmiş dal, düğüm ve yapraklardan oluşan KA elde edilir.

4.2.1. 2. Ağaç Budama

Hatalı (istenmeyen) verilerin veri setinde bulunması bütün sınıflandırma algoritmalarında sınıflandırma doğruluğuna etki ettiği gibi KA'da da aynı olumsuz etkiye yol açmaktadır. Bunun önlenmesi için öğrenme kümesindeki gürültülerden (hatalı veri) meydana gelmiş dalların KA'dan silinmesi gerekmektedir. Bu işleme budama *(pruning)* denir (Şekil 4.6).



Şekil 4.6 *Ağaç Budama ([58]'den esinlenilmiştir.)*

KA'da sınıflandırma işlemi, dallanmanın artmasıyla ve yapraklara ulaşıncaya kadar devam etmektedir. Veri örnekleminin içinde gürültü olması ise çok büyük bir KA elde edilmesine neden olur. Buna yol açmamak için veriye aşırı uyumun (overfitting) önüne geçilmesi gerekmektedir. Budama işlemi bu yüzden önemlidir. Budama işlemi ile ağaç daha sade bir hale gelir ve tahmin doğruluğu artar.

Budamaya kök ya da yapraktan başlanabilir. Budama işleminin ağaç tamamlandıktan sonra yapılması, ağacın büyümesinin erkenden durmasını engelleyeceğinden, daha doğru bir yaklaşımdır. Budama işleminde yapraktan başlanarak, o yaprağın veri örneklemi içinde en sık bulunan (en popüler) sınıftan olan her bir düğümü ağaçtan kaldırılır. Bu işleme, doğrulukta azalma meydana gelene dek devam edilir.

Çalışmada kullanılan Rapidminer KA modülünde, KA'larda kullanılan birçok özellik birlikte uygulanmış; ayırma kriteri olarak *bilgi kazancı* kullanılmış, karar ağacına hem *budama* hem de *ön budama* uygulanmıştır. Ön budama, maksimum derinlikten daha fazla sayıda durdurma koşulu meydana geldiği durumlarda, KA'nın doğruluk performansını arttırmak için kullanılmıştır.

KA'larının adım adım uygulanabilmesi ve izlenebilmesi mümkündür. KA'dan elde edilen veriler genel olarak herkes tarafından basit açıklamalar yardımıyla anlaşılabilmektedir. Gerçek hayatta karar verme süreçleri aynı KA da olduğu gibi, seçimler ve bunların sonuçlarından oluşan örgüleri içerdiği için, KA'ların yorumlanması kolaydır. KA'lar hem nominal hem de nümerik verileri hızlı bir şekilde sınıflandırabilmektedir [47].

4.3 Naive Bayes

Naive Bayes algoritması, olasılık temelli bir algoritmadır. Bayes Teoremine göre istatiksel tahmin yapar. Algoritmanın amacı sınıf üyelik olasılığını tahmin etmektir. Uygulanan işlem Eşitlik 4.2'de görülen Bayes Teoremini esas alır. Diğer sınıflandırma işlemlerinde olduğu gibi Bayes Teoremi tabanlı olarak yapılan sınıflandırmada da örneklem uzayının bir kısmı öncelikle eğitim amaçlı kullanılarak öğrenme yapılır, örneklem uzayının kalan kısmı ile öğrenilen veriler üzerinden tahmin yapılır. Bir öğenin hangi öznitelikleri kendisinde topladığı ve böylece hangi sınıfa atandığı algoritma tarafından öğrenildiğinden, test kümesi içinde sıralı birtakım özellikleri verilen öğelerin hangi sınıfa ait olduğu tahmin edilebilir [46].

$$\boldsymbol{p(C_k|x)} = \frac{p(\boldsymbol{x}|\boldsymbol{C_k})\,p(C_k)}{p(x)} \tag{4.2}$$

$p(x|C_k)$: k sınıfındaki öğelerden birinin x olma olasılığı
$p(C_k)$: tüm örneklem uzayında k sınıfının bulunma olasılığı (ilk olasılık)
$p(x)$: tüm örneklem uzayında x in bulunma olasılığı
$p(C_k|x)$: x örneklerinin içinde k sınıfından öğelerin bulunma olasılığı

Bayes Teoremi ile sınıflandırma yapılırken amaç, bir öğenin (x) hangi sınıfa sahip olduğunu bulmaktır. Bu işlem yapılırken, x öğesinin bütün sınıflar içinde ayrı ayrı bulunma olasılıkları hesap edilip, bunların içinden en yüksek olasılık tespit edilir ve x öğesi, o sınıfa dâhil edilir. x öğesinin tüm örneklem uzayında bulunma olasılığı her zaman aynı olduğu için, bir x niteliğini farklı birçok sınıfa ait olasılıkların büyüklüğünün kıyaslanmasında kullanmamak sonucu değiştirmez. Bu yüzden, toplam işlem maliyetini azaltmak adına Bayes Teoremi, Eşitlik (4.3) de belirtilen duruma indirgenir.

$$\boldsymbol{p(C_k|x)} = \boldsymbol{p(x|C_k)\,p(C_k)} \tag{4.3}$$

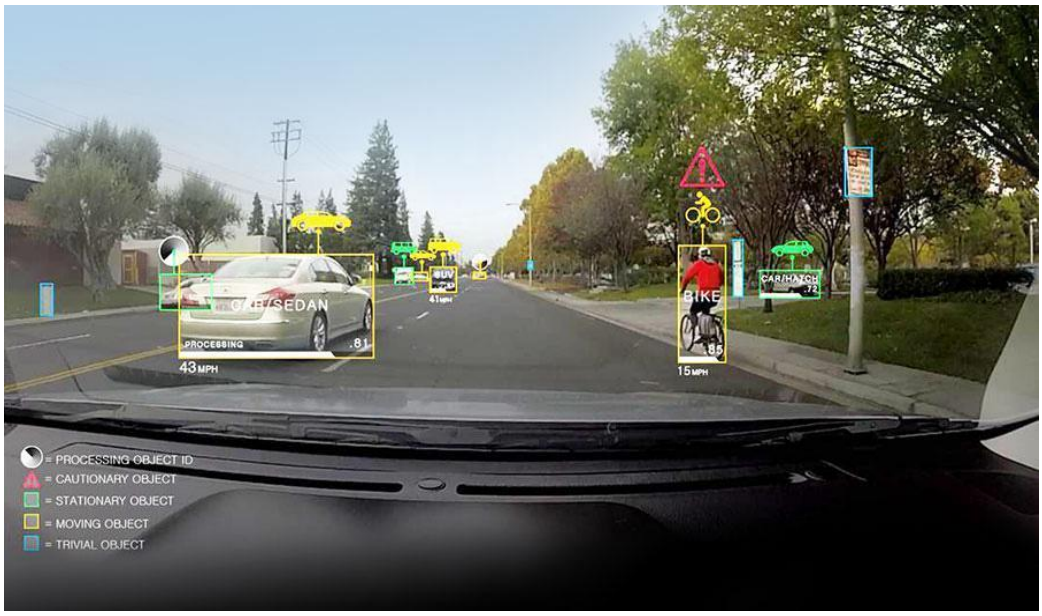
Bazı durumlarda, x birden çok birbirinden bağımsız niteliği içerebilir ve zincirleme koşullar olarak düşünülebilir. Bu durumda, her bir alt niteliğin ana sınıf

içinde bulunma durumları birbiri ile çarpılarak, x koşul kümesinin C_k sınıfı içinde olma olasılığı Eşitlik (4.4) kullanılarak bulunur.

$$p(x|C_k) = \prod_{m=1}^{n} p(x_m \mid C_k) = p(x_1 \mid C_k) \times p(x_2 \mid C_k) \times \ldots \times p(x_n \mid C_k) \qquad (4.4)$$

4.4 Derin Öğrenme

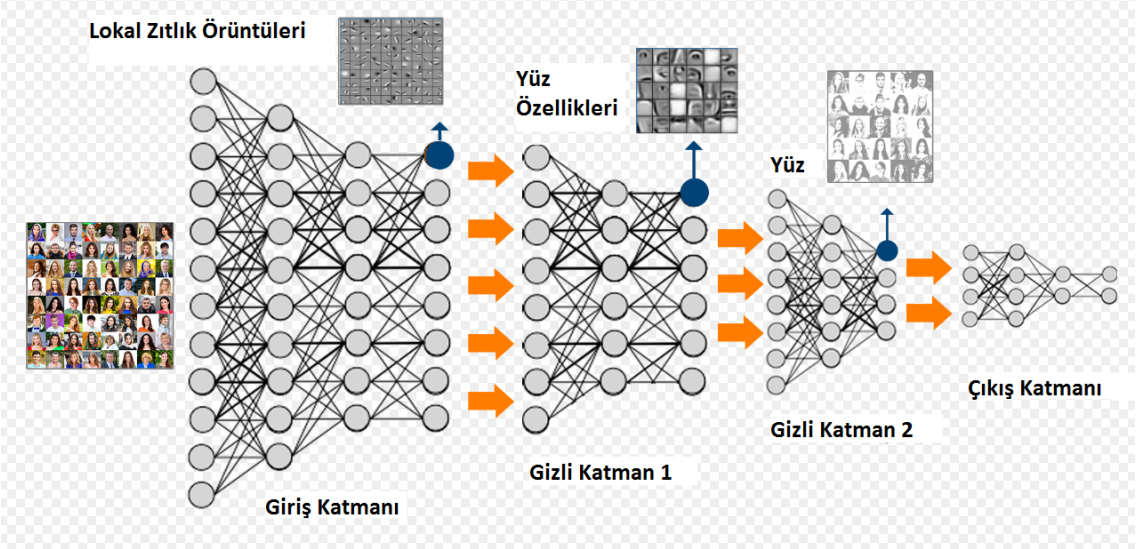
Derin öğrenme, son dönemde makine öğrenmesi ve yapay zekâ kavramları içerisinde popülerliği gitgide artan bir yöntem olmuştur. Günümüzde Derin Öğrenme, görüntü sınıflandırma, kişilerin seslerinin analizi, video işleme, doğal dil öğrenme, Canlı trafik akışı analizi (Şekil 4.7) gibi birçok sivil ve askeri alanda kullanılmakta ve özellikle büyük teknoloji firmaları ve üniversite araştırma grupları tarafından ciddi yatırımlarla desteklenmektedir.



Şekil 4.7 *Canlı trafik akışı analizi [54]*

Derin öğrenmenin klasik makine öğrenmesi yöntemlerinden en büyük farkı, önceden hazırlanmış herhangi bir matematiksel modele dayalı öznitelik çıkarımına ihtiyaç duymamasıdır. Öznitelik çıkarıcı, derin öğrenme tarafından otomatik olarak uygun şekilde öğrenildiğinden, sistem ağ katmanlarına ve tahmin işlemlerine odaklanmaktadır [50].

Derin Öğrenme algoritması, ileri beslemeli Yapay Sinir Ağları (YSA) tabanlı bir algoritmadır. Derin Öğrenme algoritmasının en küçük bileşenlerine, tüm YSA'lardaki gibi nöron adı verilir ve veri seti bir algoritma tarafından öğrenilerek test işlemine geçilir. Eğitimde, geri-yayılma uygulanarak *Olasılıksal Eğim Azalma Fonksiyonu* (OEAF) kullanılmıştır. Derin öğrenme ağı, *hiperbolik tanjant (tanh) doğrultucu* ve *maksimum çıkış aktivasyon fonksiyonlu* nöronlar içeren çok sayıda gizli katmanlardan oluşur. Derin öğrenme algoritmasının *Momentum Eğitimi* (ME), *Adaptif Öğrenme Oranı* (AÖA) gibi birçok özelliği yüksek TDP sağlamaktadır. Her hesaplama düğümü, global değişkenleri lokalde uygulayarak asenkron(paralel) programlama yapısında çalışmakta ve global modele network üzerinden model ortalaması alarak katkıda bulunmaktadır. Derin Öğrenme Algoritması Katman yapısı Şekil 4.8 de görülmektedir.



