DENGESİZ BİR DİYABET VERİ SETİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİ KULLANARAK DİYABET HASTALIĞININ TEŞHİSİ

İsmail Buğra BÖLÜKBAŞI

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ BİR DİYABET VERİ SETİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİ KULLANARAK DİYABET HASTALIĞININ TEŞHİSİ

İsmail Buğra BÖLÜKBAŞI
0000-0002-9405-0900

Prof. Dr. Betül YAĞMAHAN
(Danışman)

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2023

TEZ ONAYI

İsmail Buğra BÖLÜKBAŞI tarafından hazırlanan "Dengesiz Bir Diyabet Veri Setinde Makine Öğrenmesi Yöntemlerini Kullanarak Diyabet Hastalığının Teşhisi" adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Bursa Uludağ Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman: Prof. Dr. Betül YAĞMAHAN

Başkan : Prof. Dr. Betül YAĞMAHAN
0000-0003-1744-3062
Bursa Uludağ Üniversitesi,
Mühendislik Fakültesi,
Endüstri Mühendisliği Anabilim Dalı

Üye : Doç. Dr. Duygu YILMAZ EROĞLU
0000-0002-7730-2707
Bursa Uludağ Üniversitesi,
Mühendislik Fakültesi,
Endüstri Mühendisliği Anabilim Dalı

Üye : Doç. Dr. Aytaç YILDIZ
0000-0002-0729-633X
Bursa Teknik Üniversitesi,
Mühendislik ve Doğa Bilimleri Fakültesi,
Endüstri Mühendisliği Anabilim Dalı

Yukarıdaki sonucu onaylarım

Prof. Dr. Hüseyin Aksel EREN
Enstitü Müdürü
../../……

B.U.Ü. Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

16/01/2023

İsmail Buğra BÖLÜKBAŞI

TEZ YAYINLANMA
FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezin/raporun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma izni Bursa Uludağ Üniversitesi'ne aittir. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet hakları ile tezin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları tarafımıza ait olacaktır. Tezde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığını ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederiz.

Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge**" kapsamında, yönerge tarafından belirtilen kısıtlamalar olmadığı takdirde tezin YÖK Ulusal Tez Merkezi / B.U.Ü. Kütüphanesi Açık Erişim Sistemi ve üye olunan diğer veri tabanlarının (Proquest veri tabanı gibi) erişimine açılması uygundur.

Prof. Dr. Betül YAĞMAHAN	İsmail Buğra BÖLÜKBAŞI
16/01/2023	16/01/2023

ÖZET

Yüksek Lisans Tezi

DENGESİZ BİR DİYABET VERİ SETİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİ KULLANARAK DİYABET HASTALIĞININ TEŞHİSİ

İsmail Buğra BÖLÜKBAŞI

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Betül YAĞMAHAN

Dünya Sağlık Örgütü (DSÖ) verilerine göre diyabet hastalığına sahip kişi sayısı son zamanlarda ciddi bir artış göstermektedir. Diyabet hastalığı eğer gerekli tedbirler alınmazsa ilerleyen zamanlarda vücutta kalıcı hasarlara yol açan, hatta kişinin ölümüne neden olabilecek çok önemli bir hastalıktır. Tüm bu sebeplerden dolayı diyabet hastalığının erken ve doğru şekilde tespiti için tıp dünyasındaki çalışmaların hızla arttığı görülmektedir. Bu çalışmada tip-2 diyabet hastalığının teşhisi için gerçek hayattaki bir veri setinin analizinde, makine öğrenimi yöntemlerinden biri olan sınıflandırma yöntemi kullanılmıştır. Çalışmanın amacı, iki farklı veri bölme tekniği, üç farklı yeniden örnekleme tekniği ve altı farklı sınıflandırma yöntemi kullanarak diyabet teşhisinin en doğru şekilde sınıflandırılmasıdır. Bu çalışmada sınıflandırma modelleri KNIME programında oluşturulmuştur. Veri seti eğitim ve test verisi olarak ayrıştırılırken yüzdesel bölme (%70-30) ve k-katlı (k=5) çapraz doğrulama teknikleri kullanılmıştır. Diyabet veri setindeki sınıf dengesizliğinin giderilmesi için rastgele örneklem azaltma (RUS), rastgele aşırı örnekleme (ROS) ve sentetik azınlık aşırı örnekleme (SMOTE) tekniklerinden yararlanılmıştır. Çalışmada kullanılan sınıflandırma yöntemleri lojistik regresyon (LR), naive bayes (NB), k-en yakın komşu (k-EYK), C4.5 algoritması, rastgele orman (RO) ve çok katmanlı algılayıcıdır (ÇKA). Veri bölme tekniği, yeniden örnekleme tekniği ve sınıflandırma yöntemleri ile yapılan kombinasyonlar sonucunda 48 farklı senaryo incelenmiştir. Tüm senaryolar doğruluk, kesinlik, duyarlılık, ortalama F-ölçütü, kappa istatistiği ve AUC değeri ölçütlerine göre karşılaştırılmıştır. Yapılan deneysel çalışmalar sonucunda yüzdesel bölme ile oluşturulan senaryolar arasında en iyi sonucu %99,26 doğruluk değeriyle RUS-RO, en kötü sonucu ise %80,74 doğruluk değeriyle SMOTE-k-EYK vermiştir. K-katlı çapraz doğrulama ile oluşturulan senaryolar arasında en iyi sonucu %97,55 doğruluk değeri ile RUS-C4.5, ROS-RO ve SMOTE-RO, en kötü sonucu ise %78,62 doğruluk değeriyle RUS-EYK vermiştir.

Anahtar Kelimeler: Diyabet Teşhisi, Tip-2 Diyabet, Makine Öğrenmesi, Sınıflandırma, Dengesiz Veri Seti, Yeniden Örnekleme Yöntemleri
2023, ix + 91 sayfa.

ABSTRACT

MSc Thesis

DIAGNOSIS OF DIABETES DISEASE USING MACHINE LEARNING METHODS IN AN IMBALANCED DIABETES DATASET

İsmail Buğra BÖLÜKBAŞI

Bursa Uludağ University
Graduate School of Natural and Applied Sciences
Department of Industrial Engineering

Supervisor: Prof. Dr. Betül YAĞMAHAN

According to the data of the World Health Organization (WHO), the number of people with diabetes has increased significantly in recent years. Diabetes is a very important disease that can lead to permanent damage to the body and even death of the person in the future if the necessary precautions are not taken. For all these reasons, it is seen that studies in the medical world are increasing rapidly for the early and accurate diagnosis of diabetes. In this study, the classification method, one of the machine learning methods, was used in analyzing a real-life dataset for the purpose of diagnosing type-2 diabetes. The aim of the study is the most accurate classification of the diagnosis of diabetes using two different data-splitting techniques, three different resampling techniques, and six different classification methods. In this study, classification models were created in the software KNIME. Percentage split (70-30%) and k-fold (k=5) cross-validation techniques were used when separating the data set as training and test data. Random undersampling (RUS), random oversampling (ROS), and synthetic minority oversampling (SMOTE) techniques were used to eliminate the class imbalance in the diabetes dataset. The classification methods used in the study are logistic regression (LR), naive bayes (NB), k-nearest neighbor (k-NN), C4.5 algorithm, random forest (RF), and multilayer perceptron (MLP). As a result of combinations with data-splitting techniques, resampling techniques, and classification methods, 48 different scenarios were examined. All scenarios were compared according to criteria of accuracy, precision, recall, average F-measure, kappa statistic, and AUC value. As a result of the experimental studies, among the scenarios created with percentage split, RUS-RF gave the best result with an accuracy value of 99.26%, and SMOTE-k-NN gave the worst result with an accuracy value of 80.74%. Among the scenarios created with k-fold cross-validation, RUS-C4.5, ROS-RF, and SMOTE-RF gave the best result with an accuracy value of 97.55%, and RUS-k-NN gave the worst result with an accuracy value of 78.62%.

Key words: Diabetes Diagnosis, Type-2 Diabetes, Machine Learning, Classification, Imbalanced Dataset, Resampling Methods
2023, ix + 91 pages.

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmam boyunca akademik hayata dair tecrübelerini benden esirgemeyen ve bana vakit ayıran kıymetli danışman hocam Prof. Dr. Betül YAĞMAHAN'a içtenlikle teşekkür eder, saygılarımı sunarım.

Doğduğum günden bugüne, bana güvenen ve daima yanımda olan aileme teşekkürlerimi sunarım.

İsmail Buğra BÖLÜKBAŞI
16/01/2023

İÇİNDEKİLER

Sayfa

SİMGELER ve KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
ADASYN	Uyarlanabilir Sentetik Örnekleme (Adaptive Synthetic Sampling Approach)
AUC	ROC Eğrisinin Altında Kalan Alan (Area under the ROC Curve)
BA	Bayes Ağı (Bayesian Network)
BSMOTE	Sınıra Yakın Olan Sentetik Azınlık Aşırı Örnekleme (Borderline Synthetic Minority Oversampling Technique)
CBOS	Kümeleme Tabanlı Aşırı Örnekleme (Cluster-based Oversampling)
CS	Maliyete Duyarlı Modeller (Cost-Sensitive Models)
ÇKA	Çok Katmanlı Algılayıcı (Multi Layer Perceptron)
DN	Negatif Sınıfta Doğru Sınıflandırılan Örnek Sayısı
DÖ	Derin Öğrenme (Deep Learning)
DP	Pozitif Sınıfta Doğru Sınıflandırılan Örnek Sayısı
DSO	Doğru Sınıflandırma Oranı (Accuracy)
DVM	Destek Vektör Makineleri (Support Vector Machines)
ENN	Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbor)
FSMOTE	Uzak Komşu Sentetik Azınlık Aşırı Örnekleme (Farthest Synthetic Minority Oversampling Technique)
GA	Gradyan Arttırma (Gradient Boosting)
GAA	Gradyan Arttırımlı Ağaçlar (Gradient Boosted Trees)
GNB	Gauss Naive Bayes
KA	Karar Ağacı (Decision Tree)
k-EYK	k-En Yakın Komşu (k-Nearest Neighbors)
Kİ	Kappa İstatistiği
LR	Lojistik Regresyon (Logistic Regression)
MWMOTE	Çoğunluk Ağırlıklı Azınlık Aşırı Örnekleme (Majority Weighted Minority Oversampling Technique)
NB	Naive Bayes
OCC	Tek Sınıflı Sınıflandırma (One-Class Classification)
OSS	Tek Taraflı Sınıflandırma (One-Sided Selection)
PRC	Kesinlik-Duyarlılık Eğrisi (Precision-Recall Curve)
RA	Rastgele Ağaç (Random Tree)
RO	Rastgele Orman (Random Forest)
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristic)
ROS	Rastgele Aşırı Örnekleme (Random Oversampling)
RUS	Rastgele Örneklem Azaltma (Random Undersampling)
SL-SMOTE	Güven Seviyeli Sentetik Azınlık Aşırı Örnekleme (Safe-Level Synthetic Minority Oversampling Technique)

SMOTE	Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Oversampling Technique)
SVM-SMOTE	Destek Vektör Makinesiyle Sentetik Azınlık Aşırı Örnekleme
UCI	Kaliforniya Üniversitesi, Irvine (UCI)
XGBoost	Ekstrem Gradyan Arttırma (XGBoost)
YN	Negatif Sınıfta Yanlış Sınıflandırılan Örnek Sayısı
YP	Pozitif Sınıfta Yanlış Sınıflandırılan Örnek Sayısı
YSA	Yapay Sinir Ağı (Artificial Neural Network)

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

1. GİRİŞ

Diyabet hastalığı, pankreasın herhangi bir sebepten dolayı insülini üretemediğinde veya vücudun, pankreas tarafından üretilen insülini gerekli miktarda kullanamadığında ortaya çıkan kronik bir hastalıktır (IDF, 2022). Diyabet başta gözler, böbrekler, kalp, kan damarları ve sinirler olmak üzere çeşitli dokularda kronik hasara ve işlev bozukluğuna yol açabilmektedir (Krasteva, Panov, Krasteva, Kisselova ve Krastev, 2011). Bu hastalık ile ilgili komplikasyonlar zamanla ölümcül sonuçlar da doğurabilmektedir. Dünya Sağlık Örgütü'ne göre diyabet hastalığı, dünyadaki önde gelen ölüm nedenlerinden biridir (WHO, 2022a). Uluslararası Diyabet Federasyonu'nun verilerine göre diyabet hastalığı 2021 yılında 6,7 milyon insanın ölümüne sebep olmuştur (IDFDA, 2022).

Diyabet hastalığı yaygınlığı son birkaç on yılda hızla artış göstermektedir. (WHO, 2022a). Dünya genelinde 537 milyon erişkin (20-79 yaşlarında) diyabet hastalığı ile yaşamaktadır. 2021 yılına ait dünya geneli diyabet hasta sayısının dağılımı Şekil 1.1'de gösterilmiştir. Gerekli önlemler alınmaz ise bu hasta sayısı 2030 yılına kadar 643 milyona, 2045 yılına kadar ise 783 milyona kadar ulaşacağı tahmin edilmektedir (IDFDA, 2022).

Şekil 1.1. Dünya'da 2021 yılındaki diyabet hastası sayısının dağılımı (IDFDA, 2022)

Diyabet hastalığının teşhis edilmesi süreci, hasta olan kişinin hastanedeki ilgili birimlere başvurması ve uzman hekim tarafından yapılan tıbbi tetkikler sonrasında diyabet tanısının konması şeklindedir. Ancak son zamanlarda makine öğrenmesi yöntemleri ile ilgili çalışmalar diyabet hastalığının teşhis edilmesi konusunda oldukça başarılı performanslar göstermektedir.

Bu çalışmanın amacı, Türkiye'deki bir devlet hastanesinde iç hastalıkları, endokrin ve metabolizma hastalıkları biriminde testler yaptırıp Tip-2 diyabet tanısı alan ve almayan hastaların sınıflandırılmasında geleneksel sınıflandırma algoritmalarının performanslarının incelenmesidir. Çalışmada kullanılan makine öğrenmesi sınıflandırma algoritmaları; lojistik regresyon (LR), naive bayes (NB), k-en yakın komşu (k-EYK), C4.5, rastgele orman (RO) ve çok katmanlı algılayıcıdır (ÇKA). Sınıflandırma modellerinin performanslarının karşılaştırılmasında kullanılan ölçütler ise Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F-ölçütü (F-Measure), Kappa İstatistiği ve AUC değerleridir.

Çalışmanın ikinci bölümünde diyabet hastalığı ve sınıf dengesizliği kavramları, makine öğrenmesi türleri, literatürde diyabet hastalığı teşhisi ve sınıf dengesizliğiyle ilgili çalışmalar hakkında bilgiler verilmiştir. Bölüm 3'te veri seti hakkında bilgilere, veri önişleme adımlarına, eğitim ve test veri setlerinin oluşturulmasına, sınıflandırma modellerin oluşturulmasına ve modellerin performans ölçütlerine yer verilmiştir. Dördüncü bölümde ise orijinal eğitim veri seti ve üç farklı yeniden örnekleme yöntemiyle oluşturulan eğitim veri setleriyle eğitilen sınıflandırma modellerin performansları karşılaştırılmış ve yorumlanmıştır. Bölüm 5'te yapılan çalışma hakkında en iyi sonuçları veren kombinasyonlar hakkında bilgiler verilmiş ve gelecek çalışmalar için önerilerde bulunulmuştur.

2. KURAMSAL TEMELLER ve KAYNAK ARAŞTIRMASI

Bu bölümde diyabet hastalığı, sınıf dengesizliği, makine öğrenmesi türleri, çalışmada kullanılan makine öğrenmesi algoritmaları ile diyabet teşhisi ve sınıf dengesizliğine yönelik literatür araştırması hakkında bilgiler verilecektir.

2.1. Diyabet Hastalığı

Yaygın olarak diyabet olarak adlandırılan diyabet hastalığı, vücudun yiyecekleri enerjiye nasıl dönüştürdüğünü etkileyen kronik bir hastalıktır. Dünya çapında en önemli 10 ölüm nedeninden biri olan diyabet; yaşam tarzı, psikososyal, tıbbi durumlar, demografik ve genetik risk faktörleri arasındaki etkileşimle ilişkilidir (Ismail, Materwala, Tayefi, Ngo ve Karduck, 2022).

Diyabet hastalığı Tip-1, Tip-2 ve Gebelik (Gestasyonel) olmak üzere üç başlıkta incelenebilir.

2.1.1. Tip-1 diyabet

Tip-1 diyabet, vücudun insülin adı verilen bir hormonu üretemediği için kan şekeri (şeker) seviyesinin çok yüksek olduğu ciddi bir durumdur. Bunun nedeni, vücudun pankreastaki insülini üreten hücrelere saldırmasıdır. Bu sebepten dolayı pankreas hiç insülin üretemez. İnsülin hormonu, kanın içerisindeki glukozun hücrelere girmesini ve vücudu beslemesini sağlamak gibi çok önemli bir göreve sahiptir. Tip-1 diyabeti olan bir kişinin vücudu yiyecek ve içeceklerdeki karbonhidratı parçalar ve onu glukoza dönüştürebilir. Ancak glukoz kan dolaşımına girdiğinde, vücudun hücrelerine girmesine izin verecek insülin yoktur. Bu da kan dolaşımında giderek daha fazla glukoz birikmesine ve yüksek kan şekeri seviyelerine yol açmaktadır (DUK, 2022).

2.1.2. Tip-2 diyabet

Diyabetin en yaygın türü Tip-2 diyabet olarak bilinmektedir. Tip-2 diyabette vücutta yeterince insülin yapılamamakta veya vücut insülini iyi bir şekilde kullanamamaktadır. Kanda çok fazla biriken glukoz vücuttaki hücrelere yeterince ulaşamaması sonucunda Tip-2 diyabet hastalığı oluşmaktadır (NIDDK, 2022).

Araştırmalar Tip-2 diyabet hastalığının tedavisindeki başlıca yöntemlerin sağlıklı beslenme, düzenli fiziksel aktivite, sigara içmeme ve sağlıklı bir vücut ağırlığının korunmasının gerektiğini göstermektedir. İlerleyen zamanlarda sağlıklı bir yaşam tarzı kan şekeri düzeylerini kontrol altında tutmak için yeterli olmayabilir ve Tip-2 diyabetli kişilerin ağızdan ilaç alması gerekebilmektedir. Ağızdan alınan ilaçlar kan şekeri düzeylerini kontrol etmek için yeterli olmadığında ise Tip-2 diyabetli kişilerde insülin enjeksiyonları gerekebilmektedir (IDF, 2022).

2.1.3. Gebelik (Gestasyonel) diyabeti

Gebelik diyabeti (GD), gebeliğin en sık görülen tıbbi komplikasyonlarından bir tanesidir ve küresel olarak prevalansı artmaktadır. GD, anne ve çocuklar için gelecek dönemlerde tip-2 diyabet, obezite ve kardiyovasküler hastalıklar için ciddi bir risk faktörüdür. İlk defa gebelik sırasında ortaya çıkan bu hastalığın erken tespit edilmesi halinde bebeğin erken doğum veya ölü doğum riski önemli ölçüde azalmaktadır (Sweeting, Wong, Murphy ve Ross, 2022).

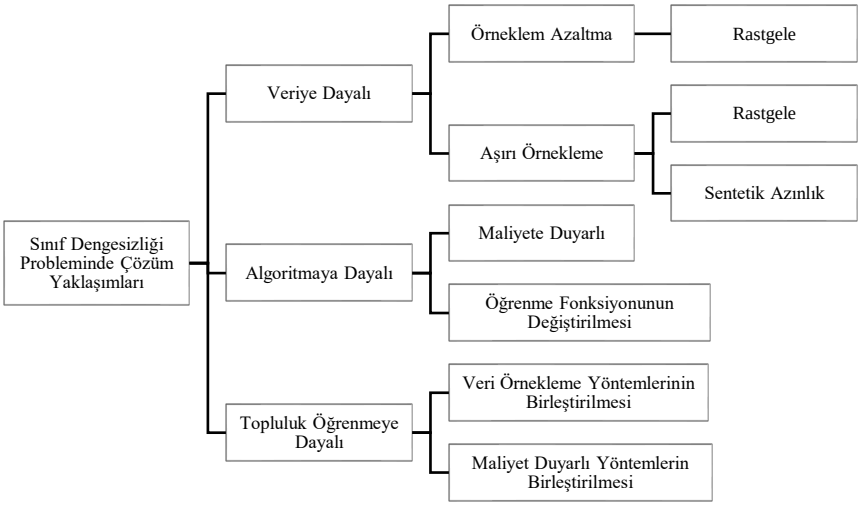
2.2. Sınıf Dengesizliği

Teknik açıdan bakıldığında sınıfları arasında eşit olmayan bir dağılım gösteren herhangi bir veri kümesi dengesiz olarak kabul edilebilir (He ve Garcia, 2009). Literatürde yapılan çalışmalar incelediğinde; bankacılık (kredi kartı dolandırılıcığı), tıp (hastalık teşhisi), telekomünikasyon (kayıp müşteri analizi) gibi birçok alanda sınıf dengesizliğine sahip veri kümeleri ile karşılaşılmaktadır.

Sınıf dengesizliğine bir örnek vermek gerekirse, elimizde ikili sınıf değişkenine (hasta / hasta değil) sahip; 980 kişi hasta, 20 kişi hasta değil olmak üzere 1000 adetlik bir sağlık veri seti olsun. Oran açısından bakıldığında hasta kişilerin sayısı, hasta olmayan kişilerin sayısına göre bariz bir üstünlük kurduğu görülmektedir. Veri setleri içerisinde sınıflardan birinin çok sayıda, diğer sınıf az sayıda veri ile temsil ediliyor ise sınıf dengesizliğinden bahsetmek mümkündür. Sınıf dengesizliği ise, sınıflandırma modellerinin doğru sınıflandırma performansını olumsuz yönde etkileyen bir durumdur (Aydın, 2020).

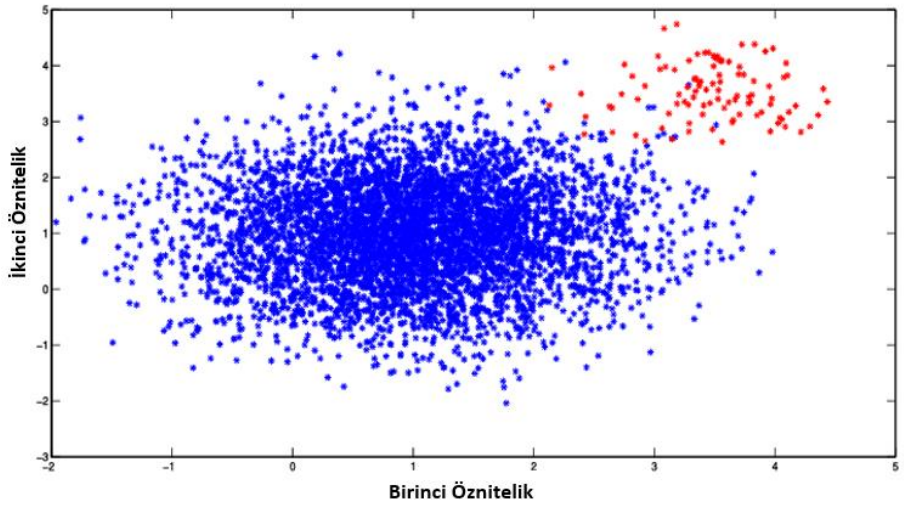
Sınıf dengesizliğinin çözümü için literatürde kullanılan yöntemleri; veri seviyelerine dayalı, algoritmalara dayalı ve topluluk öğrenmesine dayalı olarak üç başlıkta incelemek

mümkündür. Şekil 2.1'de sınıf dengesizliğinin çözümü için kullanılan yöntemlerin bir görseli verilmiştir (Devi, Biswas ve Purkayastha, 2020).



Şekil 2.1. Sınıf dengesizliği problemi çözüm yaklaşımları (Devi ve diğerleri, 2020)

Şekil 2.2'de sınıf dengesizliği probleminin bir örneği gösterilmiştir (Kaur, Pannu ve Malhi, 2019). Mavi renkli sınıf değerinin kırmızı renkli sınıf değerinden belirgin şekilde fazla olduğu gözükmektedir. Bu da veri setinde bir sınıf dengesizliği olduğunun bir göstergesidir.



Şekil 2.2. Sınıf dengesizliği problemi örneği (Kaur ve diğerleri, 2019)

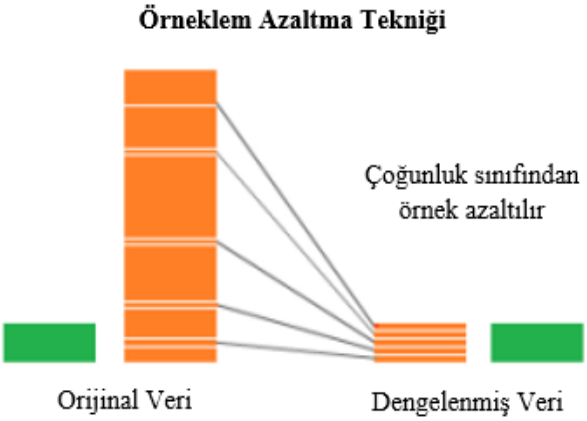
Literatürde sınıf dengesizliğini giderebilmek adına yeniden örnekleme yöntemleri sıklıkla kullanılan teknikler arasındadır. Bu yöntemler içinde dengesiz veri durumunun olumsuz etkisini azaltmak için örneklem azaltma (undersampling), aşırı örnekleme (oversampling) ve sentetik azınlık aşırı örnekleme (SMOTE) yöntemleri ön plana çıkmaktadır. Her yöntemin kendine göre avantajları ve dezavantajları mevcuttur. Örneğin örneklem azaltma yönteminde kullanılacak veri setindeki çoğunluk sınıfına ait veriler azaltılırken bilgi kaybına yol açabilmektedir. Aşırı örnekleme yönteminde ise azınlık sınıfına ait verilerin sayısı arttırılırken modelde aşırı uyum problemine yol açabilmektedir. Eğitim süreleri ve bellek açısından bir değerlendirme yapılacak olursa, örneklem azaltma yönteminde veri sayısı azaltıldığı için modelin eğitim süresi ciddi derecede azalmakta ve daha az bellek işgal edecektir. Aşırı örnekleme yönteminde ise alt örnekleme yönteminin tam tersi olarak veri sayısı arttığı için eğitim süresi artmakta ve daha fazla bellek kullanımına neden olmaktadır (Turhan, Özkan, Yürekli, Suner ve Doğu, 2020).

Bu çalışmada, rastgele örneklem azaltma (RUS), rastgele aşırı örnekleme (ROS) ve SMOTE yöntemleri kullanılmıştır.

2.2.1. Rastgele Örneklem Azaltma Tekniği (RUS)

Rastgele örneklem azaltma, yeniden örneklemeye yönelik basit bir yaklaşımdır. Eğitim setindeki çoğunluk sınıfına ait veriler, azınlık ve çoğunluk sınıfı arasındaki oran istenilen düzeye gelene kadar rastgele elenir. Teorik olarak, rastgele örneklem azaltma ile ilgili sorunlardan biri, çoğunluk sınıfı hakkında hangi bilgilerin atıldığını kontrol edememesidir. Özellikle azınlık ve çoğunluk sınıfı arasındaki karar sınırına ilişkin çok önemli bilgiler ortadan kaldırılabilir. Basit bir yöntem olmasına rağmen, rastgele örneklem azaltma deneysel olarak en etkili yeniden örnekleme yöntemlerinden biri olduğu gösterilmiştir (Liu, 2004).

Şekil 2.3'te örneklem azaltma tekniğinin görsel olarak verilmiştir (Mohammed, Rawashdeh ve Abdullah, 2020). Burada yeşil renkli ve turuncu renkli olmak üzere iki farklı sınıf mevcuttur. Örneklem azaltma tekniğine göre işlem yapıldığında turuncu renkli sınıftaki (çoğunluk sınıfı) veri sayısı yeşil renkli sınıftaki (azınlık sınıfı) veri sayısına eşit olması için örneklerin azaltılması durumu söz konusudur.

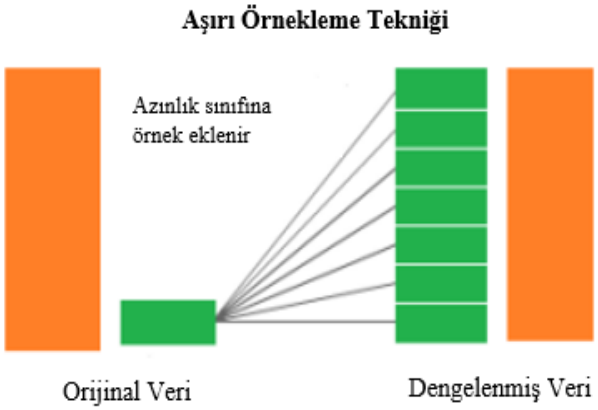


Şekil 2.3. Rastgele örneklem azaltma örneği (Mohammed ve diğerleri, 2020)

2.2.2. Rastgele Aşırı Örnekleme Tekniği (ROS)

Rastgele örneklem azaltma gibi, rastgele aşırı örnekleme de yeniden örnekleme için basit ama etkili bir yaklaşımdır. Burada azınlık sınıfından üyeler rastgele seçilir; rastgele seçilen bu üyeler daha sonra çoğaltılarak yeni eğitim setine eklenir. Rastgele aşırı örnekleme yaparken bir noktaya dikkat etmek gerekir. Yeniden örnekleme yapılırken üyeler yeni eğitim veri setinden değil orijinal veri setinden rastgele seçilmelidir. Aksi hâlde seçimin rastgeleliğinde sapma durumu söz konusu olacaktır (Liu, 2004).

