MARMARA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEKÖĞRETİM PLANLAMA VE KARAR DESTEK SİSTEMİ: TEKNOLOJİ FAKÜLTESİ ÖRNEĞİ

ESRA YILMAZ

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Yüksek Lisans Programı

DANIŞMAN

DR. ÖĞR. ÜYESİ ZEHRA AYSUN ALTIKARDEŞ

EŞ-DANIŞMAN

PROF. DR. HASAN ERDAL

İSTANBUL, 2022

MARMARA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEKÖĞRETİM PLANLAMA VE KARAR DESTEK SİSTEMİ: TEKNOLOJİ FAKÜLTESİ ÖRNEĞİ

ESRA YILMAZ

523618009

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Yüksek Lisans Programı

DANIŞMAN

DR. ÖĞR. ÜYESİ ZEHRA AYSUN ALTIKARDEŞ

EŞ-DANIŞMAN

PROF. DR. HASAN ERDAL

İSTANBUL, 2022

TEŞEKKÜR

Öncelikle sabrımdan, gayretimden, kararlılığımdan ötürü kendime,

Tez çalışmam sürecinde daima yönlendiren, destek olan, yardımını esirgemeyen danışmanım Dr. Öğr. Üyesi Zehra Aysun Altıkardeş hocama ve eş danışmanım Prof. Dr. Hasan Erdal hocama,

Tez çalışmamda Marmara Üniversitesi Teknoloji Fakültesine ait öğrenci verilerini kullanmamıza müsaade eden rektörümüz Sayın Prof. Dr. Mustafa Kurt Hocamıza,

Maddi, manevi yardımlarını esirgemeyen yârim ve yoldaşım Abdülkadir Yılmaz Bey'ime, çalıştığım vakitleri anlayışla karşılayan oğullarım Muhammed Selim ve Yusuf Talha'ya, 14 aylık kızım Ayşe Hümeyra'ya,

Kızıma bakarak yardımlarını esirgemeyen ablam Ebru Şenbahar'a, yeğenlerim Tusem ve Tuana'ya, kardeşim Mehmet Ali Şahin'e, kuzenim Ayşe Yağmur Şahin'e,

Dualarıyla yanımda olan valideme, babama, kaimvalideme, kaimbabama,

Moral veren, motive eden dostlarıma, arkadaşlarıma

Teşekkür ederim.

Haziran, 2022 **Esra YILMAZ**

İÇİNDEKİLER

TEŞEKKÜR I
İÇİNDEKİLER II
ÖZET IV
ABSTRACT V
KISALTMALAR VI
ŞEKİL LİSTESİ VII
TABLO LİSTESİ IX
BÖLÜM 1 - GİRİŞ 1
1.1. Literatür Araştırması 4
BÖLÜM 2 - MATERYAL VE YÖNTEM 11
2.1. Veri Kümesi 12
2.2. Veri Ön İşleme 15
2.3. Özellik Seçimi 17
2.4. Verinin Eğitim ve Sağlamlık Veri Kümesine Bölünmesi 21
2.5. Yöntem 23
2.5.1. Makine öğrenmesi algoritmaları 28
2.5.1.1. Karar ağaçları (decision trees - DT) 30
2.5.1.2. Destek vektör makineleri (support vector machine - SVM) 31
2.5.1.4. K-en yakın komşu (k-nearest neighbors - KNN) 33
2.5.1.5. Naive Bayes (NB) 34
2.5.1.6. Çok katmanlı algılayıcı (multilayer perceptron - MLP) 35
2.5.1.7. Rastgele orman (random forest - RF) 36
2.5.1.8. Uyarlanabilir Artırma (Adaptive Boosting - AdaBoost) 37
2.5.1.9. Gradyan ağacı artırma (gradient tree boosting - GBM) 38
2.5.1.10. Aşırı gradyan artırma (extreme gradient boosting - XGBoost) 39
2.5.1.11. Son derece rastgele ağaçlar (extremely randomized trees - ExtraTrees - ET) 40
2.5.1.12. Dengeli torbalama (Balanced bagging - BB) 41
2.5.1.13. Dengeli rastgele orman (balanced random forest - BRF) 42
2.5.1.14. Kolay topluluk (easy ensemble - EE) 42
2.5.2. Hiper parametre ayarı (hyper parameter tune) 43
2.5.3. Sınıf dengesizliği problemi 44
2.5.4. Çapraz geçerleme (cross validation - CV) 46
2.5.4.1. Holdout çapraz geçerleme 46
2.5.4.2. K-fold çapraz geçerleme 47

2.5.4.3. Tekrarlanan rastgele alt örnekleme doğrulaması 47
2.5.4.4. Tabakalı k-kat çapraz geçerleme (Stratified k-fold CV) 48
2.5.5. Sağlamlık (Robustness) ve Tutarlılık Kontrolü 49
BÖLÜM 3 - BULGULAR VE TARTIŞMA 50
3.1. Performans Ölçütleri 50
3.1.1. Accuracy / Doğruluk / Kesinlik 50
3.1.2. Recall / Duyarlılık / Geri Çağırma / Sensitivity 50
3.1.3. Precision / Hassasiyet 51
3.1.4. F1-score 51
3.1.5. Özgüllük / Specificity 51
3.1.7. ROC Area 52
3.2. Analiz Sonuçları 52
3.3. Tartışma ve Karşılaştırma 54
3.3.1. 11 özellik ile oluşturulan modellere ait sonuçlar 54
3.3.1.1. 1. yönteme ait modellerin tutarlılık değerleri 56
3.3.1.2. 2. yönteme ait modellerin tutarlılık değerleri 57
3.3.1.3. 3. yönteme ait modellerin tutarlılık değerleri 60
3.3.1.4. Sağlamlık kontrolü sonrası modele karar verilmesi 63
3.3.2. Özellik seçimi ve oluşturulan modellerin sonuçları 64
3.3.2.1. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre model sonuçları 68
3.3.2.2. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre model sonuçları 70
3.3.2.3. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre model sonuçları 71
3.3.3. Karşılaştırma 72
3.4. Çalışmanın Güçlü ve Zayıf Yönleri 73
3.5. Karar Destek Sistemi Web Uygulaması 74
BÖLÜM 4 - SONUÇLAR 76
KAYNAKLAR 79
EKLER 87
EK 1 - Excel sayfa içerikleri 87
EK 2 - 4023 öğrenciye ait öğrenci durumları 93
EK 3 - Özellik Seçimi Sonrası Kalan 21 Özelliğin İstatistiksel Bilgileri 94
EK 4 - Etik Kurul Kararı 102
ÖZGEÇMİŞ 103

ÖZET

YÜKSEKÖĞRETİM PLANLAMA VE KARAR DESTEK SİSTEMİ: TEKNOLOJİ FAKÜLTESİ ÖRNEĞİ

Eğitimsel Veri Madenciliğinin bir alt dalı olan akademik performans tahmini ile ilgili çalışmalar son yıllarda arttı. Gerçek ortamlarda eğitimsel veri kümeleri çoğunlukla sınıf dengesizliğine ve çok sınıflı hedef değişkene sahiptir. Ancak çok sınıflı ve dengesiz eğitimsel veri kümesi ile yapılan çalışmalar oldukça azdır. Bu bağlamda, bu tez çalışmasında, 23.05.2022-286783 etik no kararı ile Marmara Üniversitesi Teknoloji Fakültesi öğrencilerine ait veri kümesi kullanılarak, çok sınıflı dengesiz eğitimsel veri kümesi ile, riskli öğrencileri tespit etmek için öğrenci mezuniyet durum tahmini yapıldı. Veri ön işleme ve özellik seçimi (FS) aşamalarından sonra 4023 örneklem ve 319 özellikten, 1394 örneklem ve 11 özellik elde edildi. 2016 yılına ait 153 öğrenci sağlamlık kontrolü için kullanıldı. 7 farklı FS ile elde edilen 11, 7 ve 5 özellik içeren 3 farklı veri kümesi oluşturuldu. 9 farklı yeniden örnekleme yöntemi ve 16 farklı makine öğrenmesi algoritması kullanılarak birbirinden farklı 750 model oluşturuldu. Oluşturulan modellere sağlamlık kontrolü yapıldı. Başarı ölçütü olarak F1 Score ve Repeated Stratified 5*5 fold-CV kullanıldı. Hiper parametre ayarları GridSearchCV ile yapıldı. Sonuç olarak RandomOverSampler+RandomForest F1 Score 0.9935 değeriyle en başarılı algoritma olmasına rağmen, en başarılı ve en tutarlı modeller 7 özellikli None+ExtraTrees F1 Score 0.8187, None+MLP F1 Score 0.8255, None+Bagging_DesicionTree F1 Score 0.8302 ve None+RandomForest F1 Score 0.8341 değeriyle 4 model oldu.

Son olarak, yukarıda bahsedilen modeller kullanılarak Marmara Üniversitesi Teknoloji Fakültesini kazanarak kayıt olan öğrencilerin tercih ettikleri mühendislik bölümlerindeki mezuniyet, ilişik kesme gibi nihai durumlarını tahminleyerek, web platformunda sonucu hem yazılı hem de grafiklerle sunacak şekilde tasarlanan karar destek sistemi fakülte öğretim üyelerinin deneyimine sunulmuştur.

Keywords: Educational Data Mining, Imbalanced dataset, multiclass dataset, Prediction, Sampling Techniques, Ensemble methods.

ABSTRACT

HIGHER EDUCATION PLANNING AND DECISION SUPPORT SYSTEM: A CASE OF TECHNOLOGY FACULTY

Studies on academic performance prediction, a sub-branch of Educational Data Mining, have increased in recent years. Educational datasets in real environments often have both an imbalanced and a multiclass target variable. However, studies with multi-class and imbalanced educational datasets are very few. In this context, in this thesis study, using the data set of Marmara University (MU) Faculty of Technology (TF) students with the ethical no decision of 23.05.2022-286783, with a multi-class imbalanced educational dataset, a student graduation status estimation was made to identify risky students. A pilot decision support system was built. After data preprocessing and feature selection (FS) stages, a dataset containing 1394 samples and 11 features was obtained from 4023 samples and 319 features. 153 students from 2016 left the dataset for robustness control. Then, with 3 different datasets containing 11, 7 and 5 features obtained by comparing 7 different FS algorithms were created. Using 9 different sampling methods and 16 different machine learning algorithms, 750 different models were created. The robustness control of the created models was performed. F1 Score and Repeated Stratified 5*5 fold-CV were used as success criteria. Hyperparameter settings were made with GridSearchCV. As a result, although RandomOverSampler+RandomForest was the most successful algorithm with an F1 Score of 0.9935, the most successful and most consistent models were the 7-featured with F1 Score as None+ExtraTrees 0.8187, None+MLP 0.8255, None+Bagging_DesicionTree 0.8302 and None+RandomForest 0.8341.

Finally, using the above-mentioned models, the final status of the students who enrolled in the Marmara University Faculty of Technology, such as graduation and dismissal in their preferred engineering departments, was estimated. The decision support system, designed to present the results both in writing and in graphics, on the web platform was presented to the faculty members.

Keywords: Educational Data Mining, Imbalanced dataset, multiclass dataset, Prediction, Sampling Techniques, Ensemble methods.

KISALTMALAR

MÜ	:	Marmara Üniversitesi
FBEM	:	Fen Bilimleri Enstitü Müdürlüğü
TF	:	Teknoloji Fakültesi
EDM	:	Eğitimsel Veri Madenciliği / Educational Data Mining
ID	:	Dengesiz / Imbalanced
MC	:	Çok Sınıflı / Multiclass
SVM	:	Support Vector Machine
RF	:	Random Forest
DT	:	Decision Tree
NB	:	Naive Bayes
KNN	:	K-Nearest Neighbor
LR	:	Logistic Regression
FS	:	Feature Selection
GBM	:	Gradient Boosting
AB	:	Adaptive Boosting - AdaBoost
MLP	:	Multilayer Perceptron
XGboost	:	Extreme Gradient Boosting
BSMOTE1	:	BorderlineSMOTE1
BSMOTE2	:	BorderlineSMOTE2
ANN	:	Artificial Neural Network
AdaBoost	:	Adaptive Boosting
CV	:	Çapraz Geçerleme / Cross Validation

ŞEKİL LİSTESİ

Şekil 1.1. EDM disiplini.....1

Şekil 1.2. EDM, MC ve ID disiplinler arası çalışmaların yıllara göre dağılımı.....3

Şekil 1.3. EDM, MC ve ID disiplinler arası yapılan çalışmaların yıllara göre dağılımı.....4

Şekil 1.4. Bilimsel veri kaynakları ve EDM, MC ve ID disiplinler arası çalışmaların yıllara göre dağılımı.....5

Şekil 2.1. Tez çalışmasında kullanılan donanımsal ve yazılımsal sistem mimarisi.....12

Şekil 2.2. MÜ TF 2021 Şubat ayına kadar bölümlere ait ilk öğrenci alma ve mezun yılları...13

Şekil 2.3. MÜ TF bölümlerine ait öğrenci mezuniyet durum dağılımları.....14

Şekil 2.4. Ham veri kümesi öğrenci mezuniyet durum dağılımları.....15

Şekil 2.5. MÜ'den alınan verinin işlenme süreci.....15

Şekil 2.6. Veri önişleme aşamaları ve yapılan işlemler.....16

Şekil 2.7. Sağlamlık verisi hedef özelliğine ait durum dağılımları.....21

Şekil 2.8. 2016 yılında kayıt olan öğrencilerin mezuniyet durumları.....22

Şekil 2.9. Temizlenmiş veri kümesinde mezuniyet durum dağılımları.....22

Şekil 2.10. Temizlenmiş veri kümesinde, bölümlere göre durum dağılımları.....23

Şekil 2.11. Tez çalışması makine öğrenmesi genel akış şeması.....24

Şekil 2.12. Tez çalışması makine öğrenmesi genel akış şeması 2.....25

Şekil 2.13. 1. ve 2. yönteme ait ayrıntılı akış şeması.....26

Şekil 2.14. 3. yönteme ait ayrıntılı akış şeması.....27

Şekil 2.15. Tez çalışmasında kullanılan makine öğrenmesi algoritmaları.....28

Şekil 2.16. Bir veri kümesi örneği ve karar ağacı yapısı.....30

Şekil 2.17. Destek Vektör Makineleri hiperdüzlem ayarı.....32

Şekil 2.18. Sigmoid Fonksiyonunun Grafiği.....33

Şekil 2.19. Çeşitli k değerleri için k-en yakın komşu tahmini.....34

Şekil 2.20. Naive Bayes sınıflandırması.....35

Şekil 2.21. MLP yapısı.....36

Şekil 2.22. Random Forest sınıflandırıcısı.....37

Şekil 2.23. Adaboost sınıflandırıcısı.....38

Şekil 2.24. Gradient Boosting Karar Ağacının mimarisi.....39

Şekil 2.25. Karar-ağacı algoritmalarının XGBoost'a Evrimi.....40

Şekil 2.26. XGBoost mimarisi.....40

Şekil 2.27. ExtraTrees mimarisi.....41

Şekil 2.28. Balanced Bagging mimarisi.....41

Şekil 2.29. Easy ensemble sınıflandırıcısı.....42

Şekil 2.30. Modifiye edilmiş dengeli rastgele orman mimarisi.....43

Şekil 2.31. Yeniden örnekleme yöntemleri.....45

Şekil 2.32. Hold out geçerleme yöntemi.....47

Şekil 2.33. K-fold geçerleme iterasyon örneği.....47

Şekil 2.34. Stratified k-fold geçerleme iterasyon örneği.....48

Şekil 2.35. Repeated stratified k-fold geçerleme iterasyon örneği.....48

Şekil 3.1. chi-2 ile yapılan özellik seçimi değerleri.....66

Şekil 3.2. ET ile yapılan özellik seçimi değerleri.....66

Şekil 3.3. RF ile yapılan özellik seçimi değerleri.....66

Şekil 3.4. 11 özelliğin korelasyon sıcaklık tablosu.....67

Şekil 3.5. En başarılı ve en tutarlı modellerin test ve sağlamlık F1 Score yüzdeleri.....72

Şekil 3.6. Karar Destek Sistemi giriş paneli.....74

Şekil 3.7. Karar Destek Sistemi kullanılarak öğrenci bilgilerinin girildiği sayfa.....75

Şekil 3.8. Karar Destek Sistemi kullanılarak yapılan tahmin sonuçları.....75

TABLO LİSTESİ

Tablo 1.1. Sistematik literatür araştırmasında incelenen makalelerin gruplandırılması.....6

Tablo 2.1. Veri ön işlem sonrası kalan öğrenci mezuniyet durumları.....17

Tablo 2.2. Tez çalışmasında kullanılan özelliklerin veri türleri.....19

Tablo 2.3. Tez çalışmasında kullanılan özelliklerin değerleri ve açıklamaları.....19

Tablo 2.4. Yöntemler sonucunda oluşan model sayıları.....25

Tablo 2.5. Makine öğrenmesi algoritma türleri.....28

Tablo 3.1. Karmaşıklık matrisi.....50

Tablo 3.2. 11 özellik içeren veri kümesi eğitim modeli F1 Score sonuçları.....55

Tablo 3.3. 11 özellik içeren veri kümesi sağlamlık kontrolü F1 Score sonuçları.....56

Tablo 3.4. 11 özellik ve 1. yöntem ile oluşturulan modellerin tutarlılık oranları.....57

Tablo 3.5. 11 özellik ve 2. yöntem %75 üzeri F1 Score ve tutarlılık oranları.....57

Tablo 3.6. 11 özellik içeren 1. ve 2. yöntem ile elde edilen modellerden web uygulamasında kullanılabilecek en tutarlı ve en başarılı algoritmalar.....60

Tablo 3.7. 11 özellik ve 3. yöntem %75 üzeri F1 Score ve tutarlılık oranları.....60

Tablo 3.8. 11 özellik içeren 1. ve 3. yöntem ile elde edilen modellerden web uygulamasında kullanılabilecek en tutarlı ve en başarılı algoritmalar.....63

Tablo 3.9. Özelliklerin feature=7 olduğunda karşılaştırmalı tablosu.....64

Tablo 3.10. Özelliklerin feature=5 olduğunda karşılaştırmalı tablosu.....65

Tablo 3.11. chi2 ile yapılan özellik seçimi değerleri.....65

Tablo 3.12. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre sonuçları.....69

Tablo 3.13. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre sonuçları.....70

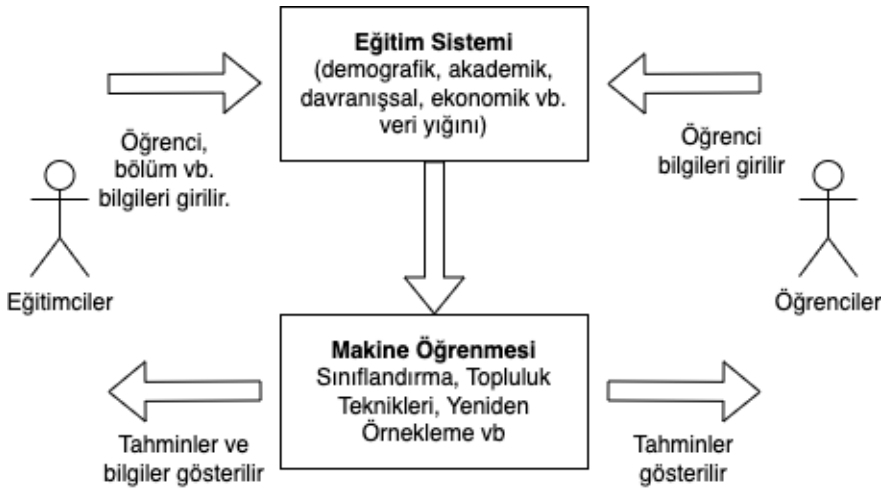
Tablo 3.14. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre sonuçları.....71

Tablo 3.15. KDS Web Uygulamasında kullanılan en tutarlı ve en başarılı algoritmalar.....72

BÖLÜM 1 - GİRİŞ

"Yükseköğretimin amacı, öğrencileri, ilgi ve yetenekleri yönünde yurt kalkınmasına ve ihtiyaçlarına cevap verecek, aynı zamanda kendi geçim ve mutluluğunu sağlayacak bir mesleğin bilgi, beceri, davranış ve genel kültürüne sahip, vatandaşlar olarak yetiştirmektir." [Yükseköğretim Kanunu. Madde 4/a/7][1]. Bu bağlamda, öğrencilerin doğru meslek seçmesi, akademik performansının geliştirilmesi, iyileştirilmesi ve takibinin yapılması çok önemlidir. Bölümü bırakma ihtimali olan öğrencileri önceden tespit etmek, ihtiyaçlarını gidermek, öğrencilere akademik başarısının artması yönünde destek vermek ve okuduğu bölüme geri kazandırmak yükseköğretimin amaçlarındandır. Yanlış bölüm seçiminden dolayı bölümü bırakma ihtimali olan öğrencilerin önceden tespit edilerek, yükseköğretime giriş aşamasında doğru mesleğe yönlendirilmesi ise öğrencinin, yükseköğretimin ve ülkenin lehinedir.

"Eğitim teknolojisini üretmek, geliştirmek, kullanmak, yaygınlaştırmak Yükseköğretim Kurumlarının görevlerindendir."[Yükseköğretim Kanunu. Madde 12/h][1]. Eğitim Teknolojisi, eğitimde performans artışını sağlamak için teknolojinin eğitim alanında kullanılmasıdır[2]. Son yıllarda, eğitim alanında kullanılan Eğitsel Veri Madenciliği (Educational Data Mining) (EDM) teknolojisi kullanımında artış vardır. Uluslararası Eğitim Veri Madenciliği Derneği, EDM'yi, "eğitim ortamlarından gelen benzersiz ve giderek daha büyük ölçekli verileri keşfetmek için yöntemler geliştirmek ve bu yöntemleri öğrencileri ve içinde öğrendikleri ortamları daha iyi anlamak için kullanmakla ilgilenen, gelişmekte olan bir disiplindir." şeklinde tanımlar[3]. EDM, Şekil 1'de görüldüğü gibi, eğitim verilerini analiz ederek, makine öğrenme yöntemleri ile öğrenme deneyimini ve kurumsal etkinliği geliştirmek için modeller geliştirir.



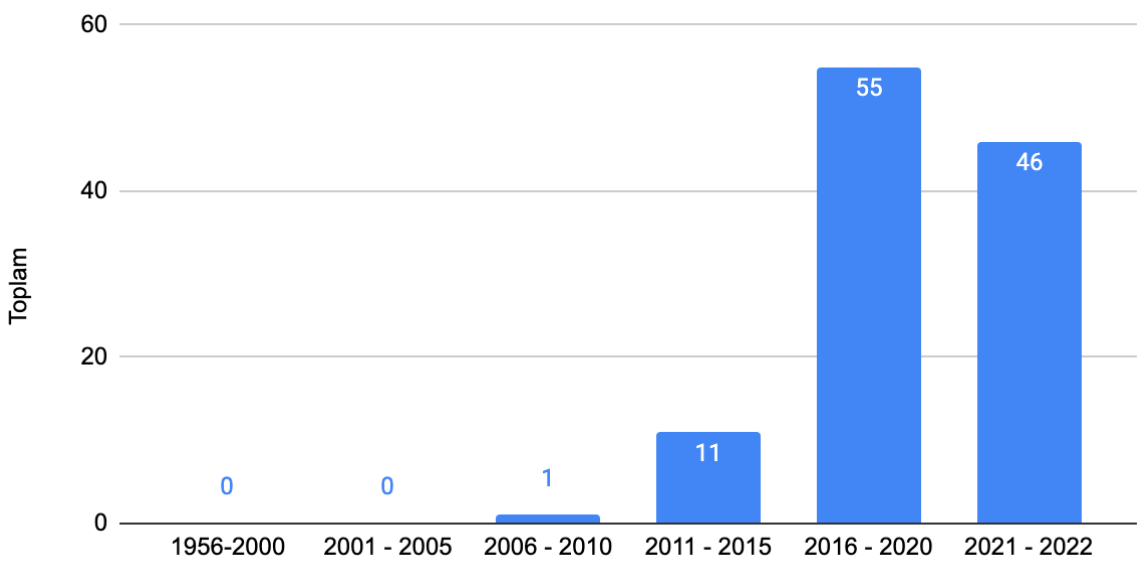
Şekil 1.1. EDM disiplini

EDM kullanarak geliştirilen modelin başarısı, bir çok kritere bağlıdır. Makine öğrenmesi algoritmalarını seçmeden önce kullanılacak verinin doğru analiz edilmesi büyük önem taşır. Bu nedenle eğitim verilerinin büyüklüğü, karmaşıklığı göz önüne alındığında, eğitim verisinin ön işlem aşamasında çok iyi organize edilmesi gerekir. Ayrıca veri kümesindeki hedef değişkene ait sınıf dağılımındaki oranlar arasında büyük farklar varsa, veri kümesinde dengesizlik oluşur[4]. Bu tür veri kümelerine dengesiz (imbalanced)(ID) veri kümesi denir. Veri kümesinde bu şekilde sınıf dengesizliği probleminin olması model başarısında oldukça etkilidir. Çok küçük oranda bulunan sınıf, azınlık sınıf olarak adlandırılır. Gerçek yaşam verilerinde, genellikle azınlık sınıfı daha önemlidir ve azınlık sınıfına ait verinin tanınma oranını iyileştirmek için farklı yöntemler uygulanır[5]. Çünkü çoğu sınıflandırma algoritmasının öğrenme süreci çoğunlukla çoğunluk sınıfı örneklerine karşı önyargılıdır, yani azınlık sınıfına ait olanlar modelleme sürecinde doğru şekilde modellenemez[6]. Bu problemi çözmek ve azınlık sınıfına ait olan verinin de sınıflandırma doğruluğunu artırmak için, veri düzeyinde yeniden örnekleme ve algoritma düzeyinde modifiye edilmiş öğrenme algoritmaları şeklinde iki ana yaklaşım vardır[7].