Şekil 4.8 *Derin Öğrenme Algoritması Katman yapısı [55]*

Çalışmada, derin öğrenme algoritması olarak tek düğümlü H2O kümeleme algoritması kullanılmıştır. Diğer derin öğrenme algoritmalarında olduğu gibi birden çok iş parçacığı H2O algoritmasında eşzamanlı olarak işlenebilmekte ve işlenecek iş parçacığı Rapidminer yazılımı aracılığıyla belirlenebilmektedir. Alternatif bir değer girilmediği takdirde Rapidminer yazılımı eşzamanlı olarak tek iş parçacığı işlemektedir. Rapidminer Derin öğrenme operatörü standart değerler olarak 50 nörona 2 gizli katman

olacak şekilde yapılandırılmış olup, kullanıcı tarafından gizli katman sayısı ve katmanlara düşecek nöron sayıları ayarlanabilmektedir [51].

Çalışmada, Derin Öğrenme operatöründe bulunan *Adaptif Öğrenme Oranı* (AÖA) seçilmiştir. AÖA, epsilon ve *rho* değerleri üzerinden hesaplanır. AÖA'nın *epsilon* ve *rho* üzerinden otomatik olarak hesaplanması, optimizasyon işleminde düzlemin çok yüksek eğime sahip olması ya da eğimin çok düşük değerde olması durumları arasında bir denge kurularak ideal öğrenme oranlarının elde edilmesini sağlar.

5. UYGULANAN HİBRİD MODEL

5.1 Veri Setinin Nitelikleri ve Hazırlanması

Bu çalışmada, önemli bir veri madenciliği yazılımı olan Rapidminer yazılımı kullanılmıştır. Rapidminer, gönüllü bir topluluk olan Rapidminer Vakfı tarafından sürekli geliştirilmekte olan oldukça kapsamlı bir veri kıymetlendirme yazılımıdır. Rapidminer yazılımı, araştırmacılara tüm veri işleme sürecinde büyük görsel kolaylık sağlamakta ve sonuçları değerlendirme/kıyaslamada birçok alternatifler sunmakta, WEKA gibi diğer bazı yazılımlara ait kütüphaneleri de eklenti olarak barındırmaktadır [52].

Bu çalışmada, Kaliforniya Üniversitesi Makine Öğrenmesi Veri Havuzunda (University of California, Irvine, USA) bulunan "Türkiye Öğrenci Kıymetlendirme Veri Seti (Turkiye Student Evaluation Data Set)" isimli açık kaynak bir veri seti (dosya: turkiye-student-evaluation_generic.csv) kullanılmıştır [53]. Veri seti, 2013 yılında Gazi Üniversitesi (Ankara)'de yapılmış ve 2850 öğrencinin, kendilerine toplam 13 ders veren 3 adet eğitmeni değerlendirdiği bir çalışmadır. Veri setinde eksik, hatalı, sıra dışı veri bulunmamaktadır. Yapılan tüm değerlendirmeler Likert ölçeğinde (1-5 arası değerler verilerek) yapılmıştır. Veri setinde toplam 33 adet niteliğin, 5 adedi derslerin isimlerini, derslerin zorluklarını, derslerin öğrenci tarafından alınma sayılarını ve öğretmenlerin isimlerini sembolize etmekte diğer 28 öznitelikte ise öğrenciler eğitmenlerine dair sorulmuş çeşitli soruları cevaplamaktadır. 33 adet niteliğin açıklamaları aşağıda belirtilmiştir:

instr: Eğitmeni ifade eder; {1,2,3} değerlerinden birini alır.
class: Alınan dersin ismini ifade eder, {1-13} aralığındaki değerlerden birini alır.
repeat: Öğrenci tarafından ilgili dersin kaç kez alındığını ifade eder; {1,2,3} değerlerinden birini alır.
attendance: Öğrencinin derse katılım seviyesi; {0, 1, 2, 3, 4} değerlerinden birini alır.
difficulty: Öğrencinin yorumuna göre dersin zorluk seviyesi; {1,2,3,4,5} değerlerinden birini alır.

Q1: Dönem başında bildirilen ders içeriği ve öğretim metodu ve kıymetlendirme sistemi

Q2: Dönem başında kursun amaç ve kazanımları açıkça bildirilmiştir.

Q3: Kurs, kendisine atanan kredi düzeyine karşılık verebilecek niteliktedir.

Q4: Kurs, ders programında belirtildiği şekilde işlendi.

Q5: Sınıf içi tartışmalar, verilen ödevler, yapılan uygulama ve çalışmalar yeterli düzeydeydi.

Q6: Kullanılan tüm dokümanlar güncel ve yeterliydi.

Q7: Ders için yeterli sınıf, laboratuvar ve tartışma alanı hazırlanmıştı.

Q8: Quiz, ödev, proje ve örnekler eğitime katkısı sağladı.

Q9: Derslere katılmaktan çok zevk aldım ve hevesliydim.

Q10: Dönem başındaki beklentilerim, dönem sonunda karşılandı.

Q11: Kurs kişisel gelişimim için faydalı oldu.

Q12: Kurs, hayata yeni bir perspektifte bakmamı sağladı.

Q13: Eğitmenin bilgisi yeterli ve günceldi.

Q14: Eğitmen sınıfa hazır geldi.

Q15: Eğitmen, anlatılan ders planına uydu.

Q16: Eğitmen, dersini adanmış ve anlaşılır bir şekilde sundu.

Q17: Eğitmen, derslere zamanında geldi.

Q18: Eğitmenin konuşması akıcı ve düzgündü.

Q19: Eğitmen, ders saatlerini efektif kullandı.

Q20: Eğitmen, dersi öğrencilere sevdirme konusunda istekliydi.

Q21: Eğitmen, öğrencilere karşı pozitif bir yaklaşım sergiliyordu.

Q22: Eğitmen, öğrencilerin yorumlarına karşı açık ve saygılıydı.

Q23: Eğitmen, öğrencileri derse katılma konusunda cesaretlendirirdi.

Q24: Eğitmen, öğrencilere derslerine katkı sağlayacak ödev/projeler verdi.

Q25: Eğitmen, öğrencilerin kurs ile ilgili sorularını her ortamda cevapladı.

Q26: Eğitmenin kursu kıymetlendirme adına yaptığı tüm sınavlar, kursun amacına yönelikti.

Q27: Eğitmen, sınavların çözümlerini öğrencilerle paylaştı ve tartıştı.

Q28: Eğitmen, bütün öğrencilere eşit ve tarafsız davrandı.

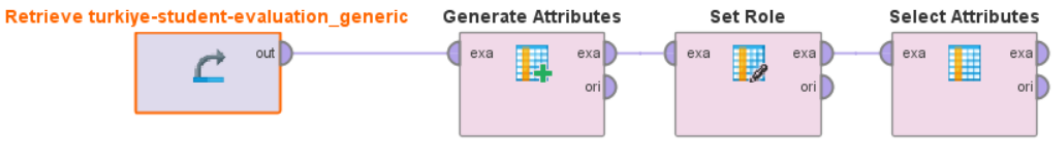
Q1-Q28 aralığındaki tüm öznitelikler Likert tipinde olup {1,2,3,4,5} değerlerinden birini almaktadır.

Çalışmada kullanılan veri setinde her satır, bir öğrencinin aldığı bir dersle ilgili, eğitmeni hakkında yaptığı değerlendirmeleri ifade etmektedir. Ancak bu özniteliklerin sonucunda öğrencinin o dersten öğretmenine verdiği bir puan olmadığı için, öğrencinin kanaati somut olarak ifade edilmemiştir. Bu durumu çözmek adına öğrencinin öğretmeni hakkında çeşitli özellikler üzerinden puan verdiği Q1-Q28 arası özniteliklerin toplanarak ortalamalarının alındığı "*SCORE*" adlı 34. öznitelik oluşturulmuştur.

Veri setindeki tüm öğretmenleri ders bazlı genel şekilde kategorize edebilmek için ise başarı durumunu sözel olarak ifade edecek *"SUCCESS"* niteliği, 35. öznitelik olarak oluşturulmuştur. SUCCESS niteliğinin değerinin hesaplanması için Rapidminer'da (programa özel formatta) Eşitlik 5.1 oluşturulmuş, 4'ten küçük değerler KÖTÜ, 4'e eşit değerler İYİ, 4'ten büyük değerler ise MÜKEMMEL olarak belirlenmiştir.

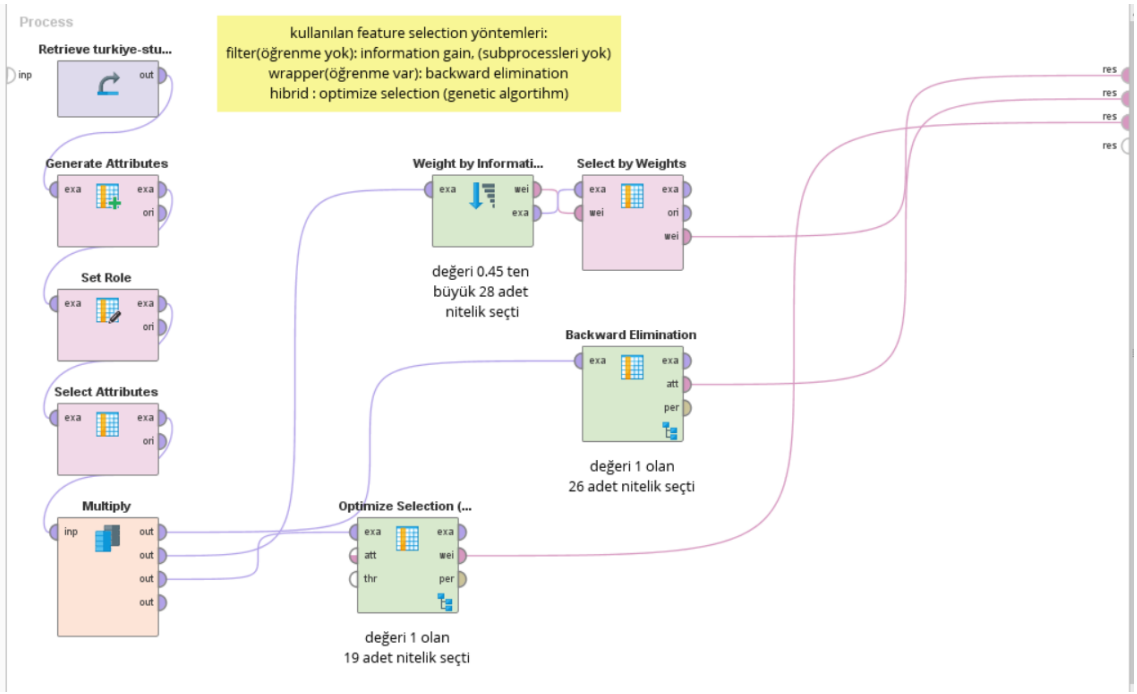
$$SUCCESS==if(SCORE<=3,"BAD",if(SCORE==4,"GOOD","PERFECT")) \quad (5.1)$$

Şekil 5.1 de yapılan veri ön işleme adımları görülmektedir. Veri setinin Rapidminer yazılımına alınması sonrasında, *generate attributes* modülü ile SCORE ve SUCCESS öznitelikleri oluşturulmuştur. Kullanılan veri setinde, bir hedef öznitelik önceden etiketlenmiş *(labeled)* olmadığı için, kullanılacak makine öğrenme yöntemlerine hedef belirtmek adına *"set role"* modülü ile SCORE niteliği etiketli öznitelik olarak atanmıştır. Hem işlem maliyetinin azaltılması, hem de makine öğrenme yöntemlerinin erken öğrenerek gerçek durum tahminlerinden uzaklaşmasını engellemek için SCORE niteliği select *attribute* modülü *kullanılarak* kaldırılmıştır. Sonuç olarak, öznitelik indirgeme bölümüne 1 niteliği etiketlenmiş 34 öznitelikli bir veri seti gönderilmiştir.