Şekil 2.4'te aşırı örnekleme tekniğinin görsel olarak ifade edilmiş hâli verilmiştir (Mohammed ve diğerleri, 2020). Burada yeşil renkli ve turuncu renkli olmak üzere iki farklı sınıf mevcuttur. Aşırı örnekleme tekniğine göre işlem yapıldığında yeşil renkli sınıftaki (azınlık sınıfı) veri sayısı turuncu renkli sınıftaki (çoğunluk sınıfı) veri sayısına eşit olması için örnekler üretilerek sınıflardaki veri sayıları dengelenmektedir.



Şekil 2.4. Rastgele aşırı örnekleme örneği (Mohammed ve diğerleri, 2020)

2.2.3. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE)

Sentetik Azınlık Örneklem Arttırma Tekniği (SMOTE), 2002 yılında Chawla ve arkadaşları tarafından önerilen bir veri önişleme tekniğidir. Bu yöntemde azınlık sınıfının, yerine koyma ile aşırı örneklemden ziyade "sentetik" örnekler oluşturarak aşırı örneklendiği bir aşırı örnekleme yöntemi olduğunu belirtmişlerdir (Chawla, Bowyer, Hall ve Kegelmeyer, 2002).

SMOTE, dengesiz sınıflandırmada araştırma topluluğu için öncü olan bir ön işleme tekniği olmuştur. Literatürde kullanılmaya başlandığı zamandan bu yana, farklı senaryolar altında performansını artırmak için birçok uzantı ve alternatif önerilmiştir. Popülerliği ve etkisi nedeniyle SMOTE, makine öğrenimi ve veri madenciliğinde en etkili veri ön işleme/örnekleme algoritmalarından biri olarak kabul edilir (Fernández, García, Herrera ve Chawla, 2018).

Bu yöntemin temel mantığında ise k-EYK algoritması kullanılır ve rastgele seçilen bir azınlık örneğinin kullanıcı tarafından belirlenen en yakın k adet komşusuna bakılır. Algoritma rastgele seçilen örnek ile diğer azınlık örnekleri arasındaki en yakın noktayı belirledikten sonra interpolasyon yöntemi ile yeni bir sentetik veri üretmiş olur. Eğer üretilecek olan sentetik veri sayısı orijinal veri kümesi sayısından küçük olur ise orijinal bir veri kullanılmaktadır. Diğer bir durumda, üretilecek olan sentetik veri sayısı, orijinal veri kümesi sayısından büyük olur ise, algoritma önceden belirlenmiş olan aşırı

örnekleme oranına göre yinelemeli şeklinde sentetik verileri oluşturmaktadır (Harman, 2021).

SMOTE algoritmasının adımları şu şekilde ifade edilebilir (Yavaş, Güran ve Uysal, 2020):

Adım 1: Azınlık sınıfındaki her bir verinin (x_i) k en yakın komşusuna bakılır,

Adım 2: Azınlık sınıfındaki veri (x_i) ile k en yakın komşusundaki gözlemin (x_j) farkı alınır,

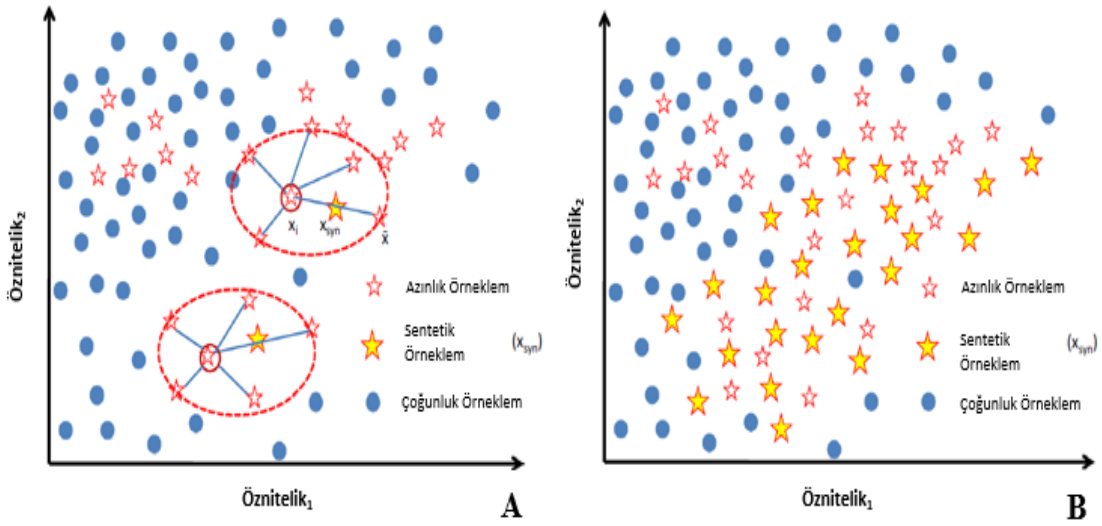
Adım 3: 0 ile 1 arasında rastgele belirlenen bir sayı ile Adım 2'deki değer çarpılır,

Adım 4: Denklem 2.1'deki eşitlik sayesinde yeni sentetik veriler (x_{syn}) üretilir.

$$x_{syn} = x_i + (x_j - x_i) * \text{rand}(0,1) \qquad (2.1)$$

Adım 5: Belirlenen sayıda veri elde edilebilmesi için Adım 1-4 tekrarlanır.

SMOTE yönteminde k=5 için sentetik veri üretiminin bir örneği Şekil 2.5 (A) ve Şekil 2.5 (B)'de gösterilmiştir (Raghuwanshi ve Shukla, 2021).



Şekil 2.5. SMOTE, Rastgele seçilen bir azınlık örneğinin k=5 değeri için en yakın komşularına göre atanarak sentetik veri üretilmesi **A)** Orijinal veri seti dağılımı **B)** SMOTE sonrası dağılım (Raghuwanshi ve Shukla, 2021)

2.3. Makine Öğrenmesi

Makine öğrenmesi; bir bilgisayarı, örnek verileri kullanarak veya deneyime dayalı olarak bir performans ölçütünü optimize edecek şekilde programlamaktır. Yapay zekâ uygulamalarının aksine verilerdeki gizli kalıpların öğrenilmesi ve problem ile ilgili bir durumu sınıflandırmak veya tahmin etmek için kalıpların kullanılması makine öğrenmesinin araştırma alanına girmektedir (Alpaydın, 2014).

Yapay zekânın bir alt dalı olan makine öğrenmesi günümüzde neredeyse her alanda kullanılmaktadır. Bankacılık ve finans, dijital medya, eğitim, e-ticaret, otomotiv, sağlık vb. sektörler makine öğrenmesinin kullanıldığı alanlara örnek olarak gösterilebilir.

Spam olmayan e-postaları, spam e-postalardan ayırma işlemi makine öğrenmesine örnek verilebilir. Bu süreçte e-postalara dahil edilen bazı belgeler veya kelimeler girdiyi temsil ederken, e-postanın sırasıyla spam veya spam olmadığını gösteren evet veya hayır çıktıyı temsil etmektedir. Ancak spam e-postaları doğru bir şekilde tanımlamak için bir algoritma yoktur. Makine öğrenmesi; bu görev için el ile spam değil veya spam olarak etiketlenen e-postalara örnekler verilen ve programın otomatik olarak aralarında ayrım yapmayı öğrenebileceği bir çözüm sunar (Alpaydın, 2014).

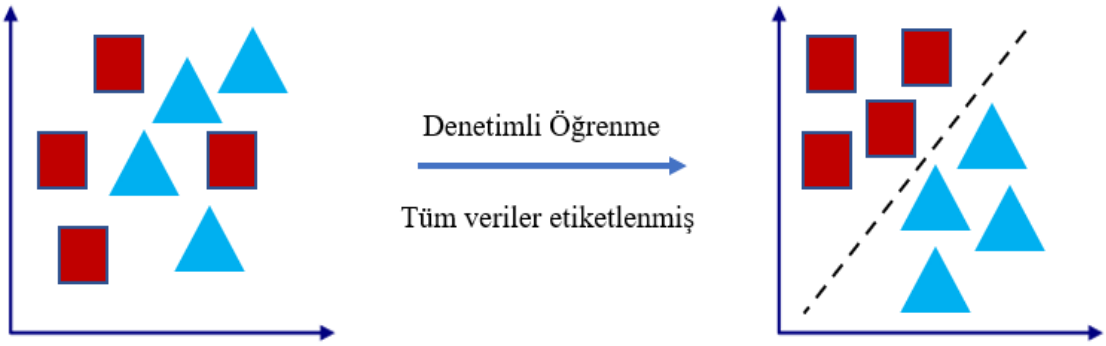
Makine öğrenmesi yöntemleri denetimli, denetimsiz ve pekiştirmeli olmak üzere üç başlıkta incelenebilir.

2.3.1. Denetimli öğrenme

Denetimli makine öğrenimi, kategorik veya sürekli etiketli veri örnekleriyle ilgilenmektedir. Veri örneğinin etiketi kategorik bir değer olduğunda, denetimli öğrenme, proses verilerinin sınıflandırılması için kullanılabilir. Bir diğer durumda, veri örneğinin etiketi sürekli bir değer olduğunda ise tahmin amacıyla regresyon modelleri oluşturulabilir (Ge, Song, Ding ve Huang, 2017).

Sağlık alanında hastalık teşhisi, bankacılık ve telekomünikasyon alanında müşteri sadakatinin tahmini, perakende sektöründe satış tahmini denetimli makine öğrenmesine

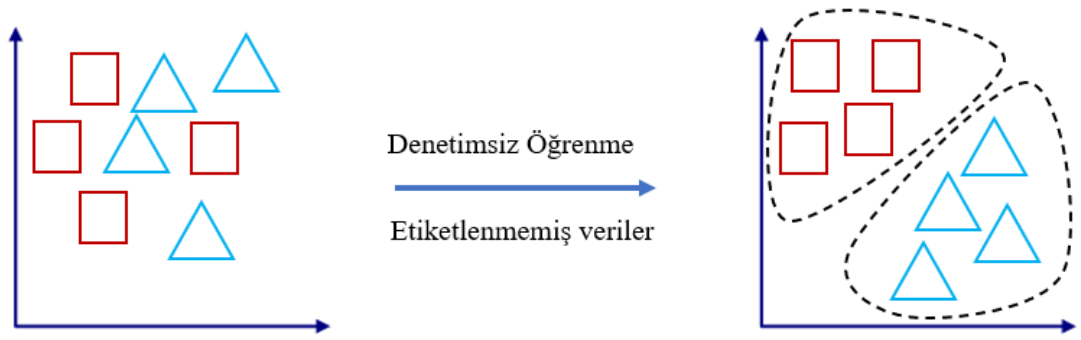
örnek olarak gösterilebilir. Denetimli makine öğrenmesinin bir örneği Şekil 2.6'da gösterilmiştir (Şenol, Canbay ve Kaya, 2021).



Şekil 2.6. Denetimli öğrenme yapısı (Şenol ve diğerleri, 2021)

2.3.2. Denetimsiz öğrenme

Denetimsiz makine öğrenimi, bir hedef özniteliğin katılımı olmadan örüntü tanımayı içerir. Yani analizde kullanılan tüm değişkenler girdi olarak kullanılmakta ve yaklaşım nedeniyle teknikler kümeleme ve birliktelik tekniklerine uygundur (Alloghani, Al-Jumeily, Mustafina, Hussain ve Aljaaf, 2020). Pazarlama alanında müşteri segmentasyonu, dijital platformlardaki film ve müzik önerileri denetimsiz makine öğrenmesine örnek olarak gösterilebilir. Denetimsiz makine öğrenmesinin bir örneği Şekil 2.7'de gösterilmiştir (Şenol ve diğerleri, 2021).



Şekil 2.7. Denetimsiz öğrenme yapısı (Şenol ve diğerleri, 2021)

2.3.3. Pekiştirmeli öğrenme

Pekiştirmeli öğrenme, yapılan hata miktarının bir önceki adımda yapılan hata miktarından az mı yoksa fazla mı olduğu dikkate alınarak bir ödül veya ceza değeri oluşturan bir tekniktir. Hesaplanan bu ödül veya ceza puanına göre sistem öğrenimine devam edilir ve en iyi kontrolü sağlayacak parametreler elde edilmeye çalışılır. Bu parametrelerin belirlenmesi sonucunda bir durum eylem grafiği elde edilir, bu grafik kontrol sağlamak için kullanılır (Emer ve Özbek, 2021). Pekiştirmeli öğrenme yöntemleri, insansız hava araçları kontrolü, hisse senedi portföy optimizasyonu, oyun programlama gibi farklı alanlarda kullanılmaktadır.

2.4. Makine Öğrenmesi Algoritmaları

Bu bölümde çalışmada kullanılan makine öğrenmesi algoritmalarından bahsedilmiştir.

2.4.1. Lojistik regresyon (LR)

LR algoritması, bağımlı değişkenin kategorik olduğu ikili sınıf etiketine sahip problemlerde kullanılan bir sınıflandırma modelidir (Sevli, 2022). Sınıflandırma için kullanılan iyi bilinen bir başka denetimli öğrenme algoritmasıdır. LR, yalnızca iki değere (0 veya 1) sahip olabilen bir sonucun olasılığını tahmin eder. Tahmin, bir veya daha fazla tahmincinin (sayısal ve kategorik) kullanımına dayanmaktadır. Bu sınıflandırıcı, bağımsız özellikler ile hedef öznitelik arasındaki ilişkiyi sağlar. Lojistik regresyon modeli, 0 ile 1 arasındaki değerlerle sınırlı bir lojistik eğri üretir (Mohammed ve diğerleri, 2020).

İkili lojistik regresyon yönteminde kullanılan matematiksel model ise denklem 2.2'deki gibi ifade edilebilir (Yazgan ve Erol, 2016):

$$P(y_j = 1) = \frac{e^{\beta_0+\beta_1 x_{j1}+\beta_2 x_{j2}+ +\beta_3 x_{j3}+\ldots+\beta_k x_{jk}}}{1+ e^{\beta_0+\beta_1 x_{j1}+\beta_2 x_{j2}+ +\beta_3 x_{j3}+\ldots+\beta_k x_{jk}}} \quad (2.2)$$

Burada, n birim sayısı (j=1,2, …, n) olmak üzere diğer notasyonların açıklamaları Çizelge 2.1'de gösterilmiştir.

Çizelge 2.1. Lojistik regresyon modelindeki notasyonların açıklamaları

Notasyon	Açıklama
$P(y_j=1)$	j. birimin incelenen kategoriye eşit olma olasılığı ya da incelenen olay ile ilgili pozitif cevap verme olasılığı
β_0	Bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin değeri
$\beta_1, \beta_2, \ldots, \beta_k$	Bağımsız değişkenlerin regresyon katsayılar
$x_1, x_2, \ldots, x_k$	Bağımsız değişkenler
k	Bağımsız değişken sayısı
e	2,71

2.4.2. Naive bayes (NB)

NB algoritması, bir veri kümesindeki değerlerin sıklığını ve kombinasyonlarını sayarak bir dizi olasılığı hesaplayan basit olasılıksal bir sınıflandırıcıdır. Algoritma özünde Bayes teoremini kullanır ve sınıf değişkeninin değeri göz önüne alındığında tüm değişkenlerin bağımsız olduğunu varsayar. Bu koşullu bağımsızlık varsayımı, gerçek hayat uygulamalarında nadiren geçerlidir, bu nedenle "naif" olarak nitelendirilir, ancak algoritma, çeşitli kontrollü sınıflandırma problemlerinde hızlı bir şekilde öğrenme eğilimindedir (Dimitoglou, Adams ve Jim, 2012).

Bayes teoremine ilişkin matematiksel ifade denklem 2.3'te gösterilmiştir:

$$P(H\backslash D) = \frac{P(H)\,P(D\backslash H)}{P(D)} \tag{2.3}$$

Burada, D durumunun gerçekleşme olasılığı bilindiği durumda H durumunun gerçekleşme olasılığı P(H\D); H durumunun gerçekleşme olasılığı ve H durumunun olasılığı bilindiği durumda D durumunun gerçekleşme olasılığı ile çarpılır ve D durumunun gerçekleşme olasılığına bölünerek hesaplanır.

Denetimli makine öğrenmesinde ise matematiksel ifade denklem 2.4'teki ifade edilmiştir (Alloghani ve diğerleri, 2020):

$$P(H\backslash D) = P(x_i, \ldots\ldots\ldots\ldots, x_n|H) = \prod_i P(x_i|H) \qquad (2.4)$$

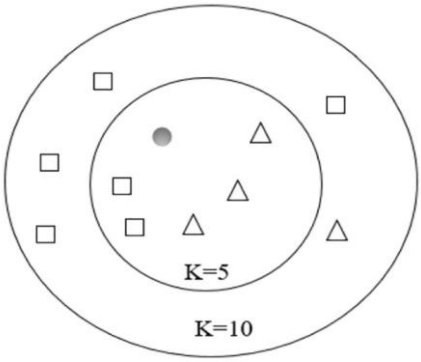
Burada, x_i , ... , x_n koşullu olasılıkların, eğitim veri kümesindeki hedef değişkenlerin bilinen olasılıklarına göre hesapladığı girdi niteliğini temsil etmektedir (Alloghani ve diğerleri, 2020).

2.4.3. K-en yakın komşu (k-EYK)

K-EYK algoritması, 1951'de iki istatistikçi tarafından literatüre kazandırılmıştır (Fix ve Hodges, 1951). Sonrasında ise (Cover ve Hart, 1967) tarafından kapsamı genişletilen algoritma günümüzde sınıflandırma ve regresyon için kullanılmaktadır.

EYK sınıflandırmasının altında yatan mantık oldukça basittir: örnekler, en yakın komşularının sınıfına göre sınıflandırılmaktadır. Birden fazla komşuyu hesaba katmak genellikle yararlıdır, bu nedenle teknik daha yaygın olarak k-EYK şeklinde adlandırılır. Burada k en yakın komşular sınıfının belirlenmesinde kullanılır (Cunningham ve Delany, 2021). k-EYK algoritması mesafeye dayalı bir algoritma olduğu için Öklid, Manhattan, Minkowski vb. uzaklık ölçütlerini kullanarak sınıflandırma işlemini yapmaktadır.

Algoritmanın mantığının daha iyi anlaşılabilmesi için Şekil 2.8'de bir örneğin görseli verilmiştir (Wang, Wang, Wan ve Song, 2020). Burada, "•", "□" ve "Δ" olmak üzere üç farklı şekil görülmektedir. Belirli test örnekleri için, bir eğitim setinden alınan k-EYK örneklerini belirli bir benzerlik ölçüsüne göre alır ve bir test örneği için öngörülen etiketi belirlemek üzere k-EYK örneklerinin etiketlerine oy verir. Şekil 2.8'de gösterildiği gibi, test örneği "•", oylama mekanizmasına göre K=5 olduğunda "Δ" olarak sınıflandırılacaktır. Ancak, K=10 olduğunda "□" olarak sınıflandırılır (Wang ve diğerleri, 2020).



Şekil 2.8. K değerine göre örneğin sınıfının belirlenmesi (Wang ve diğerleri, 2020)

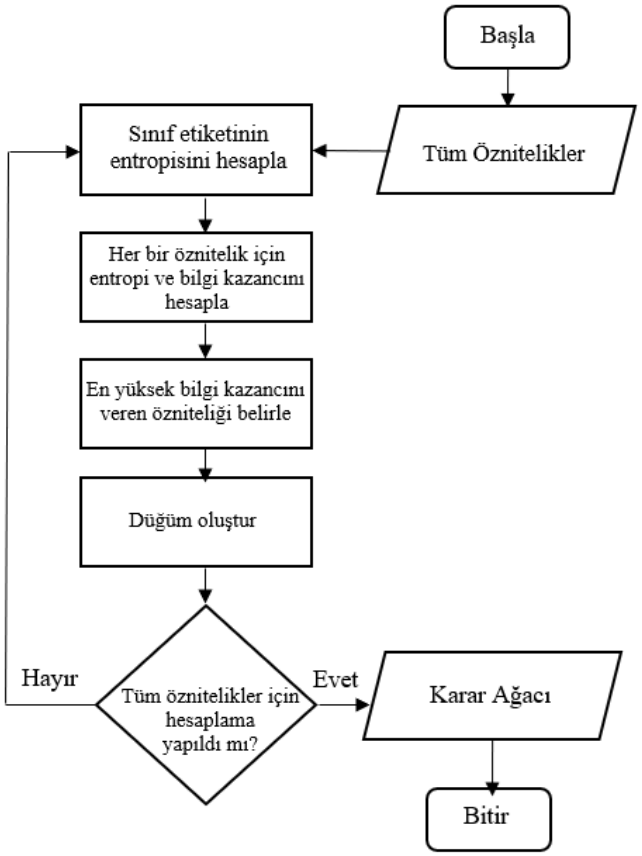
2.4.4. C4.5 algoritması (C4.5)

Quinlan (1993) tarafından literatüre kazandırılan C4.5 algoritması, verileri sınıflandırırken karar ağacı yapısından yararlanır. Sınıflandırma işlemi yaparken karar ağaçlarının tercih edilmesi kullanıcı açısından daha anlaşılır ve kolay olmaktadır (Çalış, Kayapınar ve Çetinyokuş, 2014)

C4.5 algoritması ID3 algoritmasının geliştirilmiş bir versiyonudur. ID3 algoritması sadece kategorik veriler ile çalışırken C4.5 algoritması hem kategorik hem de sürekli veri ile çalışabilmektedir.

Karar ağaçları oluşturulurken ağacın dalının hangi kriterlere göre yapılacağının belirlenmesi önem arz eden bir konudur. Böylece ağaç yapısının hangi özellik değerlerine göre oluşturulacağı belirlenir (Kavzoğlu ve Çölkesen, 2010).

C4.5 algoritma ağaç yapısındaki dallanma belirlenirken bilgi kazancı kullanılmaktadır. En yüksek bilgi kazancına sahip öznitelikten karar ağacı dallanmaktadır. Algoritmanın adımlarını gösteren akış diyagramı Şekil 2.9'da gösterilmiştir (Anwar, Pranolo ve Kurnaiwan, 2018).



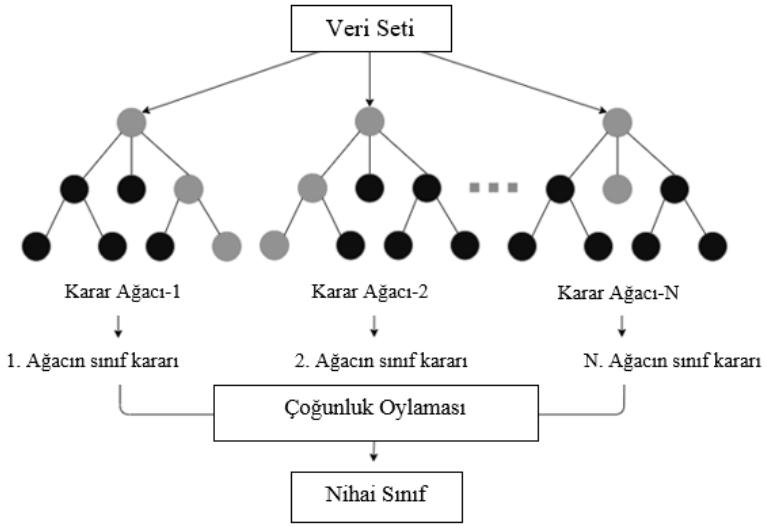
Şekil 2.9. C4.5 algoritması akış diyagramı (Anwar ve diğerleri, 2018)

2.4.5. Rastgele orman (RO)

Breiman (2001) tarafından literatüre kazandırılan RO algoritması, çeşitli uygulama alanlarında makine öğrenimi ve veri bilimi alanında kullanılan bir topluluk sınıflandırma tekniği olarak bilinmektedir (Sarker, 2021).

RO algoritması problem çeşidine göre sınıflandırma yapabilme özelliğinin yanı sıra regresyon için de kullanılabilen bir denetimli makine öğrenmesi yöntemidir. Bu yöntem, birden fazla sayıda karar ağacını rastgele biçimde birleştiren ve her bir karar ağacının tahminlerini toplamakta ve çoğunluğun oyuna göre karar vermektedir. RO algoritması, değişken sayısının gözlem sayısından çok daha büyük olduğu ortamlarda mükemmel performans gösterdiği belirtilmektedir (Biau ve Scornet, 2016).

Şekil 2.10'da RO algoritmasının mantığı daha iyi anlaşılabilmesi için bir örnek verilmiştir.



Şekil 2.10. Birden çok karar ağacını dikkate alan bir rastgele orman yapısı örneği (Sarker, 2021)

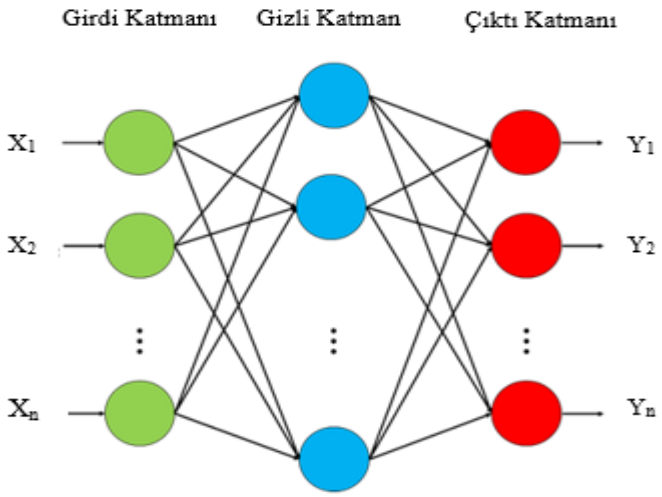
2.4.6. Çok katmanlı algılayıcı (ÇKA)

Yapay Sinir Ağı (YSA), beynin bilgi işleme yeteneğinden ilham alınarak ortaya çıkarılmıştır. Bir bilgisayardaki bir model nöron ağını simüle ederek yapay bir sinir ağı oluşturulabilir. Gerçek nöronların süreçlerini taklit eden algoritmalar uygulayarak, ağın birçok problemi çözmeyi 'öğrenmesini' sağlayabilmektedir (Krogh, 2008).

Bir YSA (veya basitçe sinir ağı), bir giriş nöron katmanından (veya düğümlerden, birimlerden), bir veya iki (hatta üç) gizli nöron katmanından ve son bir çıkış nöron katmanından oluşur (Wang, 2003).

Tek katmanlı algılayıcılar ile başlayan YSA çalışmaları doğrusal olayları çözmekte başarılı olsa da doğrusal olmayan olayları öğrenemediği için ileri beslemeli olan ÇKA geliştirilmiştir. ÇKA ile mühendislik problemlerinde sınıflama, tahmin etme, tanıma vb. konularda başarılı sonuçlar elde edilmiştir (Öztemel, 2003).

Şekil 2.11'de ÇKA algoritmasının mantığı daha iyi anlaşılabilmesi için bir örnek verilmiştir (Konakoğlu, 2020). Şekil 2.10'da verilen girdi katmanındaki $X_1, X_2, ..., X_n$ değerleri nörona verilen girdiyi, çıktı katmanındaki $Y_1, Y_2, ..., Y_n$ değerleri ise algılayıcının vermiş olduğu çıktıyı ifade etmektedir.



Şekil 2.11. ÇKA yapay sinir ağı örneği (Konakoğlu, 2020)

2.5. Literatür Araştırması

Bu bölümde diyabet hastalığı teşhisi ve sınıf dengesizliği ile ilgili literatürde yer alan çalışmalar verilmiştir.

2.5.1. Diyabet hastalığı teşhisi ile ilgili çalışmalar

Bu bölümde diyabet hastalığı sınıflandırılırken literatürde kullanılan veri setleri, sınıflandırma yöntemleri ve modellerin performans ölçütleri hakkında bilgiler verilmiştir. İncelenen çalışmaların detaylı tablosu Çizelge 2.2'de gösterilmiştir.

Das, Naik ve Behera (2018), diyabet hastalığını tahmin etmek için sınıflandırma algoritmalarının en uygun yöntemlerden biri olduğunu öne sürmüşlerdir. Bu bağlamda diyabet teşhisinde kullanmak üzere bir üniversite hastanesinde 200 kişiye yönelik 7 soruluk bir anket çalışması yapmışlardır. Elde edilen verilere J48 ve NB tekniklerini uygulamışlardır. Yapılan analiz sonucunda NB yöntemi model çalışma süresi ve tahmin doğruluğu açısından J48 algoritmasından daha iyi sonuçlar vermiştir.