Veri kümesindeki hedef değişken kategorik ve sayısal olabilir. Kategorik değişkenler, sınıflanabilir(nominal) ve sıralanabilir(ordinal) şeklinde; sayısal değişkenler ise kesikli ve sürekli olarak ikiye ayrılırlar[8]. Bu şekilde kategorik değişkenlerden sınıflanabilir veya sıralanabilir olan hedef değişkenler ikili ve çok sınıflı (multiclass) (MC) olarak ayrılır. Hedef değişken sayısal değerler olduğunda "regresyon problemi", ikili nominal sınıflı değerler olduğunda "sınıflandırma problemi" ve çok sınıflı değerler olduğunda ise "çok sınıflı sınıflandırma problemi (multiclass classification problem)" olarak adlandırılır. Perceptron, Lojistik Regresyon ve Destek Vektör Makineleri gibi sınıflandırma algoritmalarında en yaygın ayar yalnızca iki sınıfı içerecek şekildedir, bu nedenle çok sınıflı sınıflandırmayı doğal olarak desteklemez. Bu problemi çözmek üzere tasarlanan çok sınıflı veri kümesini ikili sınıflı veri kümesine bölerek model oluşturulan One-vs-Rest ve One-vs-One şeklinde yaklaşım vardır[9].

Ulusal ve uluslararası eğitimsel veri kümeleri incelendiğinde, eğitimsel veri kümelerinin çok boyutlu ve büyük olduğu, aynı zamanda iki sınıflı, çok sınıflı, kategorik, sayısal, dengeli ya da çoğunlukla dengesiz bir sınıf dağılımına sahip olduğu görülür. Bu bağlamda, "Literatür Araştırması" bölümünde ayrıntısıyla anlatılan literatür araştırmasında, EDM alanındaki en büyük boşluğun MC ve ID veri kümesi ile yapılan öğrenci performans tahmin çalışmalarının oldukça az olduğudur. Bazı çalışmaların sınıf dengesizliğini göz ardı ettikleri veya çalışma kapsamı dışında tuttukları görüldü. Ancak bu durum gerçek eğitimsel veri kümelerindeki sınıf

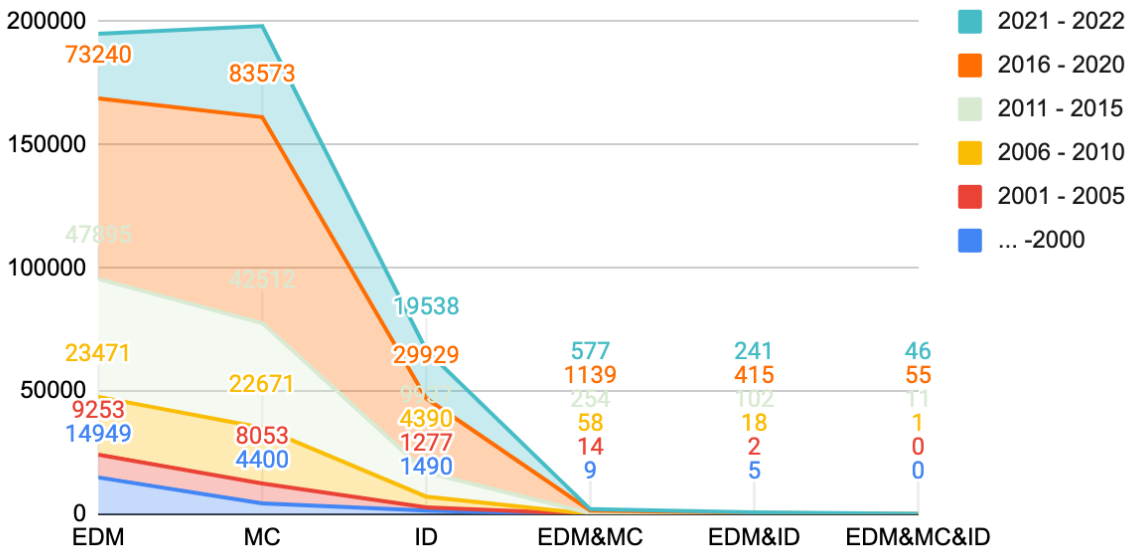
dengesizliği gerçeğini değiştirmemektedir. Bu nedenle bu tez çalışmasında MC ID eğitimsel veri kümesi ile öğrenci mezuniyet tahmini yapılması amaçlandı. Bu durumda bu tez çalışmasının diğer çalışmalardan en önemli farkı, veri kümesi olarak hem MC hem de ID olan bir eğitimsel veri kümesi kullanmasıdır. Ulusal tez tarama merkezinde EDM MC ID ile ilgili bir çalışmaya rastlanmadı. Ayrıca uluslararası aramalarda da MC ID eğitimsel veri kümesi ile özellikle 2011 yılından itibaren Şekil 1.2'de görüldüğü üzere sadece 113 çalışma yapıldığı görüldü[10]. Bu durum bu çalışmanın birinci motivasyonunu oluşturdu. Bu bağlamda, bu tez çalışmasının en önemli özgünlüğü veri kümesinin özgün olması ve çok sınıflı dengesiz eğitimsel veri kümesi ile yapılan çalışmaların oldukça az olmasıdır. Ayrıca, bu çalışmada, Marmara Üniversitesi(MÜ) Fen Bilimleri Enstitü Müdürlüğü (FBEM)'nün 23.05.2022-286783 etik no kararı ile MÜ Teknoloji Fakültesi(TF) mühendislik öğrencilerine ait veri kümesi kullanıldı. Bu bağlamda pilot bir karar destek sistemi web uygulaması yapılarak tez çalışmasının Marmara Üniversitesi Teknoloji Fakültesi Öğretim Üyeleri için kullanılabilir olmasının sağlanması, çalışmanın 2. önemli özgün özelliğidir. Yöntem olarak makine öğrenmesi algoritmalarının ve yeniden örnekleme yöntemlerinin farklı türlerinin, farklı şekillerde uygulanmasıyla elde edilen karşılaştırmalı sonuçlarla literatüre farklı çözümler sunulmaya çalışıldı.



Şekil 1.2. EDM, MC ve ID disiplinler arası çalışmaların yıllara göre dağılımı

1.1. Literatür Araştırması

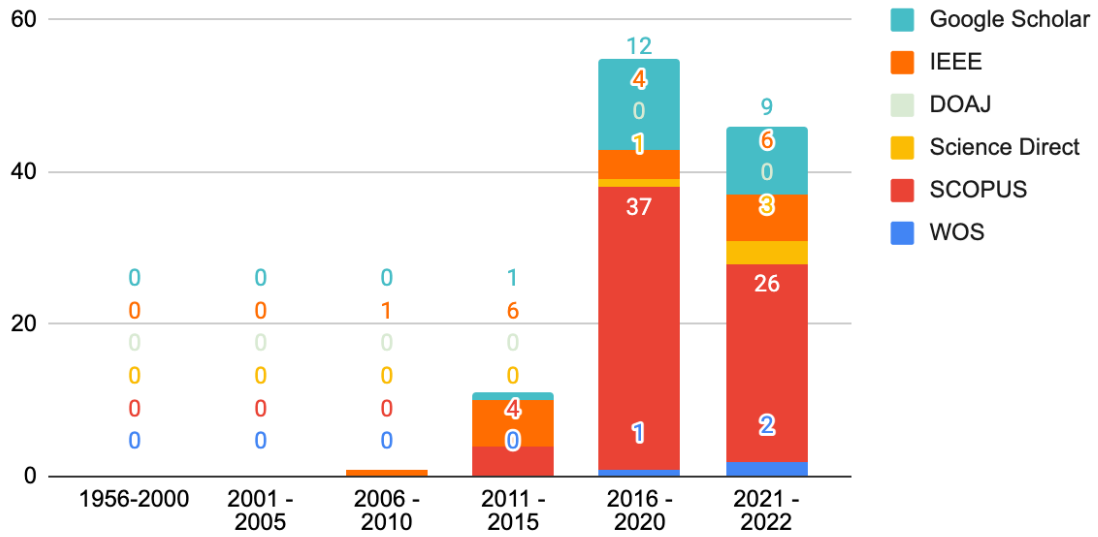
EDM'nin araştırma alanı 11 geniş alt alana bölünür ve öğrenci performans tahmini bunlardan yalnızca biridir[11]. Bu bağlamda yükseköğretim veri kümeleri ile sınıflandırma yapmak, öğrencilerin ve yükseköğretimin lehine olduğu için önemlidir. Tez çalışmasından önce sistematik literatür taraması yapıldı[10]. Şekil 1.3.'te EDM, ID ve MC disiplinleri ile ayrı ayrı yapılan çalışmaların, 2000'li yıllardan önce başladığı görüldü. EDM, ID ve MC disiplinler arası yapılan çalışmaların ise son yıllarda arttığı görüldü. Yapılan çalışmalardaki veri kümeleri incelendiğinde çoğunlukla, derse/bölüme kayıtlanmış öğrencilerin performans tahminini yaptığı görüldü. Henüz üniversite tercihinde bulunmamış ve üniversiteye yerleşmemiş aday öğrencilerin ders notu gibi akademik bilgileri yoktur. Bu şekilde henüz üniversiteye yerleşmediği ve hiç ders almadığı için ders notu olmayan bir öğrencinin, öğrenci performans tahminini yapan çalışma yok denecek kadar azdır. Bu durum bu çalışmanın ikinci motivasyonunu oluşturdu. Ayrıca yükseköğretim verilerinde görülen iki önemli problem vardır: çok sınıflı ve dengesiz veri kümesi problemi. Mevcut çalışmalar, öğrencilerin öğrenci performans tahminini geliştirmeye odaklansa da, MC ID eğitimsel veri kümesi (Multiclass imbalanced educational dataset)(MC-ID-ED-DS) ile öğrenci performans tahmini üzerinde yapılan çalışmalar Şekil 1.4'te görüldüğü üzere oldukça azdır.



Şekil 1.3. EDM, MC ve ID disiplinler arası yapılan çalışmaların yıllara göre dağılımı

Bu bağlamda, bu tez çalışmasındaki araştırma sorusu "Çok sınıflı ve dengesiz eğitimsel veri kümesi içeren makalelerde hangi makine öğrenmesi algoritmaları, yeniden örnekleme yöntemleri ve/ya metodolojileri kullanılmıştır, başarı ölçütleri nelerdir ve en başarılı

algoritmalar hangileridir?" şeklindedir. Araştırma sorusu için literatür taraması sırasında belirli anahtar kelimeler ile Marmara Üniversitesinin erişimi olan Web of Science (WoS), Science Direct, Scopus, Google Scholar, IEEE Xplore, DOAJ - Direct Open Access Journal ve Ulusal Tez Tarama[12] indexler tarandı.



Şekil 1.4. Bilimsel veri kaynakları ve EDM, MC ve ID disiplinler arası çalışmaların yıllara göre dağılımı

Belirtilen veri kaynaklarında öncelikle, "Öğrenci Performans Tahmini", "Karar Destek Sistemleri", "Makine Öğrenmesi", "Sınıflandırma Algoritmaları" şeklinde anahtar kelimeler ile tarama yapıldı. İlk tarama neticesinde, literatürde eğitim alanındaki araştırmaların EDM disipliniyle yapıldığı görüldü. Bu nedenle araştırma EDM anahtar kelimesi ile genişletildi. Bu şekilde genellikle ikili sınıf (binary-class) ve dengeli veri kümesi ile yapılan çalışmalarla karşılaşıldı. Bu nedenle anahtar kelime olarak çok sınıflı veri kümesi içeren çalışmalara ulaşmak için MC anahtar kelimeleri ile, dengesiz veri kümesi içeren çalışmalara ulaşmak için ise ID anahtar kelimeleri ile tarama yapıldı. Sonuç olarak, bu tez çalışmasında kullanılan arama tanımlayıcıları "Educational Data Mining", "Multi-class Classification" ve "Imbalanced Dataset" şeklindedir.

Yapılan sistematik literatür araştırmasında incelenen makalelerde oluşturulan sistem şu şekilde özetlenebilir: Geleneksel sınıflandırma metotları ve yeniden örnekleme işlemleri ile topluluk metotları karşılaştırmalı olarak incelenerek, kullanılan veri kümesinde en başarılı F1 Score bilgisini veren yöntem seçilmiştir. Bu çalışmalar Tablo 7'de belirtilen metot türlerine göre gruplandı.

Tablo 1.1. Sistematik literatür araştırmasında incelenen makalelerin gruplandırılması

Kodu	Makalelerde kullanılan metot türleri
SAT	Sınıf ayrıştırma teknikleri
TTY	Topluluk tabanlı yaklaşımlar
GSA	Geleneksel sınıflandırma algoritmaları
VSY & GSA	Veri seviyesinde yeniden örnekleme & Geleneksel sınıflandırma algoritmaları
VSY & TTY	Veri seviyesinde yeniden örnekleme & Topluluk tabanlı yaklaşımlar
TTY & GSA	Topluluk tabanlı yaklaşımlar & Geleneksel sınıflandırma algoritmaları
VSY & TTY & GSA	Veri seviyesinde yeniden örnekleme & Topluluk tabanlı yaklaşımlar & Geleneksel sınıflandırma algoritmaları

Bu çalışmalardan MC ID eğitimsel veri kümesi kullanan bazı çalışmalarla ilgili bilgiler aşağıdaki gibidir:

2018, Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification [13]

UCI Machine Learning Repository üzerindeki veri kümesi kullanıldı. Veri kümesi 403 örneklem ve 1 hedef değişken olmak üzere toplam 5 özellik içerir. Dengesiz bir veri kümesidir. Hedef değişkenler “very low” 50, “low” 129, “middle” 122 ve “high” 130 olmak üzere toplam 4 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & GSA metot türünden bir makale çalışmasıdır. SMOTE and OSS kombinasyonu ile çalışıldı. Support Vector Machine (SVM) makine öğrenme algoritması kullanıldı. Başarı metriği olarak accuracy, sensitivity, specificity, g-mean kullanıldı. Sunulan yeni yöntem başarılı oldu.

2021, Multiclass Prediction Model for Student Grade Prediction Using Machine Learning [14]

Malezya Politekniklerinden birinde Bilgi ve İletişim Teknolojileri (JTMK) Bölümüne ait veri kümesi kullanıldı. Veri kümesi 1282 örneklem ve 1 hedef değişken olmak üzere 10 özellik içerir. Dengesiz bir veri kümesidir. Hedef değişkenler "exceptional” 63, “excellent” 377, “distinction” 635, “pass” 186, “fail” 21 olmak üzere 5 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & GSA metot türünden bir makale çalışmasıdır. Random Forest (RF), Decision Tree (DT), SVM, Naïve Bayes (NB), K-Nearest Neighbor (KNN), Logistic Regression (LR) makine öğrenmesi algoritmaları kullanıldı. Başarı metriği

olarak, F measure kullanıldı. Yeniden örnekleme yöntemi olarak SMOTE kullanıldı. Ayrıca WrapperSubsetEval, ClassifierSubsetEval ve InfoGainAttributeEval olmak üzere 3 farklı Özellik Seçimi (Feature Selection FS) yöntemi kullanıldı. Python diliyle yazıldı. RF + SMOTE modeli en başarılı sonucu verdi.

2020, Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques [15]

İran ve Portekiz eğitim kurumları olmak üzere 2 farklı eğitim kurumundan veri kümesi kullanıldı. 1. veri kümesi 650 ve 2. veri kümesi 394 örneklem ve her iki veri kümesi 1 hedef değişken olmak üzere 19 özellik içerir. FS işlemleri elle yapıldı. Normalizasyon ve standardization yapıldı. Hedef değişkenler 1. veri kümesinde "Medium" 40%, "Good" 40% , "Poor" 11%, "Excellent" 9%" ve 2. veri kümesinde "Poor" 15%, "Medium" 44%, "Good" 35%, "Excellent" 6%" olmak üzere 4'er tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot türünden bir makale çalışmasıdır. eXtreme Gradient Boosting (XGBoost), RF, KNN, Artificial Neural Network (ANN), SVM, DT, LR, and NB makine öğrenmesi algoritmaları kullanıldı. Başarı metriği olarak, Accuracy, Recall, Precision, and F1-Score kullanıldı. Yeniden örnekleme yöntemi olarak Borderline SMOTE, Random Over Sampler (ROS), SMOTE, SMOTE-ENN, SVM-SMOTE ve SMOTE-Tomek kullanıldı. Random hold-out and Shuffle 5-fold cross-validation kullanıldı. Test olarak Friedman test, Anderson–Darling normality yapıldı. Python ile yazıldı. RF + SVM SMOTE modeli en başarılı sonucu verdi.

2021, Imbalanced Class handling and Classification on Educational Dataset [16]

Kaggle.com üzerinden açık kaynak Kalboard 360 LMS veri kümesi kullanıldı. Veri kümesinde 480 örneklem ve 1 tane hedef değişken olmak üzere 16 özellik içerir. Hedef değişkenler "Low" 127, "Medium" 211, "High" 142 olmak üzere 3 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot türünden bir makale çalışmasıdır. Stacking ensemble method, RF, LR, KNN, DT CART, SVM makine öğrenmesi algoritmaları kullanıldı. Başarı metriği olarak Accuracy, G-Mean kullanıldı. Yeniden örnekleme yöntemi olarak SMOTE, Borderline SMOTE, SMOTE - Tomek kullanıldı. Yeniden örnekleme yönteminin tüm dataya uygulandığı görüldü. 10-fold cross-validation kullanıldı. Python ile yazıldı. RF ve SMOTE-Tomek %86 F1 Score ile en başarılı model oldu. Smote ile RF ise %83 F1 Score başarısına sahip ikinci en başarılı model oldu.

2013, Imbalanced educational data classification: an effective approach with resampling and random forest [17]

Vietnam'daki Bilgisayar Bilimi ve Mühendisliği Fakültesi, Ho Chi Minh City Teknoloji Üniversitesine ait veri kümesi kullanıldı. Veri kümesi 1334 örneklem ve 1 hedef değişken olmak üzere 43 özellik içermektedir. Hedef değişkenler "very poor, poor, fair, nearly good, good, very good, excellent" olmak üzere 7 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot türünden bir makale çalışmasıdır. RF, Boosting, Bagging, NB, SVM, NN, LR, KNN, DT C4.5 makine öğrenmesi algoritmaları kullanıldı. Başarı metriği olarak Accuracy, ROC kullanıldı. 10 fold cross validation (CV) kullanıldı. Yeniden örnekleme yöntemi kullanılmadı. "Hibrit Yeniden Örnekleme Şeması ve RF" şeklinde ayrıca bir yöntem önerildi. Weka kullanıldı. RF en başarılı model oldu.

2021, Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine-Learning Classification Algorithm [18]

Universitas Terbuka'da, UT-LMS platformunda Moodle veri kümesi kullanıldı. Veri kümesi, 101617 örneklem ve 19 özellik içerir. RF ile FS yapıldı ve 3 tanesi önemli bulundu. Label Encoder and One Hot Encoding yapıldı. Hedef değişkenler, "Poor Moderate Good" olmak üzere 3 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot türünden bir makale çalışmasıdır. RF, Gradient Boosting (GBM), and Adaptive Boosting (AdaBoost) ve DT makine öğrenmesi algoritmaları kullanıldı. Yeniden örnekleme yöntemi olarak SMOTE kullanıldı. Veri kümesi Holdout %70 - %30 olarak bölündü. Başarı metriği olarak Accuracy kullanıldı. Python dili kullanıldı. SMOTE ve RF %85.03 accuracy başarı oranı ile en başarılı model oldu.

2021, Learning from Imbalanced Educational Data Using Ensemble Machine Learning Algorithms [19]

Martin Luther Christian, Meghalaya, Hindistan Üniversitesine ait veri kümesi kullanıldı. Veri kümesi, 497 örneklem ve 1 hedef değişken olmak üzere 18 özellik içerir. Hedef değişken çok sınıflı olmakla birlikte, makalede isimleri bulunamadı. TTY & GSA türünden bir makale çalışmasıdır. RF, Adaboost, stochastic gradient boosting (gbm), XGBoost, C4.5 makine öğrenmesi algoritmaları kullanıldı. Yeniden örnekleme yöntemleri kullanılmadı. Başarı metriği olarak, precision, specificity, recall, kappa metrics, balanced accuracy ve F-score kullanıldı. 10 fold CV ve holdout %70-%30 kullanıldı. R language ile yazıldı. RF ve Adaboost F-score

değeri 66.67%, balanced accuracy değeri of 75%, ve accuracy değeri 96.94% olarak en başarılı model oldu.

2021, Application of Machine Learning Techniques to the Prediction of Student Success [20]

Kaggle.com üzerinden açık kaynak Kalboard 360 LMS veri kümesi kullanıldı. Veri kümesinde 480 örneklem ve 1 tane hedef değişken olmak üzere 16 özellik içerir. Hedef değişkenler "Low = 127, Medium = 211, High = 142" olmak üzere 3 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & GSA metot türünden bir makine öğrenmesidir. Linear Discriminant Analysis, Logistic Regression, Classification and Regression Tree, K-Nearest Neighbour, Naive Bayes Classifier, and Support Vector Machines makine öğrenmesi algoritmaları kullanıldı. Yeniden örnekleme yöntemi olarak SMOTE kullanıldı. Yeniden örnekleme işlemi hem tüm dataya hem de sadece eğitim veri kümesine uygulandı. FS işlemleri için Information Gain kullanıldı. Başarı metriği olarak The confusion matrix, accuracy, precision, recall, and fl-score kullanıldı. 5 fold ve 10 fold CV kullanıldı. Linear Regression %86 accuracy ile en başarılı sonucu veren algoritma oldu.

2021, Oversampling technique in student performance classification from engineering course [21]

Rajamagala Teknoloji Üniversitesi, Tayland Üniversitesine ait veri kümesi kullanıldı. Veri kümesi 6882 örneklem ve 1 hedef değişken toplam 15 özellik içerir. Hedef değişkenler, “high” 860, “Medium” 5908, “Low” 114 olmak üzere 3 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot bir makine öğrenmesidir. GBM, AdaBoost, RF ve multilayer perceptron (MLP) makine öğrenmesi algoritmaları kullanıldı. Yeniden örnekleme yöntemlerinden SMOTE, Borderline-SMOTE, SVMSMOTE ve ADASYN kullanıldı. MLP + Borderline SMOTE modeli en başarılı sonucu verdi.