Şekil 5.1 *Veri Önişleme adımları*

5.2 En İyi Öznitelik Seçme Metodunun Tespit Edilmesi



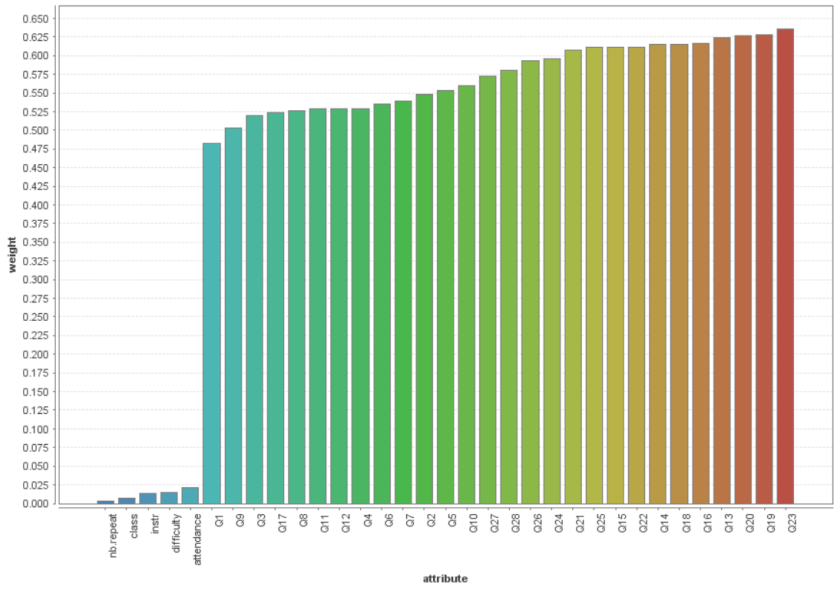
Şekil 5.2 *En iyi Özniteliklerin seçimi*

Öznitelik seçme işleminde, farklı kategorilerde bulunan öznitelik seçme algoritmalarının performanslarını kıyaslayabilmek adına, her bir öznitelik seçme kategorisinden birer olmak üzere (Filtreleme kategorisi = Bilgi Kazancı, Sarmal kategorisi = Geri Yönlü Eliminasyon, Hibrid Kategori = Genetik Algoritma) toplam 3 adet algoritma kullanılmıştır. Şekil 5.1 de veri işlemesi tamamlanmış 35 öznitelikli veri setinin en iyi özniteliklerin seçilmesi için 3 farklı öznitelik seçme algoritmasına gönderildiği görülmektedir.

Rapidminer yazılımı, Bilgi Kazancı algoritması modülünde çıktı olarak yalnızca seçilen ağırlıklar ve orijinal veri setini verdiği için ilave olarak *"Select by Weights"* modülü kullanılarak, indirgenmiş veri seti elde edilmekte ve sonuç kısmına gönderilmektedir. Geri Yönlü Eliminasyon ve Genetik Algoritma modülleri ise çıktı olarak ilave bir modüle gereksinim duymadan indirgenmiş veri setlerini sonuca gönderebilmektedir.

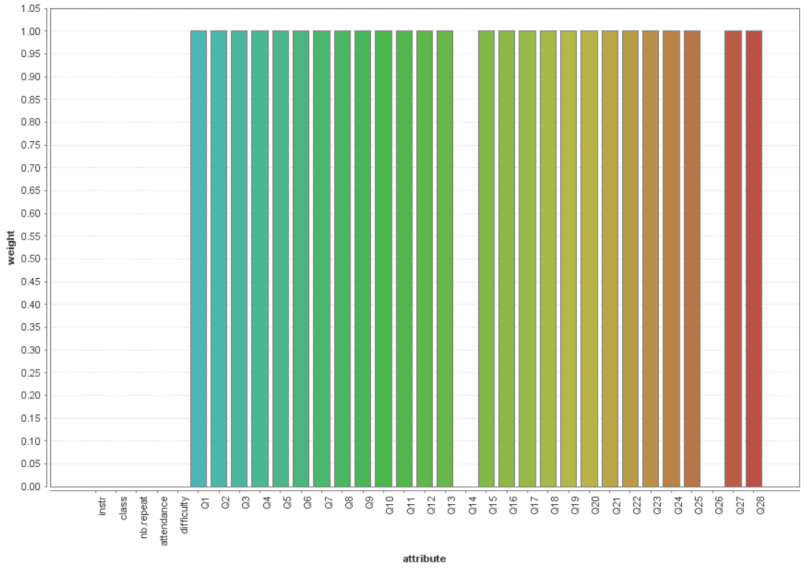
Bilgi Kazancı algoritması tarafından sıralanan öznitelikler Şekil 5.1.2 de görülmektedir. Ağırlık değeri 0,45 ten büyük olan değere sahip 28 adet öznitelik seçilmiş ve indirgenmiş veri seti, çıktı olarak tahmin kıymetlendirme kısmına gönderilmiştir. Burada sınır olarak tespit edilen 0,45 değeri, yaptığımız gözlem sonrası

5 niteliğin ağırlık değerinin diğerlerinden çok düşük olduğu görsel olarak tespit edilerek belirlenmiştir.



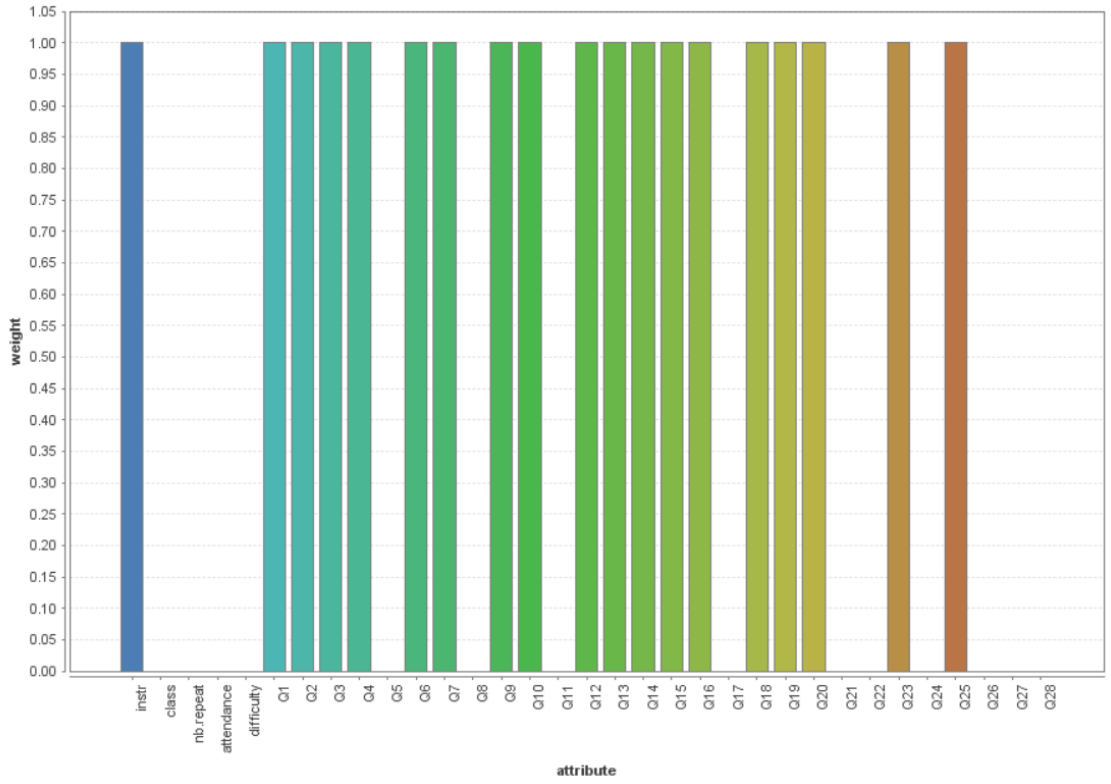
Şekil 5.2.2 *Bilgi Kazancı Yöntemiyle Ağırlıklandırılmış Veri Seti Öznitelikleri*

Geri Yönlü Eliminasyon ile öznitelik seçme yapılırken, Rapid miner "Backward Elimination" modülü kullanılmış, *maksimum eliminasyon sayısı*=10, *spekülatif tur sayısı=0*, *duruş davranışı=azalma* olacak şekilde belirlenmiş ve sonuçta ağırlıkları 1 değerine eşit olan ve 26 adet öznitelikten (Şekil 5.1.3) oluşan indirgenmiş veri seti elde edilmiştir.



Şekil 5.3 *Geri Yönlü Eliminasyon Yöntemi ile seçilen Öznitelikler*

Genetik Algoritma ile öznitelik seçimi yapmak için *Optimize Selection(Evoluotionary)* modülü kullanılmış, *minimum öznitelik sayısı=1, popülasyon boyutu=5, jenerasyonların maksimum sayısı=30* olarak belirlenmiş, 0.25 değerli turnuva yöntemi kullanılmış, p ilk=0,5, p mutasyon=-1,0, p çaprazlama=0,5, çaprazlama tipi: tekdüze olarak belirlenmiş ve ağırlıklar normalize edilmiştir. Algoritma sonuç olarak, ağırlık değeri 1 olan 19 adet öznitelik seçmiş (Şekil 5.1.4) ve çıktıları sonuç kısmına göndermiştir.