Çizelge 2.2. Diyabet hastalığı teşhisinin sınıflandırılması ile ilgili literatür araştırması

No	Yazarlar ve Yıllar	Veri Seti	Veri Sayısı	Değişken Sayısı	Hedef Değişken	Yöntemler
1	Das vd. (2018)	Gerçek Hayat Verisi	200	8	Diyabet (Pozitif/Negatif)	J48, NB
2	Chen ve Pan (2018)	Gerçek Hayat Verisi	35669	30	Diyabetli, Diyabetli Değil	LR, RO, AdaBoost.M1, LogitBoost
3	Zou vd. (2018)	Pima Yerlileri Diyabet Veri Seti Sylhet Hastanesi Diyabet Veri Seti	768 520	9 17	Diyabet (Pozitif/Negatif) Diyabet (Pozitif/Negatif)	RO, J48,YSA
4	Akyol ve Şen (2018)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	AdaBoost.M1, RO, GAA
5	Alehegn vd. (2019)	Pima Yerlileri Diyabet Veri Seti ABD'deki 130 Hastahane Veri Seti	768 70000	9 48	Diyabet (Pozitif/Negatif) Diyabetli, Diyabetli değil	J48, k-EYK, NB, RO, Stacking
6	Birjais vd. (2019)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	GA, NB, LR
7	Jakka ve Rani J. (2019)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	k-EYK, KA, NB, DVM, RO, LR
8	Tigga ve Garg (2020)	Pima Yerlileri Diyabet Veri Seti Gerçek Hayat Verisi	768 952	9 18	Diyabet (Pozitif/Negatif) Diyabet (Pozitif/Negatif)	LR, k-EYK, DVM, NB, KA, RO
9	Alpan ve Ilgi (2020)	Sylhet Hastanesi Diyabet Veri Seti	520	17	Diyabet (Pozitif/Negatif)	Bayes Ağı, NB, J48, RA, RO, k-EYK, DVM
10	Islam vd. (2020)	Sylhet Hastanesi Diyabet Veri Seti	520	17	Diyabet (Pozitif/Negatif)	NB, LR, J48, RO
11	Naz ve Ahuja (2020)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	YSA, NB, KA, DÖ
12	Daghistani ve Alshammari (2020)	Gerçek Hayat Verisi	66325	18	Diyabetli, Diyabetli Değil	RO, LR
13	Kumari vd. (2021)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	LR, k-EYK, NB, DVM, KA, RO, Soft Voting Classifier, AdaBoost.M1, Bagging, GA, XGBoost, CatBoost
14	Chaves ve Marques (2021)	Sylhet Hastanesi Diyabet Veri Seti	520	17	Diyabet (Pozitif/Negatif)	YSA, NB, AdaBoost.M1, k-EYK, RO, DVM
15	Cihan ve Coşkun (2021)	Pima Yerlileri Diyabet Veri Seti	768	9	Diyabet (Pozitif/Negatif)	LR, k-EYK, Gauss NB, DVM, KA, RO, YSA

Chen ve Pan (2018), diyabet teşhisi için, Çin'de bulunan bir üniversitesinin endokrin bölümünden elde ettikleri Temmuz 2004-Nisan 2014 arasındaki 35669 hasta kaydı ve 30 öznitelik kullanarak gerçek hayat verisi çalışmışlardır. Elde edilen verilere LR, RO, Adaboost.M1 ve LogitBoost yöntemleri uygulandığında; AdaboostM1 ve LogitBoost yöntemleri hesaplama zamanı ve doğruluk oranı yönünden çok daha iyi sonuçlar verdiği anlaşılmıştır. LogitBoost algoritması %93,93 doğruluk oranı ile Adaboost.M1 algoritmasına göre daha iyi sonuç göstermiştir. Ayrıca LogitBoost ve Adaboost.M1 algoritmaları için veri seti, 10-katlı çapraz doğrulama tekniği kullanılarak eğitim ve test verisi olarak ayrıştırıldığında ise LogitBoost algoritmasının doğru sınıflandırma oranı %95,30'a ulaşmıştır.

Akyol ve Şen (2018), diyabet hastalığı teşhisi için PIMA yerlileri veri seti üzerinde çalışmışlardır. Çalışmanın ilk aşamasında en önemli öznitelikleri bulabilmek için öznitelik seçimi veya ağırlıklandırma yöntemlerini analiz etmişlerdir. İkinci aşamada ise AdaBoost, Gradyan Arttırımlı Ağaçlar (GAA), RO algoritmalarının performanslarını değerlendirmişlerdir. Ayrıca veri seti üç farklı oranda eğitim ve test verisi olarak bölünmüştür. Yapılan çalışmadaki deneysel sonuçlara göre öznitelik seçimi Stability Selection (SS) ve sınıflandırma yöntemi Adaboost algoritması seçildiğinde %73,88 doğruluk oranına ulaşarak en iyi sonucu vermiştir.

Zou ve diğerleri (2018), diyabet hastalığını sınıflandırmak için Kaggle veri tabanından almış oldukları PIMA Yerlileri veri seti ve Kaliforniya Üniversitesi (UCI) veri tabanından almış oldukları Sylhet Hastanesi veri seti üzerinde çalışmışlardır. Verileri analiz ederken her iki veri seti içinde RO, J48 Algoritması ve Yapay Sinir Ağı (YSA) yöntemlerini kullanmışlardır. Yapılan analizler sonucunda iki veri seti için de en iyi sonucu RO algoritması vermiştir.

Alehegn, Raghvendra ve Mulay (2019), diyabet hastalığı teşhisi için PIMA Yerlileri ve ABD'de bulunan 130 hastanenin 1999-2008 yılları arasındaki veri setleri üzerinde çalışmışlardır. Verileri analiz etmek için RO, k-EYK, NB, J48 ve bu dört yöntemden elde edilen tahminleri bir kümede birleştiren istifleme (stacking) topluluk öğrenmesi yöntemini kullanmışlardır. Dört yöntemi birleşimini içeren istifleme topluluk yönteminin doğru sınıflandırma oranları PIMA Yerlileri veri seti için %93,6, ABD veri seti için ise %88,56 olarak en iyi sonuçları göstermiştir.

Birjais, Mourya, Chauhan ve Kaur (2019), diyabet hastalığını teşhis etmek için PIMA Yerlileri veri seti üzerinde çalışmışlardır. Veri setindeki eksik verileri gidermek için kesikli, sürekli, kategorik ve sıralı verileri gidermede daha iyi olduğunu belirttikleri k-EYK tekniğini kullanmışlardır. Veri setinin bölünmesi aşamasında %70 eğitim ve %30 test verisi olarak ayrıştırmışlardır. Çalışmada Gradyan Artırma (GA), LR ve NB olmak üzere üç farklı makine öğrenmesi tekniğini kullanmışlardır. GA test verilerinin %86'sını, NB %77, LR yöntemi ise %79'unu doğru olarak sınıflandırmıştır. Bu nedenle GA yöntemi bir kişinin diyabet hastası olup olmadığını tahmin etme sürecinde NB ve LR algoritmalarından daha iyi olduğu sonucuna varmışlardır.

Jakka ve Rani (2019), diyabet hastalığı teşhisi için PIMA Yerlileri veri setini kullanarak, k-EYK, Karar Ağacı (KA), NB, Destek Vektör Makineleri (DVM), LR ve RO algoritmalarının performanslarını karşılaştırmışlardır. Altı farklı yöntem; doğru sınıflandırma oranı, duyarlılık, F1 skoru, yanlış sınıflandırma oranı ve alıcı işletim karakteristiği (ROC) performans metrikleri açısından değerlendirilmiştir. LR yönteminde doğru sınıflandırma oranı %77,6 olarak elde edilmiştir. Ona en yakın doğru sınıflandırma performansı gösteren ikinci algoritma ise NB algoritması olmuştur. Tüm performans kriterlerinde LR diğer beş algoritmaya göre daha iyi sonuçlar vermiştir.

Tigga ve Garg (2020), diyabet hastalığı teşhisi üzerine yapmış oldukları çalışmada PIMA Yerlilerine ait 768 örnek ve anket yöntemiyle toplamış oldukları 952 örnek olmak üzere iki farklı veri seti üzerinde çalışmışlardır. Ankette hastalara sağlık, yaşam tarzı ve aile geçmişi ile ilgili 18 adet soru yöneltmişlerdir. Çalışmada iki veri seti için de LR, k-EYK, DVM, NB, KA ve RO algoritması olmak üzere altı yöntemin performansları karşılaştırılmıştır. PIMA Yerlileri veri seti için en yüksek doğru sınıflandırma oranını %75 değeri ile RO algoritması vermiştir. Ona en yakın performans gösteren ikinci algoritma ise %74,4 doğruluk değeri ile LR ve DVM algoritmalarıdır. Anket yöntemi ile toplanan veri setinde en yüksek doğruluk oranına sahip algoritma %94,1 değeri ile RO algoritmasıdır. Ona en yakın performansı gösteren ikinci algoritma ise %86,5 doğruluk değeri DVM'dir.

Alpan ve Ilgi (2020), diyabet hastalığını teşhis etmek için UCI veri tabanında bulunan Bangladeş'teki Sylhet Diyabet Hastanesi'ne ait bir veri seti üzerinde çalışmışlardır. Bu veri seti her biri 17 özniteliğe sahip 520 örnek içermektedir. Yapılan çalışmada Bayes

Ağı (BA), NB, J48, RA, RO, k-EYK ve DVM olmak üzere yedi farklı sınıflandırma algoritması kullanılmıştır. K-EYK algoritması %98,07 doğru sınıflandırma oranı ile diğer altı algoritmaya göre daha iyi sonuç vermiştir.

Islam, Ferdousi, Rahman ve Bushra (2020), diyabet hastalığını sınıflandırmak için Sylhet Diyabet Hastanesi'ne ait bir veri seti üzerinde çalışmışlardır. Veri setini eğitim ve test verisi olarak ayrıştırırken 10-katlı çapraz doğrulama ve %80-%20 olmak üzere iki farklı teknikten yararlanmışlardır. Yapılan çalışmada NB, LR, J48 ve RO sınıflandırma algoritmalarının performanslarını kıyaslamışlardır. RO algoritması 10-katlı çapraz doğrulama tekniğinde %97,4 doğru sınıflandırma oranı, %80-%20 tekniğinde ise %99 doğru sınıflandırma oranı ile en başarılı algoritma olmuştur.

Naz ve Ahuja (2020), diyabet teşhisi için UCI veri tabanında bulunan PIMA yerlileri diyabet veri seti üzerinde çalışmışlardır. Çalışmada YSA, NB, KA ve Derin Öğrenme (DÖ) olmak üzere dört farklı sınıflandırma tekniğinin performanslarını analiz etmişlerdir. Yapılan analizler sonucunda en iyi sınıflandırma başarısını %98,07 ile DÖ yöntemi vermiştir.

Daghistani ve Alshammari (2020), diyabet hastalığını teşhis etmek için Suudi Arabistan'da bulunan bir sağlık kuruluşundan elde etmiş oldukları veri seti üzerinde çalışmışlardır. Bu veri seti 2013-2015 yılları arasındaki 18 öznitelik ve 66325 hasta kaydını içermektedir. Yapılan çalışmada sınıflandırma algoritmalarından LR ve RO algoritmalarının performansları analiz edilmiştir. RO yönteminin sırasıyla %88, %18,8, %88,3, %88, %94,4, %87,6 ile en iyi doğruluk, hata, kesinlik, duyarlılık, AUC ve F1 skor değerlerini elde ettiği görülmüştür.

Kumari, Kumar ve Mittal (2021), hastalık teşhisinde bulunmak için literatürde yaygın olarak kullanılan PIMA Yerlileri veri seti ve Meme Kanseri veri seti üzerinde çalışmışlardır. Çalışmada önerilen Soft Voting Classifier yöntemi ikili sınıflandırma yapan bir algoritma olup, üç makine öğrenmesi algoritmasının (LR, NB ve RO) birlikte karar vermesini sağlayan bir topluluk öğrenmesi yöntemdir. Ayrıca hastalık teşhisi için önerilen topluluk öğrenmesi yönteminin yanında performansı gelişmiş yöntemler olan AdaBoost, Torbalama (Bagging), GA, Ekstrem Gradyan Arttırma (XGBoost), CatBoost, LR, DVM, RO, NB, k-EYK algoritmaları ile karşılaştırılmıştır. Önerilen topluluk

yöntemi, PIMA Yerlileri diyabet veri setinde sırasıyla %79,04, %73,48, %71,45 ve %80,6 ile en yüksek doğruluk, kesinlik, duyarlılık ve F1 skor değerlerini vermiştir. Ayrıca, önerilen metodolojinin etkinliği de meme kanseri veri seti ile karşılaştırılmış ve analiz edilmiştir. Önerilen topluluk öğrenmesi yöntemi, meme kanseri veri setinde %97,02 doğruluk sağlamıştır.

Chaves ve Marques (2021), diyabet hastalığını teşhis etmek için UCI veri tabanında bulunan Bangladeş'teki Sylhet Diyabet Hastanesi'ne ait bir veri seti üzerinde çalışmışlardır. Bu veri seti her biri 17 özniteliğe sahip 520 örnek içermektedir. Veri önişleme kısmında 0-1 normalizasyon tekniğini, veri bölme kısmında ise 10-katlı çapraz doğrulama tekniğini kullanmışlardır. Yapılan çalışmada YSA, k-EYK, NB, RO, DVM ve Adaboost olmak üzere altı farklı makine öğrenmesi yönteminin performansları karşılaştırılmıştır. Sonuçlar, diyabeti tahmin etmek için YSA kullanılması gerektiğini göstermiştir. Önerilen sinir ağı modeli, veri setinde sırasıyla %98,1, %98,4, %98,4 ile en yüksek doğru sınıflandırma oranı, F1 skoru, özgüllük değerlerine ulaşmıştır.

Cihan ve Coşkun (2021), diyabet teşhisi için Kaggle veri tabanında bulunan PIMA Yerlileri veri seti üzerinde çalışmışlardır. Yapmış oldukları çalışmada LR, k-EYK, DVM, Gauss Naive Bayes (GNB), KA, RO ve YSA olmak üzere yedi farklı makine öğrenmesi algoritmasını kullanmışlardır. Veri setini eğitim ve test verisi olarak ayırırken %70-%30 ve 10-katlı çapraz doğrulama olmak üzere iki farklı veri seti ayırma tekniği kullanmışlardır. %70-%30 ayırma tekniğinde yedi model arasında %86,7 ROC, %80,8 kesinlik, %81,3 duyarlılık ve %87,6 kesinlik-duyarlılık eğrisi (PRC) değeri ile en iyi sonucu LR yöntemi vermiştir. 10-Katlı çapraz doğrulama tekniğinde de %83,7 ROC, %76,8 kesinlik, %77,3 duyarlılık ve %83,5 PRC değeri ile en iyi sonucu yine LR yöntemi vermiştir

2.5.2. Sınıf dengesizliği ile ilgili çalışmalar

Sınıflandırma problemlerinde sıklıkla karşılaşılan dengesiz veri setleri ile ilgili literatürde çok sayıda çalışma mevcuttur. İncelenen çalışmaların detaylı tablosu Çizelge 2.3'te gösterilmiştir.

Makki ve diğerleri (2019) yapmış oldukları çalışmada, kredi kartı dolandırıcılığı tespitinde dengesiz bir veri setinde deneysel çalışma yürütmüşlerdir. Sınıflandırma modelleri oluştururken sekiz farklı makine öğrenmesi tekniği üzerinde çalışmışlardır. Ayrıca veri setindeki sınıf dengesizliğini giderebilmek için ROS, tek sınıflı sınıflandırma (OCC) ve maliyete duyarlı yöntemleri (CS) kullanmışlardır. Sonuç olarak, dengesiz sınıflandırma yaklaşımlarının, özellikle veriler oldukça dengesiz olduğunda etkisiz kaldığını belirtmişlerdir.

Hassan ve Amiri (2019), yapmış oldukları çalışmada, PIMA yerlileri diyabet veri seti için altı makine öğrenme algoritması (LR, KA, k-EYK, NB, DVM ve YSA) kullanarak tip-2 diyabet hastalarını analiz etmeyi, teşhis etmeyi ve sınıflandırmayı amaçlamışlardır. Ayrıca veri setindeki sınıf dengesizliğini giderebilmek için SMOTE tekniğinden faydalanmışlardır. Yapılan deneysel çalışmalar sonucunda bazı sınıflandırma modellerinde %20'ye varan bir iyileştirme olduğunu tespit etmişlerdir.

Kabir ve Ludwig (2019) yapmış oldukları çalışmada, meme kanseri teşhisinde dengesiz bir veri setini kullanmışlardır. Sınıflandırma modelleri oluştururken KA, RO ve XGBoost yöntemleriyle çalışmışlardır. Ayrıca veri setindeki dengesizliği gidermek için RUS, ROS, SMOTE, düzenlenmiş en yakın komşu (ENN), SMOTE + ENN, SMOTE + Tomek Link tekniklerini kullanmışlardır. Yapılan deneysel çalışmalar sonucunda, yeniden örnekleme yönteminin kullanıldığı zaman özellikle azınlık sınıfı için önemli bir gelişme olduğunu belirtmişlerdir.

Gosain ve Sardana (2019), sınıf dengesizliği problemini gidermek için en uzak sentetik azınlık aşırı örnekleme tekniğini (FSMOTE) önermişlerdir. Önermiş oldukları FSMOTE tekniğini; SMOTE, ADASYN, sınıra yakın olan sentetik azınlık aşırı örnekleme (BSMOTE) ve SL-SMOTE teknikleri ile performanslarını karşılaştırmışlardır. Yapmış oldukları çalışmada NB ve DVM yöntemlerini kullanarak yedi gerçek hayat veri kümesi

Çizelge 2.3. Sınıf dengesizliği ile ilgili literatür araştırması

No	Yazarlar ve Yıllar	Veri Seti Dengeleme Yöntemleri	Dengesizlik Oranı	Veri Seti
1	Makki vd. (2019)	ROS, OCC, CS	15,77	Fraud detection
2	Hassan ve Amiri (2019)	SMOTE	1,86	PIMA
3	Kabir ve Ludwig (2019)	RUS, ROS, SMOTE, ENN, SMOTE + ENN, SMOTE + Tomek Link	15,70	Breast Cancer SC
4	Gosain ve Sardana (2019)	SMOTE, BSMOTE, SL-SMOTE, ADASYN, FSMOTE	1,86	PIMA
			1,90	Breast cancer
			1,25	Heart
			1,79	Ionosphere
			1,54	Spam base
			2,33	German
			42,01	Mammography
5	Shuja vd. (2020)	SMOTE	1,41	Gerçek hayat diyabet veri seti
6	Mohammed vd. (2020)	SMOTE	-	Gerçek hayat diyabet veri seti
7	Turhan vd. (2020)	RUS, ROS, SMOTE	-	Gerçek hayat diyabet veri seti
8	Xiao vd. (2021)	RUS, ROS, OSS, CBOS, SMOTE, ENN, SMOTE + ENN, ADASYN, SL-SMOTE, MWMOTE	2,33	German
			1,25	Australia
			2,79	UK-thomas
			4,06	Give-credit
			13,39	PAKDD
			28,09	IFCD10
9	Mesquita vd. (2021)	ADASYN, BSMOTE, FSMOTE, ROS, SMOTE, SVM-SMOTE	1,86	PIMA
10	Wang vd. (2021)	SMOTE	1,86	PIMA
			3,31	WPBC
			1,68	WDBC
			1,79	Ionosphere
			1,86	Breast-cancer-wisconsin

üzerinde çalışmışlardır. Sonuç olarak NB ve DVM yöntemlerinde FSMOTE tekniğinin diğer dengeleme yöntemlerine göre daha iyi performans gösterdiğini tespit etmişlerdir.

Shuja, Mittal ve Zaman (2020), Kaşmir'de bulunan bir laboratuvardan alınan dengesiz bir diyabet veri seti üzerinde çalışma yapmışlardır. Çalışmaları iki aşamadan oluşmaktadır. Birinci aşamada SMOTE tekniğini kullanarak veri önişleme kısmını gerçekleştirmişlerdir. İkinci aşamada ise beş farklı sınıflandırma tekniğini içeren (Torbalama, DVM, ÇKA, LR ve KA) bir sınıflandırma modeli kullanmışlardır. En iyi sonucu SMOTE tekniğiyle birlikte çalışan KA algoritması vermiştir.

Mohammed ve diğerleri (2020) yapmış oldukları çalışmada, yeni bir gerçek diyabet veri kümesi için altı makine öğrenme algoritması (k-EYK, KA, NB, LR, DVM ve YSA) kullanarak diyabet hastalarını analiz etmeyi, teşhis etmeyi ve sınıflandırmayı amaçlamışlardır. Çalışma yaptıkları veri setinde dengesiz sınıf problemi söz konusu olduğu için SMOTE tekniğini kullanmışlardır. Ayrıca çalışmada üç farklı normalizasyon tekniğini (min-max, z-skor, L2) kullanarak algoritmaların performanslarını incelemişlerdir. Yapılan analizler sonucunda, SMOTE yeniden örnekleme yöntemi ve farklı normalleştirme yöntemleri ile sınıflandırma algoritmalarının performansını önemli ölçüde arttırmanın mümkün olduğunu ve bu nedenle dengesizlik sınıfı problemleri için daha iyi çözümler edilebileceğini tespit etmişlerdir.

Turhan ve diğerleri (2020) yapmış oldukları çalışmada, dengesiz bir veri setinde diyabet hastalığı teşhisi için topluluk öğrenmesi yöntemlerini kullanmışlardır. Sınıf dengesizliğini gidermek için RUS, ROS ve SMOTE olmak üzere üç farklı yeniden örnekleme tekniği kullanmışlardır. Orijinal veri seti ile yapılan diyabet teşhisi sınıflandırmada yeniden örnekleme yapılarak dengeli hâle getirilen veri setlerinden daha düşük performans gösterdiğini tespit etmişlerdir. Sonuç olarak, eğer veri setinde sınıf dengesizliği varsa veri setinin öncelikle yeniden örnekleme teknikleriyle dengelenmesi gerektiğini ardından sınıflandırma algoritmalarının kullanılmasının daha uygun olacağını belirtmişlerdir.

Xiao, Wang, Chen, Xie ve Huang (2021) çalışmalarında, altı kredi skorlama veri setinde sırasıyla 10 yeniden örnekleme yönteminin ve dokuz sınıflandırma modelinin performansını deneysel olarak karşılaştırmışlar ve bunların en uygun kombinasyonlarını

analiz etmişlerdir. Çalışmada kullanılan yeniden örnekleme teknikleri; RUS, ROS, tek taraflı seçim (OSS), kümeleme tabanlı aşırı örnekleme (CBOS), SMOTE, ENN, SMOTE + ENN, uyarlanabilir sentetik örnekleme (ADASYN), güven seviyeli sentetik azınlık aşırı örnekleme (SL-SMOTE) ve çoğunluk ağırlıklı azınlık aşırı örneklemedir (MWMOTE). Deneysel sonuçlara bakıldığında en iyi sonuçları geleneksel bir yöntem olan ROS ve SMOTE ile birlikte kullanılan ENN yöntemi (SMOTE + ENN) göstermiştir. Kredi skorlama için tüm kombinasyonlar arasında en iyi sonucu ise ROS yöntemi ile çalışan rastgele alt uzay (bir topluluk öğrenme yöntemi) vermiştir.

Mesquita ve diğerleri (2021), çalışmalarında PIMA yerlileri veri seti üzerinde on farklı makine öğrenmesi ve altı farklı aşırı örnekleme tekniğini kullanmışlardır. Çalışmada kullanılan yeniden örnekleme teknikleri; ADASYN, BSMOTE, FSMOTE, ROS, SMOTE ve destek vektör makinesiyle sentetik azınlık aşırı örneklemedir (SVM-SMOTE). Diyabet teşhisi için tüm kombinasyonlar arasında en iyi sonucu SVM-SMOTE yöntemiyle birlikte çalışan Adaboost algoritmasının verdiğini tespit etmişlerdir.

Wang, Dai, Shen ve Xuan (2021), UCI platformundan elde etmiş oldukları beş farklı dengesiz veri setinde sınıflandırma problemi üzerinde çalışmışlardır. Yapmış oldukları çalışmada sınıflandırma yöntemlerinden RO algoritmasını kullanmışlardır. Ayrıca sınıf dengesizliğini gidermek için normal dağılım fikrine dayalı yeni bir SMOTE algoritması önermişlerdir. Orijinal SMOTE algoritmasında azınlık sentetik verileri oluşturulurken düzgün dağılıma sahip rastgele (0,1) katsayılar ile sentetik veriler üretilirken; önerilen yöntemde, normal dağılıma sahip değerler kullanılarak azınlık verilerinin merkezine daha yakın bir dağıtım yapılarak daha etkin bir şekilde sentetik veriler üretilmektedir. Deneysel çalışmalar sonucunda AUC, F-ölçütü, G-Ortalama değerleri açısından; beş veri setinde de önerilen yeni SMOTE algoritmasının, ham veri ve orijinal SMOTE verisi ile çalışan modellere göre daha iyi sonuçlar verdiği görülmüştür.

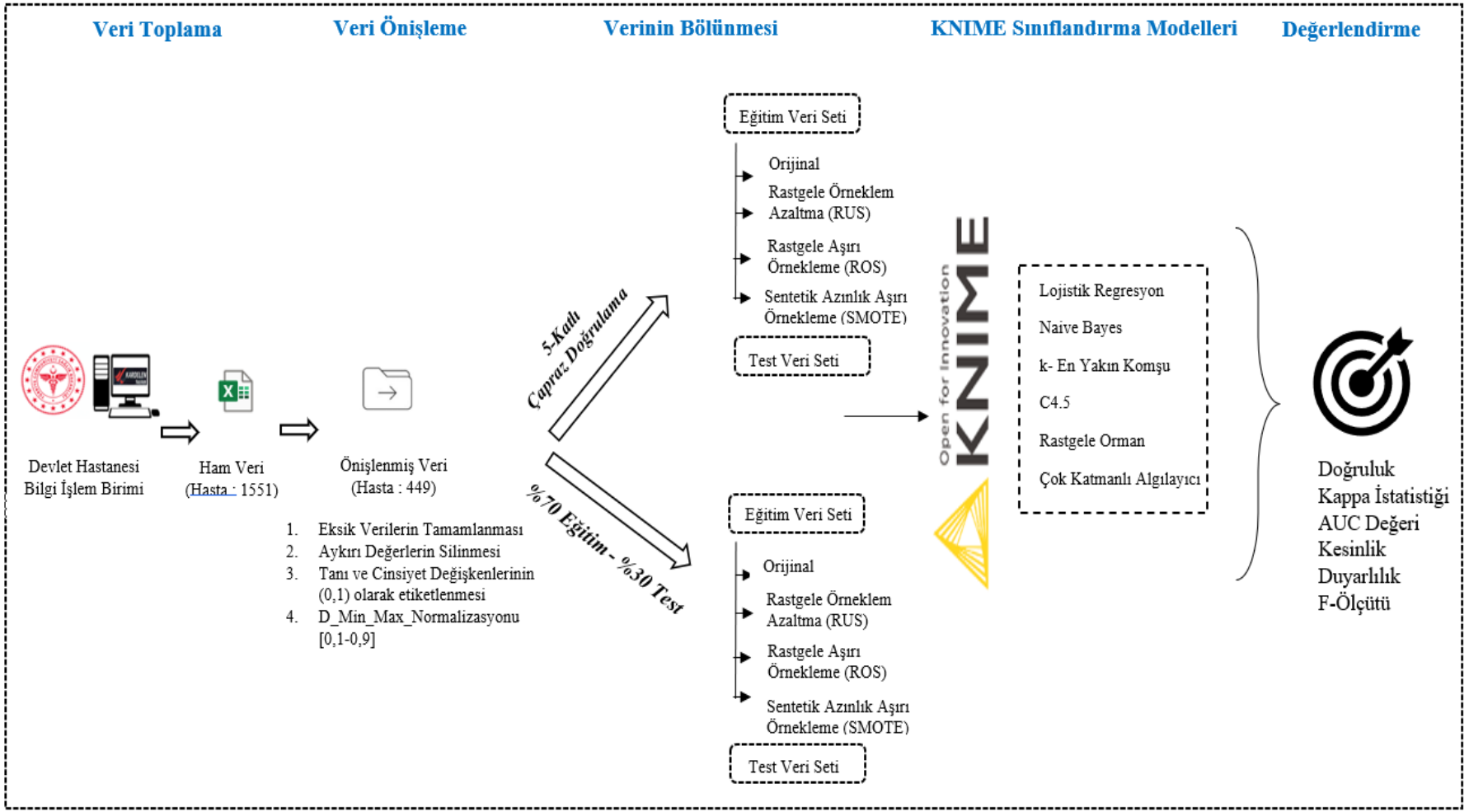
3. MATERYAL ve YÖNTEM

Bu bölümde kullanılan veri seti ve çalışmanın mimarisi, veri önişleme süreci, veri setinde kullanılan bölme yöntemleri, KNIME programında oluşturulan modeller ve modellerin performans ölçütleri hakkında bilgiler verilmiştir.

3.1. Veri Seti ve Çalışmanın Mimarisi Hakkında

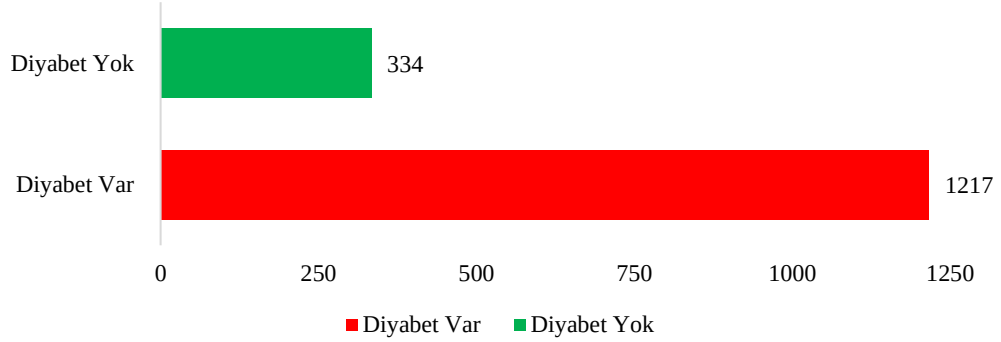
Bu çalışmada kullanılan veri seti Türkiye'deki bir devlet hastanesinin bilgi işlem biriminden elde edilmiştir. Çalışmanın yapılabilmesi için Bursa Uludağ Üniversitesi Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu'ndan alınan etik kurul kararı ve toplantı tutanağı EK-1'de sunulmuştur.

Bu çalışma için oluşturulan mimari Şekil 3.1'de gösterilmiştir. Devlet hastanesinin bilgi işlem biriminden alınan ham veri ilk olarak veri önişleme sürecinden geçirilmiştir. Veri seti incelendiğinde protein, ürik asit, albümin, LDL kolesterol gibi özniteliklerde eksik veriler olduğu için, eksik verilerin yerine ortalama değerler eklenmiştir. Ayrıca veri setinde ALT, Albumin, Hb1Ac başta olmak üzere bazı özniteliklerde aykırı değerler tespit edildiği için bu değerler silme yöntemi ile ortadan kaldırılmıştır. Bu işlemlerin ardından tanı ve cinsiyet değerleri için ikili değişken olarak 0 ve 1 değerleri atanmıştır. Son olarak veri setindeki özniteliklerin farklı aralıklarda değerler aldığı tespit edildiğinden normalizasyon işlemi uygulanmıştır. Veri önişleme sürecinin ardından veri seti eğitim ve test verisi olarak ikiye bölünmüştür. Veri seti eğitim ve test verisi olarak bölünürken yüzdesel bölme (%70-%30) ve k-katlı çapraz doğrulama (k=5) teknikleri kullanılmıştır. Önişlenmiş veri setinde sınıf dengesizliği olduğundan dolayı eğitim veri setleri üzerinde rastgele örneklem azaltma, rastgele aşırı örnekleme ve sentetik azınlık aşırı örnekleme teknikleri uygulanmıştır. Sınıflandırma modelleri eğitim veri setleriyle eğitildikten sonra test veri setlerinde performansları ölçülmüştür. Performans ölçütleri olarak doğruluk, kappa istatistiği, AUC değeri, kesinlik, duyarlılık ve F-ölçütü kullanılmıştır.