2021, Importance of Data Re-Sampling and Dimensionality Reduction in Predicting Students’ Succes [22]

Kaggle.com üzerinden açık kaynak Kalboard 360 LMS veri kümesi kullanıldı. Veri kümesinde 480 örneklem ve 1 tane hedef değişken olmak üzere 16 özellik içerir. Hedef değişkenler "Low” 127, “Medium” 211, “High” 142 olmak üzere 3 tanedir, bu nedenle çok sınıflı sınıflandırma problemine sahiptir. VSY & TTY metot türünden bir makine öğrenmesidir. GBM, SVM, KNN, CART, MLP, Linear Discriminant Analysis (LDA) makine öğrenmesi algoritmaları kullanıldı. Yeniden örnekleme yöntemlerinden RUS ROS SMOTE kullanıldı. Başarı ölçüm metriği

olarak, Confusion Matrix, Accuracy, (AUC), Kappa, Precision ve Recall kullanıldı. 10 fold CV ve holdout %70 - %30 kullanıldı. PCA kullanıldı. SVM SMOTE ve PCA uygulandıktan sonra elde edilen model %94 oranında accuracy değerini vererek en başarılı sonuç elde edildi.

Yukarıda özetlenen literatür araştırması neticesinde çalışmanın katkıları aşağıda listelenmektedir.

Çalışmanın Katkıları;

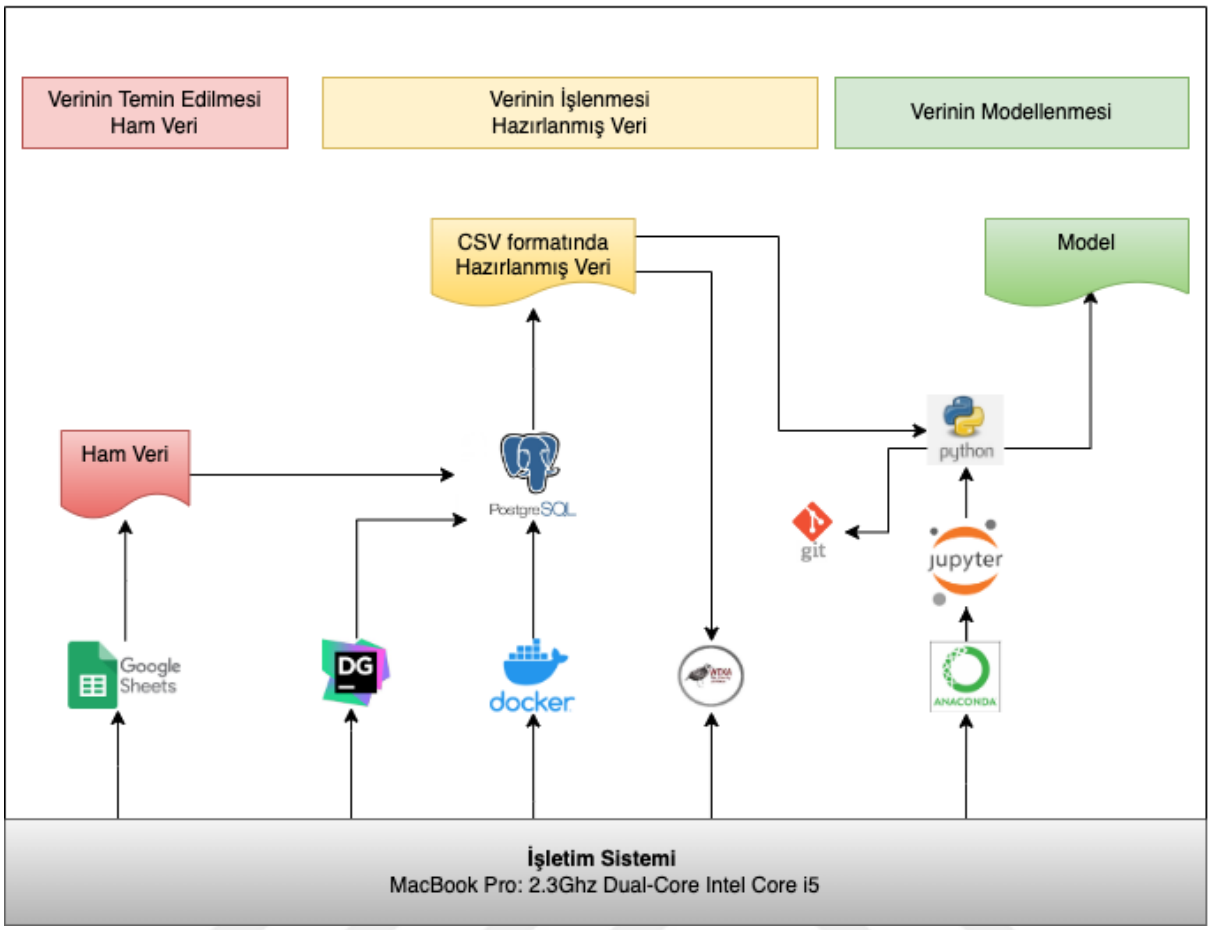
Bu çalışmadaki genel araştırma vurgusu, çoklu sınıf dengesizliğine sahip eğitim verisinin, en uygun makine öğrenmesi algoritması ve yeniden örnekleme yöntemiyle akademik performans tahminini yaparak, pilot bir karar destek sistemi web uygulamasıyla öğretim üyelerinin kullanımına sunmaktır. Böylece, üniversiteye başlamadan önce de bölüme kayıt yaptırdıkları takdirde, sınıfta kalma ihtimali, bölümü terk etme ihtimali olan riskli öğrenciler önceden tahmin edilerek, bu öğrencilerle ilgili iyileştirme çalışması yapılabilir. Literatürde yapılan çalışmalarda kullanılan veri kümeleri ayrı olduğu için genel olarak çalışmalar veri kümesine özgü olmaktadır. Hem çok sınıflı, hem de sınıf dengesizliğine sahip öğrenci verileri kullanarak yapılacak öğrenci performansını tahminlemek için bir yöntemin bulunması, EDM disiplinine yeni bir katkı sağlayabilir. Bu çalışmanın amaçları şu şekilde belirtilmiştir:

- Marmara Üniversitesi Teknoloji Fakültesi öğrenci verilerini kullanarak, öğrenci profilini çıkarmak,
- Bu öğrenci profilini çıkarırken, gelen öğrencilerin riskli olanlarını tespit etmek,
- Bu riskli öğrenciler tespit edildiği takdirde, bu öğrencilerin risklerini engelleyebilecek bir temel oluşturmak,
- Marmara Üniversitesi Teknoloji Fakültesi özelinde pilot bir karar destek sistemi web uygulaması yaparak, öğretim üyelerinin kullanımına sunmak
- Hem EDM, hem çok sınıflı, hem dengesiz bir veri kümesinin başarısını ölçümlemek için hibrit bir yöntem önermek,
- Yeniden örneklenmiş veri kümesi ve algoritmalar üzerinde bir karşılaştırma sunmak.

BÖLÜM 2 - MATERYAL VE YÖNTEM

Bu çalışmanın esasını, veri analizine dayalı makine öğrenmesi ile tahminleme modeli oluşturmak ve modeli kullanarak pilot bir web uygulaması geliştirmek oluşturur. Bunun için gerekli olan süreç verinin temin edilmesi, verinin işlenmesi ve makine öğrenmesi algoritması ile modellenmesidir. Bu bağlamda, her sürecin verimini artıracak yazılımsal ve donanımsal materyalleri kullanmak önemlidir. Şekil 2.1.'de gösterildiği gibi çalışma boyunca, donanım olarak **MacBook Pro: 2.3Ghz Dual-Core Intel Core i5** kullanıldı.

İlk olarak Şekil 2.1.'de gösterildiği gibi, temin edilen veri kümesi excel formatında olduğu için ham veri kümesiyle ilgili sayfa bilgileri, satır sayıları gibi ilk istatistikleri elde etmek için **Google Sheets** kullanıldı. **Docker**, tek işletim sistemi üzerinde, yüzlerce birbirinden bağımsız konteynerler sayesinde sanallaştırma sağlayan bir teknolojidir. Sürüm uyumsuzluklarını ortadan kaldırır. Aynı uygulamanın birden fazla farklı sürümlerinin aynı işletim sisteminde kullanımına olanak sağlar. Yönetmek ve uygulama kurulumları yapmak kolaydır. **PostgreSQL**, güçlü, açık kaynaklı bir nesne-ilişkisel veritabanı yönetim sistemidir. **DataGrip** ise bir çok farklı türde veritabanına bağlanarak, veritabanı içerisinde QRUD işlemlerinin yapılmasını sağlayan veritabanı ortamıdır[23]. Ayrıca SQL kodlarını düzenlemek için akıllı kodlama yardımı sağlayan editör içerdiği için oldukça kullanışlıdır. Bu şekilde Docker üzerinde PostgreSQL kurulumu yaparak, Google Sheets ile elde edilen ham veri, tablolar halinde PostgreSQL veritabanına aktarıldı. DataGrip ile, PostgreSQL veritabanına bağlanarak, veri ön işleme aşamasında kullanılan veri kümesini elde etmek için QRUD işlemleri yapıldı ve veri ön işleme aşamasında kullanılan CSV uzantılı veri kümesi elde edildi. **WEKA**, veri madenciliği işlemleri için önişleme (preprocess), sınıflandırma (classification), görselleştirme(visualize) gibi yöntemler kullanan, java ile geliştirilmiş açık kaynak bir programdır. Veri ön işleme aşamasında, veri görselleştirme, filtreleme, makine öğrenmesi algoritmalarının davranışlarını incelemek için WEKA kullanıldı[24]. **Python Anaconda**, Python dilini kullanarak veri bilimi, analiz ve makine öğrenmesi gibi bilimsel uygulamaların geliştirilmesini sağlayan python dağıtım sistemidir. İçerisinde 1500'den fazla açık kaynak paketi bulundurur ve aynı zamanda **"Jupyter Notebook"** geliştirici arayüzlerini de barındırır. Veri modellemesi için Jupyter Notebook geliştirici arayüzü kullanıldı[25]. Modelleme için yazılan yazılım, sürüm versiyon kontrol sistemi olan **GİT** üzerinde tutuldu ve Git repository'lerini yönetmek için **Sourcetree** uygulaması kullanıldı[26, 27]. Son olarak, Django frameworku kullanılarak, pilot bir karar destek sistemi web uygulaması geliştirildi.



Şekil 2.1. Tez çalışmasında kullanılan donanımsal ve yazılımsal sistem mimarisi

2.1. Veri Kümesi

Bu çalışmada, Marmara Üniversitesi(MÜ) Fen Bilimleri Enstitü Müdürlüğü (FBEM)'nün 23.05.2022-286783 etik no kararı ile MÜ Teknoloji Fakültesi(TF) mühendislik öğrencilerine ait veri kümesi kullanıldı.

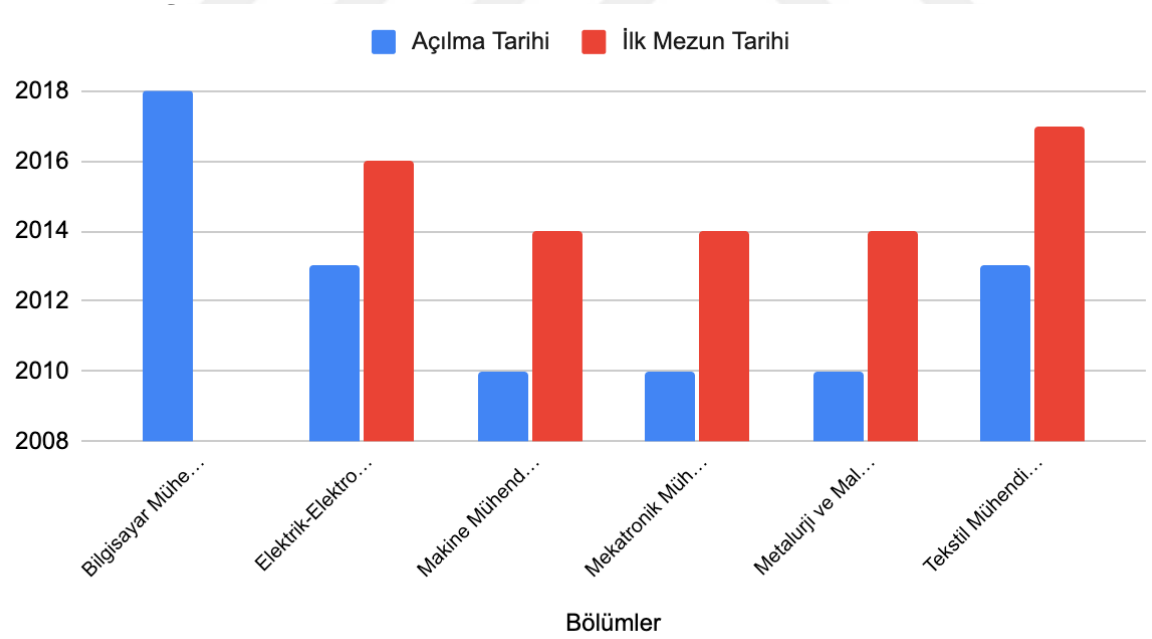
Veriler 3 ayrı excel listesi olarak alındı. Excel listelerinde sırasıyla, 13, 3 ve 11 sayfa vardır. Veri içeriğine bakılmaksızın sayfalar incelendiğinde toplam özellik sayısı 319 adettir. Her sayfanın veritabanındaki bir tabloya karşılık geldiği görüldü. Bazı sayfalar birbirinin devamı niteliğindeki aynı özelliklerden ve farklı verilerden oluşmaktadır. Bazı sayfalarda benzer içerikli özellikler bulunmaktadır. Bazı özelliklerde veri bulunmamaktadır. Bazı özelliklerde sadece tek bir değer bulunmaktadır. Ayrıca bazı özellikler tablolar arası ilişkileri gösteren sayısal id bilgilerini içermektedir. Excelde bulunan sayfa ve kolon isimleri ile satır sayıları ayrıca EK 1 Tablo 1'de gösterildi.

Genel olarak excel sayfaları incelendiğinde, öğrenciye ait genel bilgiler, ÖSYM puan bilgileri, lise puan bilgileri, üniversite ders puan bilgileri, transkript bilgileri, müfredat ders bilgileri, ders öğretim üyesi bilgileri şeklinde gruplandırılabilecek bilgiler içermektedir.

Veri kümesi üzerinde istatistiksel analizler yapıldığında Teknoloji Fakültesinde aşağıda belirtilen 6 bölüm vardır:

- Bilgisayar Mühendisliği
- Elektrik-Elektronik Mühendisliği
- Makine Mühendisliği
- Mekatronik Mühendisliği
- Metalurji ve Malzeme Mühendisliği
- Tekstil Mühendisliği

Şekil 2.2'de Mühendislik bölümlerinin ilk öğrenci kayıt yılı ve ilk öğrenci mezun yılları gösterilmiştir. Bilgisayar Mühendisliğinin 2021 Şubat ayında henüz mezun vermediği görülmektedir.

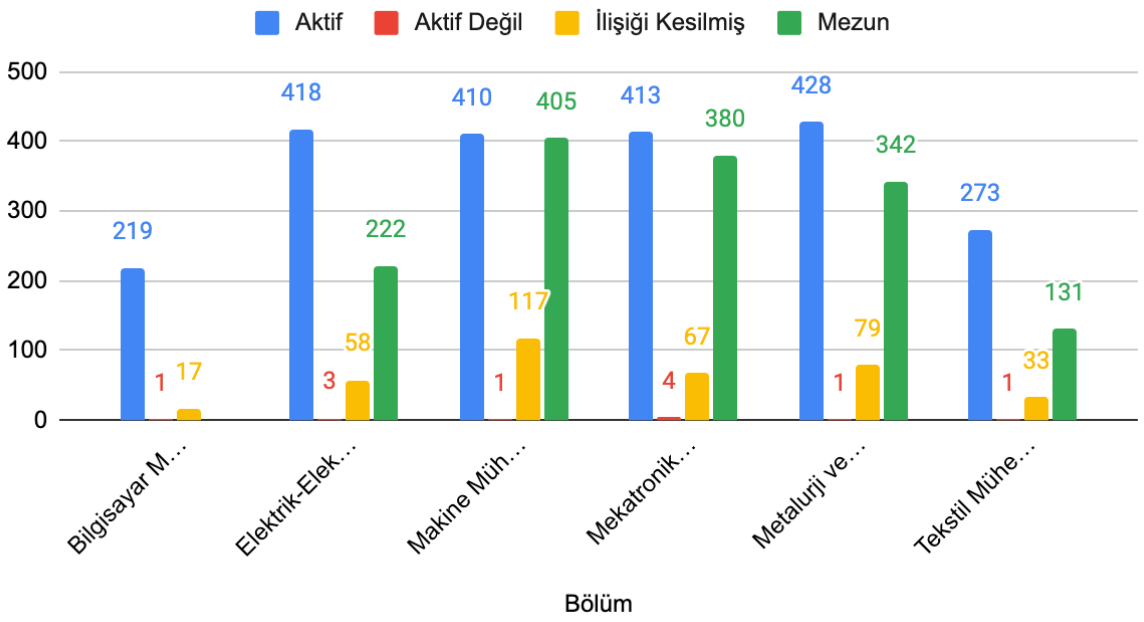


Şekil 2.2. MÜ TF 2021 Şubat ayına kadar bölümlere ait ilk öğrenci alma ve mezun yılları

Teknoloji Fakültesinde bulunan 6 bölümde toplam 4023 öğrenci vardır. Bu öğrencilerin mezuniyet durumlarını belirten toplam 4 durum vardır.

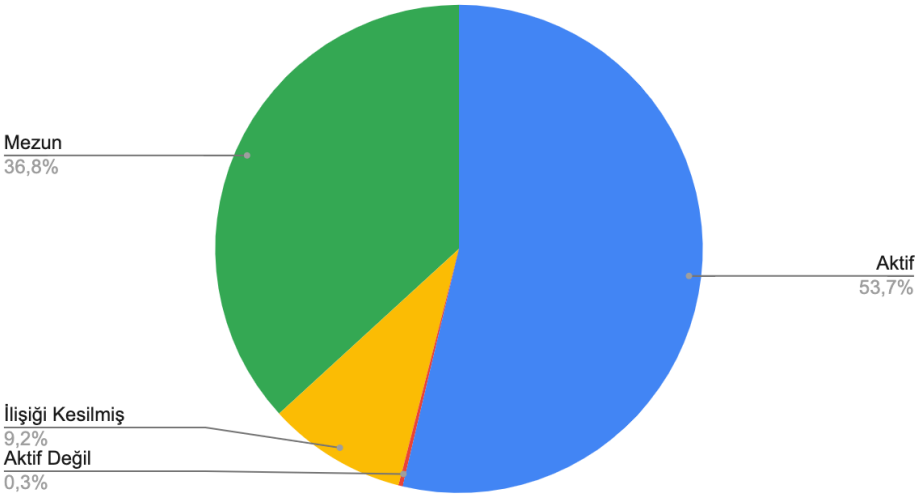
- Aktif
- Aktif Değil
- İlişiği Kesilmiş
- Mezun

Bölümlere ait öğrenci mezuniyet durum dağılımları Şekil 2.3'de gösterildiği gibidir. Ayrıca öğrencilerin mezuniyet durumları, kendi içerisinde de alt durumlara ayrılmaktadır. Örneğin "Aktif Değil" durumu "Askere Gitmek İstediğinden" ve "Kayıt Dondurma" şeklinde alt durumlara ayrılmıştır. Bütün durumlarla ilgili alt durumlar EK 2'de gösterilmiştir.



Şekil 2.3. MÜ TF bölümlerine ait öğrenci mezuniyet durum dağılımları

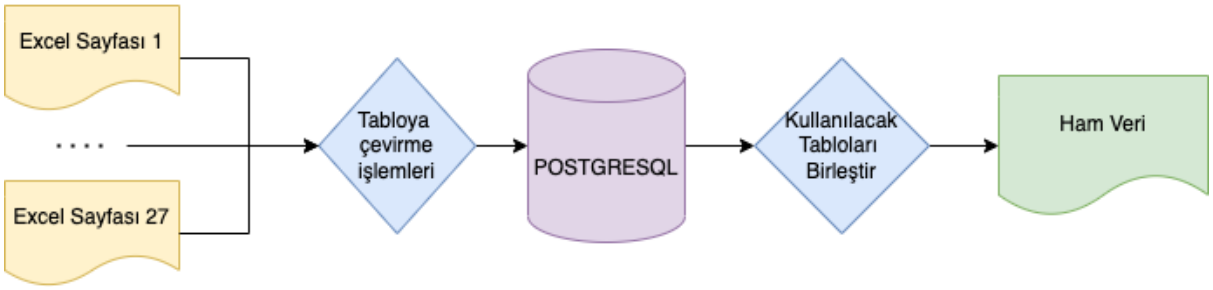
Şekil 2.3 incelendiğinde, "Mezun" durumda olan öğrenci sayısı 1480 iken, "İlişiği Kesilmiş" durumda olan öğrenci sayısı 371 ve "Aktif Değil" durumda olan öğrenci sayısı sadece 11'dir. Toplam 4023 öğrenciden sadece %9.2 si "İlişiği Kesilmiş" durumda olan öğrencilerdir. Bu durumda mezuniyet durum dağılımları arasındaki farkın çok olduğu görülmektedir. Bu nedenle Şekil 2.4'te de görüldüğü gibi, bu veri kümesinin dengesiz bir veri kümesi olduğu söylenebilir.



Şekil 2.4. Ham veri kümesi öğrenci mezuniyet durum dağılımları

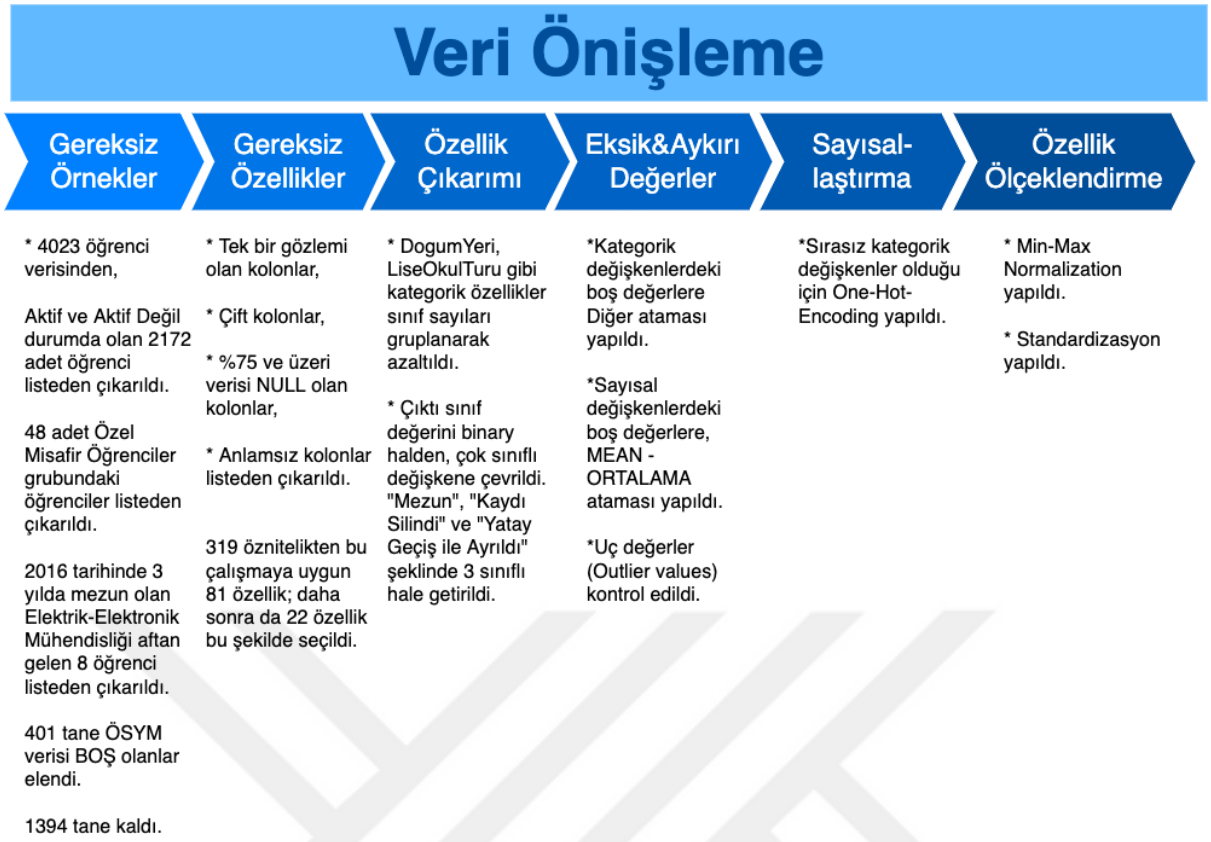
2.2. Veri Ön İşleme

MÜ TF öğrenci verileri excel listesi olarak teslim alındı. Bu nedenle ilk olarak Ek 1 Tablo 1'de gösterilen 27 tane excel sayfası, Şekil 2.5'deki gibi Docker üzerinde koşan, PostgreSQL veri tabanında 27 ayrı tabloya çevrildi.