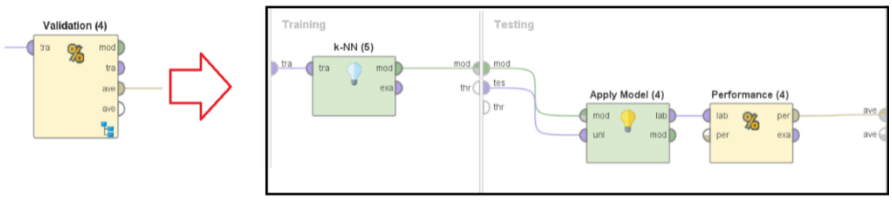


Şekil 5.4 *Genetik Algoritma ile seçilen Öznitelikler*

Öznitelik seçme işleminde görüldüğü üzere mevcut 33 öznitelik arasından sonuca etki eden en iyi 19 niteliği tespit etmesiyle Genetik Algoritma en iyi sonucu vermiş görünmektedir. Öznitelik seçme işleminde elde edilen sonucun doğruluğunu pekiştirmek ve her üç algoritmanın öznitelik seçme performansını adil bir şekilde kıyaslayabilmek için, tüm algoritmaların etiketli niteliği tahmin etmelerini kıymetlendirme adımlarında, K-En Yakın Komşuluk (K-EYK) algoritması, ilişkiye dayalı *ayrık kıymetlendirme* seçilerek kullanılmıştır. *Ayrım değeri=0,5* (veri setinin yarısı eğitim, yarısı test) olarak belirlenmiştir. K-EYK algoritmasının değerleri tüm

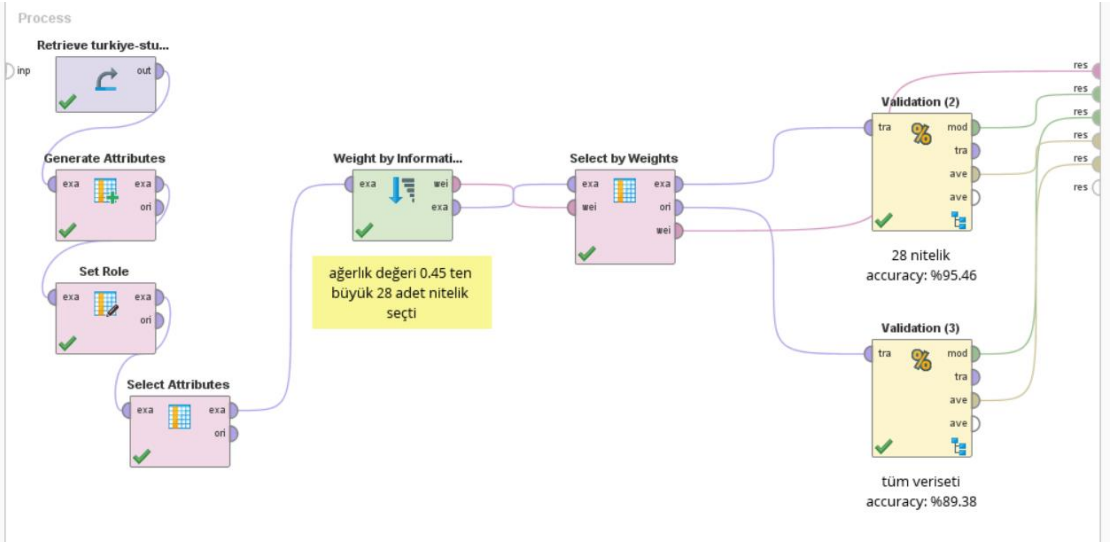
öznitelik seçme işlemleri için *k=1, ölçüm uzaklık birimi= Öklid Uzaklığı* olarak ayarlanmıştır. Bütün öznitelik seçme kıymetlendirme işlemlerinde Şekil 5.1.5'te görüldüğü gibi *kıymetlendirme (validation)* modülü içerisine, K-EYK algoritmasını eğitim ve test veri setlerine uygulayan *apply model* ve işlem performansını *doğruluk* (accuracy) (Eşitlik 5.2) kriterine göre tespit eden *performance* modülleri sıralı olarak yerleştirilmiştir.

Doğruluk = (doğru tahmin sayısı) / (toplam örnek sayısı) **(5.2)**



Şekil 5.5 *Öznitelik Seçme Kıymetlendirmesi İşlemi alt modülleri*

Filtreleme kategorisinde yer alan Bilgi Kazancı algoritması, öznitelik seçme işlemini veri setinde bulunan öğeleri ağırlıklandırarak yapar. Rapid Miner yazılımı modüllerinde bu işlem *"Bilgi Kazancı yöntemine göre ağırlıklandırma"*, *"ağırlıklara göre seçme" ve "kıymetlendirme"* modüllerin sırayla birbirine bağlanıp sonuca gidilmesiyle yapılmaktadır. Çalışmada, öznitelik indirgeme sonrası elde edilen kazanımı tespit etmek için, *ağırlıklara göre seçme (select by weights)* modülünden sonra tüm veri setine ve indirgenmiş veri setine ait olmak üzere 2 ayrı kıymetlendirme işlemi yapılmıştır (Şekil 5.6). Her iki kıymetlendirme işleminde de yine *ayrık kıymetlendirme* kullanılmış, *ayrım değeri=0,5* olarak belirlenmiştir.



Şekil 5.6 *Bilgi Kazancı Öznitelik Seçme Performansı Modülleri*

Yapılan ilk kıymetlendirme işlemi bize önemli bir konuda ışık tutmaktadır. Bu kıymetlendirmede, öznitelik indirgeme işlemi yapılmadan önce veri setinde mevcut olan 33 niteliğin tamamı kullanılarak bir tahmin işlemi yapılmış ve öznitelik indirgemenin performansa ne denli etki edeceğini anlama adına bizim için başlangıç verisi sağlamıştır. BK yöntemi ile 28 öznitelik sayısına indirgenmiş veri setinin TDP %90,10 dan %95,84 e yükselmiştir. Tüm veri seti Performans kıymetlendirmesi Zıtlık Matrisi (ZM) ile Tablo 5.1'de gösterilmiştir. ZM'lerinde SCORE niteliğinin değerleri olan BAD, PERFECT, GOOD ifadelerinin doğru tahminlerine ilişkin detaylı performans durumu gösterilmiştir. Buna göre, tüm veri setini ifade eden ZM'de en iyi tahmin %95,69 ile PERFECT, ikinci olarak %93,95 ile BAD, üçüncü olarak %70,37 ile GOOD değeri olmuştur.

Tablo 5.1 *Tüm Veri Seti kullanıldığında BK Tahmin Performans Değerleri*

TDP %90,10	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1395**	46	0	%96.81
Tahmin PERFECT	30	**1028**	17	%95.63
Tahmin GOOD	0	20	**374**	%94.92
Sınıf Geriçağrımı	%97.89	%93.97	%95.65	

ZM'ler tahmin performansının başarısını, etiketlenmiş niteliğin değer türlerini bulma başarısını göstererek detaylı şekilde açıklamaktadırlar. Performans kıymetlendirmede kullanılan *sınıfa ait tahmin başarısı (class precision),* tahminde doğru olarak bulunan değerin tüm değerlere oranını ifade etmektedir. Örneğin, BAD değerinin *sınıfa ait tahmin başarısı,* ZM'de ilk satırda yapılan BAD değerini tahmin işlemi esnasında elde edilen doğru BAD (koyu renkli) değerinin, yapılan tahminde elde edilen tüm değerlerin eleman sayılarına (aynı satırdaki tüm elemanlar) bölünmesiyle elde edilir (Eşitlik 5.3).

Class Precision (BAD): True BAD / (True BAD + False Perfect + False Good) **(5.3)**

Elde edilmiş tüm tahmin değerleri içinde, her niteliğin kendi sınıfı içine hangi oranda doğru olarak yerleştiğini tespit etmede *sınıf geri çağırma oranı (SGÇO) (class recall*) 'nı tespit etmek için Eşitlik (5.2.3) kullanılmıştır. BAD niteliğinin SGÇO ele alırsak (soldan ilk sütun) tüm BAD sonuçları incelendiğinde bunların tamamının BAD tahmini esnasında elde edilmesi gerekirken, 27 adedinin PERFECT tahmini esnasında ortaya çıktığı görülmekte ve SGÇO, tüm BAD'lerin içerisinde doğru yerde üretilen BAD'lerin oranını vermektedir (Eşitlik 5.4).

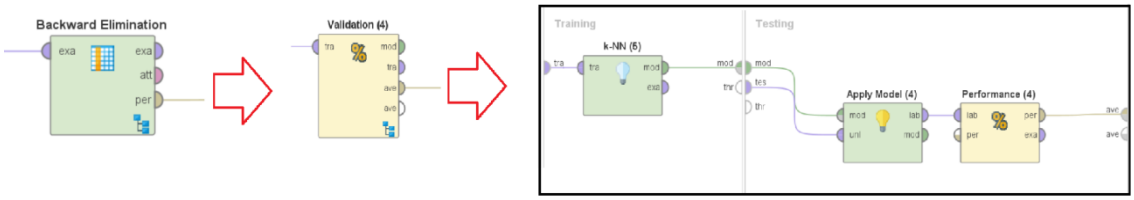
Class Recall (BAD): True BAD / (True BAD +False BAD+False BAD) **(5.4)**

BK yöntemi ile öznitelik indirgemesi yapıldıktan sonra elde edilen 28 öznitelikli veri seti ile yapılan tahmine ilişkin detaylı performans kıymetlendirmesi Tablo 5.2'de görülmektedir. Buna göre, indirgenmiş veri setini ifade eden ZM'de en iyi sınıf tahmini %97,21 ile BAD, ikinci olarak %95,26 ile PERFECT, üçüncü olarak %95,24 ile GOOD değeri olmuştur.

Tablo 5.2 *İndirgenmiş Veri Seti kullanıldığında BK Tahmin Performans Değerleri*

TDP %95,84	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1393**	40	0	%97.21
Tahmin PERFECT	32	**1024**	19	%95.26
Tahmin GOOD	0	30	**372**	%92.54
Sınıf Geriçağrımı	%97.75	%93.60	%95.14	

İkinci olarak kullanılan öznitelik seçme algoritması Geri Yönlü Eliminasyon (GYE)dur. Rapidminer yazılımı GYE ile öznitelik seçme işlemi, öznitelik indirgeme ve kıymetlendirme alt prosesi kendi içinde bulunan *Backward Elimination* modülü ile yapmaktadır (Şekil 5.7). Bu modül çıktı olarak indirgenmiş veri setine ilişkin yapılan tahminin doğruluğunu gösteren performans vektörünü (ZM) verir. GYE işleminde kıymetlendirme için yine K-EYK algoritması kullanılmış ve çıktı olarak Tablo 5.3 da görülen ZM elde edilmiştir.



Şekil 5.7 *GYE Modülü ve alt prosesleri*

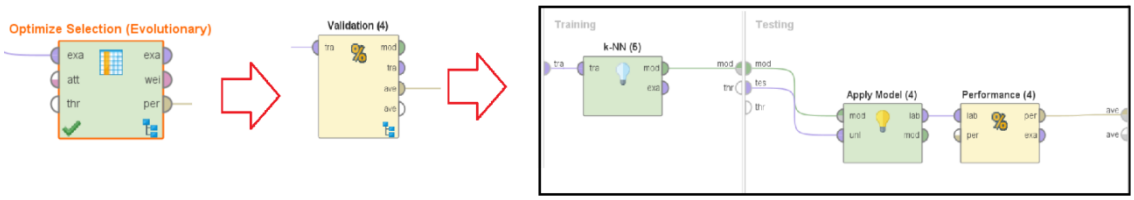
Tablo 5.3 *GYE Tahmin Performansı Değerleri*

TDP %96,98	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1400**	30	0	%97.90
Tahmin PERFECT	25	**1044**	13	%96.49
Tahmin GOOD	0	20	**378**	%94.97
Sınıf Geriçağrımı	%98.25	%95.43	%96.68	

GYE yöntemi ile öznitelik indirgemesi yapılarak 26 öznitelikli bir veri seti elde edilmiştir. Yapılan tahmine ilişkin detaylı performans kıymetlendirmesi ZM'i incelendiğinde, GYE genel tahmin doğrulu performansının %96,98, en iyi sınıf tahmininin %97,90 ile BAD, ikinci olarak %96,49 ile PERFECT, üçüncü olarak %94,97 ile GOOD değerinin yer aldığı görülmüştür. *Sınıf geri çağırma oranları (SGÇO)* kıyaslandığında BAD=98,25, GOOD=96,68 ve PERFECT=95,43 olarak sıralanmaktadır.

Üçüncü öznitelik seçme algoritması olarak, hibrid bir yapıya sahip olan Genetik Algoritma kullanılmıştır. Genetik algoritma ile öznitelik indirgeme yapmak için Rapidminer'ın "Genetik Algoritma ile Optimize Seçim (Optimize Selection –

Evolutionary- (GAOS))" modülü kullanılmıştır. GAOS modülü de öznitelik seçme ve kıymetlendirme işlemleri için dâhili alt prosesler içermektedir (Şekil 5.8). Genetik Algoritma (GA) ile öznitelik indirgemesi yapıldıktan sonra elde edilen 18 öznitelikli veri seti elde edilmiş ve tahmin işlemi yapılmıştır. Yapılan tahmine ilişkin detaylı performans kıymetlendirmesi Zıtlık Matrisi incelendiğinde (Tablo 5.4), GA genel tahmin doğrulu performansı %95,50, en iyi sınıf tahmini %97,07 ile BAD, ikinci olarak %95,30 ile PERFECT, üçüncü olarak %90,60 ile GOOD değeri olmuştur. *Sınıf geri çağırma oranları (SGÇO)* kıyaslandığında BAD=97,54, GOOD=96,16 ve PERFECT=92,60 olarak sıralanmaktadır.