Şekil 3.1. Diyabet veri seti için sınıflandırma modellerinin oluşturulma ve değerlendirilme süreci

Veri seti 01.01.2021-31.12.2021 tarihleri arasında devlet hastanesinin iç hastalıkları ile endokrin ve metabolizma hastalıkları birimine başvuran 1551 hastaya aittir. Veri seti içerisinde hastaların tam kan sayımı, hormon ve biyokimya testlerinden elde edilen 42 adet öznitelik mevcuttur. Hedef değişken olarak ise uzman doktorların kliniğe gelen hastalara koymuş oldukları tanı değişkeni kullanılmıştır. Veri seti hakkında detaylı bilgiler (değişken türü, değişkenler, KNIME adı, ölçeği, açıklama, verilerin dağılımı) EK-2'de verilmiştir. Veri setine bakıldığında 1551 kişiden 1217'sinin "diyabet var" tanısı aldığı geriye kalan 334 kişinin ise "diyabet yok" olarak tanı aldığı belirlenmiştir. Şekil 3.2'ye bakıldığında sınıf etiketlerinin dağılımı görülmektedir.



Şekil 3.2. Diyabet veri setinde sınıf etiketlerinin dağılımı

Şekil 3.2'ye bakıldığında veri setindeki sınıf dengesizliği açık şekilde gözükmektedir. Bu sınıflandırma yapılırken sadece bir sınıfa yönelik tahminde kolaylık sağlayabilir ancak modelin genel performansını düşürebilmektedir. Sınıf dengesizliğinin oranı Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Diyabet veri setindeki sınıf dengesizliği oranı

Örneklem Sayısı	Değişken Sayısı	Sürekli Değişken Sayısı	Kategorik Değişken Sayısı	Sınıf Dengesizliği Oranı (Çoğunluk Üye Sayısı /Azınlık Üye Sayısı)
1551	43	41	2	3,64

3.2. Kullanılan Program, Operatörler ve Parametreler

Bu çalışmanın tüm aşamalarında KNIME Analytics Platformu'nun 4.6.0 versiyonu kullanılmıştır (KNIME, 2022). KNIME programında sınıflandırma modelleri oluşturulurken kullanılan operatör ve parametreler Çizelge 3.2'de verilmiştir.

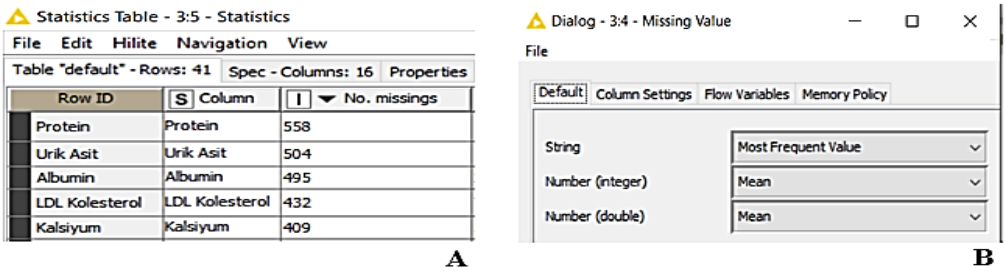
Çizelge 3.2. Kullanılan sınıflandırma algoritmaları, operatör ve parametreler

Sınıflandırma Algoritmaları	KNIME Eğitim Operatörleri	Parametreler
LR	Logistic Learner	Epsilon = 10^{-5}
k-EYK	K Nearest Neighbour	Euclidean distance; k=3
NB	Naive Bayes Learner	-
C4.5	Decision Tree Learner	Min number records per node=2
RO	Random Forest Learner	Tree depth=10; Minimum node size=1; Number of models=100
ÇKA	RProp MLP Learner	Iterations=100; Hidden layers=1; Neurons=10

3.3. Veri Ön İşleme

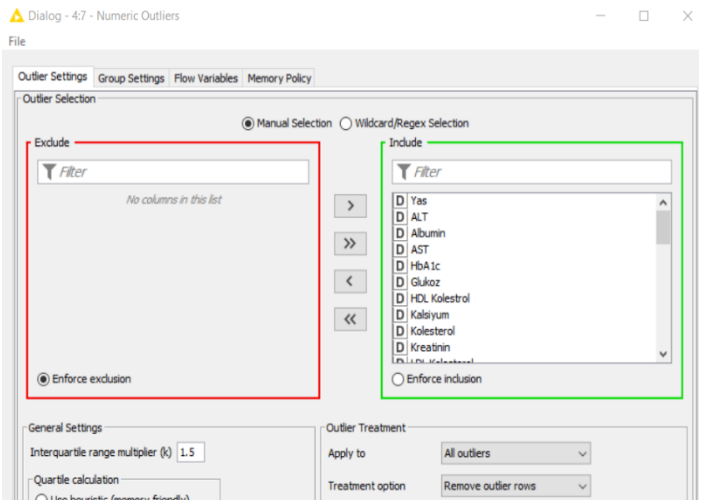
Çalışmada kullanılan veri setinde incelediğinde farklı ön işleme yöntemleri uygulanmıştır. Bu bölümde veri setine uygulanan ön işleme yöntemlerinden bahsedilmiştir.

- Veri seti "Excel Reader" operatörü ile programa aktarıldıktan sonra "Statistics" operatörü eklenmiş ve veri setinde bazı özniteliklerde eksik verilerin olduğu görülmüştür (Şekil 3.3A). Tüm özniteliklerdeki eksik veri sayısı ise EK-2'de verilmiştir. Bu eksik veriler "Missing Values" operatörü kullanılarak ilgili özniteliğin ortalama değerleri ile tamamlanmıştır (Şekil 3.3B).



Şekil 3.3. Eksik verilerin giderilmesi **A)** KNIME Statistics ekranı **B)** KNIME Missing Value ekranı

- Eksik verilerin giderilmesinin ardından veri setindeki “Cinsiyet” değişkenindeki kadın ve erkek değerleri sırası ile “0 ve 1” olarak etiketlenmiştir.
- “Tanı” değişkeninde yer alan “E11” ile başlayan kodlar, hastalıkların uluslararası sınıflandırılması kriterlerine (WHO, 2022b) göre insüline bağlı olmayan diyabet yani tip-2 diyabetin göstergesidir. Veri setinde “E11” ile başlayan kod tanısı almayan kişilerde diyabet olmadığı için “0” diğer kişiler yani “E11” ile başlayan kod tanısı alan kişiler ise diyabet hastası yani “1” olarak etiketlenmiştir.
- Veri ön işlemede bir diğer önemli aşama ise aykırı değerlerin ne yapılacağına karar verilmesi konusudur. Bu çalışmada kullanılan diyabet veri seti incelediğinde aykırı değerlerin olduğu görülmüştür. Bu aykırı değerler KNIME programındaki “Numeric Outliers” operatörü kullanılarak veri setinden çıkarılmıştır (Şekil 3.4). Numeric Outliers operatöründe aykırı veri çıkarma işlemi yapılırken IQR tekniği kullanılmış olup, bu işlemin ardından veri sayısı 1551’den 449’a düşmüştür.



Şekil 3.4. KNIME Numeric Outliers operatörü ekranı

- Veri seti incelendiğinde bazı nümerik özniteliklerin farklı aralıkta değerler aldığı görülmüştür. Örneğin “PCT” değişkeni 0,01 ile 0,66 arasında, “PLT” değişkeni 29 ile 690 arasında değerler almıştır. Bu durum modelin sınıflandırma performansını olumsuz yönde etkileyebilmektedir. Bunun için “Normalizer” operatörü kullanılarak tüm nümerik verilere normalizasyon işlemi uygulanmış ve verilerin 0,1 ile 0,9 arası değerler alması sağlanmıştır (bkz. Şekil 3.5). KNIME

programında normalizasyon işlemi yapılırken kullanılan denklem (3.1) ile gösterilmiştir (Yavuz ve Deveci, 2012):

$$x_{normalize} = 0{,}8 * \left(\frac{x_i - x_{min}}{x_{maks} - x_{min}} \right) + 0{,}1 \qquad (3.1)$$

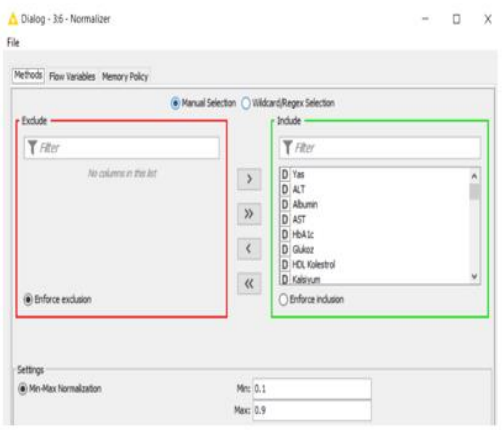
Denklem 3.1'de kullanılan değişkenlerin anlamı aşağıda verilmiştir:

$x_{normalize}$: Normalize edilen değer

x_i : Veri seti içerisindeki normalize edilecek i. girdi değeri

x_{maks} : Veri seti içerisindeki maksimum değer

x_{min} : Veri seti içerisindeki minimum değer



A

D PCT	D PLT	D RBC
0.308	0.253	0.275
0.274	0.322	0.395
0.274	0.363	0.515
0.257	0.312	0.453
0.266	0.33	0.33
0.249	0.246	0.492
0.1	0.143	0.482
0.423	0.445	0.278
0.282	0.279	0.327

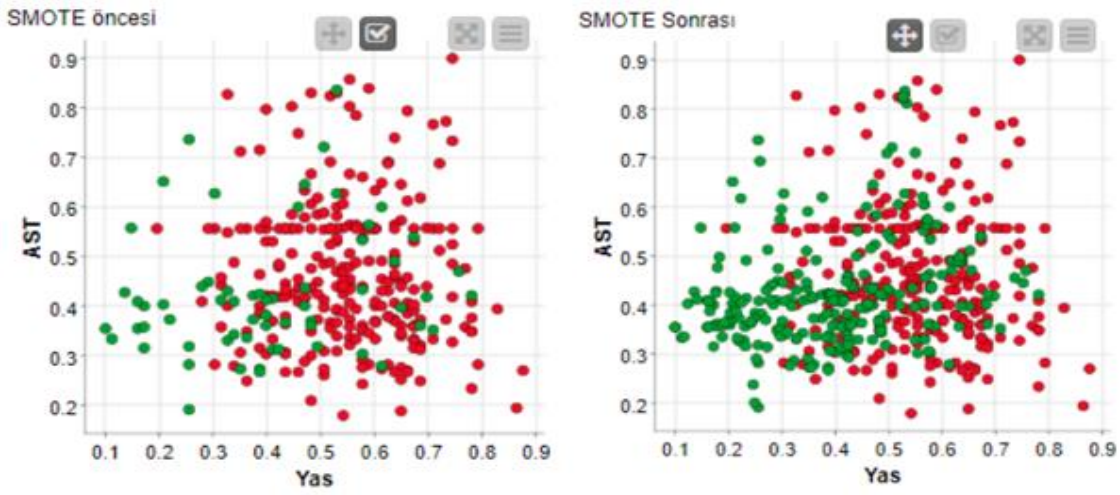
B

Şekil 3.5. Verilerin normalize edilmesi **A)** KNIME Normalizer operatörü ekranı **B)** Normalize edilmiş veri örneği

- Aydın (2020) tarafından yapılan çalışmada, literatürdeki çoğu çalışmada dengesiz veri setlerinde bölme işleminden önce dengeleme işleminin yapıldığını ancak bu yöntemin doğru bir yaklaşım olmadığını belirtilmiştir. Kullanmış olduğu veri setinde, çapraz geçerleme esnasında yeniden örnekleme yöntemlerini kullanarak dengesizliği gidermiştir. Bu yaklaşım göz önünde bulundurularak yapılan bu tez çalışmasında veri setindeki sınıf etiketleri arasındaki dengesizliğin giderilmesi için üç farklı yeniden örnekleme tekniği kullanılmıştır. Bu yöntemler; rastgele

örneklem azaltma, rastgele aşırı örnekleme ve sentetik azınlık aşırı örneklemedir. Orijinal veri seti, eğitim ve test verisi olarak ayrıldıktan sonra eğitim veri setindeki dengesizliğin giderilmesi için yeniden örnekleme yöntemleri uygulanmıştır.

- Şekil 3.6'da eğitim veri seti dengelenmeden önce ve dengelendikten sonraki durum gösterilmiştir.

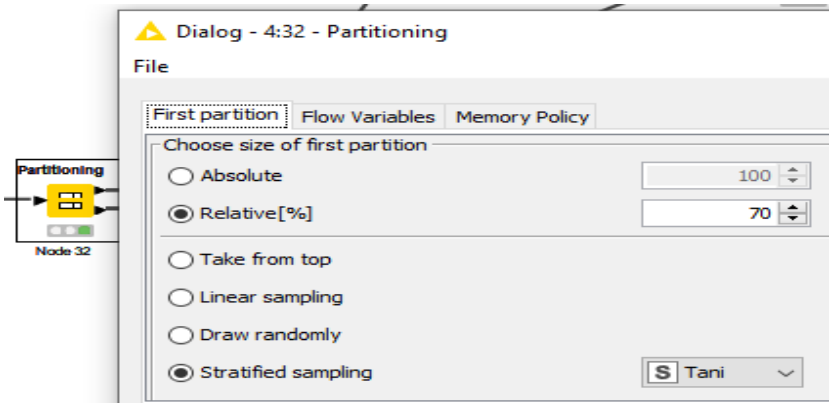


Şekil 3.6. KNIME programında SMOTE yöntemi ile eğitim veri setinin dengelenmesi

3.4. Eğitim ve Test Verisinin Ayrıştırılması

Sınıflandırma modellerinde öncelikle modelin eğitilmesi ardından test edilmesi gerekmektedir. Bu sebeple veri setinin eğitim ve test verisi olarak ayrıştırılması gerekmektedir. Literatürde en çok kullanılan veri bölme yöntemleri yüzdesel bölme ve k-katlı çapraz doğrulamadır. Yüzdesel bölme yönteminde veri seti %70 eğitim - %30 test, %80 eğitim - %20 test vb. şeklinde ayrıştırılabilmektedir. K-katlı çapraz doğrulama yönteminde ise k=5 veya k=10 gibi değerler seçilerek tüm veri seti parçalara bölünür ve parçaların her biri sırayla test veri seti olarak kullanılır. Bu çalışmada iki farklı veri bölme tekniği de kullanılmıştır.

Yüzdesel bölme yönteminde tüm veri setinin %70'i eğitim, %30'u ise test verisi olarak ayrıştırılmıştır. Ayrıca veri seti içerisindeki hedef değişkendeki sınıf etiketlerinin dağılımının korunması için tabakalı örnekleme tekniği kullanılmıştır. KNIME programında veri bölme için "Partitioning" operatörü kullanılmıştır. (bkz. Şekil 3.7)



Şekil 3.7. KNIME Partitioning operatörü ekranı

Veri önişleme işleminin ardından 449 hasta örneği KNIME programında %70 eğitim ve %30 test verisi olarak ayrıştırılarak dört farklı eğitim veri seti oluşturulmuştur.

Orijinal eğitim veri setinde, eğitime katılacak veriler üzerinde herhangi bir dengeleme işlemi yapılmamıştır. Sınıf etiketi dağılımları; 62 kişi diyabet olmayan, 252 kişi diyabet olan şeklinde ayrıldığı görülmektedir. Toplamda 314 örnek içeren bir eğitim veri seti oluşturulmuştur. Bu durumda veri setindeki sınıf dengesizliği oranı 3,64'ten 4,06'ya çıkmıştır.

Rastgele örneklem azaltılmış eğitim veri setinde, eğitime katılacak veriler üzerinde rastgele örnek azaltma işlemi uygulanmıştır. Bu işlem sonucunda sınıf etiketi dağılımları; 62 kişi diyabet olmayan, 62 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 124 örnek içeren bir eğitim veri seti oluşturulmuştur.

Rastgele aşırı örneklenmiş eğitim veri setinde, eğitime katılacak veriler üzerinde rastgele aşırı örnekleme işlemi uygulanmıştır. Bu işlem sonucunda sınıf etiketi dağılımları; 252 kişi diyabet olmayan, 252 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 504 örnek içeren bir eğitim veri seti oluşturulmuştur.

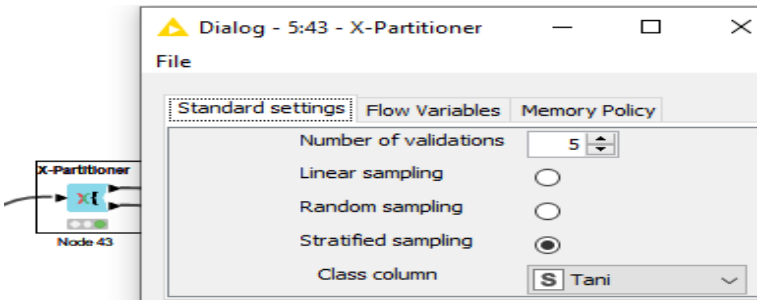
Sentetik azınlık aşırı örneklenmiş eğitim veri setinde, eğitime katılacak verilerin 5 en yakın komşusuna bakılarak sentetik veriler üretilmiştir. . Bu işlem sonucunda sınıf etiketi dağılımları; 252 kişi diyabet olmayan, 252 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 504 örnek içeren bir eğitim veri seti oluşturulmuştur.

Oluşturulan dört farklı eğitim veri seti hakkında bilgiler Çizelge 3.3'te verilmiştir.

Çizelge 3.3. Yüzdesel bölme ile oluşturulan eğitim veri setleri

Eğitim Veri Setleri	0: Negatif (Diyabet Olmayan Hasta Sayısı)	1: Pozitif (Diyabet Olan Hasta Sayısı)	Toplam Hasta Sayısı
Orijinal Eğitim Veri Seti	62	252	314
RUS Eğitim Veri Seti	62	62	124
ROS Eğitim Veri Seti	252	252	504
SMOTE Eğitim Veri Seti	252	252	504

k-katlı çapraz doğrulama yönteminde k=5 seçilerek eğitim ve test verileri oluşturulmuştur. Hedef değişkendeki sınıf etiketlerinin dağılımının korunması için bu yöntemde de tabakalı örnekleme tekniği kullanılmıştır. KNIME programında veri bölme için "X-Partitioner" operatörü kullanılmıştır (bkz. Şekil 3.8).



Şekil 3.8. KNIME X-Partitioner operatörü ekranı

Veri önişleme işleminin ardından 449 hasta örneği KNIME programında 5-katlı çapraz doğrulama tekniğiyle eğitim ve test olarak ayrıştırılarak dört farklı eğitim veri seti oluşturulmuştur.

Orijinal eğitim veri setinde, eğitime katılacak veriler üzerinde herhangi bir dengeleme işlemi yapılmamıştır. Sınıf etiketi dağılımları; 72 kişi diyabet olmayan, 288 kişi diyabet olan şeklinde ayrıldığı görülmektedir. Toplamda 360 örnek içeren bir eğitim veri seti oluşturulmuştur. Bu durumda veri setindeki sınıf dengesizliği oranı 3,64'ten 4'e çıkmıştır.

Rastgele örneklem azaltılmış eğitim veri setinde, eğitime katılacak veriler üzerinde rastgele örnek azaltma işlemi uygulanmıştır. Bu işlem sonucunda sınıf etiketi dağılımları;

72 kişi diyabet olmayan, 72 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 144 örnek içeren bir eğitim veri seti oluşturulmuştur.

Rastgele aşırı örneklenmiş eğitim veri setinde, eğitime katılacak veriler üzerinde rastgele aşırı örnekleme işlemi uygulanmıştır. Bu işlem sonucunda sınıf etiketi dağılımları; 288 kişi diyabet olmayan, 288 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 576 örnek içeren bir eğitim veri seti oluşturulmuştur.

Sentetik azınlık aşırı örneklenmiş eğitim veri setinde, eğitime katılacak verilerin 5 en yakın komşusuna bakılarak sentetik veriler üretilmiştir. Bu işlem sonucunda sınıf etiketi dağılımları; 288 kişi diyabet olmayan, 288 kişi diyabet olan şeklinde dengelenmiştir. Toplamda 576 örnek içeren bir eğitim veri seti oluşturulmuştur.

Oluşturulan dört farklı eğitim veri seti hakkında bilgiler Çizelge 3.4'te verilmiştir.

Çizelge 3.4. k-katlı çapraz doğrulama ile oluşturulan eğitim veri setleri

Eğitim Setleri	0: Negatif (Diyabet Olmayan Hasta Sayısı)	1: Pozitif (Diyabet Olan Hasta Sayısı)	Toplam Hasta Sayısı
Orijinal Eğitim Veri Seti	72	288	360
RUS Eğitim Veri Seti	72	72	144
ROS Eğitim Veri Seti	288	288	576
SMOTE Eğitim Veri Seti	288	288	576

3.5. KNIME Programında Oluşturulan Modeller

Bu bölümde KNIME programında oluşturulan modeller hakkında bilgi verilmiştir.

3.5.1. LR modelleri

LR modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-3'te verilmiştir.

EK-3'te verilen 1 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle

lojistik regresyon modelinin eğitilmesi için “Logistic Regression Learner” operatörü, test edilebilmesi için ise “Logistic Regression Predictor” operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için “ROC Curve” operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise “Scorer” operatörü kullanılmıştır.

EK-3’teki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için “Nominal Value Row Splitter” operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından “Bootstrap Sampling” operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için “Concatenate” operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-3’teki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için “Nominal Value Row Splitter” operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından “Bootstrap Sampling” operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk sınıfı ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için “Concatenate” operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-3’te verilen 4 numaralı görselde öncelikle ön işlenmiş veri “Excel Reader” operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra “Partitioning” operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise “SMOTE” operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra lojistik regresyon modelinin eğitilmesi için “Logistic Regression Learner” operatörü, test edilebilmesi için ise “Logistic Regression Predictor” operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için “ROC Curve” operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise “Scorer” operatörü kullanılmıştır.

EK-3’te verilen 5 numaralı görselde öncelikle ön işlenmiş veri “Excel Reader” operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra “X-Partitioner” operatörü ile k=5

seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle lojistik regresyon modelinin eğitilmesi için "Logistic Regression Learner" operatörü, test edilebilmesi için ise "Logistic Regression Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-3'te 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-3'te 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-3'te verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra lojistik regresyon modelinin eğitilmesi için "Logistic Regression Learner" operatörü, test edilebilmesi için ise "Logistic Regression Predictor" operatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü,

doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

3.5.2. NB modelleri

NB modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-4'te verilmiştir.

EK-4'te verilen 1 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle NB modelinin eğitilmesi için "Naive Bayes Learner" operatörü, test edilebilmesi için ise "Naive Bayes Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-4'teki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-4'teki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk sınıfı ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-4'te verilen 4 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra NB modelinin eğitilmesi için "Naive Bayes Learner" operatörü, test edilebilmesi için ise "Naive Bayes Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-4'te verilen 5 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle NB modelinin eğitilmesi için "Naive Bayes Learner" operatörü, test edilebilmesi için ise "Naive Bayes Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-4'teki 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-4'teki 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-4'te verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra NB modelinin eğitilmesi için "Naive Bayes Learner" operatörü, test edilebilmesi için ise "Naive Bayes Predictor" operatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

3.5.3. k-EYK modelleri

k-EYK modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-5'te verilmiştir.

EK-5'te verilen 1 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle k-EYK modelinin eğitilmesi ve test edilmesi için "K Nearest Neighbor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-5'teki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-5'teki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk sınıfı ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-5'te verilen 4 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra k-EYK modelinin eğitilmesi ve test edilebilmesi için "K Nearest Neighbor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-5'te verilen 5 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle k-EYK modelinin eğitilmesi ve test edilebilmesi için "K Nearest Neighbor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-5'teki 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-5'teki 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-5'te verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra k-EYK modelinin eğitilmesi ve test edilebilmesi için "K Nearest Neighbor" ooperatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

3.5.4. C4.5 modelleri

C4.5 modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-6'da verilmiştir.

EK-6'da verilen 1 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle C4.5 modelinin eğitilmesi için "Decision Tree Learner" operatörü, test edilebilmesi için ise "Decision Tree Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-6'daki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-6'daki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-6'da verilen 4 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra C4.5 modelinin eğitilmesi için "Decision Tree Learner" operatörü, test edilebilmesi için ise "Decision Tree Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-6'da verilen 5 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle C4.5 modelinin eğitilmesi için "Decision Tree Learner" operatörü, test edilebilmesi için ise "Decision Tree Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-6'daki 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-6'daki 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-6'da verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra C4.5 modelinin eğitilmesi için "Decision Tree Learner" operatörü, test edilebilmesi için ise "Decision Tree Predictor" operatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

3.5.5. RO modelleri

RO modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-7'de verilmiştir.

Şekil 3.13'te verilen 1 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra "Partitioning"

operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle RO modelinin eğitilmesi için "Random Forest Learner" operatörü, test edilebilmesi için ise "Random Forest Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-7'deki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-7'deki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-7'de verilen 4 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra RO modelinin eğitilmesi için "Random Forest Learner" operatörü, test edilebilmesi için ise "Random Forest Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-7'de verilen 5 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle RO modelinin eğitilmesi için "Random Forest Learner" operatörü, test edilebilmesi için ise "Random Forest Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-7'deki 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-7'deki 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-7'de verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra RO modelinin eğitilmesi için "Random Forest Learner" operatörü, test edilebilmesi için ise "Random Forest Predictor" operatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC

değerinin elde edilebilmesi için “ROC Curve” operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise “Scorer” operatörü kullanılmıştır.

3.5.6. ÇKA modelleri

ÇKA modelleri oluşturulurken dört farklı eğitim veri seti ve iki farklı veri bölme yöntemi kullanılmıştır. Toplamda sekiz senaryo üzerinde çalışma yapılmıştır. Senaryolar için oluşturulan modellerin tamamı EK-8’de verilmiştir.

EK-8’de verilen 1 numaralı görselde öncelikle ön işlenmiş veri “Excel Reader” operatörü ile programa aktarılmıştır. Veriler programa aktarıldıktan sonra “Partitioning” operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle ÇKA modelinin eğitilmesi için “RProp MLP Learner” operatörü, test edilebilmesi için ise “MultiLayerPerceptron Predictor” operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için “ROC Curve” operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise “Scorer” operatörü kullanılmıştır.

EK-8’deki 2 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için “Nominal Value Row Splitter” operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından “Bootstrap Sampling” operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için “Concatenate” operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-8’deki 3 numaralı görselde, 1 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için “Nominal Value Row Splitter” operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından “Bootstrap Sampling” operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için “Concatenate” operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-8'de verilen 4 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "Partitioning" operatörü ile %70 eğitim verisi, %30 test verisi olarak ayrıştırılmıştır. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra ÇKA modelinin eğitilmesi için "RProp MLP Learner" operatörü, test edilebilmesi için ise "MultiLayerPerceptron Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-8'de verilen 5 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için herhangi bir işlem yapılmadan orijinal eğitim veri setiyle RO modelinin eğitilmesi için "RProp MLP Learner" operatörü, test edilebilmesi için ise "MultiLayerPerceptron Predictor" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

EK-8'deki 6 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak çoğunluk sınıfındaki veriler azınlık sınıfı ile eşit miktarda olması için RUS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele örneklem azaltılmış eğitim veri seti oluşturulmuştur.

EK-8'deki 7 numaralı görselde, 5 numaralı görselden farklı olarak eğitim veri setindeki sınıf dengesizliğinin giderilebilmesi için "Nominal Value Row Splitter" operatörü ile sınıf etiketleri 0 ve 1 olarak ayrıştırılmıştır. Bu işlemin ardından "Bootstrap Sampling" operatörü kullanılarak azınlık sınıfındaki veriler çoğunluk ile eşit miktarda olması için ROS işlemi uygulanmıştır. Son olarak ise azınlık ve çoğunluk sınıfındaki verilerin birleştirilmesi için "Concatenate" operatörü kullanılarak rastgele aşırı örneklenmiş eğitim veri seti oluşturulmuştur.

EK-8'de verilen 8 numaralı görselde öncelikle ön işlenmiş veri "Excel Reader" operatörü ile programa aktarılmıştır. Veriler aktarıldıktan sonra "X-Partitioner" operatörü ile k=5 seçilerek beş eşit parçaya bölünmüştür. Eğitim veri setindeki sınıf dengesizliğinin giderilmesi için ise "SMOTE" operatörü kullanılarak sınıf etiketlerinin veri sayıları dengelenmiştir. Sınıf dengesizliği giderildikten sonra ÇKA modelinin eğitilmesi için "RProp MLP Learner" operatörü, test edilebilmesi için ise "MultiLayerPerceptron Predictor" operatörü kullanılmıştır. Tüm veri setindeki yapılan denemeler ve sonuçların birleştirilebilmesi için ise "X-Aggregator" operatörü kullanılmıştır. Son olarak; AUC değerinin elde edilebilmesi için "ROC Curve" operatörü, doğruluk, kesinlik, duyarlılık, F-ölçütü değerlerinin elde edilebilmesi için ise "Scorer" operatörü kullanılmıştır.