Şekil 2.5. MÜ'den alınan verinin işlenme süreci

Bu çalışmanın amacı henüz üniversiteye başlamamış öğrencilerin, yükseköğretime giriş aşamasında, bölüme kayıt yaptırdıkları takdirde mezuniyet durumlarını tespit ederek, sınıfta kalma ihtimali, bölümü terk etme ihtimali olan riskli öğrencilerin tespit edilmesini sağlamaktır. Bundan dolayı veri kümesinde mezuniyet durumu bilinen öğrenciler bu çalışmada kullanıldı. Bu nedenle Şekil 2.6'da belirtilen "Gereksiz Örnekler" aşamasında, "Aktif" ile "Aktif Değil" durumda olan, toplam 2172 öğrenci veri listesinden çıkarıldı. Elektrik-Elektronik Mühendisliği bölümüne aftan gelerek 3 yılda 2016 tarihinde mezun olan 8 öğrenci listeden çıkarıldı. ÖSYM yerleştirme puan sırası, ÖSYM tercih sırası alanları önemli olduğu için, ÖSYM Bilgileri BOŞ olan 401 öğrenci listeden çıkarıldı.



Şekil 2.6. Veri önişleme aşamaları ve yapılan işlemler

Bu çalışmada kullanılacak olan "Mezun" ve "İlişiği Kesilmiş" öğrenci mezuniyet durumlarına ait Ek 2'de belirtilen alt durumlar incelendi. "İlişiği Kesilmiş" mezuniyet durumunda birbirine benzeyen 19 alt durumunun olduğu görüldü. Bu nedenle "İlişiği Kesilmiş" öğrenci durumu kendi içerisinde aşağıdaki şekilde gruplandırıldı ve tez çalışmasının kapsamında olmayan "Özel / Misafir Öğrenciliği Sona Eren" durumundaki 48 öğrenci listeden çıkarıldı:

- İlişiği Kesilmiş
    - Kaydı silindi
    - Özel / Misafir Öğrenciliği Sona Eren
    - Yatay Geçiş / Program Değişikliği

Son durumda, hedef değişken olarak kullanılan özellik "Mezun" ve "İlişiği Kesilmiş" sınıflarını içeren ikili sınıf (binary class) yapısından; "Mezun", "Kaydı silindi" ve "Yatay Geçiş / Program Değişikliği" sınıflarını içeren çok sınıflı(multiclass) yapı haline getirildi.

Böylece, Şekil 2.6'da belirtilen "Gereksiz Özellikler" aşamasında, veri azaltılarak ve "Özellik Çıkarımı" aşamasında da hedef değişkenler gruplanarak ve genişletilerek, 4023 öğrenciden,

1158 tane “Mezun”, 110 tane “Kaydı Silindi” ve 126 tane “Yatay Geçiş / Program Değişikliği” mezuniyet durumlarını içeren toplam 1394 öğrenci Tablo 2.1.'de görüldüğü gibi çalışmada kullanılmak üzere elde edildi.

Tablo 2.1. Veri ön işlem sonrası kalan öğrenci mezuniyet sayıları

Mezuniyet Durumu Hedef Değişken	Toplam Öğrenci Sayısı
Mezun	1158
Kaydı Silindi	110
Yatay Geçiş / Program Değişikliği	126
Toplam	1394

2.3. Özellik Seçimi

Ham veri kümesinde Ek 1 Tablo 1’de belirtildiği gibi toplam 27 tablo ve 319 özellik bulunmaktadır. Bu tablolardan, bu çalışmanın amacına uygun olarak kullanılacak verileri içeren Öğrenci Bilgileri, Öğrenci ÖSYM Bilgileri, ÖSYM Detay, ÖSYM Lise, ÖSYM Lise 2014, Öğrenci Bilgiler Kalanlar ve İl İlçe tabloları Şekil 2.5’te belirtildiği şekilde birleştirildi. Bu tablolar birleştirildiğinde elde edilen özellik sayısı, 319’den 81’e indi.

Elde edilen 81 özellikten, Şekil 2.6’da belirtilen “Gereksiz Özellikler” aşamasında,

- Tek bir gözlemi olan özellikler,
- Çift özellikler,
- Anlamsız özellikler,
- %75 ve üzeri verisi boş olan özellikler elenerek özellik sayısı 81’den 20'e indirildi.

Şekil 2.6’daki “Gereksiz Özellikler” aşamasından sonra elde edilen özellikler şunlardır: BirimAdi, Cinsiyet, DogumYeriIl, DogumTarihi, KayitYili, Sinif, EgitimYariYili, Durum, MezuniyetTarihi, MezunGANO, OgrenciDurumKodID, OsymKayitYili, LiseNotu, osymOkulTuru, YerPuaniSirasi, YerBolumTercihSirasi, MtokOgrencisi, osymOkulBirincisi, KontenjanTuru, UniversiteYerlestirmeSirasi. Bu özelliklere ait korelasyon, histogram bilgileri ve istatistikî bilgiler Ek 3’tedir.

Yüksek korelasyona sahip özellikler olduğu görüldü. Bunlardan DogumTarihi, KayitYili, MezuniyetTarihi, OsymKayitYili ile; Sınıf, MezuniyetTarihi ve EğitimYarıyılı ile,

MtokOgrencisi, OsymKayitYili ile gibi özellikler birbiriyle korelasyonu yüksek çıktığı için birer tanesi bırakılarak diğer özellikler kaldırıldı.

Şekil 2.6'daki "Özellik Çıkarımı" işlemi girdi değişkenler için de uygulandı. DoğumYeriIl özelliği 82 farklı değerden bölgelere göre gruplanarak Tablo 2.3'deki "Doğum Yeri Bölge" özelliğine dönüştürüldü. "osymOkulTuru" özelliği 30 farklı değerden, türlere göre gruplanarak Tablo 2.3'deki "Lise Okul Türü" özelliğine dönüştürüldü.

Lise Okul Türü alanında, verilerden boş ve tutarsız olanlar, Excel listelerinde tekrar incelenerek ve Ösym Okul Türleri kılavuzuyla karşılaştırılarak, doğru olanlar ile güncellendi[28]. Şekil 2.6.'deki "Eksik-Aykırı Değerler" aşamasında kategorik değişkenlerden boş olanlar "Diğer" değeriyle değiştirildi. LiseNotu özelliklerinde aykırı (outliers) olan değerler görüldü. LiseNotu özniteliği 142 ile 500 aralığında iken, 1 tane 999.999, 0 ve 100.000 değerleri tespit edildi ve bunlar BOŞ olarak güncellendi. Böylece veri kümesindeki aykırı değerler temizlendi. Daha sonra tüm sayısal özelliklerin kendi içlerinde ortalaması (mean) alınarak boş olan değerler ortalama değerle güncellendi ve sayısal değerler normalize edilerek 0-1 aralığına indirildi. Min-Max Scaler ve Normalizasyon ayrı ayrı veri kümesine uygulandığında, Min-Max Scaler daha başarılı sonuçlar verdiği için Min-Max Scaler yöntemi ile normalizasyon işlemleri yapıldı.

Son durumda, Şekil 2.6'daki veri ön işleme aşamaları tamamlandığında 22 özellikten bu çalışmada kullanılacak 11 özellik elde edildi.

Tablo 2.2. Tez çalışmasında kullanılan özelliklerin veri türleri

	Öznitelik	Toplam	Boş Var mı	Dtype	Değişken
1	Birim Adı	1394	Hayır	Kategorik	Girdi
2	Cinsiyet	1394	Hayır	İkili(Boolean)	Girdi
3	Doğum Yeri Bölge	1394	Hayır	Kategorik	Girdi
4	Lise Notu	1394	Hayır	Sayısal	Girdi
5	Lise Okul Turu	1394	Hayır	Kategorik	Girdi
6	ÖSYM Yerleştirme Sırası	1394	Hayır	Sayısal	Girdi
7	Tercih Sırası	1394	Hayır	Sayısal	Girdi
8	Üniversite Yerleştirme Sırası	1394	Hayır	Sayısal	Girdi
9	MtokOgrencisi	1394	Hayır	İkili(Boolean)	Girdi
10	Okul Birincisi mi	1394	Hayır	İkili(Boolean)	Girdi
11	Mezun Durum	1394	Hayır	Kategorik	Çıktı
12	Kayıt Yılı	1394	Hayır	Sayısal	Girdi

Sonuç olarak, bu çalışmada model geliştirmek için 4023 öğrenci ve 319 özellikten, toplam 1394 öğrenci ve 11 özellik kullanıldı. Kullanılan özelliklerin içerikleri Tablo 2.3.'te belirtilmiştir.

Tablo 2.3. Tez çalışmasında kullanılan özelliklerin değerleri ve açıklamaları

	Değerler	Toplam Öğrenci Sayısı	Açıklama
1	● Bilgisayar Müh ● Elektrik-Elektronik Müh ● Makine Müh ● Mekatronik Müh ● Metalurji ve Malzeme Müh ● Tekstil Müh	14 178 365 355 345 137	Öğrencilerin tercih ettiği bölümlerdir. Bilgisayar Mühendisliği mezun vermediği için listeden çıkarılmıştır.
2	● Kadın: 0 ● Erkek: 1	316 1078	Cinsiyet alanı. Öğrencilerin cinsiyetleri.

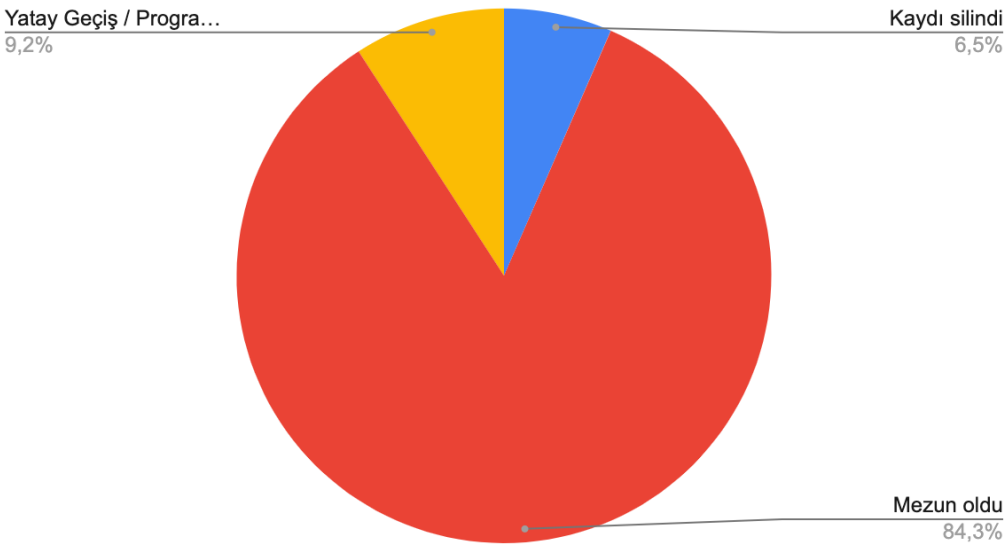
(Tablo 2.3.'ün devamı diğer sayfadadır.)

Tablo 2.3 (devam). Tez çalışmasında kullanılan özelliklerin değerleri ve açıklamaları

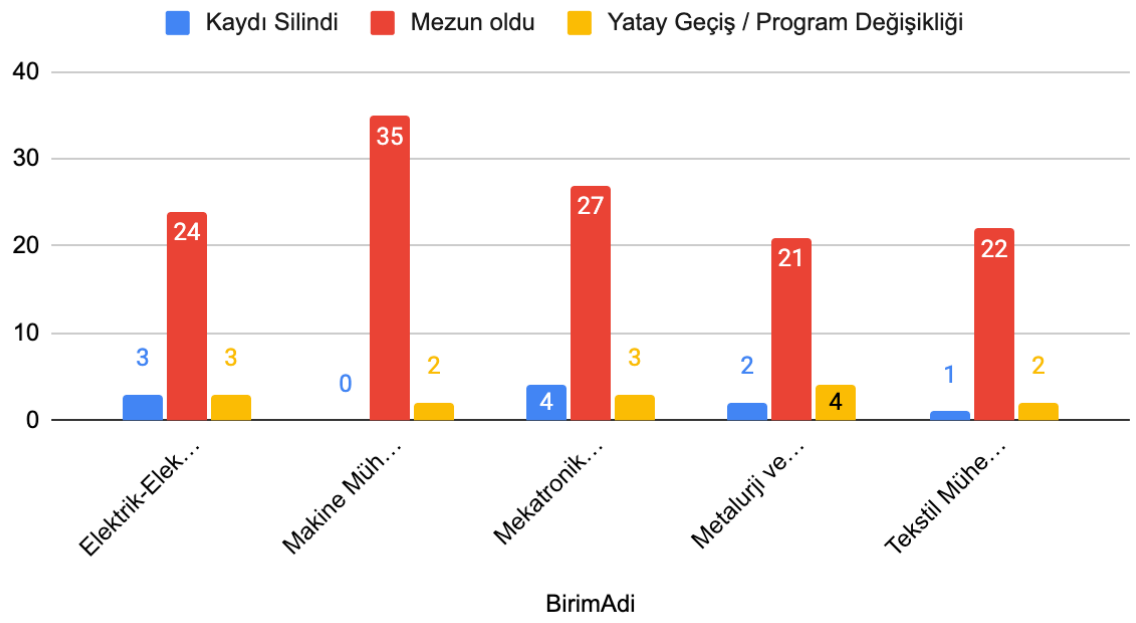
	Değerler	Toplam Öğrenci Sayısı	Açıklama
3	● Akdeniz ● Doğu Anadolu ● Ege ● Güneydoğu Anadolu ● İç Anadolu ● Karadeniz ● Marmara	89 185 91 70 194 393 372	Öğrencilerin doğum yerleri, bölgelere göre gruplandırılarak yeni özellik oluşturuldu.
4	142 <= L.N. <= 500	1394	Puan aralığıdır. Normalize edilerek 0 - 1 aralığına indirildi.
5	● Anadolu Lisesi ● Anadolu Meslek Lisesi ● Anadolu Öğretmen Lisesi ● Anadolu Teknik Lisesi ● Fen Lisesi ● İmamhatip Lisesi ● Meslek Lisesi ● Normal Lise ● Özel Lise ● Teknik Lise	533 72 87 242 59 11 118 111 69 92	Öğrencilerin kayıtlanma şekilleri 39 farklı türdedir, gruplandırılarak yeni özellik oluşturuldu.
6	15141 <= Y.P.S. <= 413695	1394	Puan sıralama. Normalize edilerek 0 - 1 aralığına indirildi.
7	1 <= T.S. <=28	1394	Tercih sırası. Normalize edilerek 0 - 1 aralığına indirildi.
8	1 <= UYS <= 56	1394	Üniversite yerleştirme sırası. Normalize edilerek 0- 1 aralığına indirildi.
9	● Evet: 1 ● Hayır: 0	474 920	MTOK öğrencisi mi?
10	● Evet: 1 ● Hayır: 0	55 1339	Okul Birincisi mi? Öğrencilerin liseden birincilikle mezun olduğunu gösteren özellik
11	● Kaydı silindi ● Mezun oldu ● Yatay Geçiş / Program Değişikliği	110 1158 126	Öğrencilerin mezuniyet durumlarını gösteren hedef değişken

2.4. Verinin Eğitim ve Sağlamlık Veri Kümesine Bölünmesi

Ön işleme aşamalarından sonra, 1394 satır 11 özellik içeren temizlenmiş (prepared) veri kümesi elde edildi. Makine öğrenmesi algoritmaları ile elde edilen son modellerin sağlamlığını kontrol etmek için, 2016 yılında kayıt yapan öğrenciler kullanıldı. 2016 yılına ait öğrenci bilgilerinin sağlamlık verisi olarak kullanılma sebebi, MÜ TF bölümlerinin aynı yılda açılmaması ve aynı yılda mezun vermeye başlamaması neticesinde her yıl bütün bölümlerin mezun vermemiş olmasıdır. 2016 yılında Şekil 2.2'de görüldüğü gibi Bilgisayar Mühendisliği bölümü hariç tüm bölümler mezun verdi. Ayrıca Şekil 2.7'de görüldüğü üzere her sınıftan veri içerdiği için 2016 yılında kayıt yapan 153 öğrenci, temizlenmiş veri listesinden sağlamlık verisi olarak ayrıldı. Bu bağlamda, 2016 yılında kayıt yapan öğrencilerin hiçbirisi, eğitim aşamasında modeller tarafından görülmemiş verilerdir. Ortak yıl olarak 2016 tarihinde kayıt yapan öğrencilerin 2021 yılındaki mezuniyet durumlarıyla ilgili bölüm bazlı sınıf bilgileri Şekil 2.8.'dedir.

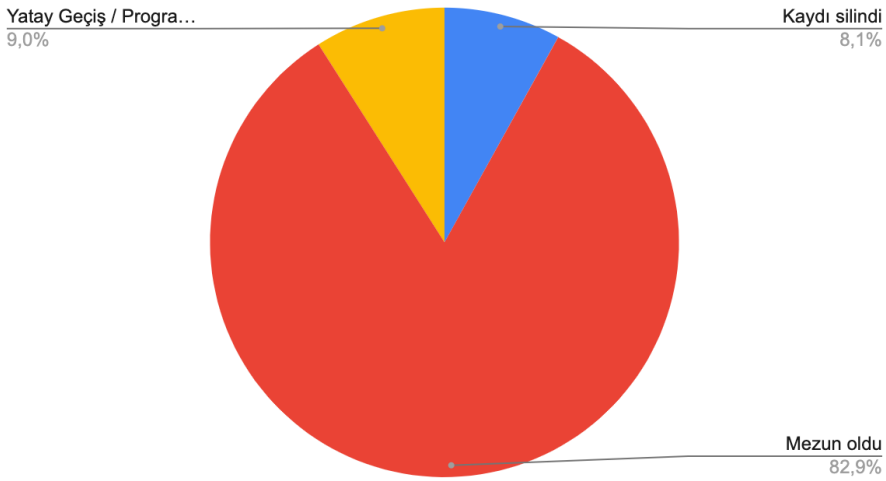


Şekil 2.7. Sağlamlık verisi hedef özelliğine ait durum dağılımları

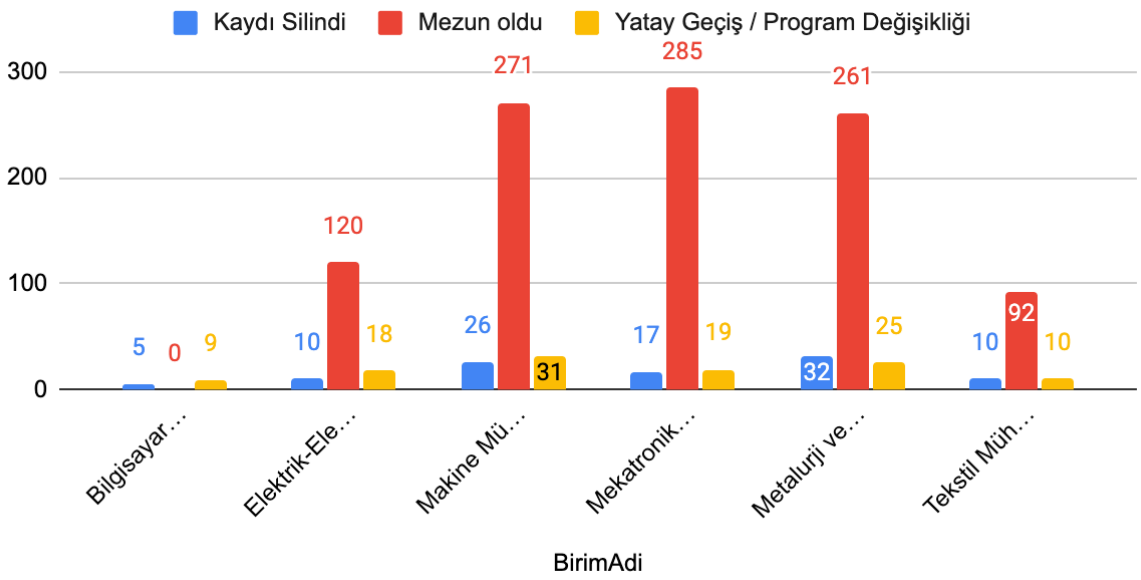


Şekil 2.8. 2016 yılında kayıt olan öğrencilerin mezuniyet durumları

Veri kümesinde 2016 yılı dışında kayıt yapan ve 1029 tane "Mezun", 100 tane "Kaydı Silindi" ve 112 tane "Yatay Geçiş / Program Değişikliği" mezuniyet durumlarını içeren toplam 1241 öğrenci algoritmaları eğitmek ve geçerleme yapmak üzere kullanıldı. Şekil 2.9.'da belirtilen oranlarda, eğitim için kullanılan veri kümesindeki sınıf dengesizliği ve çok sınıflı sınıflandırma problemleri görülmektedir. Ayrıca Şekil 2.10.'da bölümlere göre eğitim veri kümesinin öğrenci mezuniyet durum bilgileri görülmektedir.



Şekil 2.9. Temizlenmiş veri kümesinde mezuniyet durum dağılımları



Şekil 2.10. Temizlenmiş veri kümesinde, bölümlere göre durum dağılımları

2.5. Yöntem

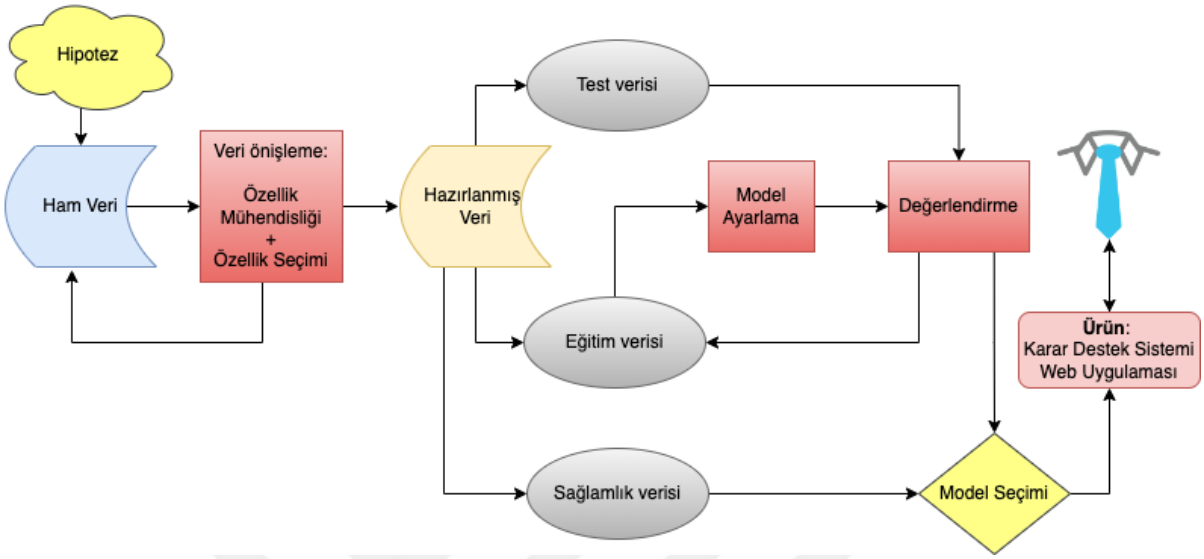
Şekil 2.11'de görüldüğü gibi veri önişleme ve özellik seçimi tamamlandıktan sonra, hazırlanmış veri sağlamlık verisi, eğitim verisi ve test verisi olarak ayrıldı ve makine öğrenme algoritmaları eğitilip, model geliştirildi ve son modellerden pilot bir karar destek sistemi web uygulaması yapıldı. Önişleme ve özellik seçiminden sonra elde edilen hazırlanmış veriden 2016 yılına ait öğrenci bilgileri sağlamlık kontrolünde kullanılmak üzere ayrıldı. Böylece, sağlamlık kontrolünde her sınıftan örneklem alınması sağlandı. Daha sonra, 2016 yılı hariç olan eğitim verisi, ayrıntısı çapraz geçerleme bölümünde anlatılan ve sınıf dengesizliği probleminde kullanımı önerilen Tekrarlanan Tabakalı 5*5 kat çapraz geçerleme (Repeated Stratified 5*5 fold cross validation) yöntemi ile eğitildi. Böylece eğitim kümesinden elde edilen her örneklemde, her sınıftan örnek olması sağlandı. Döngü tamamlanana kadar eğitim verisine, sırasıyla makine öğrenmesi algoritmaları uygulandı ve modeller oluşturuldu.