Şekil 5.8 *GAOS Modülü ve Alt Prosesleri*

Tablo 5.4 *Genetik Algoritma Tahmin Performansı Değerleri*

TDP %95,5	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1390**	42	0	%97.07
Tahmin PERFECT	35	**1013**	15	%95.30
Tahmin GOOD	0	39	**376**	%90.60
Sınıf Geriçağrımı	%97.54	%92.60	%96.16	

Sonuç olarak her 3 algoritma ile yapılan öznitelik indirgemesi işlemlerinde 19 öznitelik ile Genetik Algoritma, en değerli olan öznitelikleri bulma hedefini en az sayıda öznitelik ile elde etmiş ve işlem, %95,5 doğruluk oranıyla tamamlamıştır. GYE algoritması %96,98 doğruluk oranına ulaşmasına rağmen 26 adet öznitelik seçtiği için, çalışmanın devamında kullanılmak üzere Genetik Algoritma öznitelik indirgeme elemanı olarak belirlenmiştir.

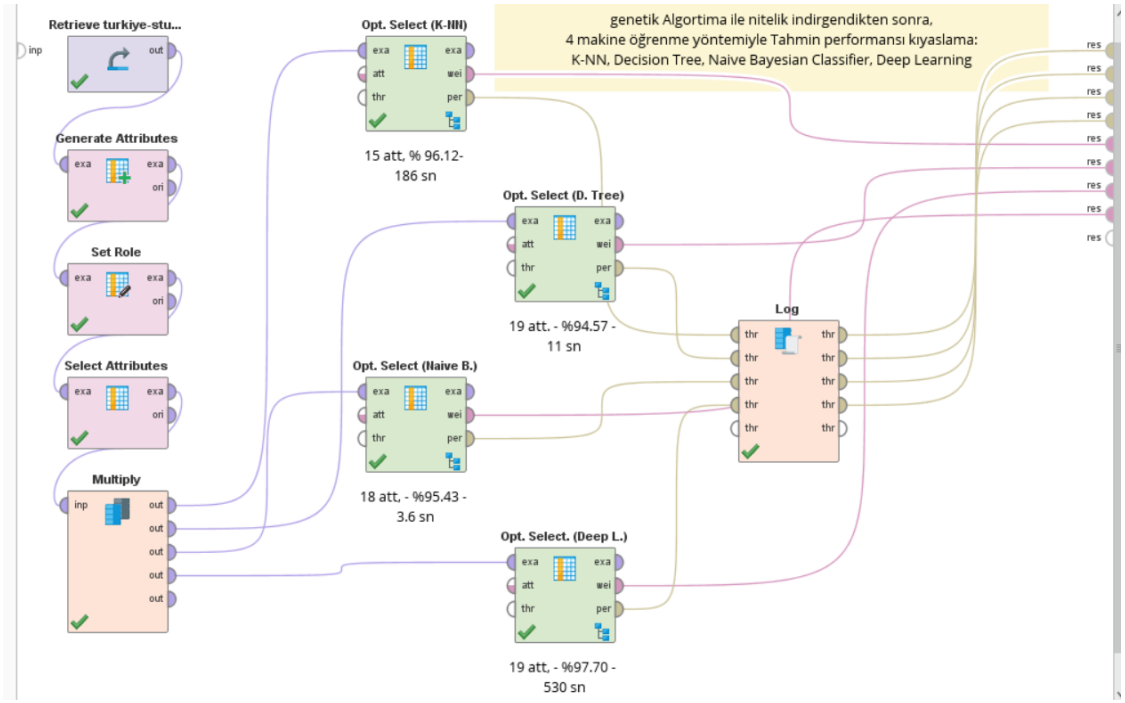
5.3 En İyi Makine Öğrenmesi Metodunun Tespit Edilmesi

Bir önceki bölümde, en etkin özniteliklerden oluşan bir indirgenmiş veri seti elde eden en iyi öznitelik seçme algoritması bulunmuştur. Bir sonraki aşama olarak, bu indirgenmiş veri seti ile en iyi tahmini yapabilecek makine öğrenmesi yöntemi tespit edilmiştir. Veri setindeki etiketli öznitelik olan SCORE niteliği, metinsel (nominal) bir yapıya sahip olduğundan, kullanılan makine öğrenmesi algoritmaları da metinsel veriyi işleyip tahmin edebilecek algoritmalardan seçilmiş ve tahmin işleminde sırasıyla K-En Yakın Komşuluk, Karar Ağacı, Naive Bayes ve Derin Öğrenme algoritmaları kullanılmıştır.

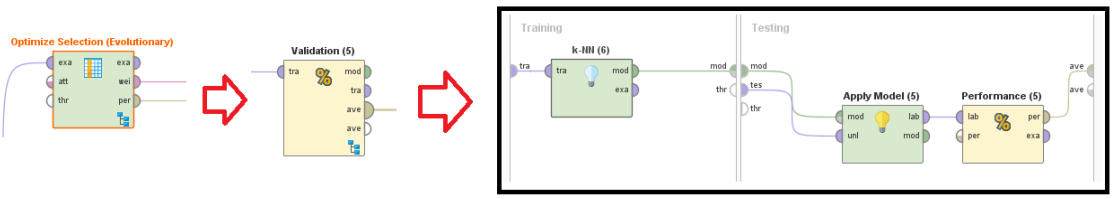
Rapid Miner yazılımında, belirtilen bu makine öğrenmesi yönteminin kullanılması Şekil 5.9 de görülmektedir. Tüm tahmin kıymetlendirme işlemleri, alt prosesine performans değerlendirme modülü yerleştirilen Genetik Algoritma modülü kullanılarak yapılmıştır. Tüm genetik algoritma modüllerinde *minimum öznitelik sayısı=1, popülasyon boyutu=5, jenerasyonların maksimum sayısı=30* olarak belirlenmiş, 0.25 değerli turnuva yöntemi kullanılmış, p ilk=0,5, p mutasyon=-1,0, p çaprazlama=0,5, çaprazlama tipi: tekdüze olarak belirlenmiş ve ağırlıklar normalize edilmiştir. Alt kıymetlendirme işlemi ana prosesinde (validation modülü) *Ayrık Kıymetlendirme* ilişkisel olarak kullanılmış ve *ayırma oranı =0,5* olarak belirlenmiş, *otomatik örnekleme* kullanılmıştır.

Genetik Algoritma (GA)nın davranışının nasıl sonuçlar meydana getireceğinin önceden öngörülemeyeceği, üçüncü bölümde belirtilmişti. Yapılan çalışma bu durumu gözler önüne sermiş, her dört tahmin işleminden önce yapılan öznitelik seçme işlemlerinde GA, farklı sayıda öznitelik seçmiştir.

Bütün veri madenciliği işlemlerinde *işlem maliyeti (total running time)* uygulanan işlemin performansının değerlendirilmesinde önemli bir ölçüt olduğundan, kullanılan 4 farklı tahmin işlemi için LOG modülü kullanılarak işlem zamanları hesaplanmıştır.

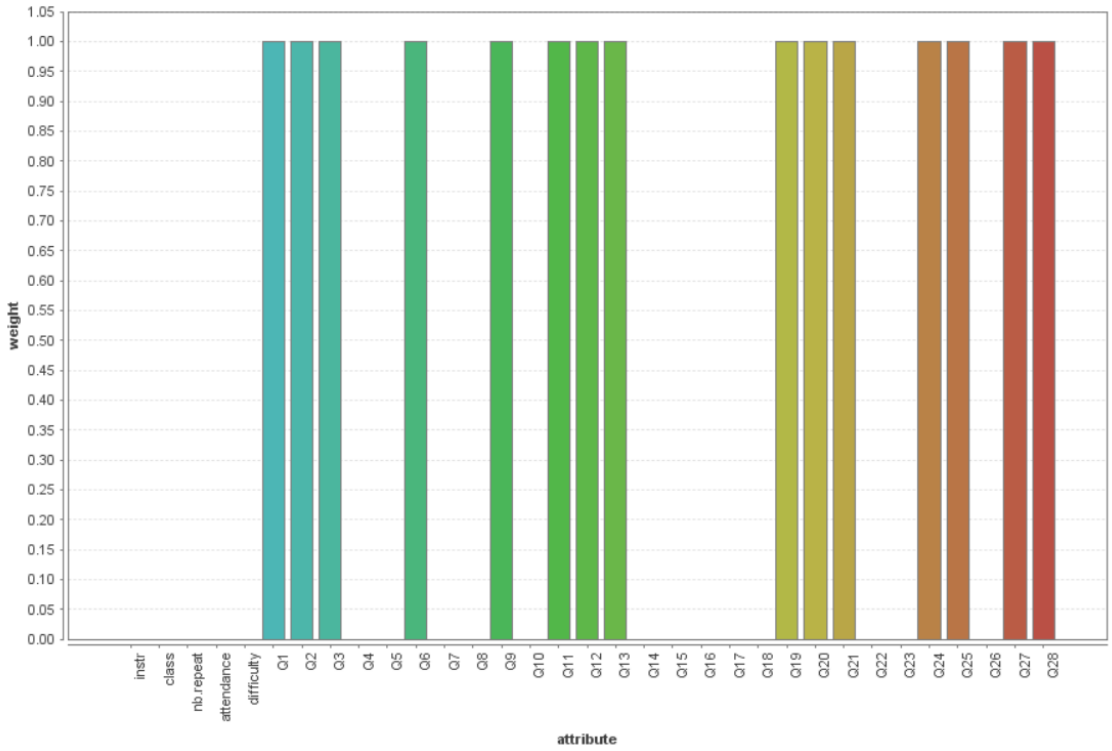


Şekil 5.9 *Çeşitli Makine Öğrenmesi yöntemleri ile tahmin yapılması*



Şekil 5.10 *K-En Yakın Komşuluk Algoritması ile tahmin yapılması*

K-En Yakın Komşuluk (K-EYK) algoritması ile yapılan tahmin işleminde (Şekil 5.10) K=1, *mesafe ölçüm yöntemi=Öklid uzaklığı* olarak belirlenmiş, öznitelik seçme işleminde (Şekil 5.11) 15 adet öznitelik seçilmiştir. K-EYK ile yapılan tahmin performansını gösteren ZM incelendiğinde (Tablo 5.5), K-EYK genel tahmin doğruluğu=%96,12, en iyi sınıflandırmanın %96,81 ile BAD niteliğine yapıldığı, en iyi sınıf geri çağrı oranını ise %97,89 ile yine BAD niteliğinin yakaladığı görülmektedir.