3.6. Modellerin Performans Ölçütleri

Bu çalışmada sınıflandırma algoritmalarının performansları değerlendirilirken literatürde kullanılan karışıklık matrisinden elde edilen değerler (doğruluk, kesinlik, duyarlılık, F-ölçütü) ile alıcı işlem karakteristik eğrisinin altında kalan alanın değeri veren AUC değeri kullanılmıştır. Karışıklık matrisinin gösterimi Çizelge 3.5'te verilmiştir.

Çizelge 3.5. Karışıklık matrisi

Gerçek Sınıf	**Tahmin Sınıfı**	
	1: Pozitif (Diyabet Var)	0: Negatif (Diyabet Yok)
1: Pozitif (Diyabet Var)	Doğru Pozitif (DP)	Yanlış Negatif (YN)
0: Negatif (Diyabet Yok)	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Doğru Pozitif (DP): Gerçekten diyabet hastalığına sahip bir kişinin sınıflandırma modeli tarafından diyabet hastası olarak sınıflandırılması

Doğru Negatif (DN): Gerçekten diyabet hastalığına sahip olmayan bir kişinin sınıflandırma modeli tarafından diyabet hastası değil olarak sınıflandırılması

Yanlış Pozitif (YP): Gerçekten diyabet hastalığına sahip olmayan bir kişinin sınıflandırma modeli tarafından diyabet hastası olarak sınıflandırılması

Yanlış Negatif (YN): Gerçekten diyabet hastalığına sahip olan bir kişinin sınıflandırma modeli tarafından diyabet hastası değil olarak sınıflandırılması

Karışıklık matrisinde elde edilen değerler ile aşağıdaki performans ölçütleri hesaplanabilmektedir:

Doğruluk (Accuracy): Bir sınıflandırma modelinde tüm değerler içerisindeki doğru sınıflandırılan değer sayısının oranını ifade etmektedir. Doğruluk oranı, denklem (3.1) ile gösterilmektedir:

$$\text{Doğruluk Oranı} = \frac{\text{DP} + \text{DN}}{\text{DP} + \text{DN} + \text{YP} + \text{YN}} \tag{3.1}$$

Kesinlik (Precision): Pozitif olarak tahmin edilen değerlerden gerçekte kaç tanesinin pozitif sınıfa ait olduğunu gösterir. Kesinlik, denklem (3.2) ile gösterilmektedir:

$$\text{Kesinlik} = \frac{\text{DP}}{\text{DP} + \text{YP}} \tag{3.2}$$

Duyarlılık (Recall): Gerçekte pozitif olarak sınıflandırılması değerler arasından kaç tanesinin doğru şekilde pozitif sınıfa atandığını göstermektedir. Aynı zamanda doğru pozitif oranı (DPR) olarak da ifade edilebilmektedir. Duyarlılık, denklem (3.3) ile gösterilmektedir:

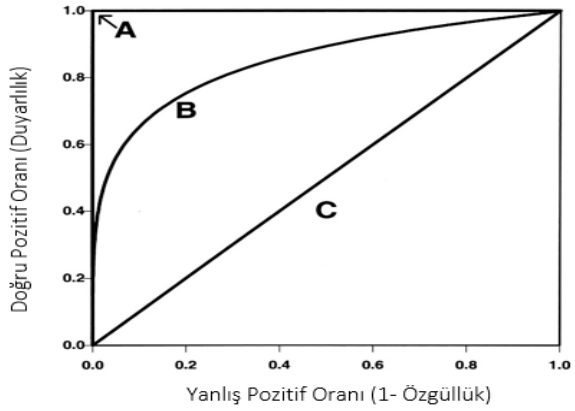
$$\text{Duyarlılık} = \frac{\text{DP}}{\text{DP} + \text{YN}} \tag{3.3}$$

F-ölçütü: F-ölçütü değeri özellikle dengesiz sınıfa sahip veri setlerinde modelin sınıflandırma performansını ölçmek için sıklıkla kullanılmaktadır. Kesinlik ve duyarlılık değerlerinin harmonik ortalaması ile hesaplanmaktadır. F-ölçütü, denklem (3.4) ile gösterilmektedir:

$$\text{F} - \text{ölçütü} \; = \; 2 * \frac{\text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{3.4}$$

Bu dört değerin dışında sınıflandırma modelinin performansını ölçerken kullanılan bir diğer ölçüt ise alıcı işlem karakteristiği eğrisinin (ROC) altında kalan alanın değerini belirten AUC (Area under the ROC Curve) değeridir.

AUC Değeri: ROC eğrisinin altında kalan alanın sayısal değerini ifade etmektedir. ROC eğrisi oluşturulurken x-ekseninde yanlış pozitiflerin oranı yer alırken y-ekseninde ise doğru pozitiflerin oranı yer almaktadır. AUC değeri 0,5 değerine eşit olduğunda model rastgele sınıflandırma yaptığını ve başarısız olduğu söylenebilmektedir. Bu değer 1 değerine ne kadar yakın olursa modelin sınıflandırma performansının o kadar iyi olduğu belirtilmektedir. Şekil 3.9'a bakılacak olursa A noktasında AUC=1, B eğrisi için AUC=0,85 ve C doğrusu için AUC=0,5 sayılarını temsil etmektedir. C doğrusundaki AUC=0,5 değeri modelin rastgele bir sınıflandırma yaptığını, AUC değerinin B eğrisine doğru evrildiğinde ise modelin performansının iyileştiği ve A noktasındaki AUC=1 değeri için ise modelin mükemmel bir sınıflandırma yaptığı söylenebilmektedir (Zou vd., 2007).



Şekil 3.9. ROC Eğrisi (Zou ve diğerleri, 2007)

Kappa İstatistiği (Kİ): Kategorik değişkenler arasındaki ilişkiyi ölçerken iki değerlendirici arasındaki uyumu ölçen istatistiki bir ölçektir. İki değerlendirici arasındaki uyumun derecesi ölçülürken Cohen'in kappa istatistiği, ikiden fazla değerlendirici arasındaki uyumun derecesi ölçülürken ise Fleiss'in kappa istatistiği kullanılmaktadır (Fleiss, 1971).

Cohen'in kappa istatistiği (k), denklem (3.5) ile gösterilmektedir (Cohen, 1960):

$$k = \frac{p_o - p_c}{1 - p_c} \tag{3.5}$$

Burada;

$\boldsymbol{p_o}$: İki değerlendiricinin gözlemleri arasındaki uyumun toplam oranı

$\boldsymbol{p_c}$: İki değerlendiricinin gözlemleri arasındaki uyumun tesadüfen olma oranı

Landis ve Koch (1977), "k" değerini yorumlarken Çizelge 3.6'daki tabloyu önermişlerdir.

Çizelge 3.6. Kappa istatistiği yorumları (Kılıç, 2015)

Kappa İstatistiği (k)	Yorum
<0	Uyum düzeyi çok kötü
0,01 – 0,20	Uyum düzeyi önemsiz
0,21 – 0,40	Uyum düzeyi zayıf
0,41 – 0,60	Uyum düzeyi orta
0,61 – 0,80	Uyum düzeyi iyi
0,81 – 1,00	Uyum düzeyi çok iyi

4. BULGULAR ve TARTIŞMA

Yapılan bu çalışmada; iki farklı veri seti bölme tekniği (yüzdesel bölme, k-katlı çapraz doğrulama) ve üç farklı yeniden örnekleme tekniği (RUS, ROS, SMOTE) kullanılarak farklı eğitim veri setleri oluşturulmuştur. Dengelenmemiş veri seti ile performans karşılaştırılması yapılabilmesi için ayrıca orijinal eğitim veri seti (dengelenmemiş) de çalışmaya dahil edilmiştir. Tüm bu eğitim veri setleri, altı farklı makine öğrenmesi (LR, NB, k-EYK, C4.5, RO, ÇKA) modelinin eğitim sürecinde kullanılmış ve sonunda test veri setlerinde sınıflandırma modellerinin performansları karşılaştırılmıştır. Toplamda 48 farklı senaryo incelenmiş olup, elde edilen bulgular bu bölümde verilmiştir.

4.1. Yüzdesel Bölme ile Elde Edilen Sonuçlar

Veri önişleme yapılmış veri seti, %70 eğitim- %30 test verisi olarak ayrılmıştır. Veri bölme işleminin ardından dengeleme yapılmamış orijinal eğitim veri seti ve dengelenmiş üç farklı eğitim veri seti üzerinde farklı makine öğrenmesi yöntemleri (LR, NB, k-EYK, C4.5, RO, ÇKA) uygulanmıştır.

4.1.1. LR yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş LR modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.1'de verilmiştir.

Çizelge 4.1. LR yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	94,81	92,59	94,81	**96,30**
AUC Değeri	0,988	0,986	0,996	**0,997**
Ortalama F-ölçütü	**0,988**	0,945	0,924	0,940
Kİ	0,821	0,792	0,848	**0,879**

Çizelge 4.1'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %94,81'den %96,30'a, AUC değeri 0,988'den 0,997'ye, Kİ değeri ise 0,821'den 0,879'a çıkmıştır. Ortalama F-ölçütü değeri orijinal eğitim veri setinde 0,988 değeri ile en iyi sonucu vermiştir.

4.1.2. NB yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş NB modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.2'de verilmiştir.

Çizelge 4.2. NB yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	**96,30**	94,81	94,81	94,07
AUC Değeri	**0,995**	0,991	0,989	0,986
Ortalama F-ölçütü	**0,941**	0,918	0,922	0,899
Kİ	**0,883**	0,836	0,844	0,798

Çizelge 4.2'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa tüm performans ölçütlerinin değerlerinin düştüğü tespit edilmiştir.

4.1.3. k-EYK yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş k-EYK modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.3'te verilmiştir.

Çizelge 4.3. k-EYK yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	82,96	82,96	**83,70**	80,74
AUC Değeri	0,778	**0,848**	0,829	0,838
Ortalama F-ölçütü	0,644	0,744	**0,764**	0,749
Kİ	0,886	**0,892**	0,781	0,840

Çizelge 4.3'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %82,96'dan %83,70'e, AUC değeri 0,778'den 0,848'e, ortalama F-ölçütü değeri 0,644'ten 0,764'e, Kİ değeri ise 0,886'dan 0,892'ye çıkmıştır.

4.1.4. C4.5 yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş C4.5 modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.4'te verilmiştir.

Çizelge 4.4. C4.5 yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	95,56	96,30	**97,04**	**97,04**
AUC Değeri	**0,986**	0,978	0,968	0,957
Ortalama F-ölçütü	0,942	0,946	**0,955**	0,952
Kİ	0,857	0,892	**0,910**	0,905

Çizelge 4.4'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %95,56'dan %97,04'e, ortalama F-ölçütü 0,942'den 0,955'e, Kİ değeri ise 0,857'den 0,910'a çıkmıştır. AUC değeri orijinal eğitim veri setinde 0,986 değeri ile en iyi sonucu vermiştir.

4.1.5. RO yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş RO modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.5'te verilmiştir.

Çizelge 4.5. RO yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	94,81	**99,26**	97,04	98,52
AUC Değeri	0,996	**0,997**	0,988	0,996
Ortalama F-ölçütü	0,915	**0,989**	0,944	0,977
Kİ	0,831	**0,977**	0,889	0,954

Çizelge 4.5'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %94,81'den %99,26'a, AUC değeri 0,996'dan 0,997'ye, ortalama F-ölçütü değeri 0,915'ten 0,989'a, Kİ değeri ise 0,831'den 0,977'ye çıkmıştır.

4.1.6. ÇKA yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti %70 Eğitim- %30 Test verisi olarak ayrıldıktan sonra, dört farklı eğitim veri seti ile eğitilmiş ÇKA modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.6'da verilmiştir.

Çizelge 4.6. ÇKA yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	**96,30**	**96,30**	93,00	94,81
AUC Değeri	**0,997**	0,993	0,974	0,987
Ortalama F-ölçütü	0,943	**0,946**	0,890	0,920
Kİ	0,886	**0,892**	0,781	0,840

Çizelge 4.6'ya bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa ortalama F-ölçütü değeri 0,943'ten 0,946'ya, Kİ değeri ise 0,886'dan 0,892'ye çıkmıştır. DSO değeri orijinal eğitim veri seti ve rastgele örneklem azaltılmış eğitim veri setinde en iyi sonuçları verirken, AUC değeri ise en iyi sonucu orijinal eğitim veri setinde göstermiştir.

4.2. K-Katlı Çapraz Doğrulama ile Elde Edilen Sonuçlar

Veri önişleme yapılmış veri seti, 5-katlı çapraz doğrulama tekniği ile eğitim ve test verisi olarak ayrıştırılmıştır. Eğitim verisi olarak ayrıştırılan her bir veri setinde üç farklı dengeleme tekniği uygulanmıştır. Eğitim verileri farklı makine öğrenmesi yöntemleriyle (LR, NB, k-EYK, C4.5, RO, ÇKA) birleştirilerek performansları karşılaştırılmıştır.

4.2.1. LR yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. LR modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.7'de verilmiştir.

Çizelge 4.7'ye bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %94,43'ten %95,10'a, ortalama F-ölçütü 0,911'den 0,925'e, Kİ değeri ise 0,821'den 0,851'e çıkmıştır. AUC değeri orijinal eğitim veri setinde 0,987 değeri ile en iyi sonucu vermiştir.

Çizelge 4.7. LR yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	94,43	89,09	94,43	**95,10**
AUC Değeri	**0,987**	0,966	0,986	0,985
Ortalama F-ölçütü	0,911	0,844	0,916	**0,925**
Kİ	0,821	0,689	0,831	**0,851**

4.2.2. NB yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. NB modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.8'de verilmiştir.

Çizelge 4.8. NB yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	95,10	94,65	**95,99**	95,10
AUC Değeri	**0,989**	0,988	0,988	0,985
Ortalama F-ölçütü	0,923	0,919	**0,939**	0,921
Kİ	0,846	0,839	**0,877**	0,842

Çizelge 4.8'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %95,10'dan %95,99'a, ortalama F-ölçütü 0,923'ten 0,939'a, Kİ değeri ise 0,846'dan 0,877'ye çıkmıştır. AUC değeri orijinal eğitim veri setinde 0,989 değeri ile en iyi sonucu vermiştir.

4.2.3. k-EYK yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. k-EYK modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.9'da verilmiştir. Çizelge 4.9'a bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa AUC değeri 0,714'ten 0,859'a, Kİ değeri ise 0,443'ten 0,504'e çıkmıştır. DSO değeri ve ortalama F-ölçütü değeri en iyi sonuçlarını orijinal eğitim veri setinde göstermiştir.

Çizelge 4.9. k-EYK yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	**86,19**	78,62	82,63	81,07
AUC Değeri	0,714	0,842	0,807	**0,859**
Ortalama F-ölçütü	**0,815**	0,713	0,751	0,745
Kİ	0,443	0,433	**0,504**	0,496

4.2.4. C4.5 yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. C4.5 modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.10'da verilmiştir.

Çizelge 4.10. C4.5 yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	96,21	**97,55**	95,55	97,10
AUC Değeri	0,951	**0,980**	0,921	0,970
Ortalama F-ölçütü	0,940	**0,963**	0,929	0,955
Kİ	0,879	**0,925**	0,857	0,909

Çizelge 4.10'a bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %96,21'den %97,55'e, AUC değeri 0,951'den 0,980'e, ortalama F-ölçütü değeri 0,940'tan 0,963'e, Kİ değeri ise 0,879'dan 0,925'e çıkmıştır.

4.2.5. RO yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. RO modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.11'de verilmiştir. Çizelge 4.11'e bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %97,10'dan %97,55'e, ortalama F-ölçütü 0,955'ten 0,962'ye, Kİ değeri ise 0,909'dan 0,925'e çıkmıştır. AUC değeri orijinal eğitim veri setinde 0,992 değeri ile en iyi sonucu vermiştir.

Çizelge 4.11. RO yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	97,10	96,44	**97,55**	**97,55**
AUC Değeri	**0,992**	0,991	0,990	0,990
Ortalama F-ölçütü	0,955	0,947	**0,962**	**0,962**
Kİ	0,909	0,894	0,924	**0,925**

4.2.6. ÇKA yönteminin dört farklı eğitim veri setindeki sonuçları

Önişleme yapılmış veri seti 5-katlı çapraz doğrulama tekniği esnasında, üç farklı dengeleme yöntemi kullanılarak sınıf dengesizliği giderilmiş ve orijinal veri seti de dâhil olmak üzere model dört farklı eğitim veri seti ile eğitilmiştir. ÇKA modellerinin sınıflandırma performanslarının sonuçları Çizelge 4.12'de verilmiştir.

Çizelge 4.12. ÇKA yönteminin dört farklı eğitim veri setindeki performansları

Performans Ölçütleri	Orijinal	RUS	ROS	SMOTE
DSO (%)	95,32	94,65	**96,44**	95,77
AUC Değeri	0,989	0,985	**0,993**	0,991
Ortalama F-ölçütü	0,926	0,921	**0,945**	0,934
Kİ	0,852	0,842	**0,890**	0,869

Çizelge 4.10'a bakıldığında eğitim veri setinde dengeleme işlemi uygulanırsa DSO değeri %95,32'den %96,44'e, AUC değeri 0,989'dan 0,993'e, ortalama F-ölçütü değeri 0,926'dan 0,945'e, Kİ değeri ise 0,852'dan 0,890'a çıkmıştır.

4.3. Tüm Senaryoların Performans Ölçütleri Açısından Değerlendirilmesi

Bu bölümde KNIME programında oluşturulan tüm senaryolar DSO, AUC değeri, ortalama F-ölçütü ve Kİ açısından karşılaştırılmıştır.

Çizelge 4.13'e bakılacak olursa yeşil renkli değerler bulundukları sütundaki en yüksek değeri, kırmızı renkli değerler bulundukları sütundaki en küçük değeri, mavi renkli değerler ise senaryolardaki iyileştirme miktarını göstermektedir. Eğitim ve test verileri oluşturulurken yüzdesel bölme (%70-%30) tekniği kullanıldığında en iyi DSO değerini %99,26 ile RUS-RO kombinasyonu, en küçük DSO değerini ise %80,74 ile SMOTE-k-EYK vermiştir. Veri setleri oluşturulurken 5-katlı çapraz doğrulama tekniği kullanıldığında en iyi DSO değerini %97,55 ile RUS-C4.5 ve ROS-RO kombinasyonları,

en küçük DSO değerini ise %78,62 ile RUS-k-EYK vermiştir. Ayrıca oluşturulan senaryolarda yüzdesel bölme tekniği yerine çapraz doğrulama tekniği kullanılırsa on bir senaryoda performans artışı gözlemlenmiştir.

Çizelge 4.13. Tüm senaryoların DSO açısından karşılaştırılması

Senaryolar	%70-%30	5-katlı	İyileştirme
Orijinal -LR	94,81	94,43	-0,38
Orijinal -NB	96,30	95,10	-1,20
Orijinal -k-EYK	82,96	86,19	**3,23**
Orijinal -C4.5	95,56	96,21	**0,65**
Orijinal -RO	94,81	97,10	**2,29**
Orijinal -ÇKA	96,30	95,32	-0,98
RUS-LR	92,59	89,09	-3,50
RUS-NB	94,81	94,65	-0,16
RUS-k-EYK	82,96	78,62	-4,34
RUS-C4.5	96,30	97,55	**1,25**
RUS-RO	99,26	96,44	-2,82
RUS-ÇKA	96,30	94,65	-1,65
ROS-LR	94,81	94,43	-0,38
ROS-NB	94,81	95,99	**1,18**
ROS-k-EYK	83,70	82,63	-1,07
ROS-C4.5	97,04	95,55	-1,49
ROS-RO	97,04	97,55	**0,51**
ROS-ÇKA	93,00	96,44	**3,44**
SMOTE-LR	96,30	95,10	-1,20
SMOTE-NB	94,07	95,10	**1,03**
SMOTE-k-EYK	80,74	81,07	**0,33**
SMOTE-C4.5	97,04	97,10	**0,06**
SMOTE-RO	98,52	97,55	-0,97
SMOTE-ÇKA	94,81	95,77	**0,96**

Çizelge 4.14'e bakılacak olursa yeşil renkli değerler bulundukları sütundaki en yüksek değeri, kırmızı renkli değerler bulundukları sütundaki en küçük değeri, mavi renkli değerler ise senaryolardaki iyileştirme miktarını göstermektedir. Eğitim ve test verileri oluşturulurken yüzdesel bölme (%70-%30) tekniği kullanıldığında en iyi AUC değerini 0,997 ile Orijinal-ÇKA, RUS-RO ve SMOTE-LR kombinasyonları, en küçük AUC değerini ise 0,778 ile Orijinal-k-EYK vermiştir. Veri setleri oluşturulurken 5-katlı çapraz doğrulama tekniği kullanıldığında en iyi AUC değerini 0,993 ile ROS-ÇKA kombinasyonu, en küçük AUC değerini ise 0,714 ile Orijinal-k-EYK vermiştir. Ayrıca

oluşturulan senaryolarda yüzdesel bölme tekniği yerine çapraz doğrulama tekniği kullanılırsa altı senaryoda performans artışı gözlemlenmiştir.

Çizelge 4.14. Tüm senaryoların AUC değeri açısından karşılaştırılması

Senaryolar	%70-%30	5-katlı	İyileştirme
Orijinal -LR	0,988	0,987	-0,001
Orijinal -NB	0,995	0,989	-0,006
Orijinal -k-EYK	0,778	0,714	-0,064
Orijinal -C4.5	0,986	0,951	-0,035
Orijinal -RO	0,996	0,992	-0,004
Orijinal -ÇKA	0,997	0,989	-0,008
RUS-LR	0,986	0,966	-0,020
RUS-NB	0,991	0,988	-0,003
RUS-k-EYK	0,848	0,842	-0,006
RUS-C4.5	0,978	0,980	**0,002**
RUS-RO	0,997	0,991	-0,006
RUS-ÇKA	0,993	0,985	-0,008
ROS-LR	0,996	0,986	-0,010
ROS-NB	0,989	0,988	-0,001
ROS-k-EYK	0,829	0,807	-0,022
ROS-C4.5	0,968	0,921	-0,047
ROS-RO	0,988	0,990	**0,002**
ROS-ÇKA	0,974	0,993	**0,019**
SMOTE-LR	0,997	0,985	-0,012
SMOTE-NB	0,986	0,985	-0,001
SMOTE-k-EYK	0,838	0,859	**0,021**
SMOTE-C4.5	0,957	0,970	**0,013**
SMOTE-RO	0,996	0,990	-0,006
SMOTE-ÇKA	0,987	0,991	**0,004**

Çizelge 4.15'e bakılacak olursa yeşil renkli değerler bulundukları sütundaki en yüksek değeri, kırmızı renkli değerler bulundukları sütundaki en küçük değeri, mavi renkli değerler ise senaryolardaki iyileştirme miktarını göstermektedir. Eğitim ve test verileri oluşturulurken yüzdesel bölme (%70-%30) tekniği kullanıldığında en iyi ortalama F-ölçütü değerini 0,989 ile RUS-RO kombinasyonu, en küçük ortalama F-ölçütü değerini ise 0,644 ile Orijinal-k-EYK vermiştir. Veri setleri oluşturulurken 5-katlı çapraz doğrulama tekniği kullanıldığında en iyi ortalama F-ölçütü değerini 0,963 ile RUS-C4.5 kombinasyonu, en küçük ortalama F-ölçütü değerini ise 0,713 ile RUS-k-EYK vermiştir.

Ayrıca oluşturulan senaryolarda yüzdesel bölme tekniği yerine çapraz doğrulama tekniği kullanılırsa on senaryoda performans artışı gözlemlenmiştir.

Çizelge 4.15. Tüm senaryoların ortalama F-ölçütü açısından karşılaştırılması

Senaryolar	%70-%30	5-katlı	İyileştirme
Orijinal -LR	0,988	0,911	-0,077
Orijinal -NB	0,941	0,923	-0,018
Orijinal -k-EYK	0,644	0,815	**0,171**
Orijinal -C4.5	0,942	0,940	-0,002
Orijinal -RO	0,915	0,955	**0,040**
Orijinal -ÇKA	0,943	0,926	-0,017
RUS-LR	0,945	0,844	-0,101
RUS-NB	0,918	0,919	**0,001**
RUS-k-EYK	0,744	0,713	-0,031
RUS-C4.5	0,946	0,963	**0,017**
RUS-RO	0,989	0,947	-0,042
RUS-ÇKA	0,946	0,921	-0,025
ROS-LR	0,924	0,916	-0,008
ROS-NB	0,922	0,939	**0,017**
ROS-k-EYK	0,764	0,751	-0,013
ROS-C4.5	0,955	0,929	-0,026
ROS-RO	0,944	0,962	**0,018**
ROS-ÇKA	0,890	0,945	**0,055**
SMOTE-LR	0,940	0,925	-0,015
SMOTE-NB	0,899	0,921	**0,022**
SMOTE-k-EYK	0,749	0,745	-0,004
SMOTE-C4.5	0,952	0,955	**0,003**
SMOTE-RO	0,977	0,962	-0,015
SMOTE-ÇKA	0,920	0,934	**0,014**

Çizelge 4.16'ya bakılacak olursa yeşil renkli değerler bulundukları sütundaki en yüksek değeri, kırmızı renkli değerler bulundukları sütundaki en küçük değeri, mavi renkli değerler ise senaryolardaki iyileştirme miktarını göstermektedir. Eğitim ve test verileri oluşturulurken yüzdesel bölme (%70-%30) tekniği kullanıldığında en iyi Kİ değerini 0,977 ile RUS-RO kombinasyonu, en küçük Kİ değerini ise 0,303 ile Orijinal-k-EYK vermiştir. Veri setleri oluşturulurken 5-katlı çapraz doğrulama tekniği kullanıldığında en iyi Kİ değerini 0,925 ile RUS-C4.5 ve SMOTE-RO kombinasyonları, en küçük Kİ değerini ise 0,433 ile RUS-k-EYK vermiştir. Ayrıca oluşturulan senaryolarda yüzdesel

bölme tekniği yerine çapraz doğrulama tekniği kullanılırsa on bir senaryoda performans artışı gözlemlenmiştir.

Çizelge 4.16. Tüm senaryoların Kİ açısından karşılaştırılması

Senaryolar	**%70-%30**	**5-katlı**	**İyileştirme**
Orijinal -LR	0,821	0,821	0,000
Orijinal -NB	0,883	0,846	-0,037
Orijinal -k-EYK	0,303	0,443	**0,140**
Orijinal -C4.5	0,857	0,879	**0,022**
Orijinal -RO	0,831	0,909	**0,078**
Orijinal -ÇKA	0,886	0,852	-0,034
RUS-LR	0,792	0,689	-0,103
RUS-NB	0,836	0,839	**0,003**
RUS-k-EYK	0,489	0,433	-0,056
RUS-C4.5	0,892	0,925	**0,033**
RUS-RO	0,977	0,894	-0,083
RUS-ÇKA	0,892	0,842	-0,050
ROS-LR	0,848	0,831	-0,017
ROS-NB	0,844	0,877	**0,033**
ROS-k-EYK	0,530	0,504	-0,026
ROS-C4.5	0,910	0,857	-0,053
ROS-RO	0,889	0,924	**0,035**
ROS-ÇKA	0,781	0,890	**0,109**
SMOTE-LR	0,879	0,851	-0,028
SMOTE-NB	0,798	0,842	**0,044**
SMOTE-k-EYK	0,508	0,496	-0,012
SMOTE-C4.5	0,905	0,909	**0,004**
SMOTE-RO	0,954	0,925	-0,029
SMOTE-ÇKA	0,840	0,869	**0,029**

5. SONUÇ

Yapılan bu çalışmada ilk olarak diyabet hastalığı, sınıf dengesizliği ve makine öğrenmesinde sınıflandırma modelleri kavramları incelenmiş olup ve bu kavramlarla ilgili literatür araştırması yapılmıştır. Bir devlet hastanesinin bilgi işlem biriminden alınan laboratuvar sonuçlarını içeren diyabet veri setinde hastalık teşhisi için altı farklı makine öğrenmesi tekniği kullanılarak sınıflandırma modelleri oluşturulmuş ve bu modellerin performansları farklı ölçütlere göre karşılaştırılmıştır. Oluşturulan sınıflandırma modellerinin performanslarını karşılaştırmak için DSO, kesinlik, duyarlılık, ortalama F-ölçütü, kappa istatistiği ve AUC değerleri hesaplanmıştır.

Sınıflandırma modelleri oluşturulurken diyabet veri setindeki sınıf dengesizliği sebebiyle farklı yeniden örnekleme tekniklerinin (RUS, ROS, SMOTE) ve farklı veri bölme tekniklerinin (yüzdesel bölme, 5-katlı çapraz doğrulama) kullanılması sonucunda 48 farklı senaryo türetilmiştir. Bu senaryolar ışığında en iyi sınıflandırma performansı veren senaryo yüzdesel bölme tekniğiyle oluşturulan eğitim veri setiyle eğitilen %99,26 DSO ile RUS-RO modeli, ona en yakın performansı gösteren senaryo ise %98,52 DSO ile SMOTE-RO modeli olmuştur. Tüm bu senaryolar arasında en kötü sonucu veren model ise 5-katlı çapraz doğrulama tekniği kullanılan %78,62 DSO oranı ile RUS-k-EYK modeli olmuştur. Ayrıca bazı senaryolarda veri bölme tekniklerinden yüzdesel bölme yerine 5-katlı çapraz doğrulama tekniği uygulanırsa performans artışlarının olduğu tespit edilmiştir.

Bu çalışmada bir gerçek hayat diyabet verisinde sınıf dengesizliği varlığında, farklı veri bölme teknikleri, yeniden örnekleme teknikleri ve makine öğrenmesi yöntemleri kullanılarak tip-2 diyabet teşhisinin başarılı bir şekilde sınıflandırılabildiği tespit edilmiştir.