3 ayrı yöntem ile eğitim yaparak modeller oluşturuldu:

1. yöntemde, 11 özellik ile algoritmalar yeniden örnekleme(resampling) yapılmadan ve hiper parametreler her algoritmaya özel belirlenerek modeller oluşturuldu.

2. yöntemde, 11 özellik ile sadece eğitim verisine yeniden örnekleme yöntemleri uygulanarak ve hiper parametreler her algoritmaya özel belirlenerek modeller oluşturuldu.

3. yöntemde, 11 özellik ile sağlamlık verisi hariç tüm veri kümesine yeniden örnekleme yöntemleri uygulanarak ve hiper parametreler her algoritmaya özel belirlenerek modeller oluşturuldu.



Şekil 2.11. Tez çalışması makine öğrenmesi genel akış şeması

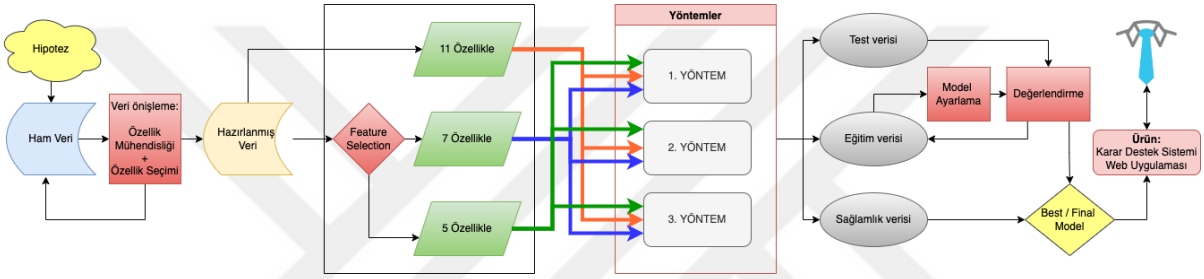
Her 3 yöntemde de her döngüde oluşan modelin F1 Score, Recall, Accuracy, ROC Area, confusion matrix gibi performans ölçüm değerleri elde edildi. Ek olarak yeniden örnekleme sonucu oluşan veri kümesi boyutu ve son olarak yürütme zamanı (execution time) bilgileri kaydedildi. Ayrıca sağlamlık kontrolünde kullanmak için, eğitim sonucunda oluşan her model, .pkl dosya formatında, dosya sistemine kayıt edildi. Repeated Stratified 5*5 fold-CV kullanıldığı için her algoritma için toplam 25 farklı model ve sonuçlar elde edildi. Döngü tamamlandıktan sonra her 3 yöntemde de, yeniden örnekleme + makine öğrenmesi algoritmalarına göre elde edilen 25 modele ait tüm bilgilerin ortalama değerleri alındı.

Bu aşamada 1. yöntemde 13 makine öğrenmesi algoritması ve 3 modifiye edilmiş dengeli topluluk metotları (balanced ensemble) olmak üzere 16 ayrı model, 2. yöntemde, 13 makine öğrenmesi algoritması * 9 yeniden örnekleme = 117 ayrı model, ve 3. yöntemde, 13 makine öğrenmesi algoritması * 9 yeniden örnekleme = 117 ayrı model olmak üzere toplam 250 farklı model elde edildi.

Daha sonra 250 ayrı model üzerinde sağlamlık testi uygulandı. Sağlamlık kontrolü, elde edilen 250 modele ayrı ayrı 153 tane satır içeren sağlamlık verisi verilerek yapıldı. Bu nedenle her model için ayrıca 250 tane ayrı sağlamlık sonuçları elde edildi. Her modelin test F1 Score değerleri ve sağlamlık F1 Score değerleri karşılaştırıldı ve tutarlılık oranları elde edildi. Şekil

2.13'de 1. ve 2. yönteme ait akış şemasının ayrıntılı modellenmesi yapıldı. Şekil 2.14'de 3. yönteme ait akış şemasının ayrıntılı modellemesi yapıldı[29].

Ek olarak, FS aşamasında elde edilen 1241 satır ve 11 özellik içeren hazırlanmış veri kümesi, Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF, Anova F Test şeklindeki FS yöntemleri kullanılarak tekrar özellik seçimi yapıldı. 7 FS yöntemine göre özelliklerin önem sırası karşılaştırıldı. Elde edilen sonuçlara göre, 1., 2. ve 3. yöntem, 11 tane özellik içeren veri kümesi haricinde, 7 tane en çok kullanılan özellik ve 5 tane en çok kullanılan özellik ile çalıştırılarak ayrıca 250'şer tane model oluşturuldu. Bu modellere de ayrıca sağlamlık kontrolü yapıldı ve 250'şer tane sağlamlık sonucu elde edildi. Ve ek olarak modellerin test F1 score değerleri ile karşılaştırılarak, tutarlılık oranları elde edildi.

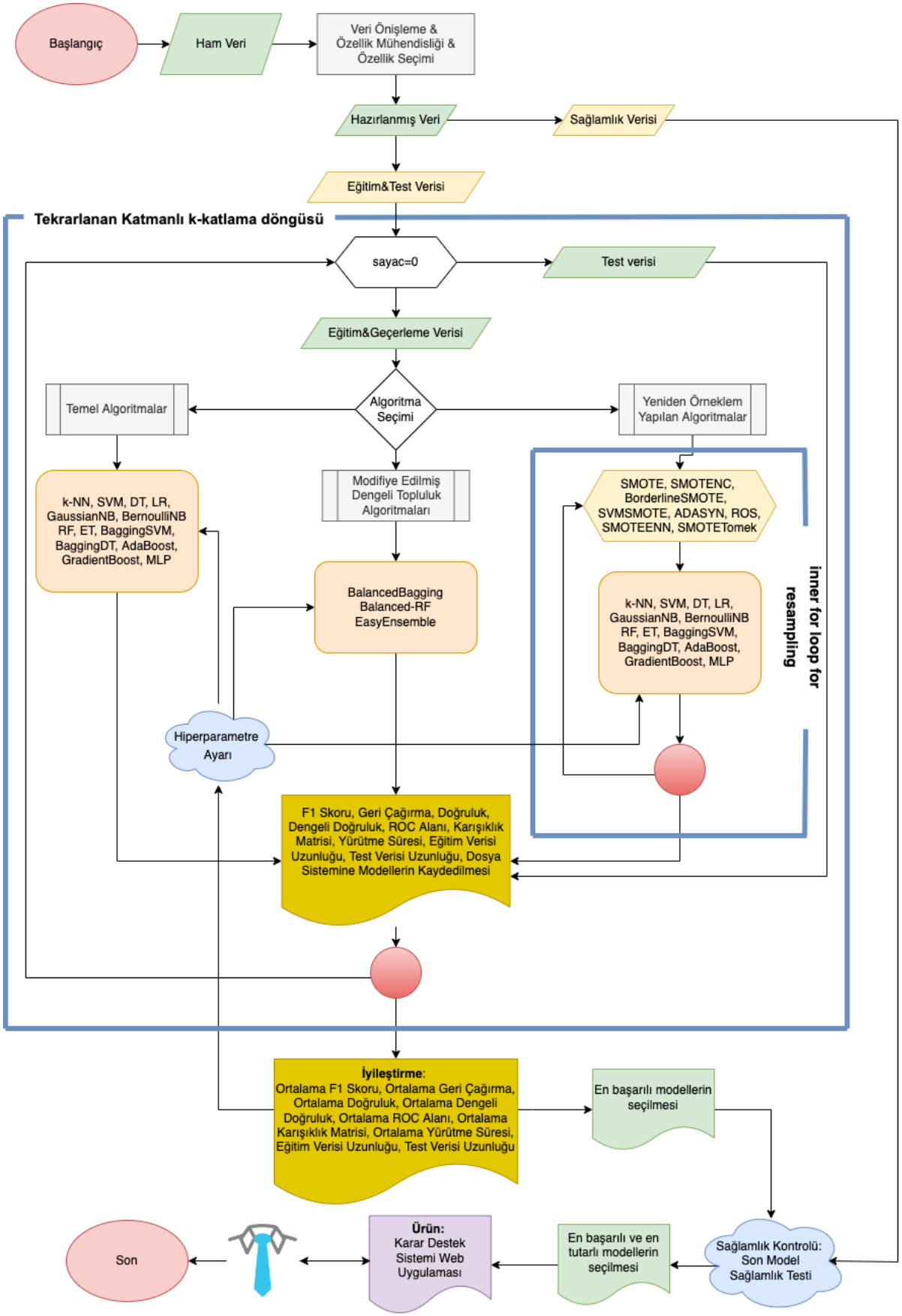


Şekil 2.12. Tez çalışması makine öğrenmesi genel akış şeması 2

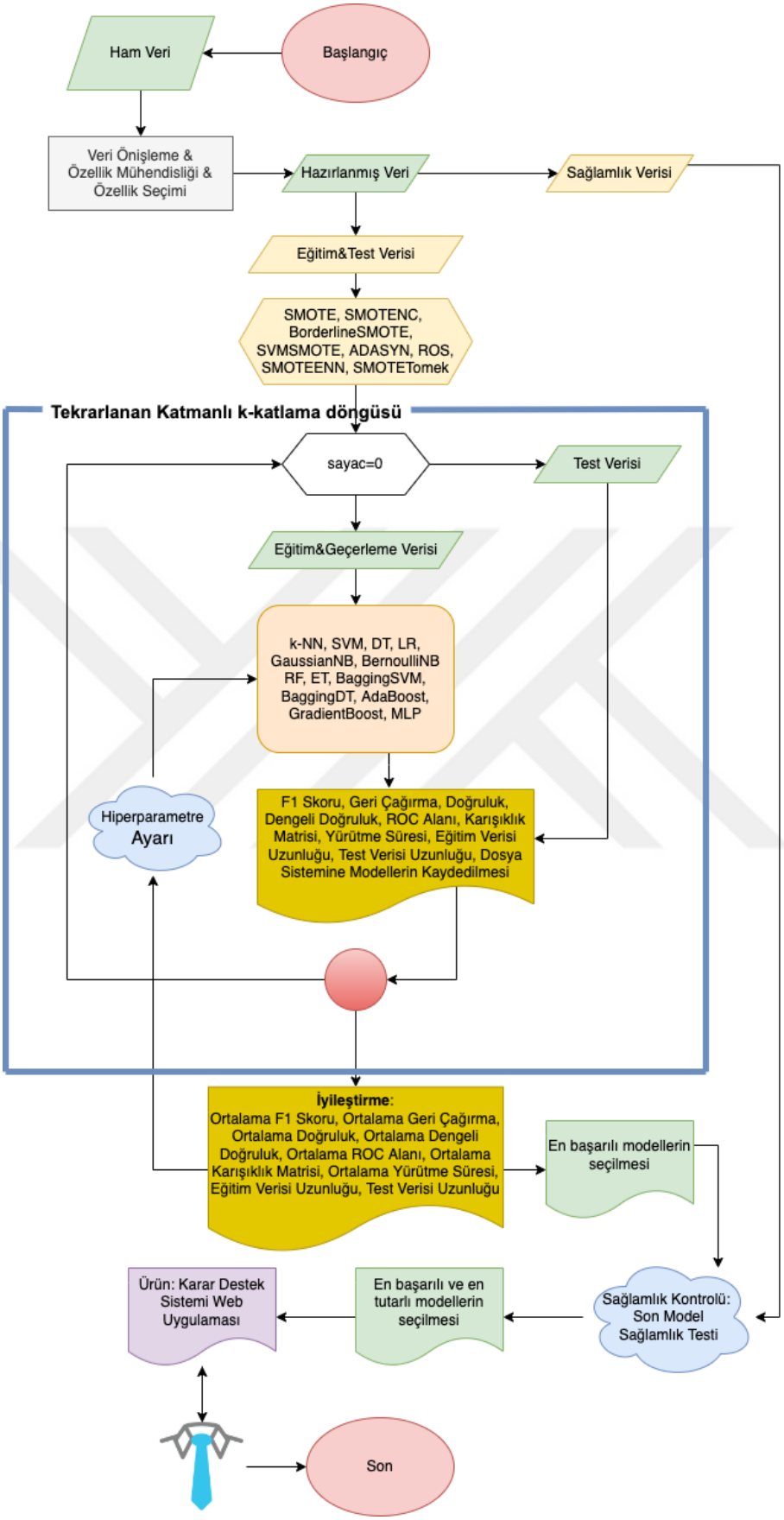
Böylece Tablo 2.4.'te gösterildiği gibi, 3 farklı yöntem ve 3 farklı özellik seçimi ile toplam 9 çeşit model türü içeren 750 model elde edildi. Son olarak, 11 özellikli 250 model ve sağlamlık verisi, 7 özellikli 250 model ve sağlamlık verisi, 5 özellikli 250 model ve sağlamlık verisi karşılaştırıldı. En başarılı ve en tutarlı ilk 4 model seçildi ve Django-Python frameworku kullanılarak, pilot bir karar destek sistemi web uygulaması yapıldı.

Tablo 2.4. 1., 2. ve 3. yöntem sonucunda oluşan model sayıları

	11 özellikle		**7 özellikle**		**5 özellikle**	
	Model Sayısı	**Sağlamlık kontrolü yapılan model sayısı**	**Model Sayısı**	**Sağlamlık kontrolü yapılan model sayısı**	**Model Sayısı**	**Sağlamlık kontrolü yapılan model sayısı**
1. yöntem	16	16	16	16	16	16
2. yöntem	117	117	117	117	117	117
3. yöntem	117	117	117	117	117	117
Toplam	**250**	**250**	**250**	**250**	**250**	**250**



Şekil 2.13. 1. ve 2. yönteme ait akış şeması



Şekil 2.14. 3. yönteme ait akış şeması

2.5.1. Makine öğrenmesi algoritmaları

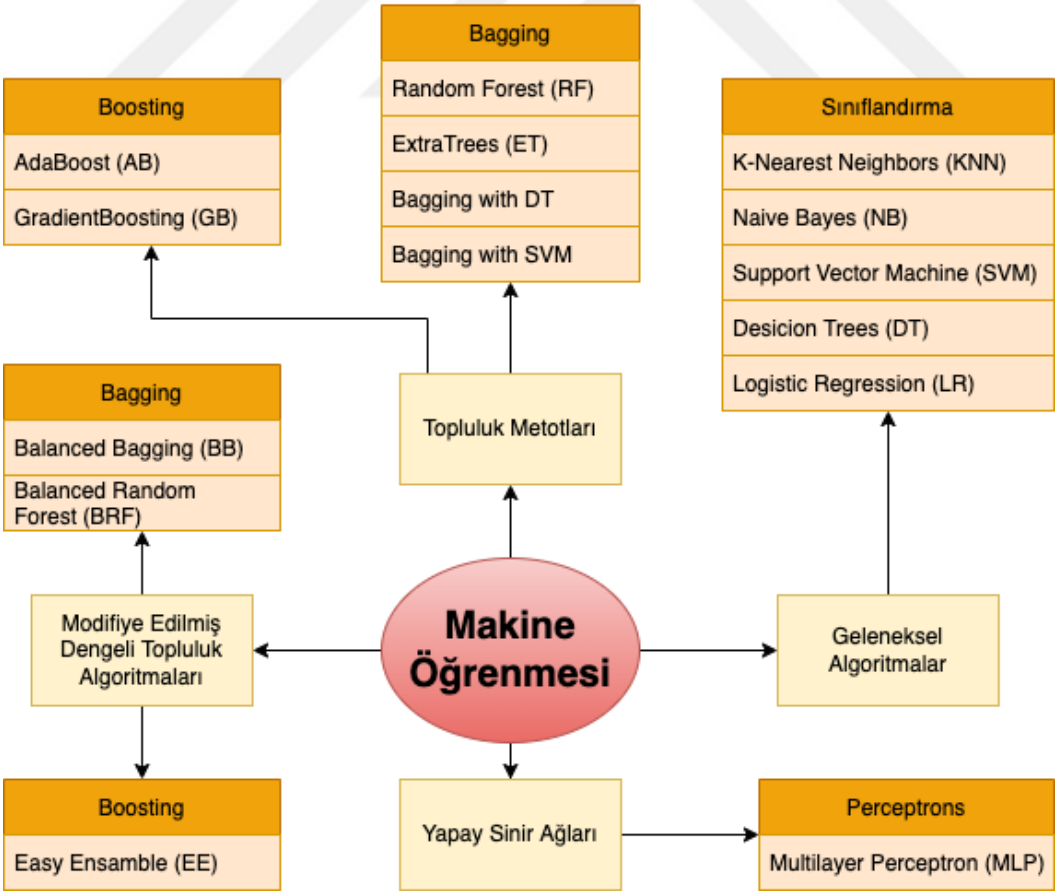
Bu çalışmada kullanılan veri kümesi çok sınıflı kategorik hedef değişkenine sahiptir. Ayrıca etikete sahiptir. Bu nedenle, Tablo 2.5.'te türlere göre gruplandırılan makine öğrenmesi algoritmalarından, veri kümesi üzerinde sınıflandırma algoritmaları ile çalışıldı.

Tablo 2.5. Makine öğrenmesi algoritma türleri

ML Algoritmaları	Denetimli	Denetimsiz
Kategorik	Sınıflandırma	Birliktelik Kuralı Madenciliği
Sayısal (Devamlı)	Regresyon	Kümeleme Boyutsal Küçülme

Denetimli makine öğrenmesi algoritmalarından, Şekil 2.15'de gösterilen makine öğrenmesi algoritmaları kullanıldı.

Sınıflandırma, veri kümesindeki belirli özellikleri kullanarak, hedef sınıfı kategorilere ayırmamızı sağlar. Makine öğrenmesi algoritma türleri denetimli, denetimsiz ve takviyeli öğrenme şeklinde üçe ayrılır.



Şekil 2.15. Tez çalışmasında kullanılan makine öğrenmesi algoritmaları

Denetimli makine öğrenmesi, etiketlenmiş veri kümesini kullanarak, verileri sınıflandırmayı veya sonuçları doğru tahmin etmek için algoritmaları eğitmeyi amaçlar. Denetimli makine öğrenmesinde, veri kümesinde istenen çıktıyı elde etmek üzere kullanılan özellikler ve doğru çıktıları içeren hedef bulunmaktadır. Ayrıca yapay öğrenmenin ana teorisi istatistiğe dayanır ve sınıflandırma istatistikte ayırtaç (discriminant) analizi olarak adlandırılır. İstatistik, matematiksel olduğu için, modeller matematiksel bakımdan analiz edilebilirdir. Mühendislikte ise sınıflandırma, örüntü tanıma olarak adlandırılır ve yaklaşım teoriden daha gözlemseldir.

Topluluk (ensemble) metotları, tekil öğrenicilerin ve kararlarının birleştirilmesi ile oluşan öğrenci topluluğuna kolektif öğrenme denir. Birkaç temel modeli(veya zayıf öğrenenleri) birleştirerek daha iyi bir tahmin modeli üretmek için kullanılan bir makine öğrenme tekniğidir. Torbalama, Artırma, İstifleme (Bagging, boosting, stacking) türlerine ayrılır. Ana hipotez, "zayıf modeller doğru bir şekilde birleştirildiğinde daha doğru modeller elde edilebilir" hipotezidir. **Torbalama (Bagging),** homojen zayıf öğrenicileri dikkate alarak, bunları birbirinden bağımsız olarak öğrenir ve sonuçları deterministik ortalama alma sürecini izleyerek birleştirir. **Artırma (Boosting),** homojen zayıf öğrenicileri dikkate alarak, birbirleriyle uyumlu bir şekilde sıralı olarak öğrenir. Bir sonraki model, bir önceki modele bağlıdır ve deterministik bir strateji izleyerek bunları birleştirir[4]. Kısaca zayıf öğrenicileri(weak learner) güçlü öğreniciye(strong learner) dönüştürür. Bunu iterasyonlar ile aşamalı olarak yapar. Artırma algoritmaları arasındaki fark genellikle zayıf öğrenicilerin eksikliğini nasıl tanımladıklarıdır[30, 31]. **İstifleme (Stacking),** birden fazla modelden tahminleri kullanarak öğrenir. Bu model, test setinde öngörülerde bulunmak için kullanılır[32].

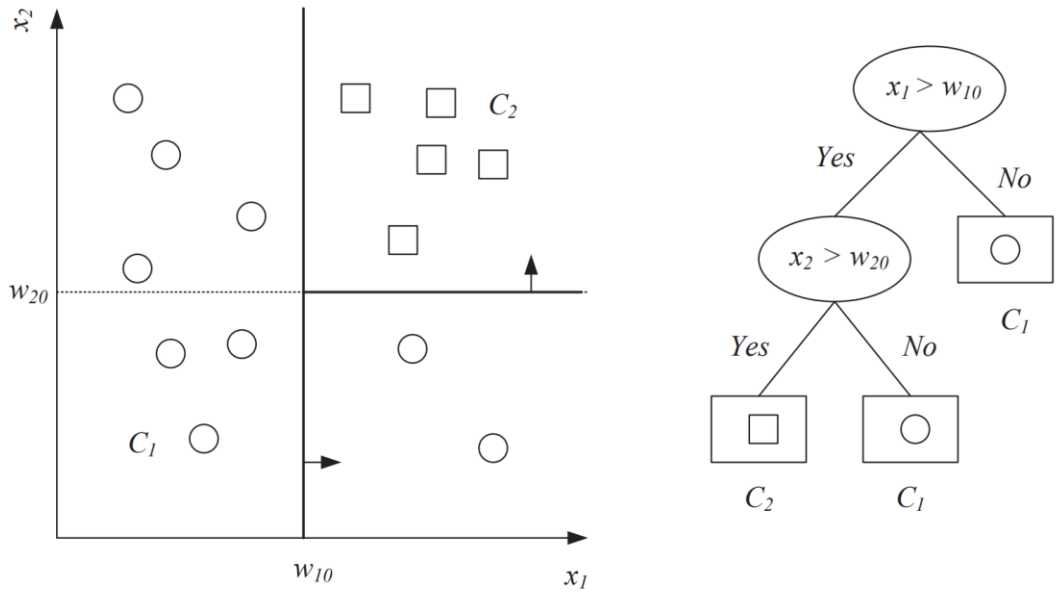
Modifiye Edilmiş Dengeli Topluluk Metotları (balanced ensemble), dengesiz veri kümelerinde kullanmak üzere geliştirilen modifiye topluluk metotlarıdır. Normal torbalama yönteminde eğitilmek üzere kullanılan örneklem her sınıftan veri içermeyebilir, bu nedenle çoğunluk sınıflarını tercih edebilir. Modifiye edilmiş dengeli topluluk metotları ise yeniden örnekleme stratejilerini kullanarak önce veriyi dengeler ve dengelenmiş veri kümesi üzerinde tahminleme yapar[33].