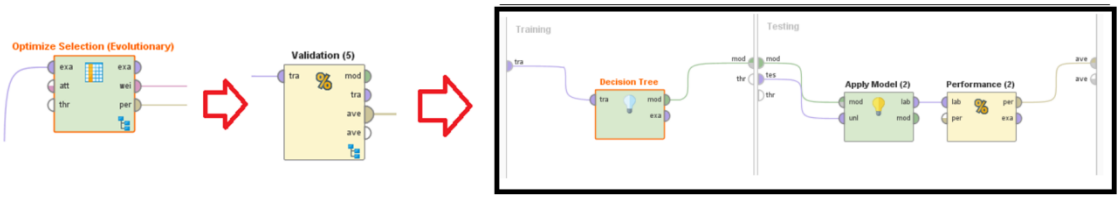


Şekil 5.11 *K-EYK Kullanılmasından Önce GA tarafından Seçilen Öznitelikler*

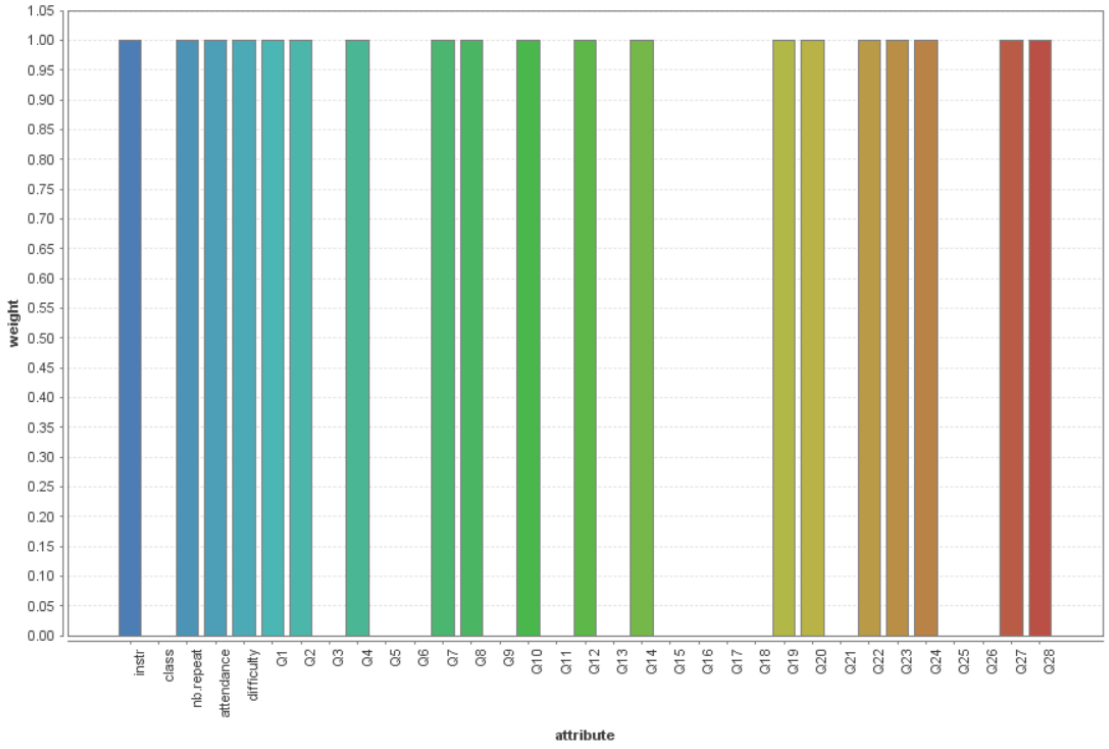
Tablo 5.5 *K-En Yakın Komşuluk Algoritması ile yapılan tahminlerin performansı*

TDP %96,12	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1395**	46	0	%96.81
Tahmin PERFECT	30	**1028**	17	%95.63
Tahmin GOOD	0	20	**374**	%94.92
Sınıf Geriçağrımı	%97.89	%93.97	%95.65	

Karar Ağacı algoritmasıyla yapılan tahminde (Şekil 5.12), *kriter=kazanç oranı, maksimal derinlik=20, güven=0,25, minimal kazanç=0,1, minimal yaprak sayısı=2, minimal ayrım boyutu=4* olarak ayarlanmış, *budama ve ön budama işlemleri* yapılmıştır. Öznitelik indirgeme işleminde GA tarafından 19 öznitelik seçilmiştir (Şekil 5.13). Karar Ağacı ile yapılan tahmin performansını gösteren ZM incelendiğinde (Tablo 5.6), Karar Ağacı genel tahmin doğruluğunun %94,57, en iyi sınıflandırmanın %96,77 ile BAD niteliğine yapıldığı, en iyi sınıf geri çağrı oranına ise %96,63 ile yine BAD niteliğinin sahip olduğu görülmektedir.



Şekil 5.12 *Karar Ağacı Algoritması ile Tahmin Yapılması*

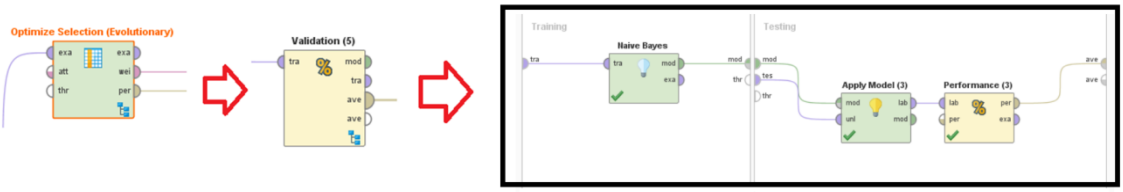


Şekil 5.13 *Karar Ağacı Algoritması Kullanımı Öncesinde Seçilen Öznitelikler*

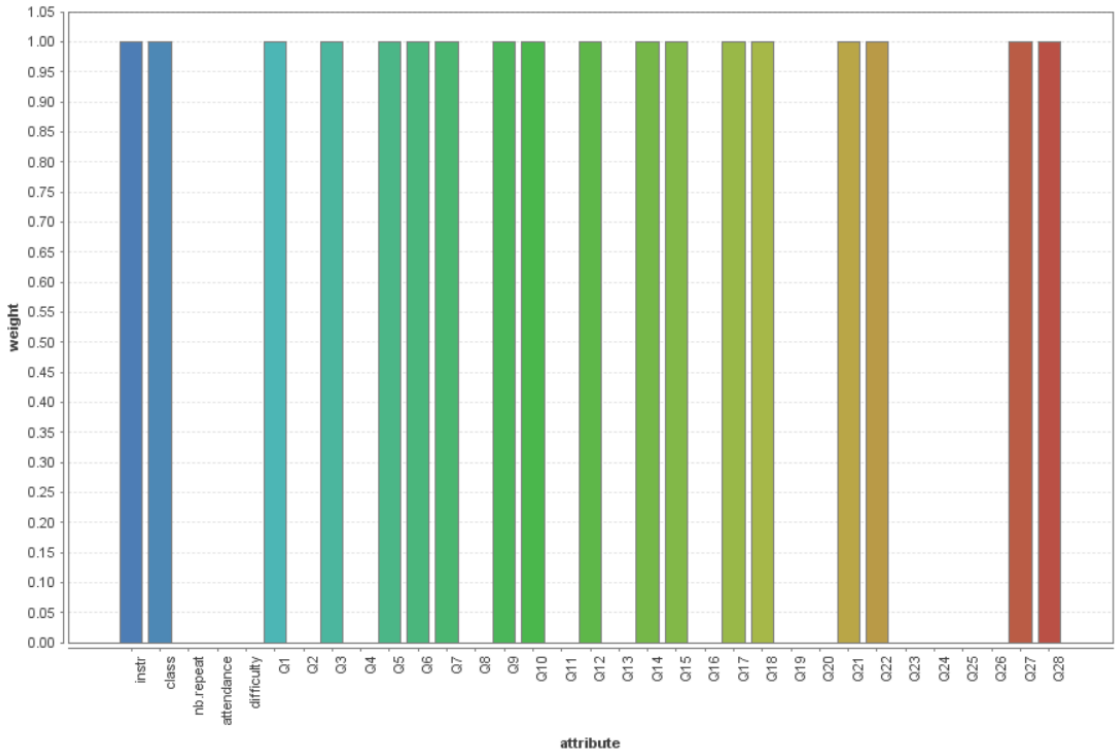
Tablo 5.6 *Karar Ağacı Algoritması ile yapılan tahminlerin performansı*

TDP %94,57	**Doğru BAD**	**Doğru PERFECT**	**Doğru GOOD**	**Sınıf Tahmini**
Tahmin BAD	**1377**	46	0	%96.77
Tahmin PERFECT	48	**1000**	16	%93.98
Tahmin GOOD	0	48	**375**	%88.65
Sınıf Geriçağrımı	%96.63	%91.41	%95.91	

Naive Bayes algoritması ile yapılan tahmin işleminde Eşitlik (4.3.3) te belirtilen Naive Bayes teoremi ve *Laplace Doğrulaması* uygulanmış (Şekil 5.14) ve 18 adet öznitelik seçilmiştir (Şekil 5.15). Naive Bayes ile yapılan tahmin performansını gösteren ZM incelendiğinde (Tablo 5.7), Naive Bayes genel tahmin doğruluğu=%95,43, en iyi sınıflandırmanın %97,72 ile PERFECT niteliğine yapıldığı, en iyi sınıf geri çağrı oranını ise %99,44 ile BAD niteliğinin yakaladığı görülmektedir.



Şekil 5.14 *Naive Bayes Algoritması ile Tahmin Yapılması*

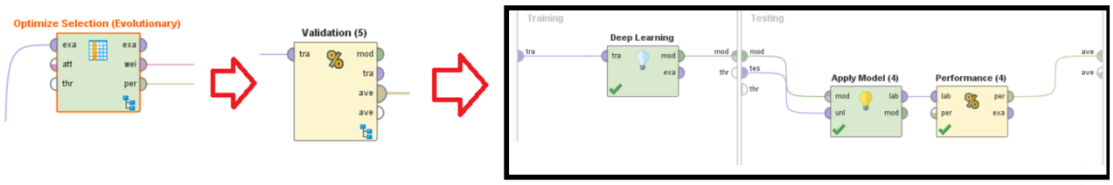


Şekil 5.15 *Naive Bayes Algoritması Kullanımı Öncesinde Seçilen Öznitelikler*

Tablo 5.7 *Naive Bayes Algoritması ile yapılan tahminlerin performansı*

TDP %95,43	Doğru BAD	Doğru PERFECT	Doğru GOOD	Sınıf Tahmini
Tahmin BAD	**1417**	85	0	%94.34
Tahmin PERFECT	8	**984**	15	%97.72
Tahmin GOOD	0	25	**376**	%93.77
Sınıf Geriçağrımı	%99.44	%89.95	%96.16	

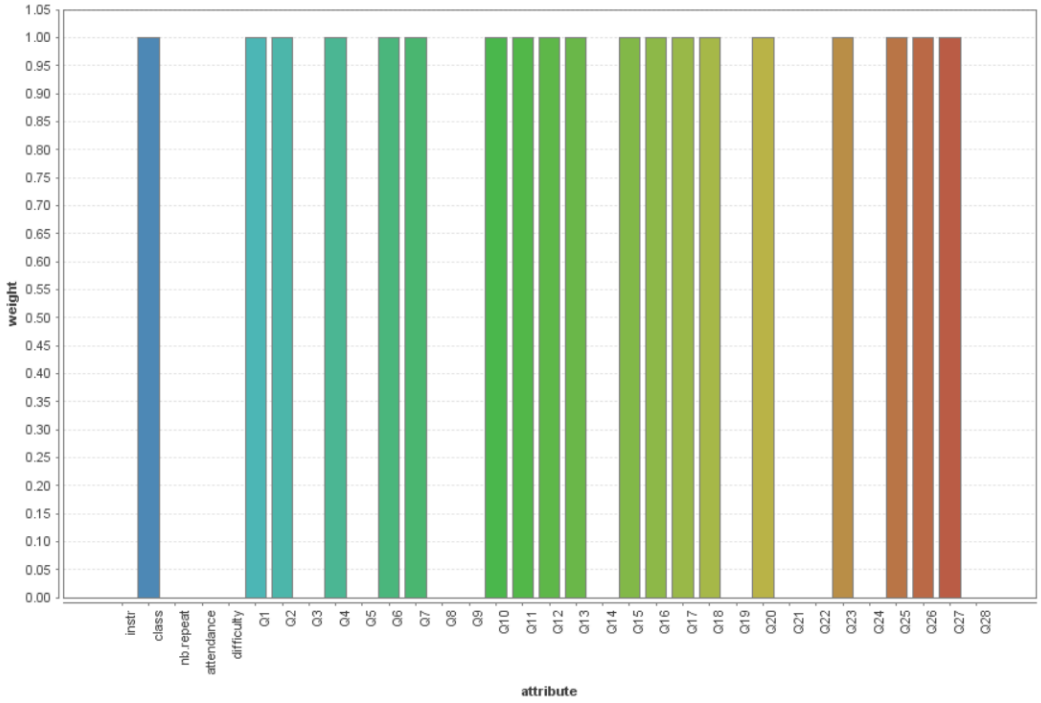
Derin Öğrenme algoritması ile yapılan tahmin işleminde *aktivasyon=doğrultucu, gizli katman sayısı=50, peryot=10* olarak belirlenmiş, öznitelik seçme işleminde (Şekil 5.16) 19 adet öznitelik seçilmiştir (Şekil 5.17). Derin Öğrenme ile yapılan tahmin performansını gösteren ZM incelendiğinde (Tablo 5.8) Derin Öğrenme genel tahmin doğruluğu=%97,70, en iyi sınıflandırmanın %98,19 ile BAD niteliğine yapıldığı, en iyi sınıf geri çağrı oranını ise %99,02 ile BAD niteliğinin yakaladığı görülmektedir.