Bu çalışma için elde edilen diyabet veri setinde hastanın ailesindeki diyabet öyküsü, sigara içip içmediği, boy, kilo vb. bilgilerin olmadığı görülmüştür. Hâlbuki bu bilgiler literatürdeki çalışmalarda sıklılıkla kullanılan öznitelikler arasında yer almaktadır. Bu nedenle bahsedilen bu öznitelikler de veri setine dahil edilerek çalışmanın kapsamı genişletilebilir. Ayrıca gelecek çalışmalarda topluluk öğrenmesi sınıflandırma yöntemleri kullanılarak sınıf dengesizliği altında bu algoritmaların performansları ölçülebilir.

Topluluk öğrenmesi modellerinin performansları bu çalışmada incelenen senaryoların performansları karşılaştırabilir.

Son olarak, doktorların diyabet hastalığı teşhisi karar verme süreçlerini kolaylaştırabilmek ve koydukları teşhislerdeki doğruluk payını artırmak amacıyla bir yazılım programında arayüz geliştirilip, karar destek sistemi tasarlanabilir.

KAYNAKLAR

Akyol, K., & Şen, B. (2018). Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms. *International Journal of Modern Education and Computer Science*, *10*(6), 10-16. https://doi.org/10.5815/ijmecs.2018.06.02

Alehegn, M., Raghvendra Joshi, R., & Mulay, P. (2019). Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach. *International Journal of Scientific & Technology Research*, *8*(09). www.ijstr.org

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Içinde *Supervised and Unsupervised Learning for Data Science* (ss. 3-21). Springer. https://doi.org/10.1007/978-3-030-22475-2_1

Alpan, K., & Ilgi, G. S. (2020). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. *4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2020 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ISMSIT50672.2020.9254720

Alpaydın E. (2014), Introduction to Machine Learning , MIT Press, Cambridge, United States.

Anwar, N., Pranolo, A., & Kurnaiwan, R. (2018). Grouping the community health center patients based on the disease characteristics using C4.5 decision tree. *IOP Conference Series: Materials Science and Engineering*, *403*(1). Institute of Physics Publishing. https://doi.org/10.1088/1757-899X/403/1/012084

Aydın, M. A. (2020). Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi. *Journal of Polytechnic*. https://doi.org/10.2339/politeknik.734916

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Applied Sciences*, *1*(9). https://doi.org/10.1007/s42452-019-1117-9

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32.

Chaves, L., & Marques, G. (2021). Data mining techniques for early diagnosis of diabetes: A comparative study. *Applied Sciences (Switzerland)*, *11*(5), 1-12. https://doi.org/10.3390/app11052218

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Içinde *Journal of Artificial Intelligence Research* (C. 16).

Chen, P., & Pan, C. (2018). Diabetes classification model based on boosting algorithms. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2090-9

Cihan, P., & Coskun, H. (2021). Performance comparison of machine learning models for diabetes prediction. *SIU 2021 - 29th IEEE Conference on Signal Processing and Communications Applications, Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/SIU53274.2021.9477824

Cohen, J. (1960). A Coefficient of Agreement For Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/https://doi.org/10.1177/001316446002000104

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3459665

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. *Endüstri Mühendisliği Dergisi*, *25*(3-4), 2-19. Geliş tarihi gönderen https://dergipark.org.tr/tr/pub/endustrimuhendisligi/issue/46771/586362

Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, *11*(2), 78-83. https://doi.org/10.12720/jait.11.2.78-83

Das, H., Naik, B., & Behera, H. S. (2018). Classification of diabetes mellitus disease (DMD): A data mining (DM) approach. *Advances in Intelligent Systems and Computing*, *710*, 539-549. Springer Verlag. https://doi.org/10.1007/978-981-10-7871-2_52

Devi, D., Biswas, S. K., & Purkayastha, B. (2020). A Review on Solution to Class Imbalance Problem: Undersampling Approaches. *2020 International Conference on Computational Performance Evaluation (ComPE)*, 626-631. Shillong, India: IEEE. https://doi.org/10.1109/ComPE49325.2020.9200087

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability Index

Terms-Data mining, mining methods and algorithms, text mining. *Journal of Computing*, *4*(8), 1-9. https://doi.org/10.48550/arXiv.1206.1121

DUK (Diabetes United Kingdom Community). (2022, Ağustos 05). Type 1 diabetes. https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-1

Emer, N., & Özbek, N. S. (2021). Control of Attitude Dynamics of an Unmanned Aerial Vehicle with Reinforcement Learning Algorithms. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1021970

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Içinde *Journal of Artificial Intelligence Research* (C. 61).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Içinde *Psychological Bulletin* (C. 76). https://doi.org/https://doi.org/10.1037/h0031619

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, *5*, 20590-20616. https://doi.org/10.1109/ACCESS.2017.2756872

Gosain, A., & Sardana, S. (2019). Farthest SMOTE: A Modified SMOTE Approach. *Advances in Intelligent Systems and Computing*, *711*, 309-320. Springer Verlag. https://doi.org/10.1007/978-981-10-8055-5_28

Harman, G. (2021). Destek Vektör Makineleri ve Naive Bayes Sınıflandırma Algoritmalarını Kullanarak Diabetes Mellitus Tahmini. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1041186

Hassan, M. M., Njmh Amiri, N., Muhammed Hassan, M., & Amiri, N. (2019). Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms. *IV.International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE)*, 50-55. https://www.researchgate.net/publication/336672231

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

IDF (International Diabetes Federation). (2022, Haziran 06). About Diabetes. https://www.idf.org/aboutdiabetes/what-is-diabetes.html

IDFDA (International Diabetes Federation Diabetes Atlas). (2022, Haziran 08). Diabetes around the world in 2021. https://diabetesatlas.org/

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Advances*

in Intelligent Systems and Computing, *992*, 113-125. Springer Verlag. https://doi.org/10.1007/978-981-13-8798-2_12

Ismail, L., Materwala, H., Tayefi, M., Ngo, P., & Karduck, A. P. (2022). Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. *Archives of Computational Methods in Engineering*, *29*(1), 313-333. https://doi.org/10.1007/s11831-021-09582-x

Jakka, A., & Vakula Rani, J. (2019). Performance evaluation of machine learning models for diabetes prediction. *International Journal of Innovative Technology and Exploring Engineering*, *8*(11), 1976-1980. https://doi.org/10.35940/ijitee.K2155.0981119

Kabir, M. F., & Ludwig, S. (2019). Classification of Breast Cancer Risk Factors Using Several Resampling Approaches. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1243-1248. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICMLA.2018.00202

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, C. 52. Association for Computing Machinery. https://doi.org/10.1145/3343440

Kavzoğlu, T., & Çölkesen, İ. (2010). Classification of Satellite Images Using Decision Trees: Kocaeli Case. *Electronic Journal of Map Technologies*, *2*(1), 36-45. www.tuik.gov.tr

Kılıç, S. (2015). Kappa test. *Journal of Mood Disorders*, *5*(3), 142. https://doi.org/10.5455/jmood.20150920115439

KNIME (Konstanz Information Miner). (2022, Haziran 15). Download KNIME Analytics Platform. https://www.knime.com/downloads/download-knime

Konakoğlu, B. (2020). Çok Katmanlı Algılayıcı Yapay Sinir Ağı ile Jeodezik Elipsoidal Koordinatların (φ, λ, h) 3 Boyutlu Global Kartezyen Koordinatlara (X, Y, Z) Dönüşümü. *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *10*(3), 702-710. https://doi.org/10.17714/gumusfenbil.712100

Krasteva, A., Panov, V., Krasteva, A., Kisselova, A., & Krastev, Z. (2011, Şubat). Oral cavity and systemic diseases - Diabetes mellitus. *Biotechnology and Biotechnological Equipment*, C. 25, ss. 2183-2186. https://doi.org/10.5504/bbeq.2011.0022

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*, 195-197. http://www.r-project.org/

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier.

International Journal of Cognitive Computing in Engineering, *2*, 40-46. https://doi.org/10.1016/j.ijcce.2021.01.001

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. https://doi.org/https://doi.org/10.2307/2529310

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*, *7*, 93010-93022. https://doi.org/10.1109/ACCESS.2019.2927266

Mesquita, F., Mauricio, J., & Marques, G. (2021). Oversampling Techniques for Diabetes Classification: A Comparative Study. *2021 9th E-Health and Bioengineering Conference, EHB 2021*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/EHB52898.2021.9657542

Mohammed, A. J. (2020). Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3161-3172. https://doi.org/10.30534/ijatcse/2020/104932020

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243-248. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICICS49469.2020.239556

Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, *19*(1), 391-403. https://doi.org/10.1007/s40200-020-00520-5

NIDDK (National Instute of Diabetes and Digestive and Kidney Diseases). (2022, Ağustos 05). Type 2 Diabetes. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes

Öztemel, E. (2003). Yapay sinir ağları. Papatya Yayıncılık, İstanbul.

Quinlan, R. (1993). *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Raghuwanshi, B. S., & Shukla, S. (2021). Classifying imbalanced data using SMOTE based class-specific kernelized ELM. *International Journal of Machine Learning and Cybernetics*, *12*(5), 1255-1280. https://doi.org/10.1007/s13042-020-01232-1

Sarker, I. H. (2021, Mayıs 1). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, C. 2. Springer. https://doi.org/10.1007/s42979-021-00592-x

Sevli, O. (2022). Diyabet hastalığının farklı sınıflandırıcılar kullanılarak teşhisi. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*. https://doi.org/10.17341/gazimmfd.880750

Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. Içinde *Advances in Computing and Intelligent Systems* (ss. 195-211). Springer. https://doi.org/10.1007/978-981-15-0222-4_17

Sweeting, A., Wong, J., Murphy, H. R., & Ross, G. P. (2022). A Clinical Update on Gestational Diabetes Mellitus. *Endocrine Reviews*, *43*(5), 763-793. https://doi.org/10.1210/endrev/bnac003

Şenol, A., Canbay, Y., & Kaya, M. (2021). Makine Öğrenmesi Yaklaşımlarını Kullanarak Salgınları Erken Evrede Tespit Etme Alanındaki Eğilimler. *Bilişim Teknolojileri Dergisi*, *14*(4), 355-366. https://doi.org/10.17671/gazibtd.878089

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, *167*, 706-716. Elsevier B.V. https://doi.org/10.1016/j.procs.2020.03.336

Turhan Sultan, Özkan, Y., Yürekli, B. S., Suner, A., & Doğu, E. (2020). Comparison of Ensemble Learning Methods for Disease Diagnosis in Presence of Class Unbalanced: Case of Diabetes. *Turkiye Klinikleri Journal of Biostatistics*, *12*(1), 16-26. https://doi.org/10.5336/biostatic.2019-66816

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-03430-5

Wang, S.-C. (2003). Artificial Neural Network. Içinde *Interdisciplinary Computing in Java Programming* (C. 743). Kluwer Academic Publishers.

Wang, Z. W., Wang, S. K., Wan, B. T., & Song, W. W. (2020). A novel multi-label classification algorithm based on K-nearest neighbor and random walk. *International Journal of Distributed Sensor Networks*, *16*(3). https://doi.org/10.1177/1550147720911892

WHO, 2022a. (World Health Organizations). (2022, Haziran 06). Diabetes. https://www.who.int/health-topics/diabetes#tab=tab_1

WHO, 2022b. (World Health Organizations). (2022, Ağustos 19). International statistical classification of diseases and related health problems.

https://cdn.who.int/media/docs/default-source/classification/icd/cause-of-death/icd10volume2_en_2016.pdf

Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. *Information Sciences, 569,* 508-526. https://doi.org/10.1016/j.ins.2021.05.029

Yavaş, M., Güran, A., & Uysal, M. (2020). Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. *European Journal of Science and Technology,* 258-264. https://doi.org/10.31590/ejosat.779952

Yavuz, S., & Deveci, M. (2012). İstatiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 40,* 167-187.

Yazgan, E., & Erol, D. (2016). Sivil Pilot Adayları için Seçim Kriterlerinin Belirlenmesi. *Niğde Üniversitesi Mühendislik Bilimleri Dergisi, 5*(2), 97-104.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*(5), 654-657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics, 9.* https://doi.org/10.3389/fgene.2018.00515

EKLER

EK 1 Bursa Uludağ Üniversitesi Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu'ndan alınan etik kurul kararı ve toplantı tutanağı
EK 2 Veri seti hakkında detaylı bilgiler tablosu
EK 3 KNIME programında oluşturulan LR modelleri
EK 4 KNIME programında oluşturulan NB modelleri
EK 5 KNIME programında oluşturulan k-EYK modelleri
EK 6 KNIME programında oluşturulan C4.5 modelleri
EK 7 KNIME programında oluşturulan RO modelleri
EK 8 KNIME programında oluşturulan ÇKA modelleri

EK 1 Bursa Uludağ Üniversitesi Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu'ndan alınan etik kurul kararı ve toplantı tutanağı

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Hukuk Müşavirliği
Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu Başkanlığı

Sayı: E-92662996-044-71240 31.08.2022
Konu: İsmail Buğra BÖLÜKBAŞI'ın Etik Kurul Kararı

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 26.07.2022 tarihli ve E-31037515-300-15879 sayılı yazınız.

Enstitünüz, Endüstri Mühendisliği Anabilim Dalı öğretim üyelerinden Doç. Dr. Betül YAĞMAHAN'ın tez danışmanlığını yürüttüğü yüksek lisans öğrencisi İsmail Buğra BÖLÜKBAŞI'nın "Diyabet Hastalığının Teşhisinde Makine Öğrenmesi Tekniklerinin Geçerliliği" konulu tez çalışması, BUÜ Araştırma ve Yayın Etiği Kurulları Başkanlığı (Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu)' nın 22 Ağustos 2022 tarih ve 2022-06 sayılı oturumunda görüşülmüş olup, alınan karar ekte gönderilmektedir.

Bilgilerinizi ve gereğini rica ederim.

Prof. Dr. Ferudun YILMAZ
Rektör a.
Rektör Yardımcısı

Ek:
Karar Örneği (1 Sayfa)

EK 1 Bursa Uludağ Üniversitesi Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu'ndan alınan etik kurul kararı ve toplantı tutanağı (devam)

BURSA ULUDAĞ ÜNİVERSİTESİ
ARAŞTIRMA VE YAYIN ETİK KURULLARI
(Fen ve Mühendislik Bilimleri Araştırma ve Yayın Etik Kurulu)
TOPLANTI TUTANAĞI

OTURUM TARİHİ
22 AĞUSTOS 2022

OTURUM SAYISI
2022-06

KARAR NO 2: Fen Bilimleri Enstitüsü Müdürlüğü'nden alınan Endüstri Mühendisliği Anabilim Dalı öğretim üyelerinden Doç. Dr. Betül YAĞMAHAN'ın tez danışmanlığını yürüttüğü yüksek lisans öğrencisi İsmail Buğra BÖLÜKBAŞI'nın "Diyabet Hastalığının Teşhisinde Makine Öğrenmesi Tekniklerinin Geçerliliği" konulu tez çalışması kapsamında hastalara ait laboratuvar tahlil sonuçlarını istenilmesinin değerlendirilmesine geçildi.

Yapılan görüşmeler sonunda; Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı öğretim üyelerinden Doç. Dr. Betül YAĞMAHAN'ın tez danışmanlığını yürüttüğü yüksek lisans öğrencisi İsmail Buğra BÖLÜKBAŞI'nın "Diyabet Hastalığının Teşhisinde Makine Öğrenmesi Tekniklerinin Geçerliliği" konulu tez çalışması kapsamında kişisel verileri içermeyen labrotuvar sonuçlarının incelenmesinin etik açıdan bir sakınca bulunmamaktadır. Çalışmanın fikri, hukuki ve telif hakları bakımından metot ve ölçeğine ilişkin sorumluluğu başvurucuya ait olmak üzere uygun olduğuna oybirliği ile karar verildi.

Prof. Dr. Feridun YILMAZ
Kurul Başkanı

Prof. Dr. İlhan TURGUT
Üye

Prof. Dr. Asim OLGUN
Üye

Prof. Dr. M.İhsan KARAMANGİL
Üye

Prof. Dr. Recep EREN
Üye

Prof. Dr. Adnan GERÇEK
Üye

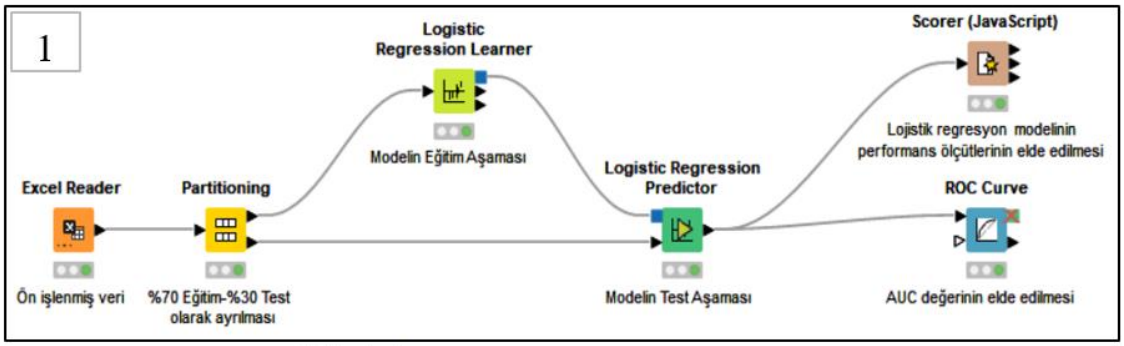
Prof. Dr. Fahri VATANSEVER
Üye

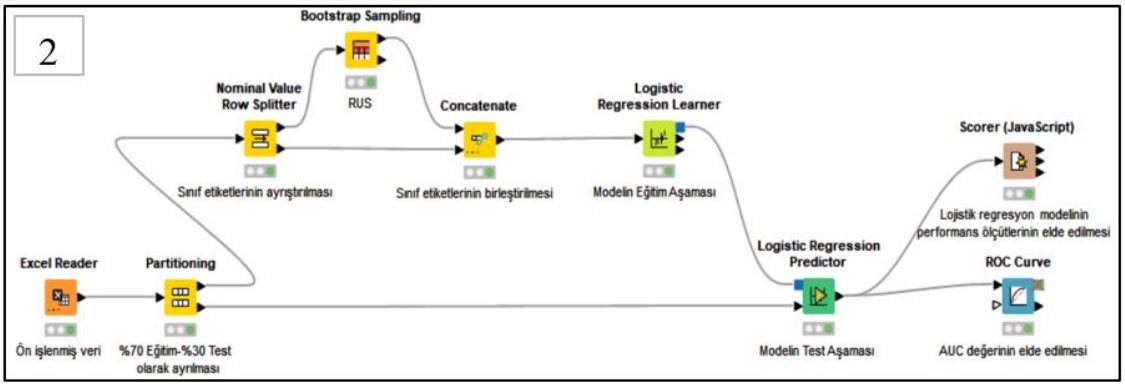
EK 2 Veri seti hakkında detaylı bilgiler tablosu

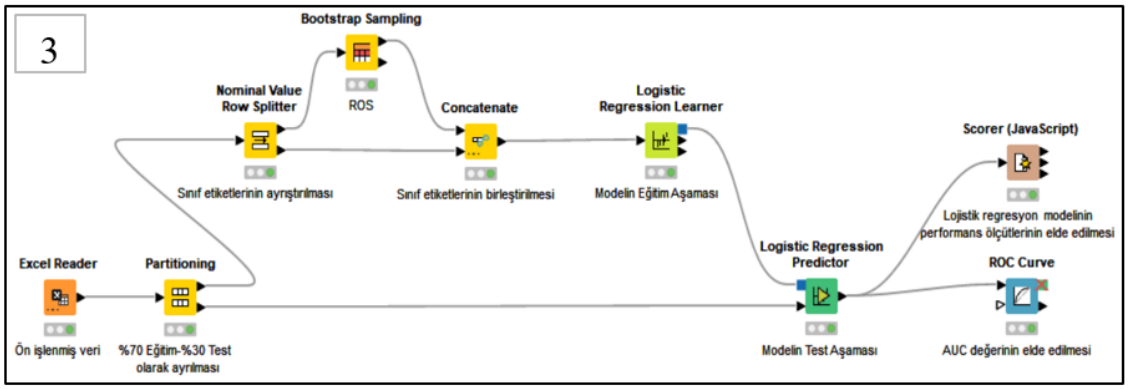
Değişken Türü		Değişkenler	KNIME Adı	Ölçeği	Açıklama	Verilerin Dağılımı				Eksik Veri Sayısı
Bağımsız (Girdi)	**1**	Cinsiyet	Cinsiyet	Kategorik-Binary	Hastanın cinsiyeti	0: Kadın (931 kişi) , 1: Erkek (620 kişi)				-
	2	Yaş	Yas	Nümerik-Sürekli (yıl)	Hastanın yaşı	Min: 17,00	Maks: 93,00	Ort: 57,05	SS: 13,81	-
	3	Alanin aminotransferaz (ALT)	ALT	Nümerik-Sürekli (IU/L)	Karaciğerde üretilen bir enzim türü	Min: 2,50	Maks: 1763,80	Ort: 23,79	SS: 47,83	45
	4	Albümin	Albumin	Nümerik-Sürekli (g/dL)	Karaciğerde sentezlenen bir protein türü	Min: 2,66	Maks: 5,53	Ort: 4,51	SS: 0,31	495
	5	Aspartat aminotransferaz (AST)	AST	Nümerik-Sürekli (IU/L)	Karaciğerde üretilen bir enzim türü	Min: 5,20	Maks: 1642,50	Ort: 20,86	SS: 43,68	81
	6	Glike hemoglobin (Hb A1c)	HbA1c	Nümerik-Sürekli (%)	Son üç aydaki ortalama kan şekeri seviyesi	Min: 4,20	Maks: 16,80	Ort: 7,39	SS: 2,09	23
	7	Glukoz	Glukoz	Nümerik-Sürekli (mg/dL)	Kandaki şeker seviyesi	Min: 62,70	Maks: 512,20	Ort: 154,25	SS: 69,27	0
	8	HDL kolesterol	HDL Kolestrol	Nümerik-Sürekli (mg/dL)	Vücuttaki dokulardan karaciğere kolesterol taşıyan yüksek yoğunluklu lipoprotein	Min: 16,60	Maks: 127,40	Ort: 49,29	SS: 13,83	143
	9	Kalsiyum (Ca)	Kalsiyum	Nümerik-Sürekli (mg/dL)	Vücut fonksyinoları için gerekli olan bir kimyasal element türü	Min: 5,84	Maks: 11,43	Ort: 9,56	SS: 0,47	409
	10	Kolesterol	Kolesterol	Nümerik-Sürekli (mg/dL)	Kanda bulunan bir çeşit yağ dokusu	Min: 84,40	Maks: 464,30	Ort: 197,40	SS: 45,35	63
	11	Kreatinin	Kreatinin	Nümerik-Sürekli (mg/dL)	Kaslarda enerji olarak depolanan kreatin adlı bileşenin yıkım ürünü	Min: 0,32	Maks: 4,98	Ort: 0,86	SS: 0,34	38
	12	LDL kolesterol (Direkt)	LDL Kolesterol	Nümerik-Sürekli (mg/dL)	Vücuttaki dokulardan karaciğere kolesterol taşıyan düşük yoğunluklu lipoprotein	Min: 7,48	Maks: 400,00	Ort: 125,59	SS: 43,49	432
	13	Total Protein	Protein	Nümerik-Sürekli (g/dL)	Kan dolaşımında bulunan albumin ve globulin adlı proteinlerin toplam miktarı	Min: 1,43	Maks: 9,12	Ort: 7,23	SS: 0,52	558
	14	Serbest t4	Serbest T4	Nümerik-Sürekli (ng/dL)	Tirodi fonksiyonlarının ölçülmesinde kullanılan	Min: 0,06	Maks: 6,63	Ort: 1,27	SS: 0,30	360
	15	Trigliserid	Trigliserid	Nümerik-Sürekli (mg/dL)	Vücutta enerji kaynağı olarak kullanılan bir lipit türü	Min: 1,44	Maks: 1425,70	Ort: 176,50	SS: 112,46	62
	16	TSH	TSH	Nümerik-Sürekli (µIU/mL)	Tiroid bezinin çalışmasını düzenleyen peptid yapıda bir hormon türü	Min: 0,01	Maks: 259,60	Ort: 2,38	SS: 7,56	181
	17	Üre	Ure	Nümerik-Sürekli (mg/dL)	Vücuttaki proteinlerin yıkımı ile meydana gelen azotlu bir bileşik türü	Min: 8,80	Maks: 325,20	Ort: 33,30	SS: 18,47	94
	18	Ürik Asit	Urik Asit	Nümerik-Sürekli (mg/dL)	Kandaki pürin maddesinin parçalanması ile ortaya çıkan bir bileşik türü	Min: 1,80	Maks: 23,20	Ort: 4,93	SS: 1,65	504
	19	Eos#	Eos#	Nümerik-Sürekli (µL)	Eozinofil sayısı	Min: 0,00	Maks: 1,45	Ort: 0,21	SS: 0,16	1
	20	EOS%	Eos%	Nümerik-Sürekli (%)	Eozinofil yüzdesi	Min: 0,00	Maks: 16,10	Ort: 2,70	SS: 1,93	-
	21	MCHC	MCHC	Nümerik-Sürekli (g/dL)	Kırmızı kan hücrelerindeki hemoglobin yoğunluğu	Min: 28,30	Maks: 176,70	Ort: 33,57	SS: 3,82	-
	22	Mon#	Mon#	Nümerik-Sürekli (µL)	Monosit sayısı	Min: 0,00	Maks: 1,39	Ort: 0,48	SS: 0,16	-
	23	Mon%	Mon%	Nümerik-Sürekli (%)	Monosit yüzdesi	Min: 0,00	Maks: 14,90	Ort: 6,22	SS: 1,55	-
	24	Neu#	Neu#	Nümerik-Sürekli (µL)	Nötrofil sayısı	Min: 0,11	Maks: 15,12	Ort: 4,67	SS: 1,72	-
	25	Neu%	Neu%	Nümerik-Sürekli (%)	Nötrofil yüzdesi	Min: 12,00	Maks: 93,50	Ort: 58,41	SS: 8,71	-
	26	Lymph#	Lymph#	Nümerik-Sürekli (µL)	Lenfosit sayısı	Min: 0,41	Maks: 24,70	Ort: 2,52	SS: 1,01	63
	27	Lymph%	Lymph%	Nümerik-Sürekli (%)	Lenfosit yüzdesi	Min: 5,70	Maks: 86,90	Ort: 32,35	SS: 8,08	10
	28	RDW-CV	RDW-CV	Nümerik-Sürekli (%)	Eritrosit dağılımı genişliği (yüzde)	Min: 11,70	Maks: 27,10	Ort: 13,85	SS: 1,49	-
	29	RDW-SD	RDW-SD	Nümerik-Sürekli (fL)	Eritrosit dağılımı genişliği	Min: 31,90	Maks: 69,70	Ort: 42,11	SS: 3,50	-
	30	HCT	HCT	Nümerik-Sürekli (%)	Kırmızı kan hücrelerinin hacminin, dolaşımdaki kanın hacmine oranı	Min: 22,50	Maks: 55,10	Ort: 41,18	SS: 4,63	1
	31	HGB	HGB	Nümerik-Sürekli (g/dL)	Kırmızı kan hücresinde depolanan bir protein türü (Hemoglobin)	Min: 7,70	Maks: 18,60	Ort: 13,89	SS: 1,69	29
	32	MCV	MCV	Nümerik-Sürekli (fL)	Kırmızı kan hücresi hacmi	Min: 55,70	Maks: 120,50	Ort: 85,64	SS: 6,20	4
	33	MPV	MPV	Nümerik-Sürekli (fL)	Ortalama trombosit hacmi	Min: 7,10	Maks: 13,60	Ort: 9,89	SS: 1,07	-
	34	PCT	PCT	Nümerik-Sürekli (ng/mL)	Kandaki enfeksiyon durumunu gösteren bir peptit türü	Min: 0,01	Maks: 0,66	Ort: 0,27	SS: 0,07	-
	35	PLT	PLT	Nümerik-Sürekli (µL)	Kanın pıhtılaşmasında görev alan renksiz kan hücresi	Min: 29,00	Maks: 676,00	Ort: 272,43	SS: 73,42	-
	36	RBC	RBC	Nümerik-Sürekli (µL)	Kırmızı kan hücresi (Eritrosit)	Min: 0,38	Maks: 7,31	Ort: 4,82	SS: 0,56	-
	37	WBC	WBC	Nümerik-Sürekli (µL)	Beyaz kan hücresi (Lökosit)	Min: 0,19	Maks: 28,47	Ort: 7,88	SS: 2,27	-
	38	Bas#	Bas#	Nümerik-Sürekli (µL)	Bazofil sayısı	Min: 0,00	Maks: 0,34	Ort: 0,04	SS: 0,02	-
	39	Bas%	Bas%	Nümerik-Sürekli (%)	Bazofil yüzdesi	Min: 0,00	Maks: 4,00	Ort: 0,49	SS: 0,27	-
	40	PDW	PDW	Nümerik-Sürekli (fL)	Kandaki trombositlerin büyüklüğünün dağılımı	Min: 14,90	Maks: 17,40	Ort: 16,06	SS: 0,38	-
	41	MCH	MCH	Nümerik-Sürekli (pg)	Kırmızı kan hücresi başına ortalama hemoglobin kütlesi	Min: 16,80	Maks: 39,40	Ort: 28,70	SS: 2,57	-
	42	NLR	NLR	Nümerik-Sürekli (µL)	Kanda bulunan nötrofil miktarının lökosit miktarına oranı	Min: 0,10	Maks: 25,60	Ort: 2,11	SS: 1,54	-
Bağımlı (Çıktı)	**43**	Tanı	Tani	Kategorik-Binary	Hastanın diyabet hastası olup olmadığı durumu	0: Diyabet Yok (334 kişi) , 1: Diyabet Var (1217 kişi)				-