Yapay Sinir Ağları (YSA), insan beyninin bilgi işleme tekniğinden esinlenerek geliştirilmiş bilgi işlem teknolojisidir. YSA ile biyolojik sinir sisteminin çalışma şekli taklit edilir. Yani biyolojik nöron hücrelerinin ve bu hücrelerin birbirleri ile arasında kurduğu sinaptik bağın dijital olarak modellenmesidir[34]. **Algılayıcı (Perceptron)** ise tek katmanlı bir yapay sinir ağının temel birimidir. Eğitilebilecek tek bir yapay sinir hücresinden oluşmaktadır. Denetimli bir öğrenme algoritmasıdır[35]. Çok Katmanlı Algılayıcılar (Multi Perceptron - MLP) ise XOR

problemini çözmek için yapılan çalışmalar sonucu ortaya çıkmıştır ve özellikle sınıflandırma yapma durumlarında etkin çalışır[36].

2.5.1.1. Karar ağaçları (decision trees - DT)

Karar ağaçları, denetimli bir sınıflandırma algoritmasıdır. Yapısı, Eğer-O zaman (If-Then) kurallarını uygulayan bir akış şemasına benzer ve kök, düğüm, dal ve yapraklardan oluşur. Ağacın yapısına dahil olan her düğüm, bir özelliktir. Her dal ise testin sonucunu temsil eder ve her yaprak düğümü bir sınıf etiketini temsil eder. Bir karar düğümünün en az iki dalı vardır. Bu nedenle çok sınıflı sınıflandırma problemlerinde uygundur. Kök düğüm, en iyi tahmin ediciye karşılık gelen en üstteki düğümdür. En büyük avantajı, hem kategorik hem de sayısal verileri işleyebilir(Sharma & Kumar, 2015)[37]. Seçilen kök, veri kümesini en çok açıklayan özelliktir. Köke karar verirken Gini ve Entropi değerleri kullanılır. Gini, alt kümenin saflık değeridir. Her düğüm için Gini değeri hesaplanır. Gini değeri 0 ile 1 arasında bir sonuç alır ve sonuç 0'a ne kadar yakınsa veri kümesini o kadar iyi temsil eder ve kök olarak seçilir. Bu şekilde tüm özelliklerin GİNİ değeri hesaplanarak, ağaç oluşturulur[38].



Şekil 2.16. Bir veri kümesi örneği ve karar ağacı yapısı.

Avantajları;

- Ağaçlar görselleştirilir.
- Uygulaması ve yorumlaması basittir.
- Kategorik ve sayısal verilerde çalışabildiği için veri hazırlığı aşaması kısadır.

- Çok sınıflı sınıflandırma yapısına uygundur.

Dezavantajları;

- Aşırı öğrenme gibi problemlere sebep olabilir. Bunu önlemek için budama yöntemi, yaprak sayısını azaltmak gibi verileri küçültmeye gitmek yapılabilir.
- Verideki küçük değişikliklerden çok fazla etkilenebilir, bu nedenle kararsız olabilirler.
- Sınıf dengesizliği problemlerinde önyargılı ağaçlar oluşabilir. Bundan dolayı hedef değişkenlerin düzenlenmesi bunu önlemek için yapılabilir.
- Karmaşık ağaç üretilebilir.

2.5.1.2. Destek vektör makineleri (support vector machine - SVM)

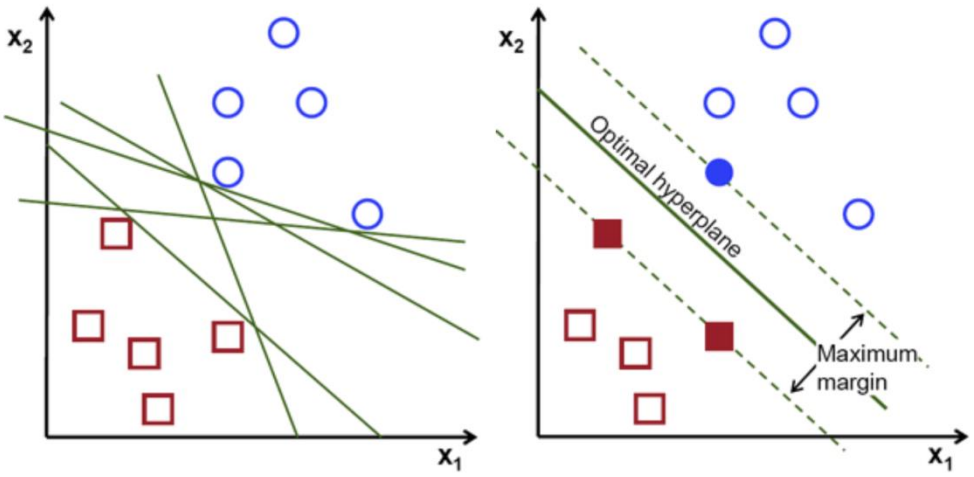
Destek Vektör Makineleri, istatistiksel öğrenme teorisine dayalı, çekirdek (kernel) tabanlı bir denetimli sınıflandırma algoritmasıdır. VC-teorisi, istatistiksel öğrenme, maksimum marj, optimal hiperdüzlem, çekirdek hileleri vb. gibi farklı fikirleri birleştirir[40]. n boyutlu bir uzayda, sınıfları birbirinden ayırarak, belirgin şekilde sınıflandırma yapan bir hiper düzlem bulunur. İki veri noktasını birbirinden ayırabilecek bir çok düzlem içinden, her iki sınıfın veri noktaları arasındaki maksimum marja sahip hiper düzlemi bulmak temel amaçtır ve bu en büyük toleransa (offset) sahip olan düzlem olur[41, 42].

Avantajları [43]:

- Küçük veya orta büyüklükteki veri kümeleri için uygundur.
- Aşırı öğrenme (overfitting) sorunu yaşanmaz.
- Karmaşık sınırları, farklı çekirdek yöntemleri kullanarak modelleyebilir.
- Bağımsız değişken sayısının çok olması algoritmayı etkilemez.

Dezavantajları [44]:

- Yüksek eğitim süresi vardır. Bu nedenle büyük boyutlu veri kümelerinde uygun değildir(maliyet).
- Büyük bellek gereksinimi bulunmaktadır (maliyet).
- Kullanılan çekirdek türüne duyarlıdır.
- Çakışan sınıflarla kötü çalışır.



Şekil 2.17. Destek Vektör Makineleri hiperdüzlem ayarı

2.5.1.3. Lojistik regresyon (logistic regression - LR)

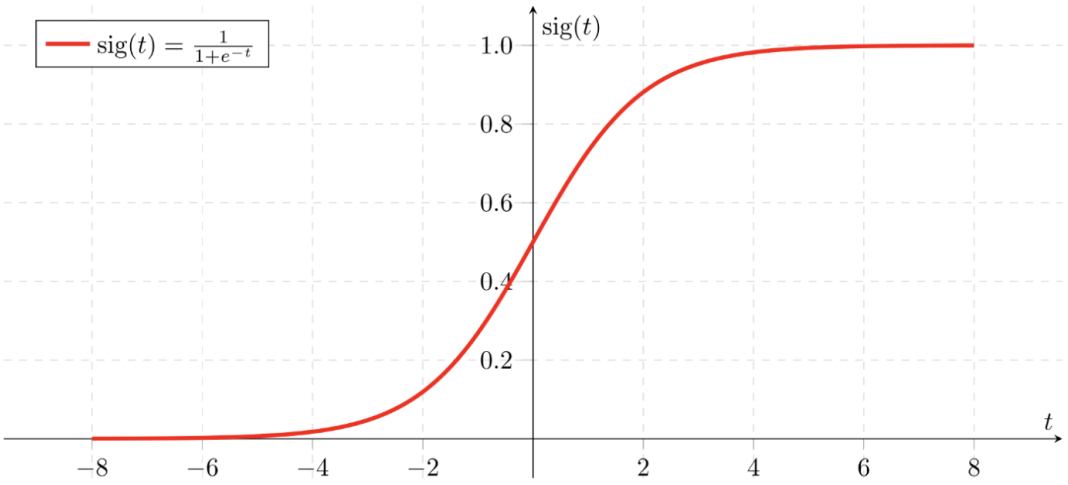
Lojistik Regresyon, denetimli bir sınıflandırma algoritmasıdır. Sürekli sayısal, kesikli sayısal, nitelik şeklinde bağımsız değişkenlerle çalışır. Bağımlı değişken, iki terimli (binominal), çok terimli (multinominal) olabilir. Veri özellik türlerine göre uygulanacak yöntem değişmektedir. Sırasız çok sınıflı bağımlı değişkeni olan ve çeşitli bağımsız değişkeni olan veri kümelerinde çok terimli lojistik regresyon kullanılır. Bağımsız değişkenlerle, sebep sonuç ilişkisini belirlemek için kullanılır ve bu etkileri odds oranı ile açıklar[45].

Avantajları:

- Uygulaması ve yorumlaması basittir.
- Veri kümesi doğrusal olarak ayrıldığında, çok iyi performans gösterir.
- Aşırı öğrenme (overfitting) sorunu yaşanmaz.

Dezavantajları:

- Veri kümesi doğrusal olarak ayrılmadığı zaman lojistik regresyon ayrım yapamaz.
- Büyük veri kümelerinde aşırı öğrenme (overfiting) sorun olabilir[47].



Şekil 2.18. Sigmoid Fonksiyonunun Grafiği

2.5.1.4. K-en yakın komşu (k-nearest neighbors - KNN)

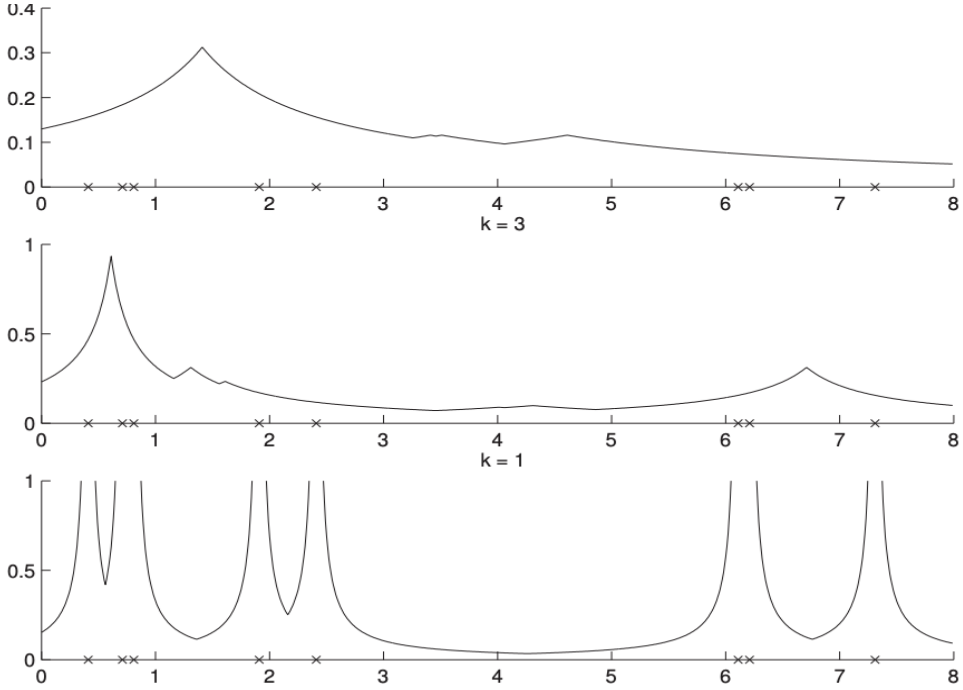
K-en yakın komşu, denetimli bir sınıflandırma algoritmasıdır. KNN algoritmasında tahmin edilen sınıf, bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşuları hangi sınıfta yoğun ise o sınıfa sınıflandırır. KNN, komşuluk sayısı ve uzaklık değerleri üzerinden tahmin yapar. Uzaklık hesapları için "Euclidean", "Manhattan", "Minkowski" gibi birçok uzaklık fonksiyonu kullanmaktadır. KNN, eski, basit ve gürültülü eğitim verilerine karşı dirençlidir yani bu eğitim verilerinde de başarılı sonuçlar verir[48].

Avantajları:

- Uygulaması ve yorumlaması basittir.
- Matematiksel hesaplamalar olduğu için süreçler takip edilebilir.
- Gürültülü eğitim verilerine karşı etkilidir.

Dezavantajları:

- Büyük verileri işlerken, büyük bellek gereksinimi bulunmaktadır (maliyet).
- Yüksek eğitim süresi vardır[50].



Şekil 2.19. Çeşitli k değerleri için k-en yakın komşu tahmini

2.5.1.5. Naive Bayes (NB)

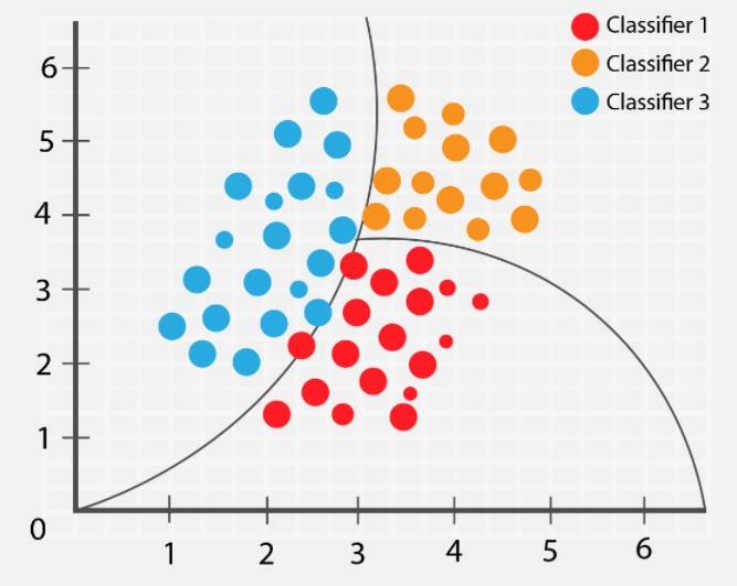
Naive Bayes, denetimli bir sınıflandırma algoritmasıdır. En iyi tahminleme işlemini kategorize etmek için olasılık ölçüsünü kullanır. 18. yüzyılda Thomas Bayes tarafından ortaya konulan Bayes Teoremine dayanır. Gaussian, Multinomial ve Bernoulli Naive Bayes türleri vardır. Multinomial Naive Bayes çok sınıflı kategorilerde de kullanılır. Bernoulli Naive Bayes sadece iki sınıflı kategorilerde kullanılır. Gaussian Naive Bayes ise, özellikler sürekli verilerden oluştuğunda kullanılır[51].

Avantajları:

- Uygulaması ve yorumlaması basittir.
- Özelliklerin birbirinden bağımsız kabul edilmesi iyi performansa sebep olur.
- Veri kümesi az olduğunda da başarılı sonuçlar verir.
- Büyük veri kümelerinde de başarılı sonuçlar verir.
- Kesikli ve sürekli verilerle çalışır.

Dezavantajları:

- Değişkenler arası ilişkiler modellenemez.
- Sıfır olasılık problemi oluşabilir.



Şekil 2.20. Naive Bayes sınıflandırması

2.5.1.6. Çok katmanlı algılayıcı (multilayer perceptron - MLP)

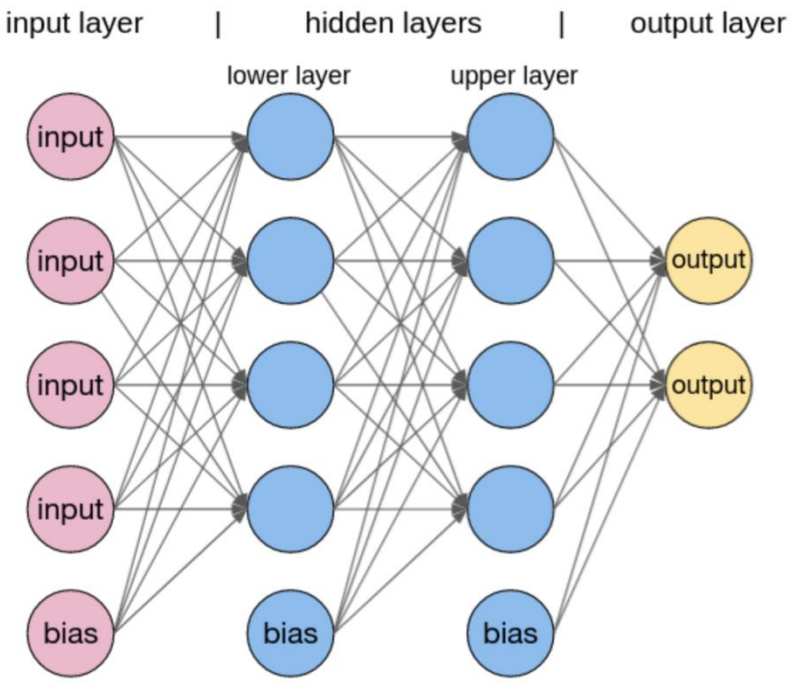
Çok katmanlı algılayıcı, denetimli bir öğrenme tekniği olarak, derin öğrenme yöntemidir. Bir dizi girdiden bir dizi çıktı üreten bir ileri beslemeli yapay sinir ağıdır. MLP, XOR Problemi'ni çözmek için yapılan çalışmalar sonucu ortaya çıkmıştır. Özellikle sınıflandırma durumlarında etkin çalışır[53].

Avantajları:

- Karmaşık doğrusal olmayan problemlerde uygulanabilir.
- Büyük veri kümelerinde ve küçük veri kümelerinde başarılı sonuçlar verir.
- Hızlıdır.

Dezavantajları:

- Değişkenler arası bağımlılıklar bilinmemektedir.
- Yüksek eğitim süresi vardır.
- Veri ön işleme aşamasının kalitesi, model performansını etkiler[54].



Şekil 2.21. MLP yapısı

2.5.1.7. Rastgele orman (random forest - RF)

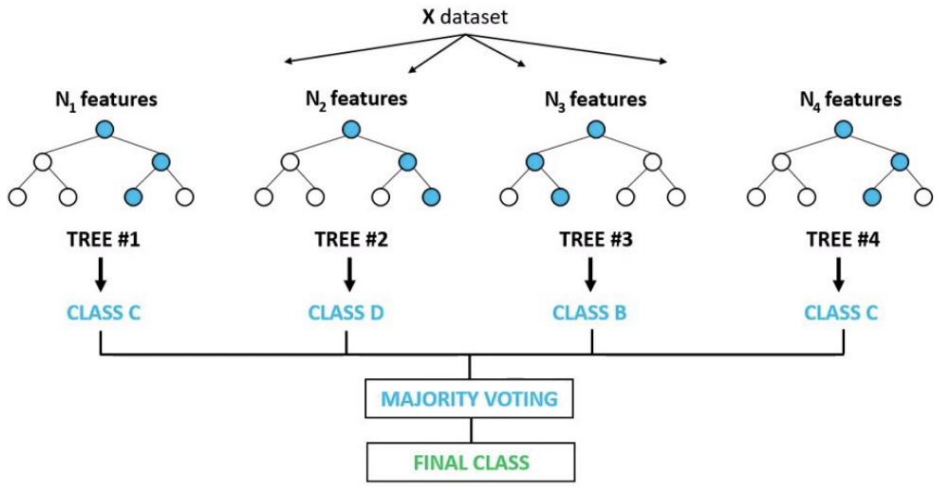
Rastgele orman, denetimli bir sınıflandırma algoritmasıdır. Topluluk öğrenmeyi kullanır. Birçok karar ağacından oluşur ve karar ağaçlarının tahminlerine göre sonucu belirler. Ağaç sayısının artması, sonucun kesinliğini artırır. Karar ağaçlarının aşırı öğrenme problemini azaltır ve hassasiyeti artırır, bu nedenle karar ağacından daha doğrudur. Böylece sonuçlardaki başarı oranı artar[55].

Avantajları:

- Uygulaması ve yorumlaması basittir.
- Büyük veri kümelerinde başarılı sonuçlar verir.
- Hiper parametre ayarı olmadan da başarılı tahmin üretir.

Dezavantajları:

- Büyük bellek gereksinimi bulunmaktadır (maliyet).
- Yüksek eğitim süresi vardır.



Şekil 2.22. Random Forest sınıflandırıcısı

2.5.1.8. Uyarlanabilir Artırma (Adaptive Boosting - AdaBoost)

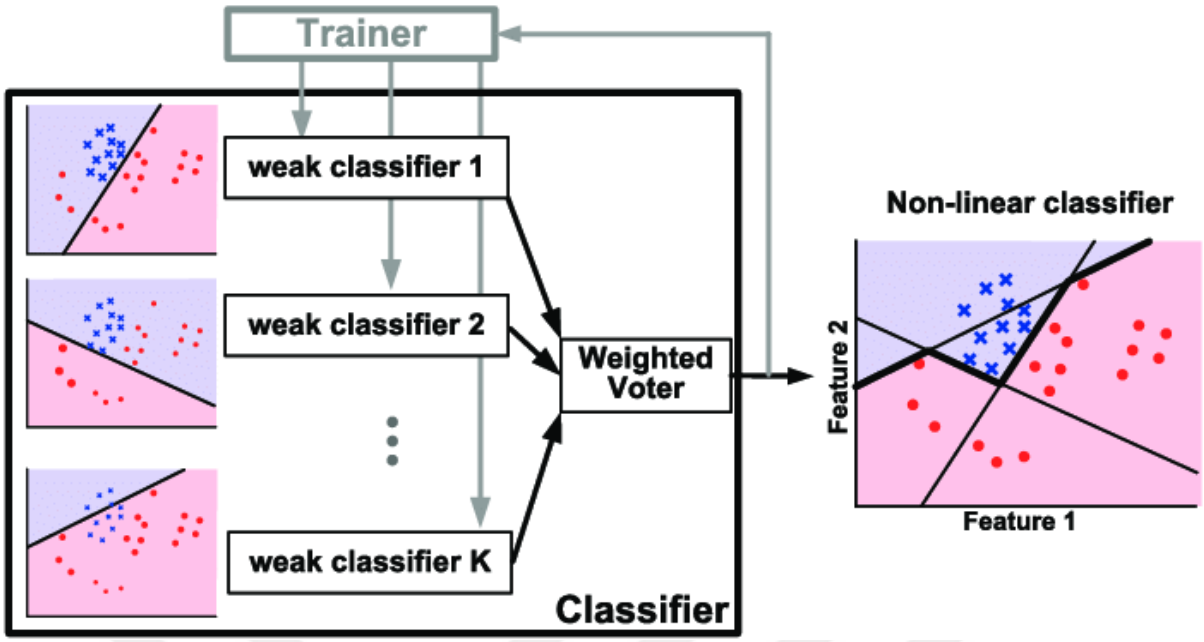
Adaboost, topluluk metot türlerinden olan, en basit artırma algoritması olarak sayılır. Temel öğrenici olarak genellikle karar ağaçları kullanılır. Freund ve Schapire (1997) tarafından önerilen algoritmada, zayıf algoritmaların bir araya getirilerek daha güçlü bir eğitim modelinin oluşturulması amaçlanır. İçerisinde daha önce geliştirilmiş olan farklı makine öğrenmesi algoritmalarını birleştirir ve öğrenme konusunda en iyi sonucun elde edilmesini hedefler(Freud ve Schapire, 1997)[57]. Ve genel olarak topluluk öğrenme metotları sınıflandırmada, tekil öğrenme algoritmalarına göre daha başarılıdır. Döngüde, yanlış tahmin edilen gözlemlere ağırlık atanır ve sonraki modelde bu değerler doğru tahmin edilmeye çalışılır[58, 59].

Avantajları:

- Uygulaması ve yorumlaması basittir.
- Hiper parametre ayarı olmadan da başarılı tahmin üretir.
- Zayıf sınıflandırıcıların doğruluğunu artırabilir.
- Önyargı ve varyansı azaltır.

Dezavantajları:

- Gürültülü veriye karşı çok hassastır.
- Hiperparametre optimizasyonu zordur.
- XGBoost'tan daha yavaştır[61, 62].