Şekil 5.16 *Derin Öğrenme Algoritması ile tahmin yapılması*

Tablo 5.8 *Derin Öğrenme Algoritması ile yapılan tahminlerin performansı*

TDP %97,70	Doğru BAD	Doğru PERFECT	Doğru GOOD	Sınıf Tahmini
Tahmin BAD	**1411**	26	0	%98.19
Tahmin PERFECT	14	**1053**	12	%97.59
Tahmin GOOD	0	15	**379**	%9619
Sınıf Geriçağrımı	%99.02	%96.25	%96.93	



Şekil 5.17 *Derin Öğrenme Algoritması Kullanımı Öncesinde Seçilen Öznitelikler*

6. DENEYSEL SONUÇLAR

6.1. Öznitelik Seçme İşlemi Sonuçları

Çalışmada ilk aşama olan öznitelik seçme işleminde, her öznitelik seçme kategorisinden birer adet olmak üzere 3 farklı öznitelik seçme algoritması kullanılmış ve bunların veri setini indirgerken elde ettikleri sonuçlar ile tüm veri setinin işleme alındığında elde edilen bulgular, önceki bölümlerde ortaya konulmuştu.

Öznitelik seçme işlemenin bizlere ne denli bir katkı sağladığı Tablo 6.1 de somut olarak görülmektedir. Hiçbir öznitelik seçme yapılamadan, 33 öznitelik kullanılarak yapılan tahmin işlemi, ancak %90,1'lik bir TDP getirmiştir. Filtreleme algoritmalarından olan Bilgi Kazancı (BK) yöntemi veri setinin öznitelik sayısını 28'e indirgerken, tahmin doğruluğunu %95,84 e çıkarmıştır. Öğrenme tabanlı çalışan Geri Yönlü Eliminasyon Yöntemi (GYE) ise, BK'na göre durumu biraz daha iyileştirerek öznitelik sayısını 26 ya, doğruluk performansını da 96,98'e taşımıştır. Hibrid yapıdaki Genetik Algoritma ise BK ve GYE metotlarına karşı öznitelik indirgeme açısından üstünlük sağlayarak, 19 öznitelik sayısı ile %95,5 lik bir TDP elde edilmiştir.

Tablo 6.1 *Öznitelik Seçme işlemlerinin kıyaslanması*

NİTELİK SEÇME ALGORİTMASI	SEÇİLEN NİTELİK ADEDİ	TAHMİN DOĞRULUK ORANI (ACCURACY) (%)
YOK	33	90,10
Bilgi Kazancı	28	95,84
Geri Yönlü Eliminasyon	26	96,98
Genetik Algoritma	19	95,5

Bölüm 5.2 de belirtildiği üzere, çalışmada kullanılan her 3 öznitelik seçme algoritması da özniteliklere yaklaşımlarında farklı davranışlar sergilemişler, ortak bazı öznitelikler seçtikleri gibi çok farkı öznitelikleri de seçmiş, aynı öznitelikleri birbirlerinden farklı sırada ve farklı ağırlıklar vererek değerlendirmişlerdir.

6.2 Makine Öğrenmesi Algoritmaları Performans Sonuçları

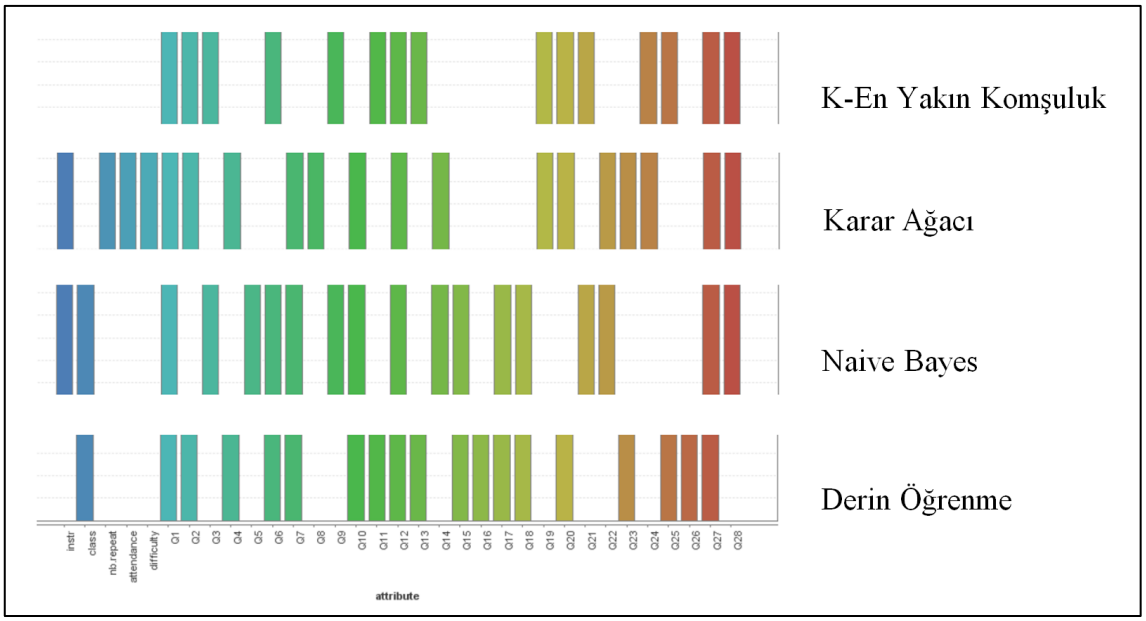
Çalışmanın ikinci aşaması olan, veri setindeki etiketli niteliğin çeşitli makine öğrenmesi yöntemleriyle tahmin edilmesi işlemlerinde ise dört farklı makine öğrenmesi yöntemi kullanılmıştır. Ön işlem olarak kullanılan Genetik Algoritma (GA) çıktılarının önceden öngörülemeyeceği, önceki bölümlerde açıklanmıştır. Tablo 6.2 de görüldüğü üzere GA, her işlemde farklı sayıda öznitelik seçmiştir. Tabloya göre, K-En Yakın Komşuluk algoritması için 15 öznitelik seçilmiş ve sonuçta %96,12 TDP elde edilmiştir. Naive Bayes %95,43 doğruluk oranı 18 öznitelik, Karar Ağacı %94,57 doğruluk oranı 19 öznitelik elde etmiştir. Derin Öğrenme algoritması ise 19 öznitelik seçmesine rağmen %97,70 ile en yüksek doğruluk performansı oranını elde etmiştir. Toplam işlem zamanı açısından bakıldığında ise 3,6 saniye işlem zamanı ile Naive Bayes algoritması diğer algoritmalara tüm işlemleri çok daha hızlı tamamlamıştır.

Toplam İşlem Zamanlarının hesaplanmasında Intel i5-4210U (1.7 GHz) işlemci, 8 GB RAM ve NVDIA GT 820M (2 GB) ekran kartı donanımlarına sahip bir bilgisayar kullanılmıştır.

Tablo 6.2 *Makine Öğrenmesi yöntemlerinin performanslarının kıyaslanması*

MAKİNE ÖĞRENMESİ ALGORİTMASI	SEÇİLEN NİTELİK ADEDİ	TAHMİN DOĞRULUK ORANI (ACCURACY) (%)	TOPLAM İŞLEM ZAMANI (milisaniye)
K-EYK	15	96,12	185640
Karar Ağacı	19	94,57	10703
Naive Bayes	18	95,43	3575
Derin Öğrenme	19	97,70	529774

GA, öznitelik seçme işleminde seçtiği tüm özniteliklere 1 ağırlık verirken, her makine öğrenmesi yöntemi için hem ortak hem de farklı öznitelikler seçebilmiştir (Şekil 6.1).



Şekil 6.1 *Dört Farklı Makine Öğrenmesi yöntemi için GA tarafından seçilen öznitelikler*

7. DEĞERLENDİRME

Bu çalışmaya başlanılması amaçlar kısaca, eğitim kıymetlendirme işleminin mümkün olduğunda doğru, hızlı ve etkin bir şekilde yapılması sağlayabilecek veriler elde etmek olmuştur. Mevcut eğitim sistemlerinin kıymetlendirilmesi yüksek sayıda katılımcıdan toplanan, yüksek sayıda niteliğin değerlendirilmesini esas almakta ve bu da işlem maliyeti, zaman, doğru öznitelik seçimi gibi birçok kısıtı beraberinde getirmekteydi.

Geçen bölümlerde yapılan çalışmalarda, elde edilen en kritik bulgu olarak Tablo 6.1 de gösterilen Genetik Algoritma'nın (GA) başarısı gösterebilir. GA ile mevcut veri setindeki 33 öznitelik arasından yalnızca 19 unu seçerek işlem maliyeti büyük oranda indirilirken, kullanılan öznitelik seçme algoritmaları içinde 95,5'lik oran yakalaması da tahmin başarısının diğer 2 algoritmaya oldukça yakın olduğu görülmüştür.

Makine öğrenmesi algoritmalarının performanslarının kıyaslanmasında ise (Tablo 6.2) %97,70 ile en iyi doğruluğu Derin Öğrenme vermesine karşılık, en hızlı algoritma olarak 3,6 sn ile Naive Bayes algoritması ön plana çıkmaktadır. Derin öğrenme algoritması ise 530 sn ile Naive Bayes algoritmasına göre oldukça yavaş kalmaktadır.

Çalışma; Genetik Algoritma ile öznitelik seçmenin, kullanılan veri seti için en iyi sonucu verdiğini göstermiş, kullanılan öznitelik sayısının ciddi oranda azaltması işlem maliyetini düşürdüğü gibi hem anket hazırlayanlara hem de anket dolduranlara – daha az veri ile daha iyi sonucu elde etme- olanağı sağlamıştır. Tahmin kısmında ise uygulayıcılara alternatifler sunulmaktadır. Tahmin doğruluğunu ilk sıraya koyanlar Derin Öğrenme algoritmasını, düşük işlem maliyetini amaçlayanlar ise Naive Bayes algoritmasını tahmin için kullanarak sonuca gidebilecek, iki farklı hibrid model iki farklı amacı yerine getirmek üzere kullanılabilecektir.