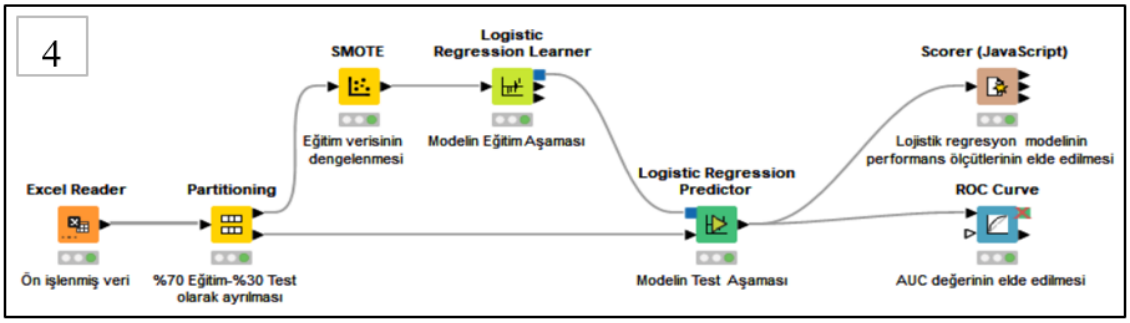
Min: Minimum Değer **Maks:** Maksimum Değer **Ort:** Ortalama Değer **SS:** Standart Sapma Değeri

EK 3 KNIME programında oluşturulan LR modelleri

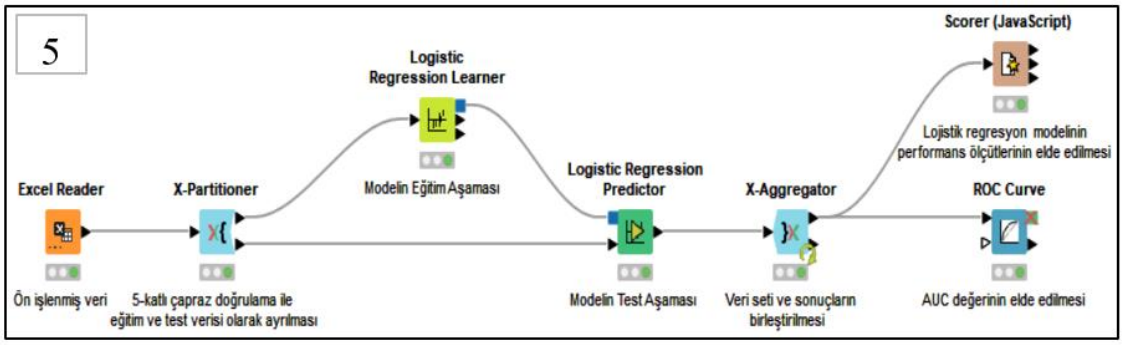


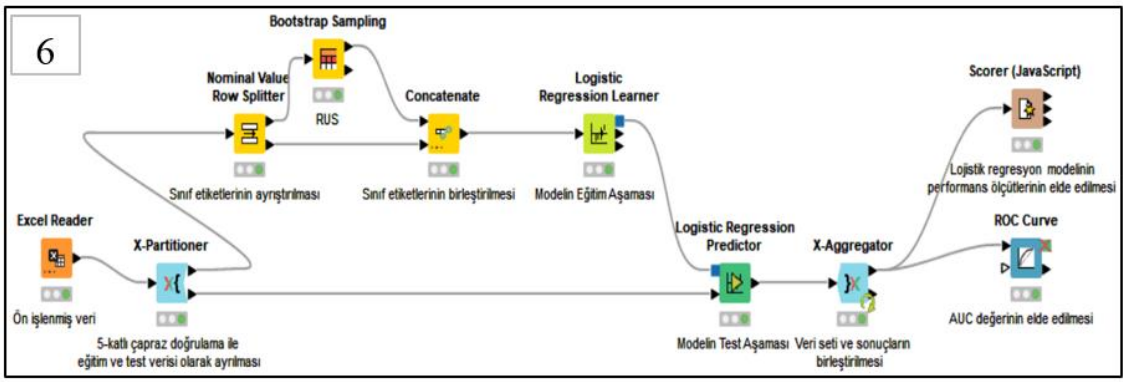


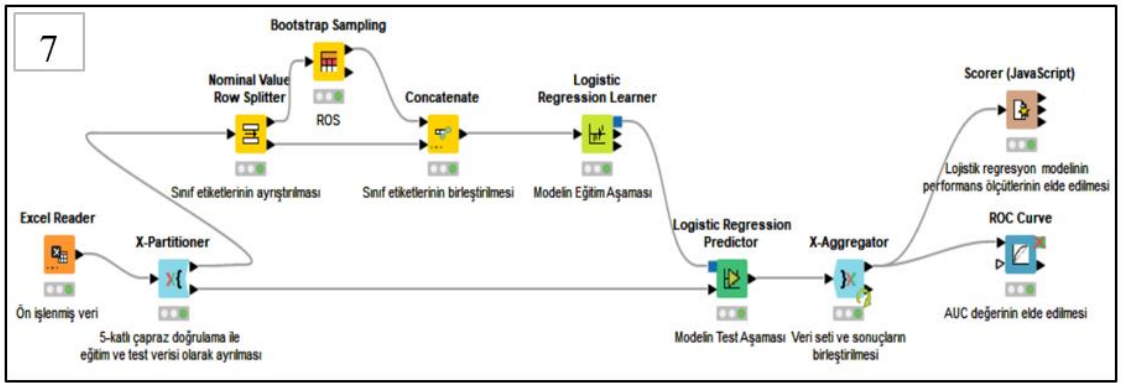


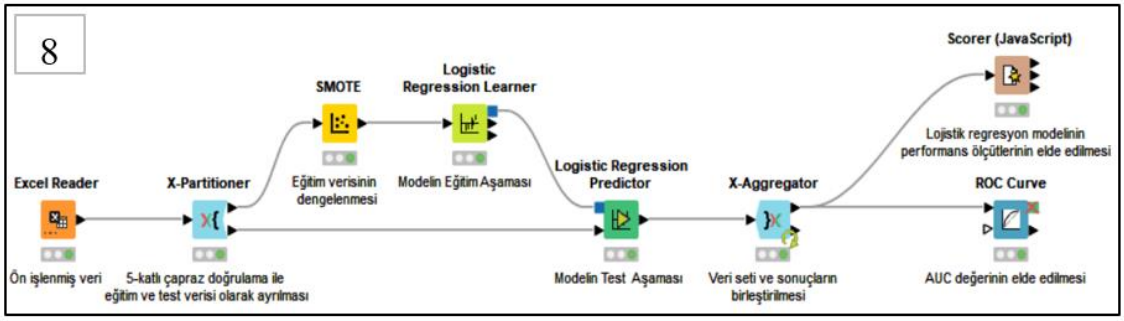


EK 3 KNIME programında oluşturulan LR modelleri (devam)

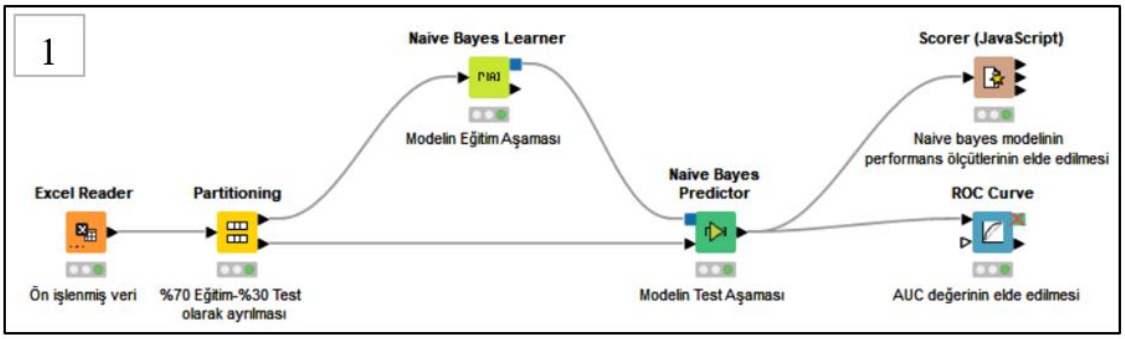


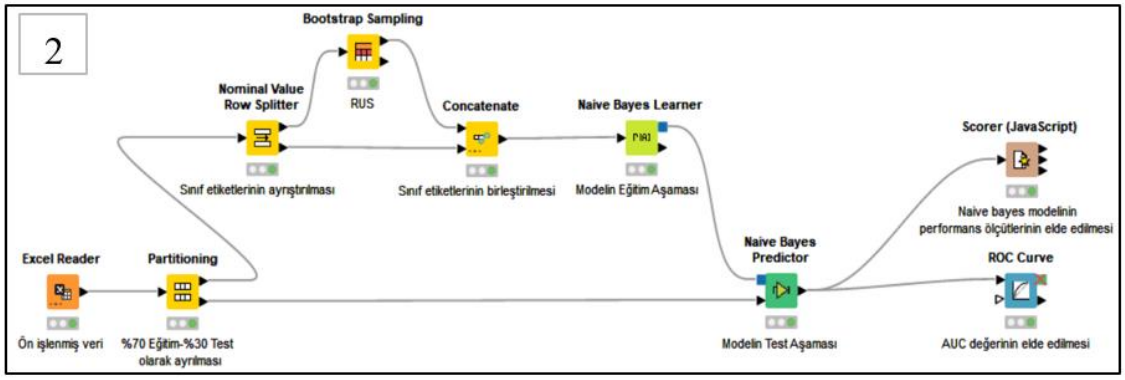


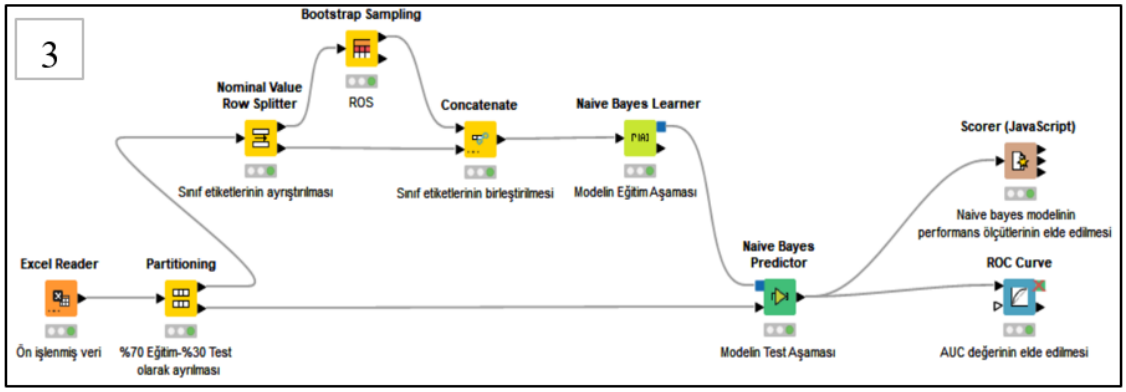


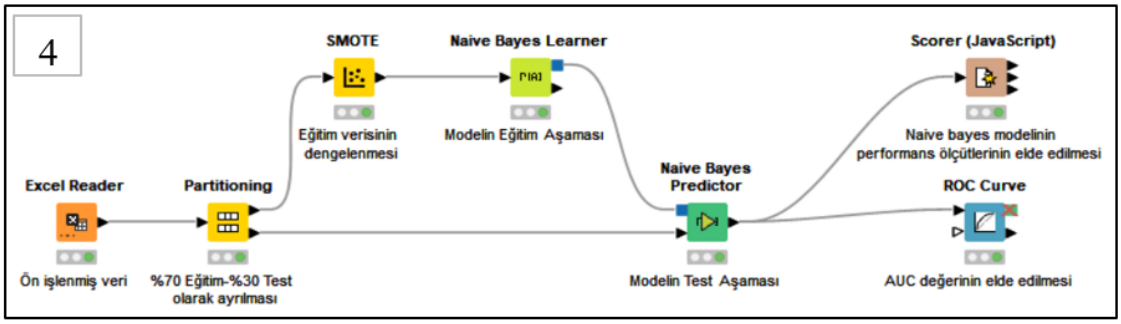


EK 4 KNIME programında oluşturulan NB modelleri

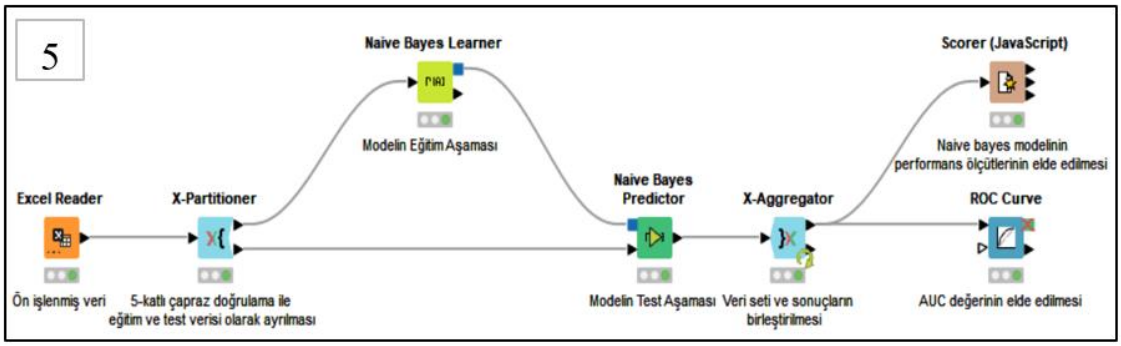


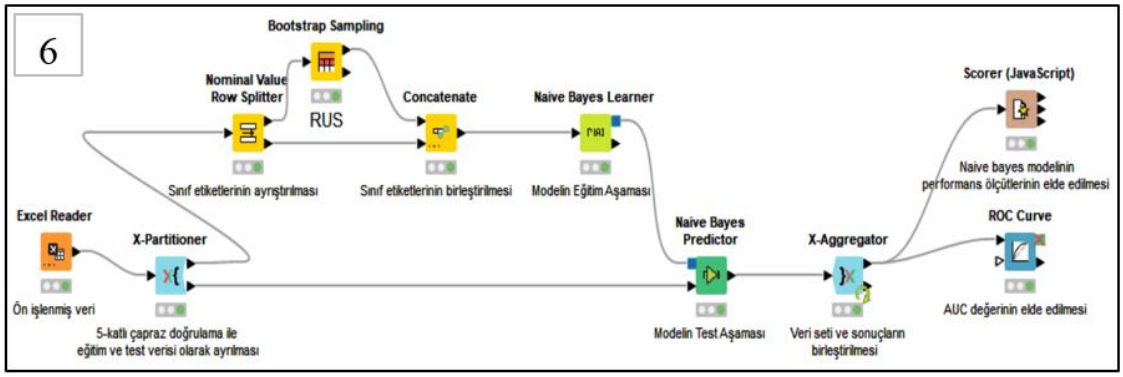


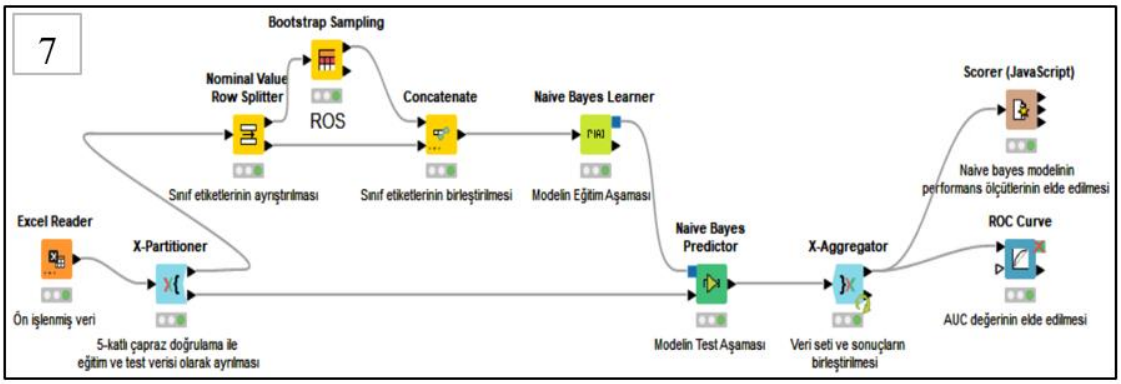


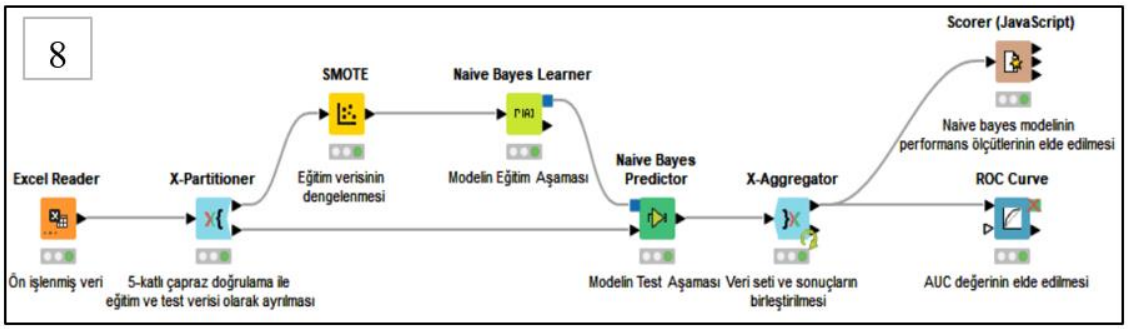


EK 4 KNIME programında oluşturulan NB modelleri (devam)

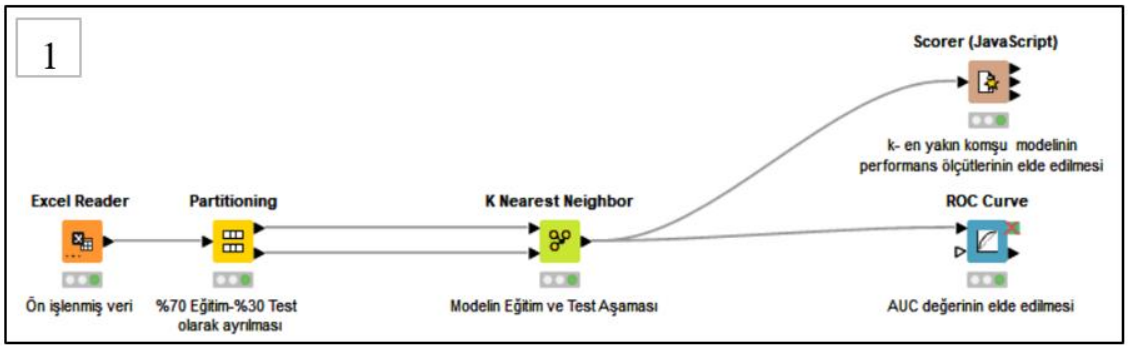


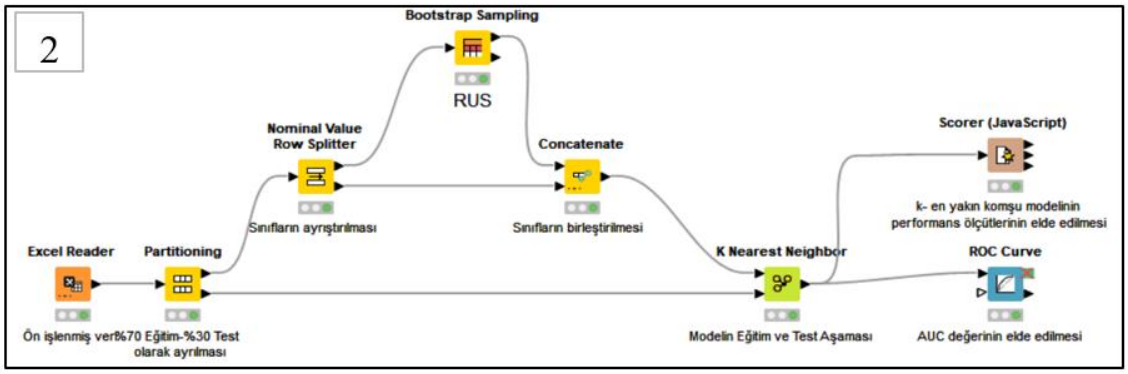


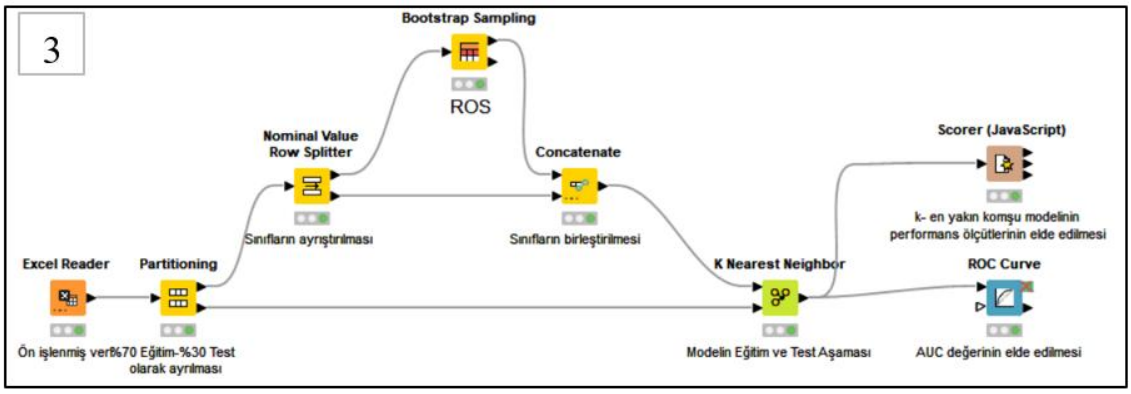


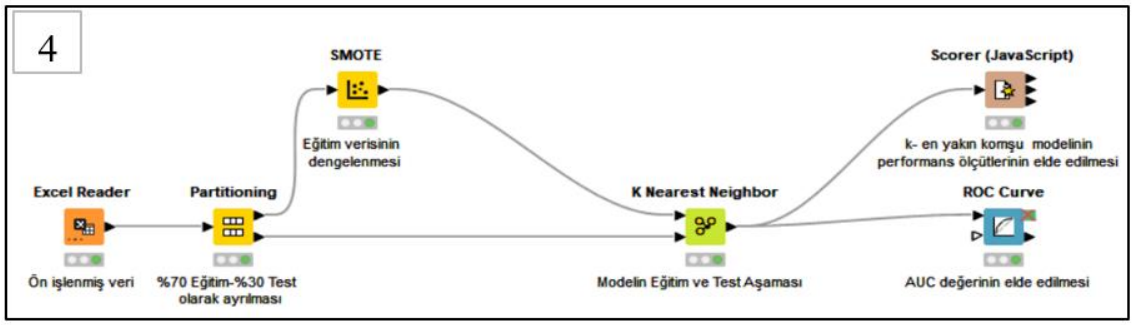


EK 5 KNIME programında oluşturulan k-EYK modelleri

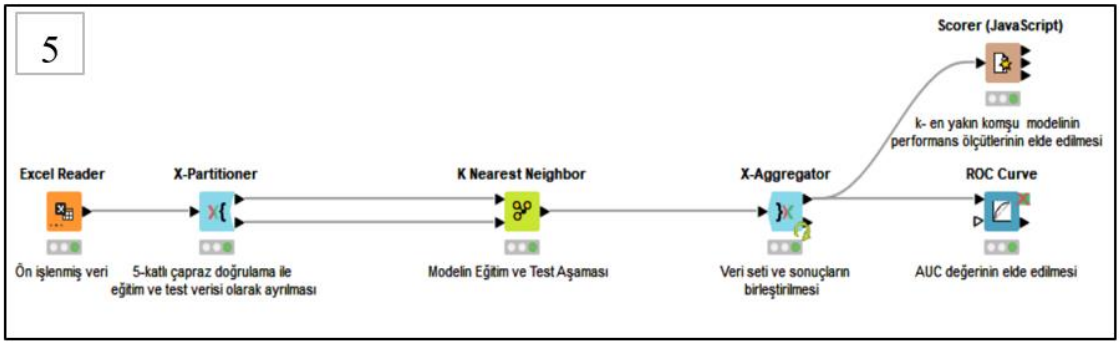


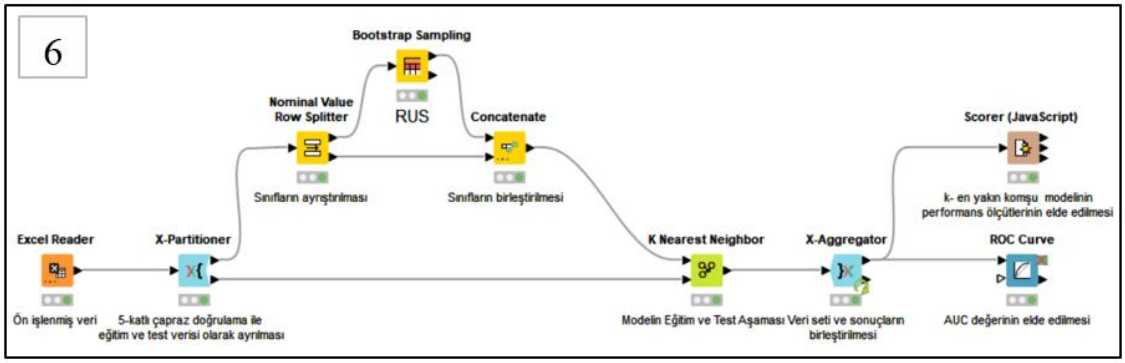


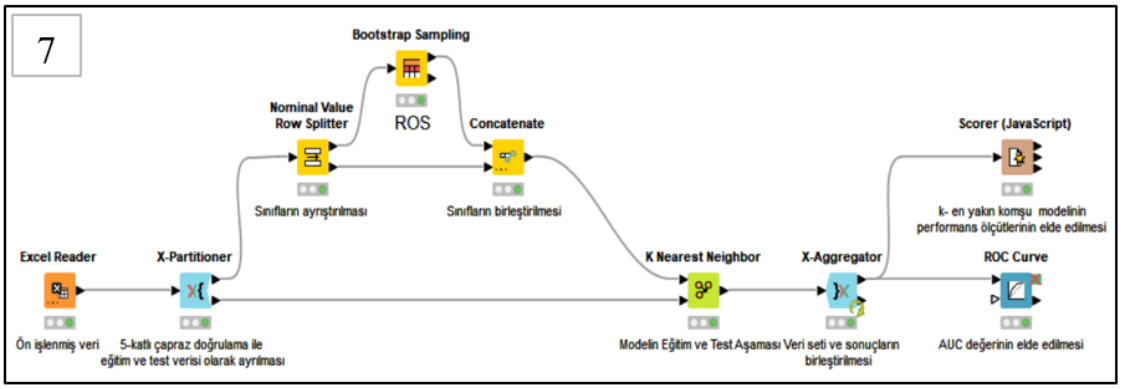




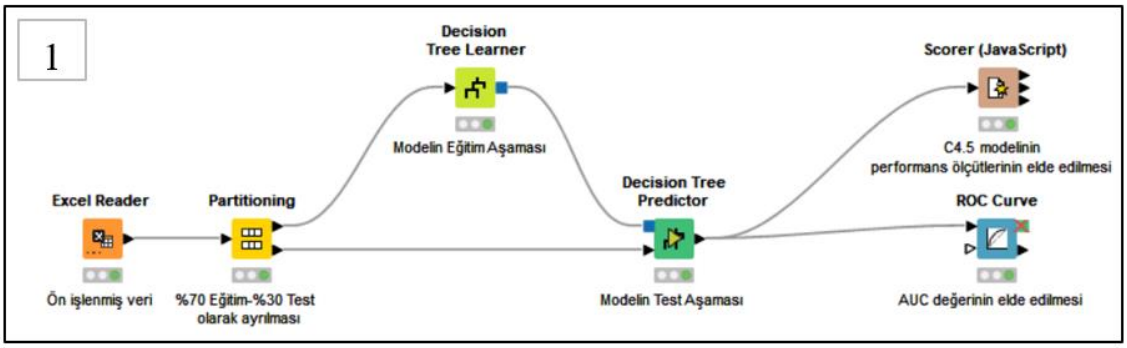
EK 5 KNIME programında oluşturulan k-EYK modelleri (devam)