Şekil 2.23. Adaboost sınıflandırıcısı

2.5.1.9. Gradyan ağacı artırma (gradient tree boosting - GBM)

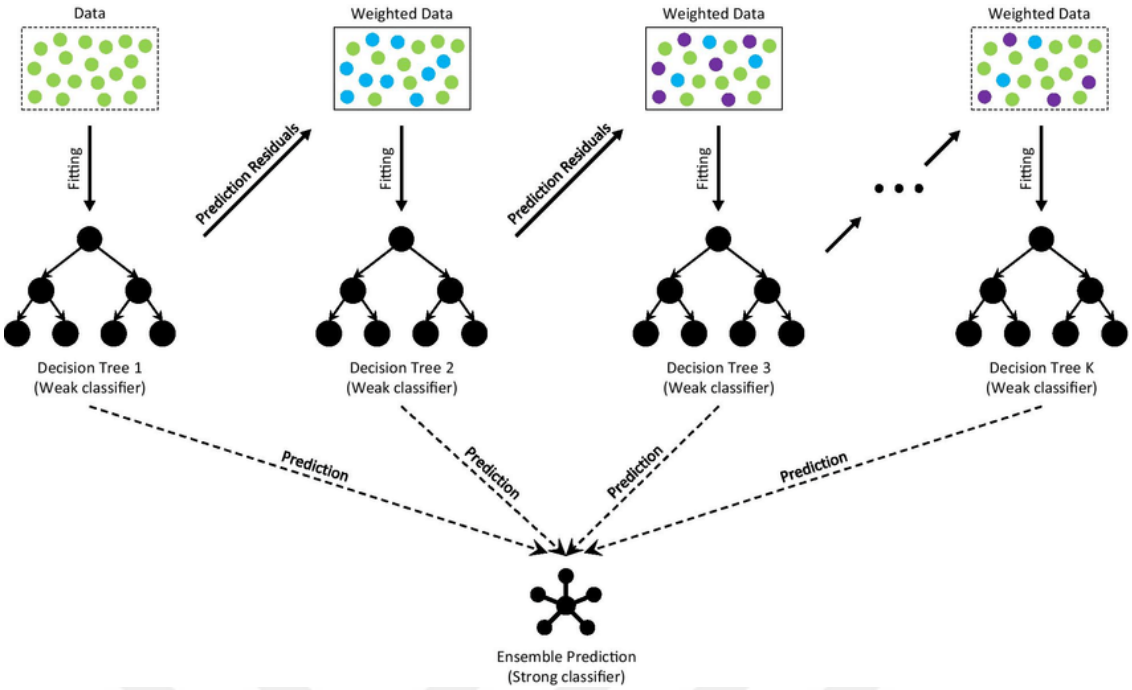
Gradyan ağacı artırma, bir topluluk makine öğrenme algoritmasıdır. "Gradient Descent" ve "Boosting" kavramlarıyla oluşturulmuştur. Temel öğrenici olarak regresyon ağaçları kullanılır. Yapısı oluşturulurken, ilk önce ilk yaprak(initial leaf) oluşturulur. Sonra tahmin hatalarına göre yeni ağaçlar oluşturulur. Bu döngü, değer olarak verilen ağaç sayısı tamamlanan veya değişmeyen başarı grafiği olana kadar devam eder. Yani bir ağaç önceki ağaç tarafından hesaplanan hatalar üzerine kuruludur[63, 64].

Avantajları:

- Farklı kayıp fonksiyonlarını destekler ve etkileşimlerle iyi çalışır.

Dezavantajları:

- Aşırı öğrenme (overfitting) problemi yaşanabilir.
- Hiperparametre optimizasyonunun iyi yapılması gerekir[66, 67].



Şekil 2.24. Gradient Boosting Karar Ağacının mimarisi

2.5.1.10. Aşırı gradyan artırma (extreme gradient boosting - XGBoost)

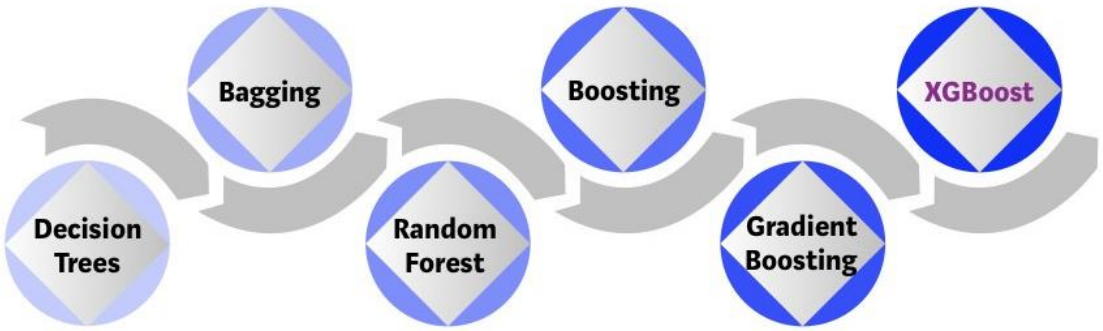
Aşırı gradyan artırma, popüler bir gradyan artırma uyarlamasıdır. Tianqi Chen ve Carlos Guestrin tarafından 2016 yılında oluşturulan, popüler bir algoritmadır. Hız ve performans için tasarlanmıştır. Temel öğrenici olarak karar ağaçlarını kullanır. Yüksek tahmin gücüne sahiptir ve gradyan artırma uyarlamalarından neredeyse 10 kat hızlıdır. Performanslı çalışacak iyileştirmeleri vardır. Paralel işlem uygular. Eksik değerleri işleyebilir[68].

Avantajları:

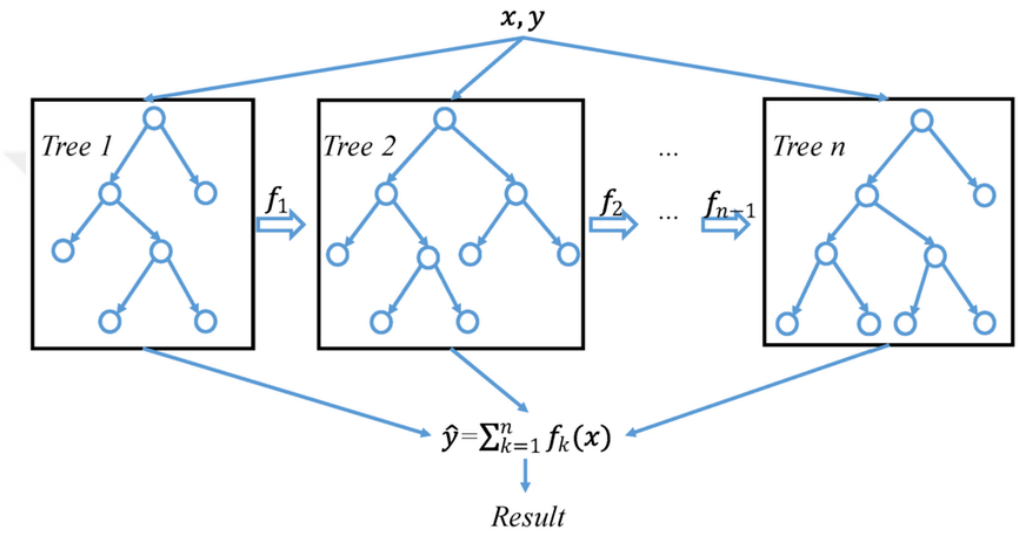
- Eksik olan verileri işlediği için, BOŞ verileri önceden doldurmak gerekmez.
- Paralel veri işleme yapar.
- Karar ağacındaki gereksiz verileri eleme yapar.
- Aşırı öğrenme (overfitting) önler.

Dezavantajları:

- Hiper parametre optimizasyonunun iyi yapılması gerekir[71].



Şekil 2.25. Karar-ağacı algoritmalarının XGBoost'a Evrimi



Şekil 2.26. XGBoost mimarisi

2.5.1.11. Son derece rastgele ağaçlar (extremely randomized trees - ExtraTrees - ET)

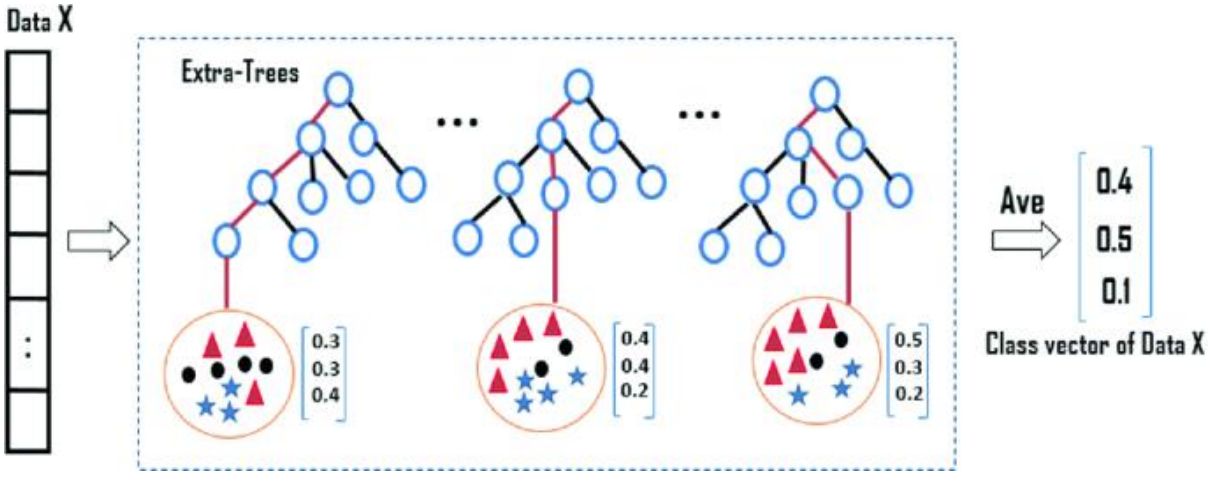
ExtraTrees, bir topluluk makine öğrenme algoritmasıdır. Ağaç üretilirken karar kuralı rastgele seçilir. RF'e göre, daha hızlı olabilirler. Önyargı-varyans analizi içerir, böylece önyargıyı azaltmayı amaçlar [72].

Avantajları:

- Önyargı ve varyansı azaltır.
- Hızlıdır.

Dezavantajları:

- Büyük bellek gereksinimi bulunmaktadır(maliyet)[74].



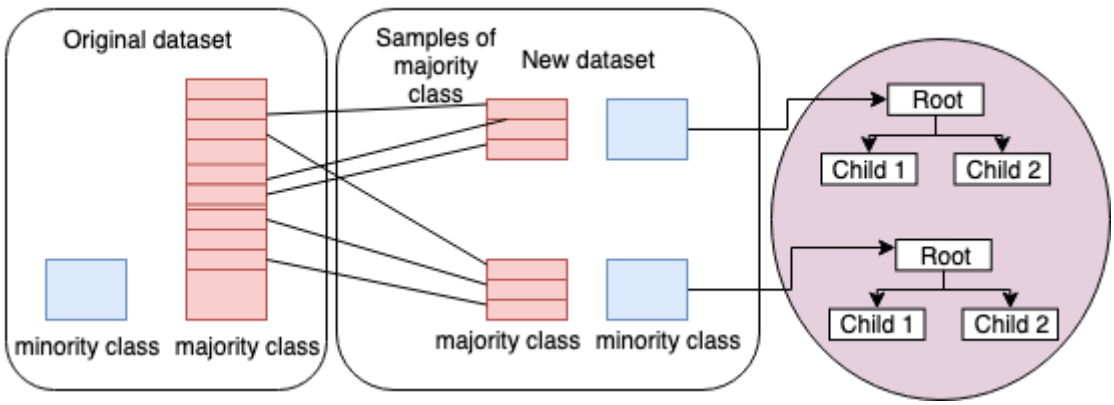
Şekil 2.27. ExtraTrees mimarisi

2.5.1.12. Dengeli torbalama (Balanced bagging - BB)

Dengesiz veri kümeleri için modifiye edilmiş, torbalama algoritmasıdır. Torbalama, topluluktaki her bir ağacı oluşturmak için veri kümesinden birçok kez örnekler çeker ancak bu örneklerde her sınıftan veri bulunmayabilir. Dengeli torbalama, bir ağacı eğitmek için kullanılan her örnekte, tüm sınıflardan veri seçerek dengeli olmasını sağlar. Özellikle torbalama sınıflandırıcısı dengesiz veri kümesinde eğitim yaparken, çoğunluk sınıfı tercih edebildiği için dengesiz veri kümelerinde, modifiye edilmiş dengeli torbalama sınıflandırıcısı kullanılır. "sampling_strategy" parametresi ile yeniden örnekleme yaparak sınıflandırır[75].

Avantajları:

- Veri kümesindeki dengesizliği gidererek uygulanır.



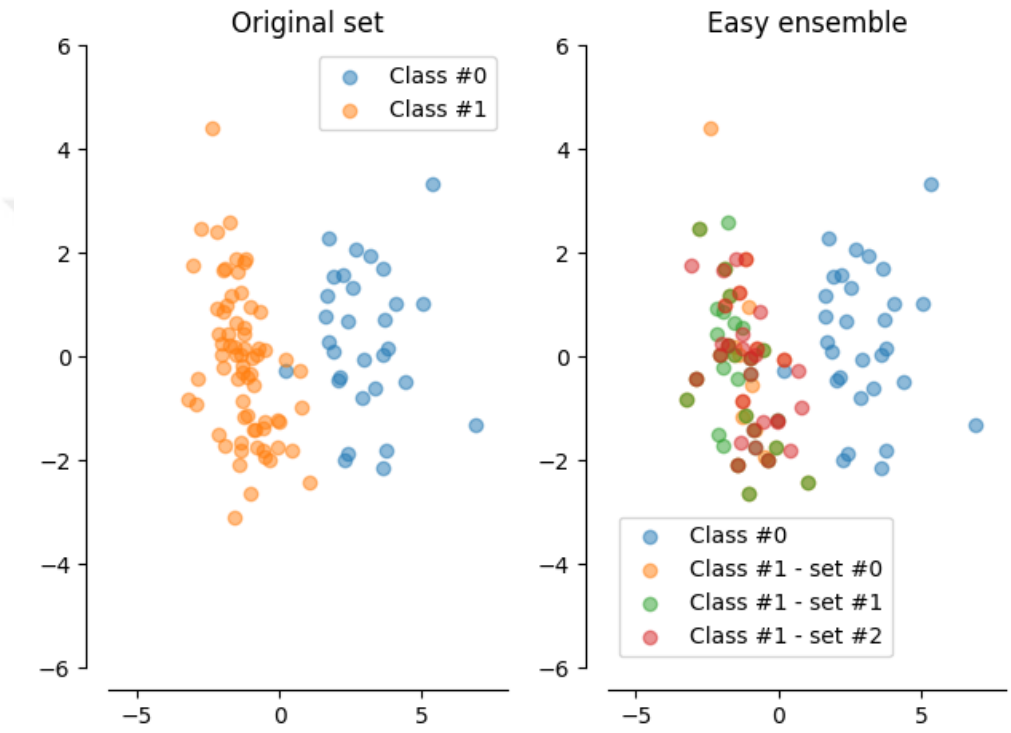
Şekil 2.28. Balanced Bagging mimarisi

2.5.1.13. Dengeli rastgele orman (balanced random forest - BRF)

Dengesiz veri kümesi için modifiye edilmiş, torbalama algoritmasıdır. RF sınıflandırıcısının tüm işlevleri bu algoritmada da vardır. Ek olarak, ormanın her ağacına dengeli bir önyükleme örneği sağlanır[77].

Avantajları:

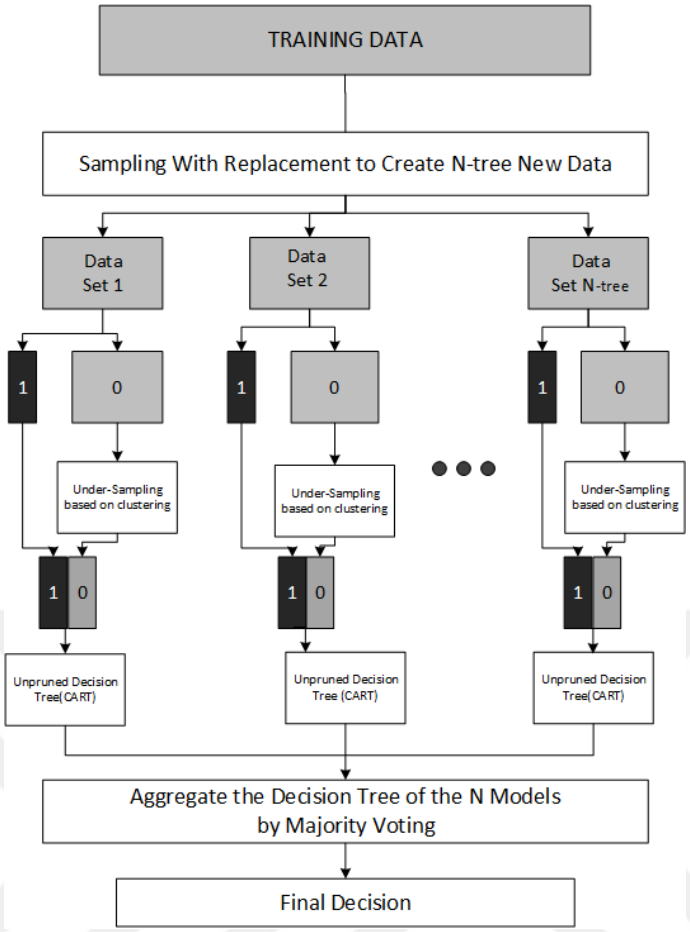
- Veri kümesindeki dengesizliği gidererek uygulanır.



Şekil 2.29. Easy ensemble sınıflandırıcısı

2.5.1.14. Kolay topluluk (easy ensemble - EE)

Dengesiz veri kümeleri için modifiye edilmiş, artırma algoritmasıdır. Modifiye edilerek eğitilmiş bir AdaBoost öğrencileri topluluğudur. Dengeleme işlemi rastgele eksik örnekleme yapılarak sağlanır. Her sınıfı bağımsız olarak örnekleyerek çok sınıflı yeniden örneklemeyi de destekler[79, 80].



Şekil 2.30. Dengeli rastgele orman mimarisi

2.5.2. Hiper parametre ayarı (hyper parameter tune)

Eğitim sürecinden önce belirlenen parametrelere hiper-parametre denir. Çoğu algoritmanın eğitilen modelinin performansı hiper-parametrelerin seçimlerinden doğrudan etkilenir. Bu nedenle seçilen algoritmayı veri kümesine uygulamadan önce, uygun bir hiper-parametre kümesi seçmek gerekir. Kümeyi seçmek için farklı hiper-parametre setleri belirlenir. Böylece aynı algoritmaya ait de farklı modeller oluşturulabilir[82]. Hiper-parametre optimizasyonu için kullanılan yöntemler aşağıdaki gibidir:

- GridSearchCV: Başlangıçta tanımlanan hiper-parametre değerleri ızgarasından, en uygun hiper-parametre kümesi üzerinde arama işlemi gerçekleştirilir ve en uygun parametreler belirlenir.
- RandomizedSearchCV: Uygun hiper-parametre kümesini bulmak için ızgara içerisinde rastgele gezinir.

Bu çalışmada kullanılan makine öğrenmesi algoritmalarının parametreleri, her model oluşturulmadan önce GridSearchCV yöntemi kullanılarak optimize edildi.

2.5.3. Sınıf dengesizliği problemi

Sınıf dengesizliği, hedef değişkendeki sınıfların eşit dağılmadığı durumlarda oluşur. Örneğin 1000 satırdan oluşan bir veri kümesinde, 1. hedef değişkene ait olan satır 950, 2. hedef değişkene ait olan satır 50 olduğunda, makine öğrenme algoritması 2. hedef değişkenin hepsini hatalı yerleştirme yapsa bile %95 doğruluk(accuracy) oranı elde edebilir. Hata matrisi(confusion matrix) kontrol edildiğinde doğru tahminlerin aslında tek bir sınıfa ait olduğu görülür. Bu durum yanıltıcıdır ve elde edilen bulgular güvenilir değildir[83].

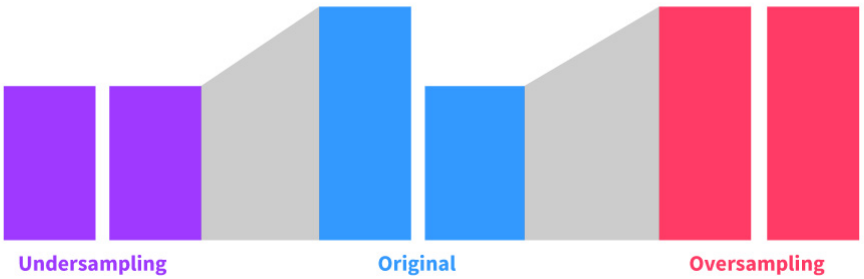
Sınıf dengesizliği probleminin çözümünde, azınlık sınıfına ait olan verinin de sınıflandırma doğruluğunu artırmak için, veri düzeyinde yeniden örnekleme ve algoritma düzeyinde modifiye edilmiş öğrenme algoritmaları şeklinde iki ana yaklaşım vardır[7]. "Makine Öğrenmesi Algoritmaları" başlıklı bölümde bu çalışmada kullanılan modifiye edilmiş makine öğrenmesi algoritmalarından bahsedildi. Ek olarak veri düzeyinde yeniden örnekleme yöntemleri de bu çalışmada kullanıldı.

Şekil 2.31.'de gösterildiği gibi yukarı yeniden örnekleme (oversampling), azınlık (minority) olan ve küçük örneklem içeren sınıfın çoğaltılması işlemidir. Aşağı örnekleme (undersampling) ise, yukarı yeniden örnekleme yönteminin tersi olarak, çoğunluk (majority) olan büyük örneklem içeren sınıfın azaltılması işlemidir. Bu nedenle, büyük verilerde aşağı örnekleme işlemi uygundur ancak, küçük veri kümelerinde performans azalmasına sebep olabilir. Aşağı örnekleme işleminde tüm nadir olaylar kullanılır, veri azaltılırken tüm nadir olayların kullanımda olduğundan emin olunur. Yukarı örnekleme yönteminde ise, nadir olayların sayısı artırılır. Nadir olayların artırılması veya azaltılması için gerekli olan örneklerin belirlenmesi için farklı yöntemler vardır[84]. Bu yöntemlerden literatürde kullanılanlardan bazıları aşağıdaki gibidir:

- Aşağı örnekleme - Undersampling
    - ENN (edited nearest neighbors) (en yakın komşular)
    - RUS (Random Under Sampling) (rastgele aşağı örnekleme)
    - TomekLinks
- Yukarı yeniden örnekleme - Oversampling
    - ROS (Random Over Sampling) (rastgele yukarı örnekleme)
    - ADASYN (Uyarlanabilir Sentetik Örnekleme Yaklaşımı)

- SMOTE (Sentetik Azınlık Aşırı Örnekleme Tekniği)

- Aşırı ve Az örneklemenin beraber kullanılması - Hibridizasyon
  - SMOTE+TOMEK
  - SMOTE + ENN
  - SMOTEBoost (SMOTE + Boosting)
  - RUSBoost (Rastgele az örnekleme + Boosting)



Şekil 2.31. Yeniden örnekleme yöntemleri

RUS, çoğunluk sınıfından olayları rastgele siler.

ENN, dağılım grafiğindeki her olay için en yakın üç komşuya bakar. Olay, çoğunluk sınıfına aitse, en yakın üç komşusuna göre yanlış sınıflandırılırsa, çoğunluk sınıfı kaldırılır. Olay azınlık sınıfına aitse, en yakın üç komşusuna göre yanlış sınıflandırılırsa, çoğunluk sınıfına ait en yakın komşuları kaldırılır[84].

TomekLinks, çok sınıflı yeniden örneklemeyi destekler. Sınıflara ait en yakın iki örneğin arasındaki çoğunluk sınıfının örnekleri kaldırılarak, iki sınıf arasındaki boşluk arttırılır[85].

ROS, mevcut örnekleri değiştirir ve rastgele yeni örnek oluşturur[83].

SMOTE ve ADASYN, yeniden örnekleme yaparak, azınlık sınıfı çoğaltır. Hem SMOTE hem de ADASYN, yeni örnekler oluşturmak için KNN sınıflandırıcısını kullanır ancak karar fonksiyonları farklıdır. Yani, ADASYN, KNN sınıflandırıcısını kullanılarak yanlış sınıflandırılan orijinal örneklerin yanında örnekler üretir. SMOTE ise KNN sınıflandırıcısını kullanarak sınıflandırılan örnsıeklerin yanlış olup olmamasıyla ilgilenmez, rastgele seçer[86]. Genel formül, azınlık sınıfından bir Xi örnek seçilir. Bunun en yakın komşularından Xzi seçilir ve arasındaki uzaklık ile ilişkili yeni bir Xnew örnek oluşturulur. Xnew = Xi + Lambda x (Xzi - Xi). Lambda 0 - 1

aralığında rastgele bir sayıdır. Bu formül, Xi ve Xzi doğrusu üzerinde yeni bir Xnew oluşturur[83].