EVM alanında -bu tez de dâhil olmak üzere- bugüne dek yapılan çalışmalarda eğitim sistemi kalitesinin kıymetlendirilmesi işlemi tek yönlü olmuştur. Örneğin, eğitmen performansını tespit için öğrencilere yapılan anketlerde, eğitmenleri hakkındaki görüşleri alınmış ve eğitmen performans kıymetlendirmesi yalnızca bu doneler üzerinde yapılmıştır. Ancak, bu verilerin yeterince objektif olamayacağı açıktır. Bu kıymetlendirme işlemlerindeki objektiflik performansını arttırmak için ise, her bir öğrencinin ilgili eğitmeninden aldığı ilgili dersin yıl sonu notlarının diğer anket veri

setlerine entegre edilmesi ya da, her bir öğrenciye ait "objektiflik katsayısı" benzeri bir çarpanının -öğrencinin geçmiş akademik kariyerinden hareketle- tespit edilerek, bu katsayının ilgili öğrencinin eğitmeni hakkındaki kanaatiyle çarpılması etkili olacaktır. Birleştirilmiş yeni veri setlerinin hem satır bazlı hem de tüm veri seti olarak yeniden değerlendirilmesiyle (verilere *objektiflik optimizasyonu* yapılmasıyla), öğrencilerin eğitmenleri hakkındaki görüşlerin ne kadar gerçekçi olduğu büyük ölçüde ortaya çıkmış olacaktır.

Çalışmada olduğu gibi, birçok eğitmene ait verilerin bulunduğu veri setlerinde, eğitmen başarılarının sıralanması için ilave bir ölçüt olarak *ağırlıklandırılmış eğitmen başarısı* kullanılabilir. Bir başka deyişle öğretmenlerin tüm öğrencilerin kaçta kaçına ders vermiş olduğu hesaplanarak, bu değerin eğitmenin ortalama başarı puanı ile çarpılması ile elde edilecek bu değer, eğitmen başarısına öğrenci sayısının da katkısı ele aldığı için, daha kaliteli bir kıymetlendirme elde edilmesini sağlayacaktır. Böylece daha kalabalık bir kitleye hitap etmesine rağmen başarı elde edebilen bir eğitmenin değeri takdir edilebilecektir.

Gelecekte bu eksende, EVM sistemindeki kaliteyi tespit adına şu çalışmalar da yapılabilir: örneğin lise son sınıf öğrencilerinin öğretmenleri hakkındaki görüşleri anket veri seti + öğrencilerin yıl sonu notları birleştirilmiş bir veri seti ile aynı öğrencilere ait Üniversite giriş sınavı veri setleri karşılaştırılarak, bu iki veri setinin birbiriyle örtüşmesi baz alınarak bu okul ve öğretmenlerinin ulusal eğitimdeki genel başarı düzeyinin neresinde olduklarına (ulusal anlamda gerçekten başarılı olup olmadıklarına) dair bir kanaat elde edilebilir. Ayrıca, eğitim yöneticilerine dair EVM alanında literatürde yapılmış bir çalışma bulunmadığından, eğitim yöneticileri kıymetlendirme işlemi, bu yöneticilerin astı konumunda bulunan bireylerden ya da üstlerinden alınan değerlendirme anket veri setleri ile yapılabilir. Tüm eğitim ekosisteminin sağlıklı olabilmesi adına bu sistemin tepesinde bulunan yöneticilere de diğer bireylere yapıldığı gibi mutlaka performans analizi yapılmalıdır.

KAYNAKÇA

[1] Abaidullah, A. M., Ahmed, N., & Ali, E. (2015). Identifying Hidden Patterns in Students" Feedback through Cluster Analysis. *International Journal of Computer Theory and Engineering*, *7*(1), 16.

[2] Mendes, R., de Voznika, F., Freitas, A., & Nievola, J. (2001). Discovering fuzzy classification rules with genetic programming and co-evolution. *Principles of Data Mining and Knowledge Discovery*, 314-325.

[3] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462.

[4] Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M. R. (2008). Data mining application in higher learning institutions. *Informatics in Education*, *7*(1), 31-54.

[5] Coburn, L. (1984). Student Evaluation of Teacher Performance.

[6] Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of psychology*, *135*(3), 259-268.

[7] Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College teaching*, *49*(1), 26-31.

[8] Andonie, R. (2010). Extreme data mining: Inference from small datasets. *International Journal of Computers Communications & Control*, *5*(3), 280-291.

[9] Chye Koh, H., & Meng Tan, T. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4), 170-178.

[10] McKinney, K. (1997). What do student ratings mean?. In *The National Teaching and Learning Forum* (Vol. 7, No. 1, pp. 5-6).

[11] Timpson, W. W., & Andrew, D. (1997). Rethinking student evaluations and the improvement of teaching: Instruments for change at the University of Queensland. *Studies in Higher Education*, *22*(1), 55-65.

[13] Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business*, *77*(5), 282-289.

[14] Ahmadi, M., Helms, M. M., & Raiszadeh, F. (2001). Business students' perceptions of faculty evaluations. *International Journal of Educational Management*, *15*(1), 12-22.

[15] Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality assurance in Education*, *11*(1), 37-46.

[16] Heiner, C., Heffernan, N., & Barnes, T. (2007, July). Educational data mining. In *Supplementary Proceedings of the 12th International Conference of Artificial Intelligence in Education*.

[17] Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the Sciences of Learning* (pp. 1075-1079). Springer US.

[18] Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American psychologist*, *52*(11), 1187.

[19] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

[20] Minaei-Bidgoli, B., & Punch, W. F. (2003, July). Using genetic algorithms for data mining optimization in an educational web-based system. In *Genetic and evolutionary computation conference* (pp. 2252-2263). Springer, Berlin, Heidelberg.

[21] Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining* (Vol. 32, p. 234).

[22] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

[23] Koutina, M., & Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. In *Artificial Intelligence Applications and Innovations* (pp. 159-168). Springer, Berlin, Heidelberg.

[24] Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications*, *41*(14), 6400-6407.

[25] Sorour, S. E., Goda, K., & Mine, T. (2015, July). Estimation of student performance by considering consecutive lessons. In *Advanced Applied Informatics (IIAI-AAI), 2015 IIAI 4th International Congress on* (pp. 121-126). IEEE.

[26] Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect the students academic performance-Data Mining Approach. *arXiv preprint arXiv:1409.2222*.

[27] Oyedotun, O. K., Tackie, S. N., Olaniyi, E. O., & Khashman, A. (2015). Data Mining of Students' Performance: Turkish Students as a Case Study. *International Journal of Intelligent Systems and Applications*, *7*(9), 20.

[28] Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, J. M. (2015). A Model-Based Approach to Predicting Graduate-Level Performance Using Indicators of Undergraduate-Level Performance. *Journal of Educational Data Mining*, *7*(3), 151-176.

[29] Kentli, F. D., & Sahin, Y. (2011). An SVM approach to predict student performance in manufacturing processes course. *Energy, Edu., Sci. Technol. B*, *3*(4), 535-544.

[30] Agaoglu, M. (2016). Predicting Instructor Performance Using Data Mining Techniques in Higher Education. *IEEE Access*, *4*, 2379-2387.

[31] Ahmed, A. M., Rizaner, A., & Ulusoy, A. H. (2016). Using data mining to predict instructor performance. *Procedia Computer Science*, *102*, 137-142.

[32] Pal, A. K., & Pal, S. (2013). Evaluation of teacher's performance: a data mining approach. *IJCSMC*, *2*(12), 359-369.

[33] Kumar, V., & Chadha, A. (2012). Mining association rules in student's assessment data. *International Journal of Computer Science Issues*, *9*(5), 211-216.

[34] Punlumjeak, W., Rachburee, N., & Arunrerk, J. (2017). Big Data Analytics: Student Performance Prediction Using Feature Selection and Machine Learning on Microsoft Azure Platform. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *9*(1-4), 113-117.

[35] Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2017). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 1-17.

[36] Sanjay, S. S., & Keshav, B. B. (2017). Teacher's Performance Analyzer. *IJETT*, *1*(1).

[37] İnternet: https://machinelearningmastery.com/an-introduction-to-feature-selection/ Erişim tarihi: 30.11.2017

[38] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

[39] Ding, S. (2009, November). Feature selection based F-score and ACO algorithm in support vector machine. In *Knowledge Acquisition and Modeling, 2009. KAM'09. Second International Symposium on* (Vol. 1, pp. 19-23). IEEE.

[40] Lee, S., Park, Y. T., & d'Auriol, B. J. (2012). A novel feature selection method based on normalized mutual information. *Applied Intelligence*, *37*(1), 100-120.

[41] Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert systems with applications*, *36*(3), 6843-6853.

[42] Kabir, M. M., Shahjahan, M., & Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing*, *74*(17), 2914-2928.

[43] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[44] Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, *2*(2), 88-105.

[45] İnternet: http://bilgisayarkavramlari.sadievrenseker.com/2008/11/17/knn-k-nearest-neighborhood-en-yakin-k-komsu/ Erişim tarihi: 14.12.2017

[46] KAŞIKÇI, T., & GÖKÇEN, H. (2013). Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi. Bilişim Teknolojileri Dergisi, 7(1).

[47] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[48] İnternet: https://cdn.edureka.co/blog/wp-content/uploads/2017/05/Deep-Neural-Network-What-is-Deep-Learning-Edureka.png Erişim tarihi: 02.02.2018

[49] İnternet: http://www.derinogrenme.com/2015/07/21/derin-ogrenme-deep-learning-nedir/

[50] Karahan, Ş., & Akgül, Y. S. (2016, May), Eye detection by using deep learning, In Signal Processing and Communication Application Conference (SIU), 2016 24th (pp. 2145-2148), IEEE.

[51] İnternet: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/deep_learning.html Erişim tarihi: 14.11.2017

[52] İnternet: https://rapidminer.com/ Erişim tarihi: 10.09.2017

[53] İnternet: http://archive.ics.uci.edu/ml/datasets/turkiye+student+evaluation Erişim tarihi: 14.10.2017

[54] İnternet: http://www.derinogrenme.com/2015/07/21/derin-ogrenme-deep-learning-nedir/ Erişim tarihi: 15.01.2018

[55] İnternet: https://www.edureka.co/blog/what-is-deep-learning Erişim tarihi: 10.01.2018

[56] İnternet: http://komhedos.com/sezgisel-en-iyileme-yontemleri/ Erişim tarihi: 03.01.2018

[57] İnternet: http://bilgisayarkavramlari.sadievrenseker.com/2008/11/17/knn-k-nearest-neighborhood-en-yakin-k-komsu/ Erişim tarihi: 02.01.2018

[58] İnternet: http://yazilimagiris.com/2017/11/karar-agaci/ Erişim tarihi: 11.12.2017

ÖZGEÇMİŞ

Adı Soyadı: Fatih ÇİFÇİ

Doğum Yeri ve Tarihi: Çarşamba/SAMSUN, 1982

E-mail: fatihcifci@anadolu.edu.tr

Eğitim

- Elektrik-Elektronik Mühendisliği, Kırıkkale Üniversitesi, Kırıkkale, Temmuz 2008
- Bilgisayar Mühendisliği Anabilim Dalı, Anadolu Üniversitesi, Eskişehir, Türkiye, Mart 2018

Meslek

• Mobil Haberleşme Sistemleri Yöneticisi, Eskişehir, 2011

• Bilgi Sistemleri Ağ Yöneticisi, Eskişehir, Eskişehir, 2012

• Aviyonik Destek Sistemleri Yöneticisi, Amasya, 2017