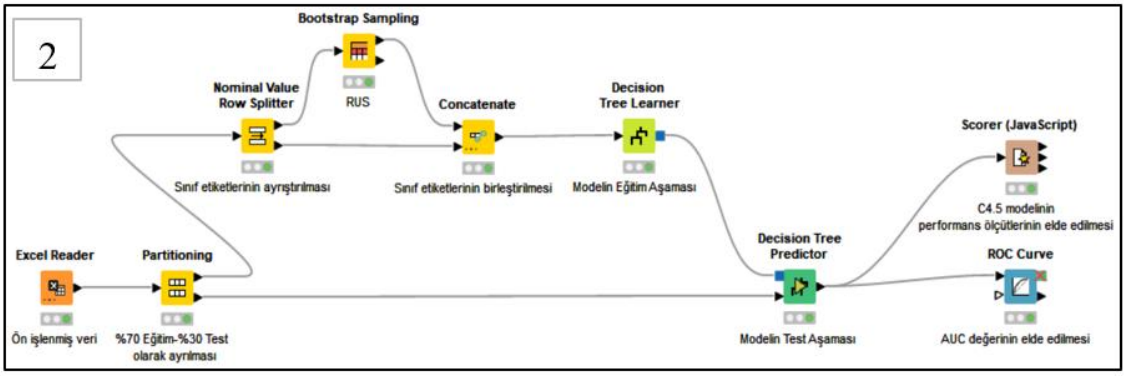


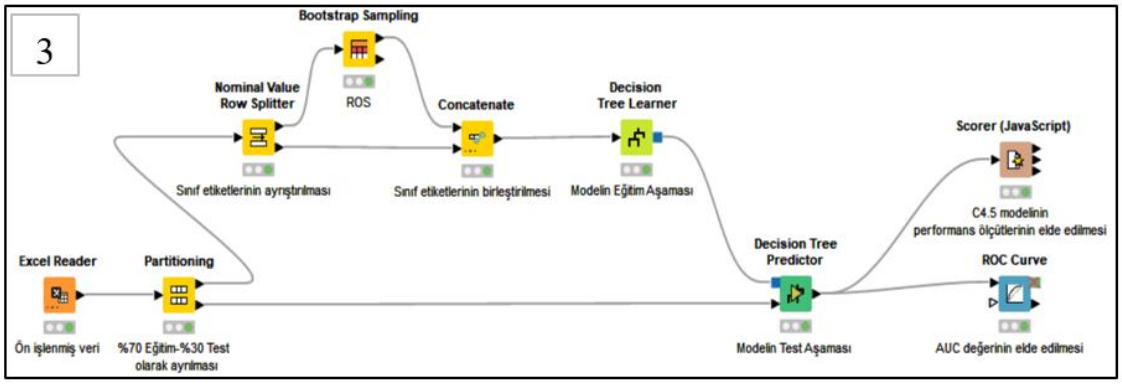


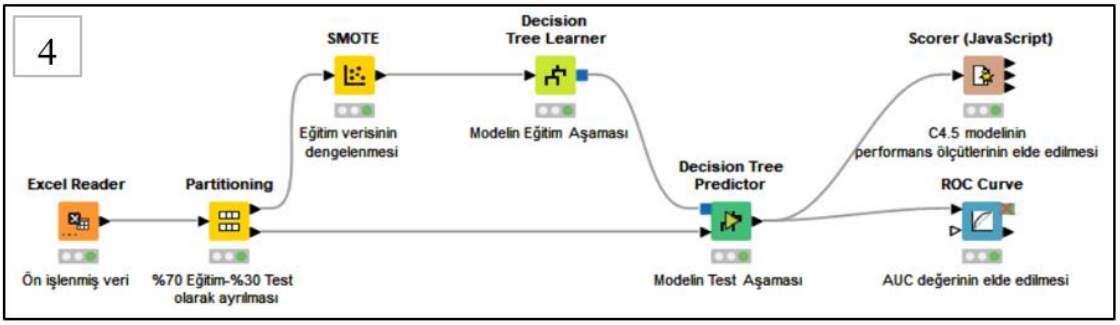


EK 6 KNIME programında oluşturulan C4.5 modelleri

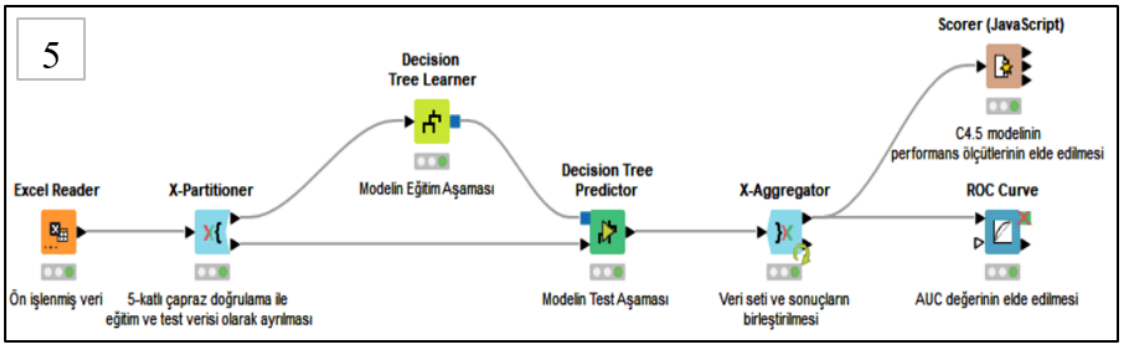


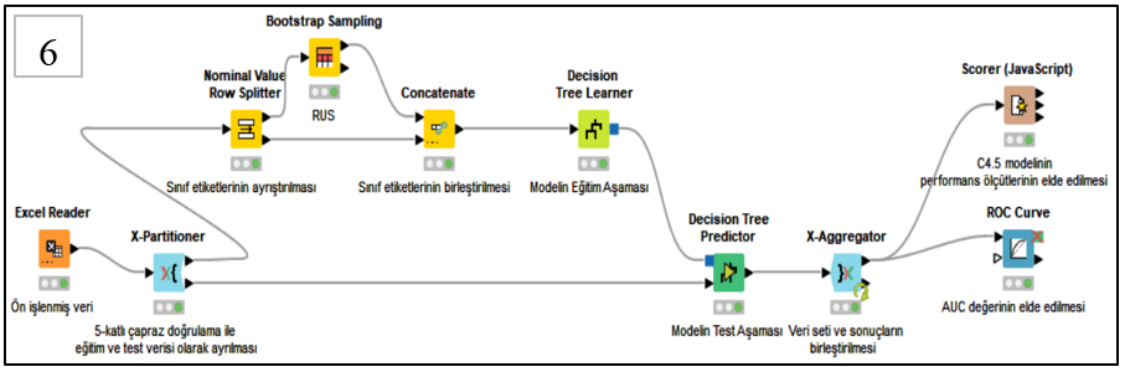


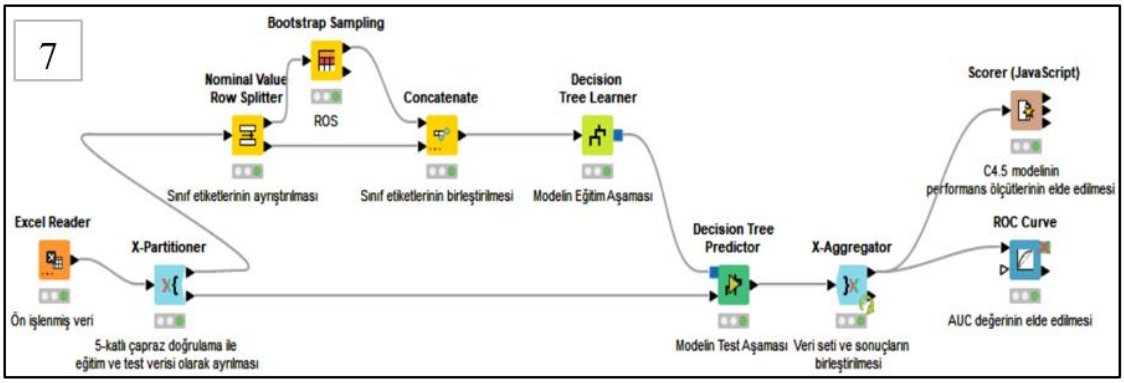


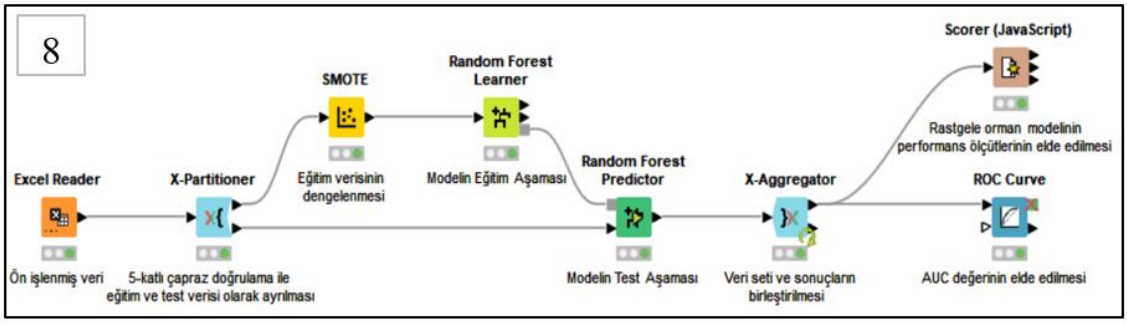


EK 6 KNIME programında oluşturulan C4.5 modelleri (devam)

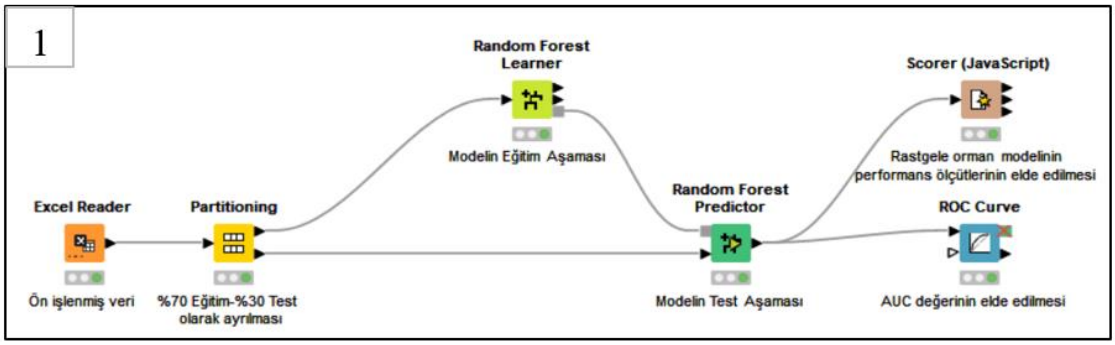


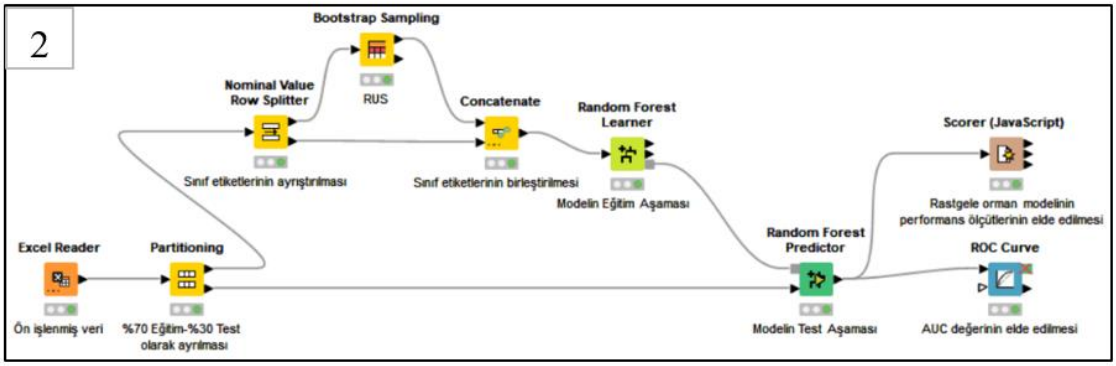


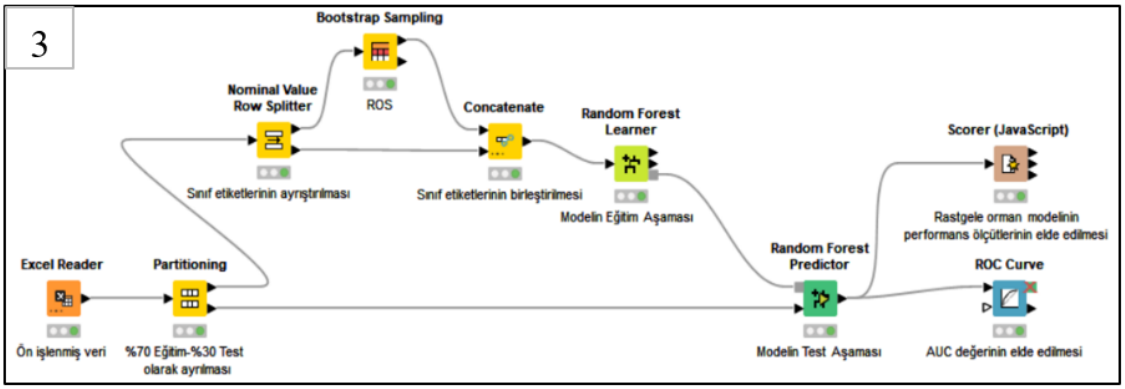


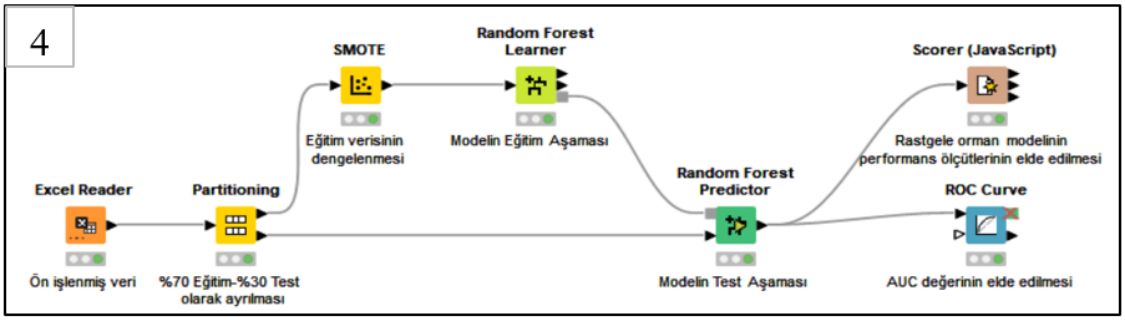


EK 7 KNIME programında oluşturulan RO modelleri

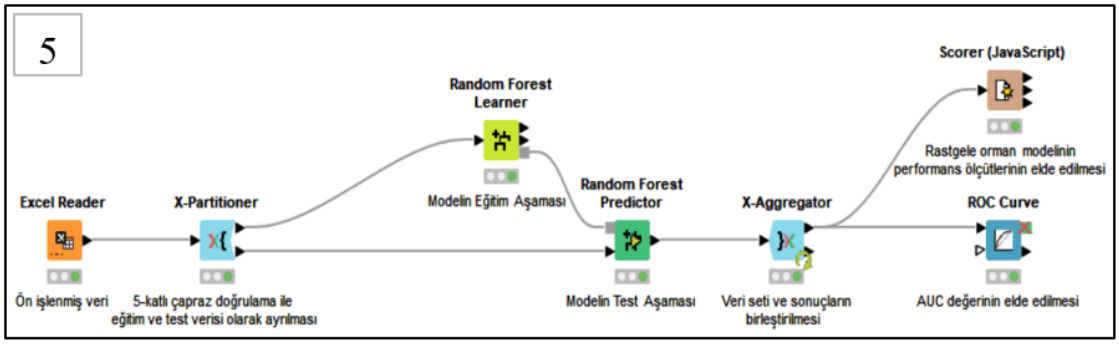


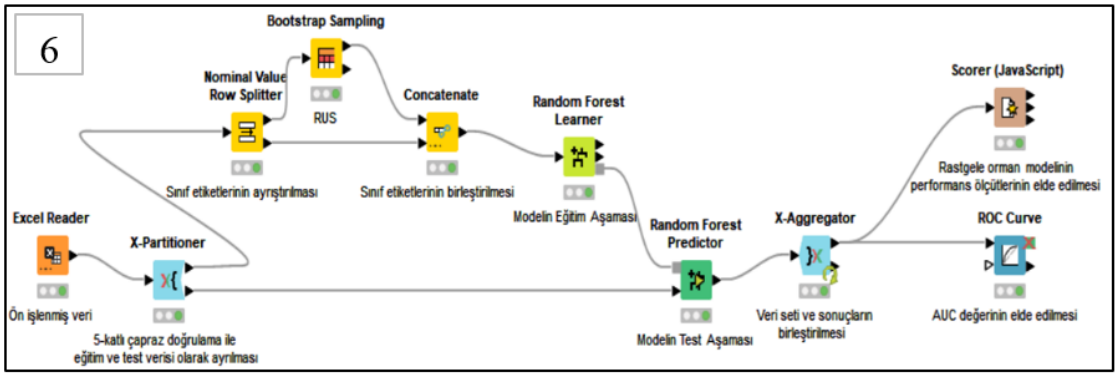


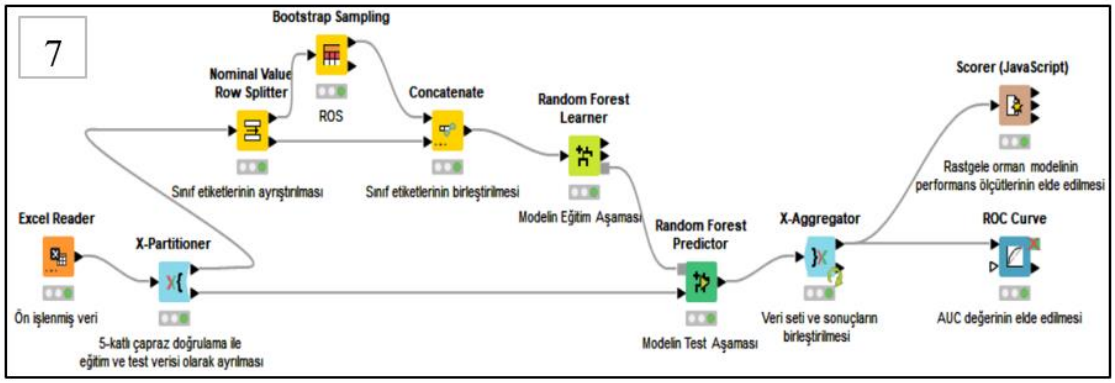


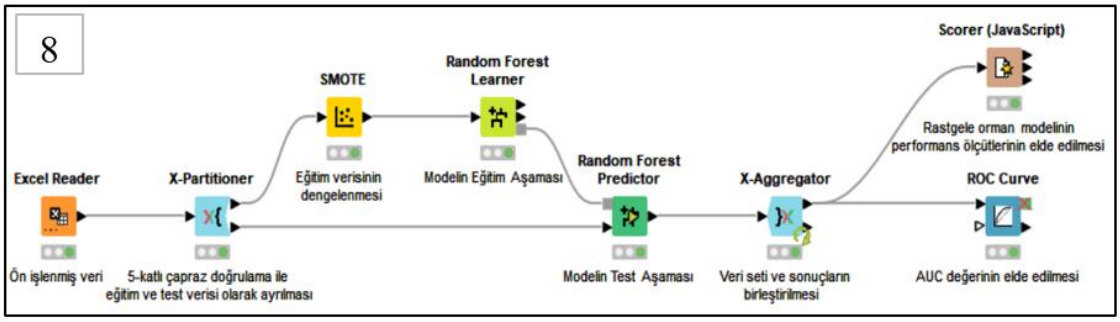


EK 7 KNIME programında oluşturulan RO modelleri (devam)

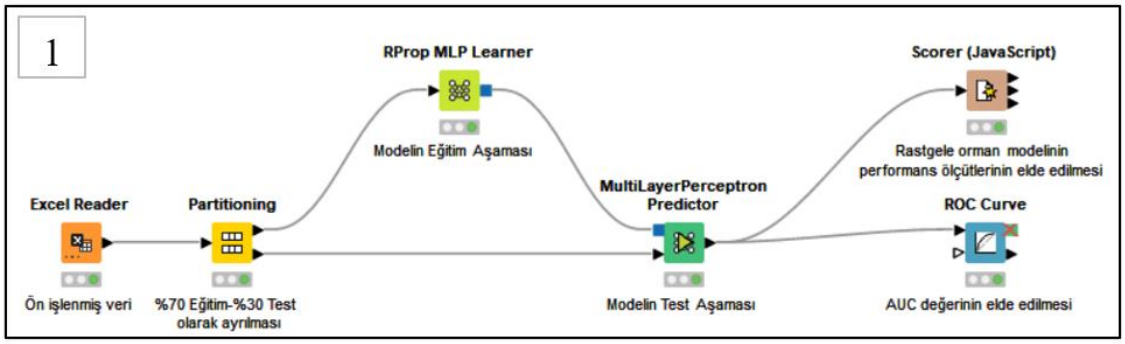


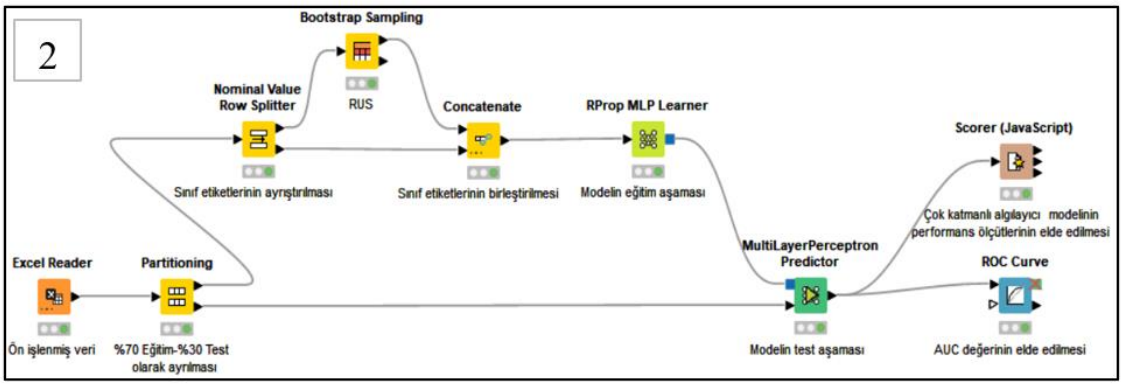


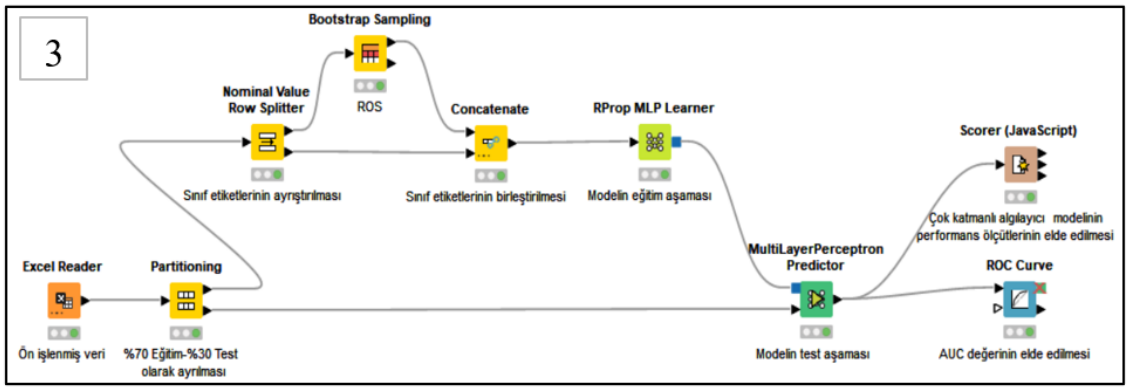


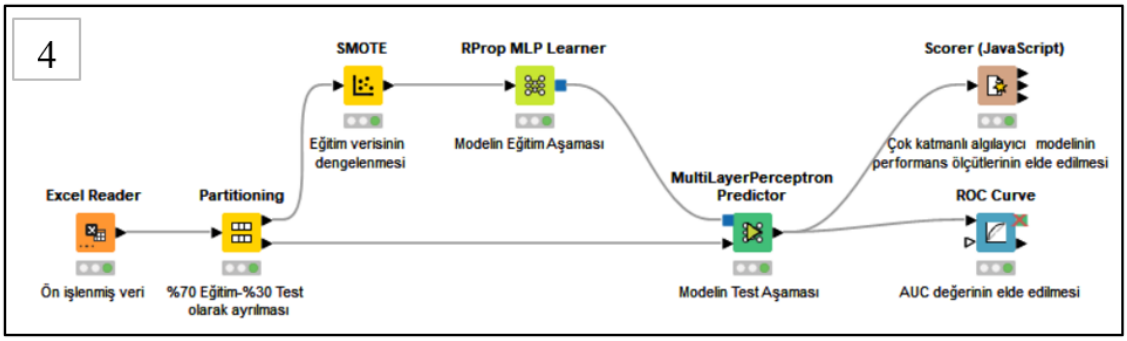


EK 8 KNIME programında oluşturulan ÇKA modelleri

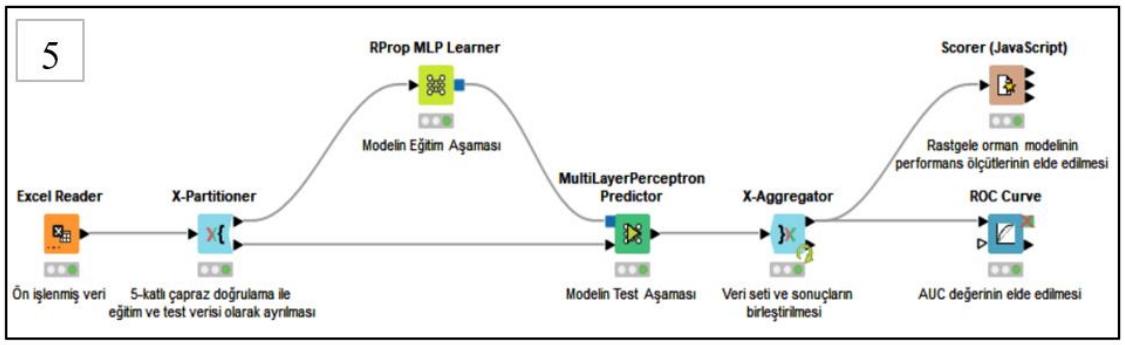


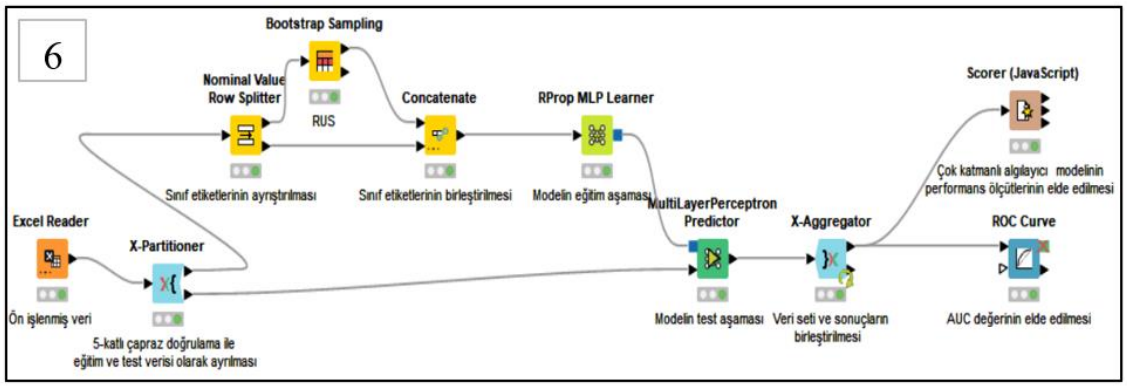


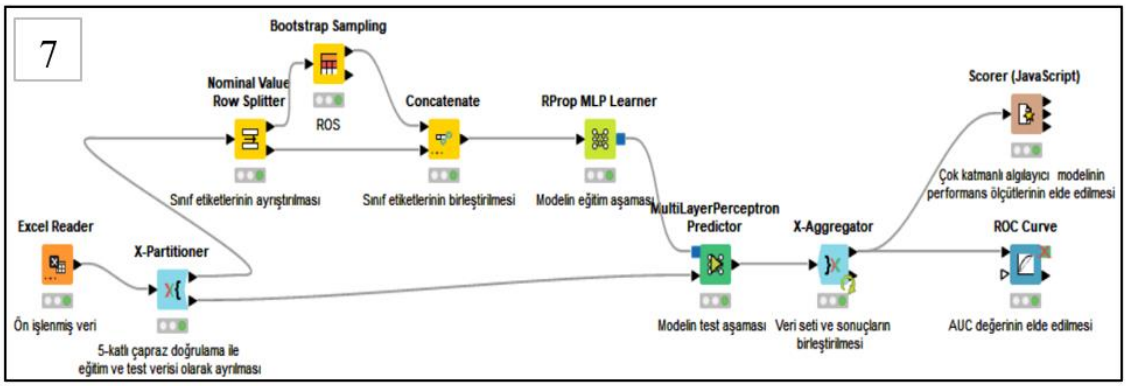


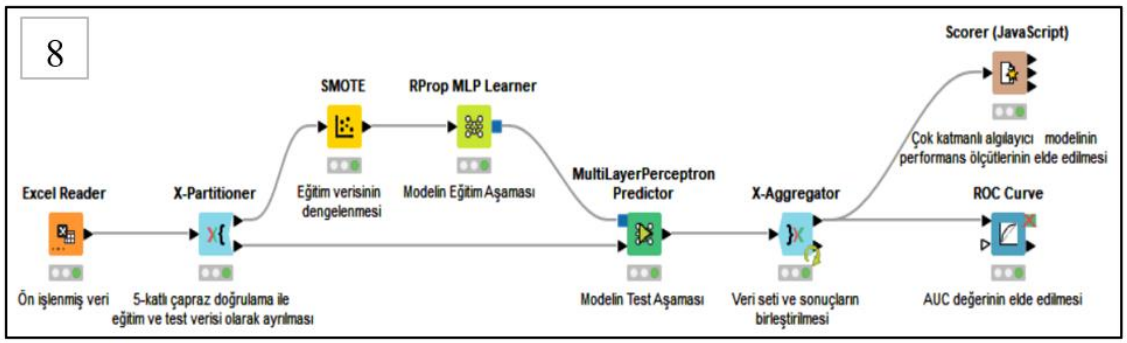


EK 8 KNIME programında oluşturulan ÇKA modelleri (devam)









ÖZGEÇMİŞ

Adı Soyadı : İsmail Buğra BÖLÜKBAŞI
Doğum Yeri ve Tarihi :
Yabancı Dil : İngilizce

Eğitim Durumu
Lise : Muharrem Hasbi Anadolu Lisesi
Lisans : Erciyes Üniversitesi-Endüstri Mühendisliği
Yüksek Lisans : Uludağ Üniversitesi-Endüstri Mühendisliği (Tezli)

Çalıştığı Kurum/Kurumlar : Yalova Üniversitesi
Yarış Kabin San. ve Tic. A.Ş.
Ergül Mobilya Ev Teks. İnş. San. ve Tic. A.Ş.

İletişim (e-posta) :

Yayınları : BÖLÜKBAŞI, İ. B., & YAĞMAHAN, B. (2022) Diagnosis of Diabetes Disease Using Machine Learning Methods in an Imbalanced Diabetes Dataset, *Cukurova 9th International Scientific Researches Conference*, Adana, Türkiye, 9-11 Ekim, ss. 330-331