BorderlineSMOTE (BSMOTE1-BSMOTE2), her Xi örneğini, güvenli (safe), tehlike(danger) ya da gürültü (noise) olarak sınıflandırır ve tehlike(danger) altındaki örnekleri yeni örnekler oluşturmak için kullanır. Borderline-1 SMOTE, Xzi yi seçerken, Xi örneğinin ait olduğu sınıftan seçer. Borderline-2 SMOTE, herhangi bir sınıftan, Xzi yi seçer. SVMSMOTE, destek vektörlerini bulmak için SVM sınıflandırıcısı kullanır. KMeans SMOTE, KMeans kümeleme yöntemiyle önce örnekleri gruplar ve küme yoğunluğuna bağlı olarak yeni örnekler oluşturur. SMOTEN yalnızca kategorik verilerle çalışır. SMOTE-NC, sadece sayısal verileri değiştirerek, kategorik verilerin aynı kalmasını sağlayan SMOTE uzantısıdır[83].

SMOTEENN, SMOTE kullanarak yukarı yeniden örnekleme yapar daha sonra da ENN kullanarak temizleme işlemi yapar. Aynı şekilde SMOTE-Tomek SMOTE kullanarak yukarı yeniden örnekleme yapar daha sonra da Tomek kullanarak temizleme işlemi yapar[83].

Örnekleme yöntemleri, veri kümesinin tümüne uygulanmaz. Veri kümesinin tümüne uygulandığında, veri sızıntısına sebep olup, bulgu ve sonuçların güvenilirliğini azaltır. Bu nedenle, örnekleme yöntemleri, yalnızca eğitim veri kümesine uygulanır. Orjinal verileri doğru şekilde temsil edebilmesi için test veri kümesi değişmeden kalır.

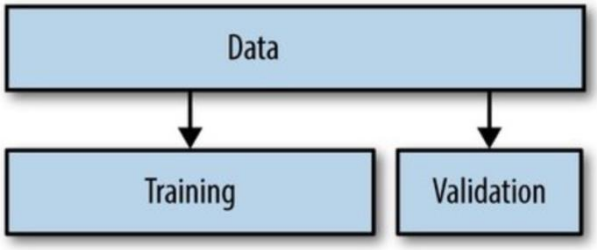
İki sınıflı hedef değişkeni olan veri kümesinde kullanıldığı gibi, çok sınıflı hedef değişkeni olan veri kümelerinde de örnekleme yöntemleri kullanılır. Ek olarak, SMOTEN gibi bazı yöntemler, sadece kategorik verilerde kullanılır. Bu nedenle, örneklem yöntemlerini belirlerken veri kümesinin özellik türleri önemlidir.

Bu çalışmada, veri kümesine uygun olan örnekleme yöntemlerinden, SMOTE, SMOTENC, BSMOTE1, BSMOTE2, SVMSMOTE, ADASYN, ROS ve hibrit yöntemlerden SMOTE - ENN, SMOTE - Tomek kullanıldı.

2.5.4. Çapraz geçerleme (cross validation - CV)

2.5.4.1. Holdout çapraz geçerleme

Veri kümesi, rastgele %80-%20 veya %70-%30 şeklinde eğitim ve doğrulama verilerine bölünür. Olumsuz yanı, verilerin bir kısmı eğitim verisi olarak kullanılmamış olur. Ayrıca sınıf dengesine bakmadan rastgele bölündüğü için, eğitim verisinin içerisinde, az örneklem içeren sınıf bulunmayabilir. Bu da, az örneklem içeren sınıfların eğitilmesine engel olur. Bu nedenle dengesiz veri kümesinde kullanımı uygun değildir.



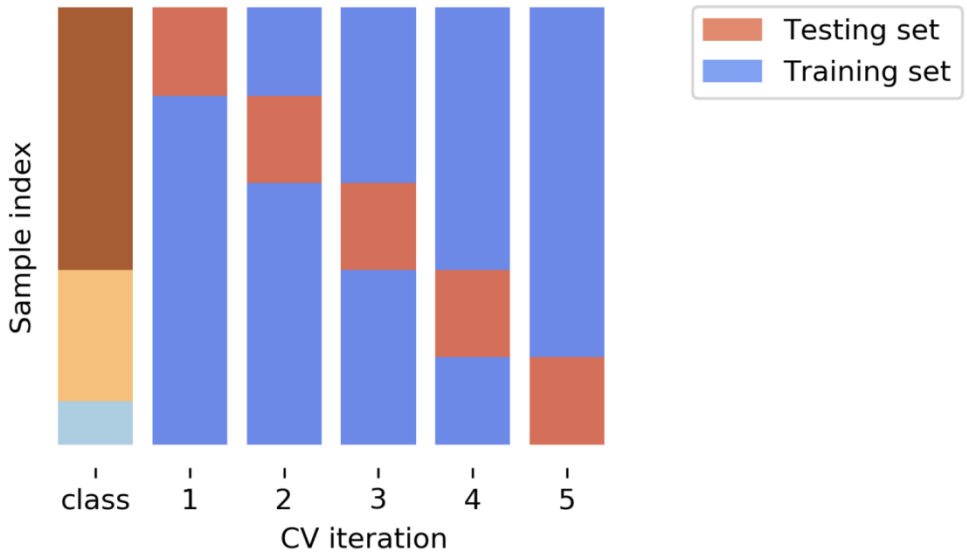
Şekil 2.32. Hold out geçerleme yöntemi

2.5.4.2. K-fold çapraz geçerleme

Orijinal veri kümesi eşit olarak k alt parçaya bölünür. Şekil 2.33'te görüldüğü gibi holdout yöntemindeki gibi sınıf dengesine bakılmadığı için, eğitim verisinin içerisinde, az örneklem içeren sınıf bulunmayabilir. Bu da, az örneklem içeren sınıfların eğitilmesine engel olur. Bu nedenle dengesiz veri kümesinde kullanımı uygun değildir.

2.5.4.3. Tekrarlanan rastgele alt örnekleme doğrulaması

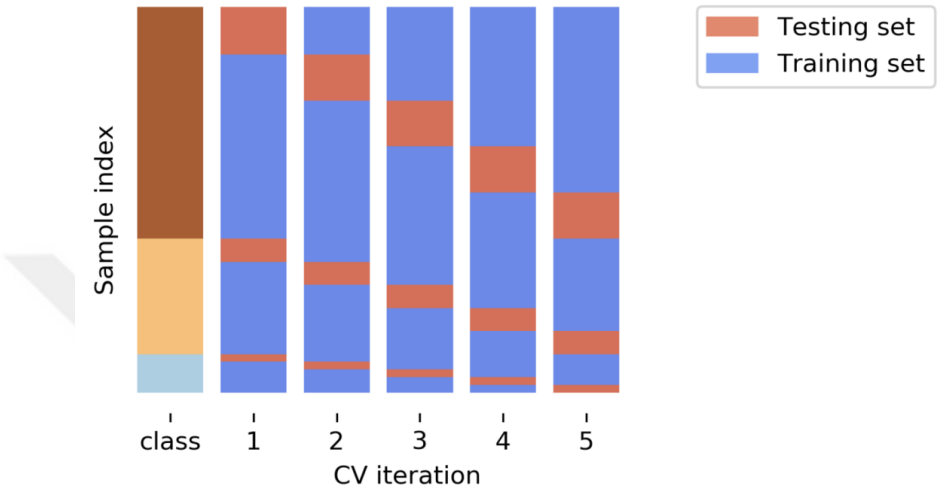
Tekrarlanan rastgele alt örnekleme doğrulamasında (Repeated random subsampling validation) her iterasyonda rastgele örneklem seçilir. Holdout yönteminde olduğu gibi sınıf dengesine bakılmadığı için, eğitim verisinin içerisinde, az örneklem içeren sınıf bulunmayabilir. Bu da, az örneklem içeren sınıfların eğitilmesine engel olur. Bu nedenle dengesiz veri kümesinde kullanımı uygun değildir.



Şekil 2.33. K-fold çapraz geçerleme iterasyon örneği

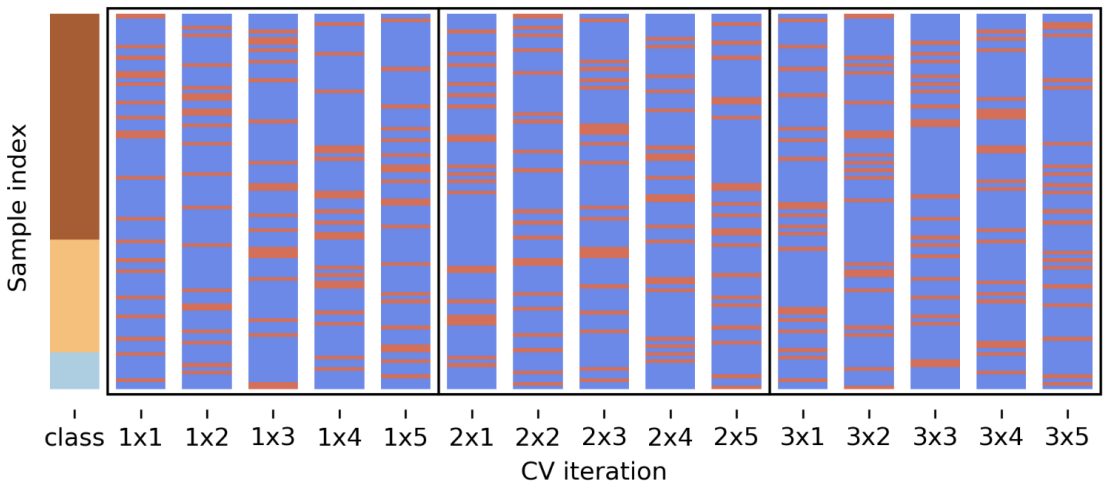
2.5.4.4. Tabakalı k-kat çapraz geçerleme (Stratified k-fold CV)

StratifiedKFold, katmanlı kıvrımları döndüren bir k-fold çeşididir. Şekil 2.34'te görüldüğü gibi her iterasyonda, her set, tam set olarak her bir hedef sınıftan yaklaşık olarak aynı yüzdede örnek içerir. Bu, özellikle dengesiz veri kümelerinde, doğrulama veya eğitim verilerinde belirli bir sınıfın fazla bulunmasını engeller. Böylece, az örneklem içeren sınıfların da eğitilmesi sağlanmış olur. Bu nedenle dengesiz veri kümesinde kullanımı uygundur.



Şekil 2.34. Stratified k-fold çapraz geçerleme iterasyon örneği

Bu çalışmada kullanılan veri kümesinde sınıf dengesizliği problemi olduğu için, Şekil 2.35'te görüldüğü gibi Tekrarlı Tabakalı k-kat çapraz geçerleme (Repeated Stratified k-fold CV) yöntemi kullanıldı.



Şekil 2.35. Repeated Stratified k-fold CV iterasyon örneği

2.5.5. Sağlamlık (Robustness) ve Tutarlılık Kontrolü

Şekil 2.12.'de belirtildiği üzere, model oluşturulduktan sonra, makine öğrenmesi algoritmalarının hiç görmediği yeni bir veri kümesi ile oluşan son modellerin sağlamlık testi yapıldı. Böylece, modelin doğruluğu, başarısı, tutarlı olup olmaması, yeni bir veri kümesi üzerinden kontrol edildi.

Tutarlılığın kontrol edilmesi için, 2016 yılında kayıt olan öğrenci verileri hazırlanmış veri kümesinden ayrıldı. Model geliştirdikten sonra, modeller üzerinde 2016 yılı verisiyle tahminleme yapıldı ve sonuçlar elde edildi. Sağlamlık sonuçları, test sonuçlarıyla karşılaştırılarak en başarılı ve en tutarlı modeller seçildi ve pilot bir karar destek sistemi web uygulamasında model olarak kullanıldı.

BÖLÜM 3 - BULGULAR VE TARTIŞMA

3.1. Performans Ölçütleri

Karmaşıklık matrisi, makine öğrenme teknikleriyle eğitilen model performansının yorumlanmasını sağlayan değerleri içerir. Tablo 3.1'de görüldüğü gibi modelin performansı ve hangi hatayı yaptığı hakkında bilgi gösterir.

Tablo 3.1. Karmaşıklık matrisi

Toplam = P + N		**Gerçek Değer**	
		Pozitif (P)	**Negatif (N)**
Tahmin Edilen Değer	**Pozitif (PP)**	True Positive (TP) Pozitif değer, pozitif olarak tahmin edilir. Doğru tahmin.	False Positive (FP) (Type 1 Error) Negatif değer, pozitif olarak tahmin edilir. Hatalı tahmin.
	Negatif (PN)	False Negative (FN) (Type 2 Error) Pozitif değer, negatif olarak tahmin edilir. Hatalı tahmin.	True Negative (TN) Negatif değer, negatif olarak tahmin edilir. Doğru tahmin.

Bir modelin performansını ölçmek için çok sayıda farklı ölçüm değeri vardır. Doğru metriği seçmek, modelin doğru değerlendirmesi için önemlidir[88].

3.1.1. Accuracy / Doğruluk / Kesinlik

Tüm veriler içinde doğru tahmin edilen veri sayısıdır. Dengesiz veri kümelerinde, tek başına kullanmak yeterli değildir[88, 89].

Accuracy = $(TP + TN) / (TP+FN+FP+TN)$

3.1.2. Recall / Duyarlılık / Geri Çağırma / Sensitivity

Tüm Gerçek Pozitif değerler arasından, doğru tahmin edilen pozitif sayıdır. Pozitif değerin önemli olduğu durumlarda kullanılır. Mümkün olduğunca yüksek olması beklenir. **False Negative** (doğruyu yanlış diye hatalı tahmin etmek) maliyetinin yüksek olduğu durumlarda kullanılan bir metriktir[88, 89].

$Recall= TP / (TP+FN)$

3.1.3. Precision / Hassasiyet

Tüm tahmin edilen Pozitif değerler içindeki gerçek pozitif sayıdır. Precision, True Positive öğesinin doğruluğunu hesaplar. Positive olarak tahmin ettiğimiz değerlerin gerçekten kaç adedinin Positive olduğunu göstermektedir. **False Positive**(yanlışı doğru diye hatalı tahmin etmek) maliyetinin yüksek olduğu durumlarda kullanılır. Mümkün olduğunca yüksek olması gereklidir[88, 89].

$$Precision = TP/(TP + FP)$$

3.1.4. F1-score

F1 puanı, Precision ve Recall arasındaki dengeyi sağlar. Genellikle sınıf dağılımı dengesiz olduğunda kullanılır[88, 89].

$$F1 = 2 * ([precision * recall] / [precision + recall])$$

3.1.5. Özgüllük / Specificity

Tüm Gerçek Negatif değerler arasından doğru tahmin edilen negatif sayıdır. Negatif olan değerler arasından kaç tanesinin doğru tahmin edildiğinin bilgisini verir. Dengesiz veri kümelerinde, negatif değer önemli ise, değerin yüksek olması beklenir.[88, 89].

$$Specificity = TN / (TN + FP)$$

3.1.6. Balanced Accuracy / Dengeli Doğruluk

İkili ve çok sınıflı sınıflandırmada kullanılır. Duyarlılık ve özgüllüğün(specificity) aritmetik ortalamasıdır. Genellikle sınıf dağılımı dengesiz olduğunda kullanılır.

$$Balanced\ Accuracy = (Sensitivity + Specificity) / 2$$

Balanced Accuracy, TN değerleriyle de ilgilenirken, F1-Score TN değerleriyle ilgilenmez. Bu nedenle, TP'ler kadar TN'ler de önemliyse, Balanced Accuracy, F1-Score'den daha iyi bir ölçümdür. TN değerinde bir azalma olduğunda F1-Score hiç değişmez ama Balanced Accuracy sonucunda azalma olur. Bu, F1 puanının pozitif veri noktalarına Balanced Accuracy'den daha fazla öncelik verdiğini göstermektedir.

Sonuç olarak, dengesiz veri kümelerinde; TN'leri dikkate almamız gerektiğinde, Balanced Accuracy F1-Score'den daha iyi sonuç verir. TP'leri dikkate almamız gerektiğinde ise F1-Score daha iyi sonuç verir. Balanced Accuracy ve F1 Score değerlerinin yüksek olması beklenir. Balanced Accuracy değerinin F1-Score'den yüksek olması, TN lerin de çok olduğunu

gösterir. Recall ve Specificity nin de yüksek olması beklenir. TP ve TN lerin ne kadar doğru tahmin edildiğini gösterir[89].

3.1.7. ROC Area

ROC eğrisi, modelin iki şeyi ne kadar iyi ayırt edebileceğini gösterir. AUC, ROC eğrisinin altındaki alandır. Bu puan bize modelin ne kadar iyi performans gösterdiğine dair iyi bir fikir verir. Çok sınıflı veri kümesinde, OvO veya OvR yöntemi kullanılarak ROC alanı ortalama olarak hesaplanır[89, 90].

Bu çalışmada, TN değerleri de, TP değerleri kadar önemlidir. Örneğin, "Kaydı Silindi" durumunda olan bir öğrencinin doğru tahminlenmesi bu çalışmada önemlidir. Bu nedenle, metrik seçimi olarak Balanced Accuracy ve F1-Score kullanıldı. Accuracy, Recall ve Specificity değerleri ayrıca gösterildi. Ayrıca Accuracy dengeli veri kümelerinde, Balanced Accuracy değerine yakındır. Bu çalışmada dengesiz sınıf dağılımı olduğu için, Accuracy değerinin başarı oranında Balanced Accuracy ve F1 Score yükseldiğinde düşüş olması ihtimali vardır.

3.2. Analiz Sonuçları

Bu çalışmada, MÜ FBEM'nün 23.05.2022-286783 etik no kararı ile kullanılan MÜ TF öğrenci veri kümesi ile yapılan çalışmanın analiz sonuçları sunuldu. Tartışma bölümünde, elde edilen sonuçlarla ilgili makine öğrenmesi algoritmaları ve yeniden örnekleme yöntemleri ile ilgili tablolar özetlendi ve karşılaştırmalar yapıldı. Sonuç bölümünde yapılan çalışma özetlenerek, elde edilen sonuçlardan, gelecek çalışmalar için araştırma fikirleri geliştirildi.

Çalışmada kullanılan öğrenci verisine öncelikle veri önişleme işlemleri ve özellik seçimi uygulandı. Python programlama diliyle geliştirilneper yapıldı. Python dilinde bulunan paketler kullanıldı. Şekil 2.12, Şekil 2.13 ve Şekil 2.14'teki model akış şemasındaki yöntemler takip edilerek bulgular elde edildi. Veri Önişleme işlemlerinden sonra elde edilen Hazırlanmış Veriden, öncelikle, 2016 yılında kayıt olan öğrenciler sağlamlık veri kümesi olarak ayrıldı. Daha sonra Repeated Stratified 5*5-fold-CV ile eğitim ve test datası olarak bölünerek 5 tekrar ve 5 kat çapraz geçerleme ile işlem yapıldı. Makine öğrenmesi algoritmaları parametreleri Hiper parametre ayarı yöntemiyle en optimum parametreler elde edilene kadar tekrarlandı. Daha sonra 3 farklı yöntem ile modeller oluşturuldu. 1. yöntem olarak algoritmaların tenel halleri modellendi. 2. yöntem olarak sadece eğitim kümesine yeniden örnekleme yöntemleri kullanılarak, makine öğrenme algoritmaları ile yeni modeller oluşturuldu. 3. yöntemde ise sağlamlık verisi hariç tüm veri kümesine yeniden örneklem uygulanarak yeni modeller

oluşturuldu. Daha sonra bu 3 yöntem, FS sonuçlarına göre 7 özellik ve 5 özellik ile tekrarlandı. Böylece 11 özellik, 7 özellik ve 5 özellik ile modellenen 9 farklı model türleri elde edildi. Ayrıca 9 farklı model türlerinin her birisine sağlamlık kontrolleri yapıldı. Kullanılan makine öğrenme algoritmaları, DT, RF, SVM, LR, KNN, GaussianNB, BernoulliNB, MLP, ET, AdaBoost, GradientBoost, Bagging-SVC, Bagging-DT algoritmaları ile BalancedBagging, BalancedRF, EasyEnsemble modifiye edilmiş topluluk metotlarıdır. Yeniden örnekleme yöntemleri olarak, SMOTE, SMOTENC, SVMSMOTE, ADASYN, BSMOTE1, BSMOTE2, ROS, SMOTEENN ve SMOTETomek yöntemleri kullanıldı. Makine öğrenmesi yöntemlerinin başarısı F1-Score, Balanced Accuracy, Accuracy, Recall, ROC Area, Confusion Matrix performans değerlendirme ölçütlerine göre değerlendirildi.

Tartışma bölümünde oluşturulan tablolara ait değerlerin tümü, her algoritmaya ve yeniden örnekleme yöntemlerine göre tekrarlanan sonuçların ortalamasını içermektedir. Yeniden örnekleme yapılmadan eğitilen makine öğrenmesi modellerinin temel / base sonuçları gösterildi. 1. yöntemde 13 makine öğrenmesi algoritması ve 3 modifiye edilmiş topluluk metotları olmak üzere 16 ayrı model, 2. yöntemde, 13 makine öğrenmesi algoritması * 9 yeniden örnekleme = 117 ayrı model, ve 3. yöntemde, 13 makine öğrenmesi algoritması * 9 yeniden örnekleme = 117 ayrı model olmak üzere toplam 250 farklı model elde edildi.

Daha sonra 250 ayrı model üzerinde sağlamlık testi uygulandı. Sağlamlık kontrolü, elde edilen 250 modele ayrı ayrı 153 örnek içeren sağlamlık verisi verilerek yapıldı. Bu nedenle her model için ayrıca 250 tane ayrı sağlamlık sonuçları elde edildi. Her modelin test F1 Score değerleri ve sağlamlık F1 Score değerleri karşılaştırıldı ve tutarlılık oranları elde edildi. Şekil 2.13'de 1. ve 2. yönteme ait akış şemasının ayrıntılı modellenmesi yapıldı. Şekil 2.14'de 3. yönteme ait akış şemasının ayrıntılı modellemesi yapıldı[29].

Ek olarak, FS aşamasında elde edilen hazırlanmış data olan 1241 satır ve 11 özellik içeren hazırlanmış veri kümesi, Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF, Anova F Test şeklindeki FS yöntemleriyle tekrar ele alındı. Elde edilen sonuçlara göre, 1., 2. ve 3. yöntem, 7 tane en çok kullanılan özellik ve 5 tane en çok kullanılan özellik ile ayrıca 250'şer tane model oluşturuldu. Bu modellere de ayrıca sağlamlık kontrolü yapıldı ve 250'şer tane sağlamlık sonucu elde edildi. Ve ek olarak modellerin test F1 score değerleri ile karşılaştırılarak, tutarlılık oranları elde edildi.

Modifiye edilmiş topluluk metotlarının eğitim (train) ve test sonuçları da ayrıca gösterildi. Modifiye edilmiş topluluk metotlarının parametreleri en iyi sonuç verecek şekilde optimize

edilerek algoritmalar çalıştırıldı. Balanced Bagging algoritmasının örnekleme tekniği olan Random Under Sampling, SMOTE ile değiştirildi.

2. yöntem olan Şekil 2.13'deki gibi, sadece eğitim verisine SMOTE, SMOTENC, SVMSMOTE, ADASYN, BSMOTE1, BSMOTE2, ROS, SMOTEENN ve SMOTETomek örnekleme yöntemleri uygulanarak, makine öğrenmesi algoritmaları yeniden çalıştırıldı. Her algoritmanın çapraz geçerleme sonuçlarının ortalamaları oluşturuldu. Tartışma bölümündeki tablolarda sadece eğitim kümesine uygulanan yeniden örnekleme yöntemleri ile elde edilen değerler de eklendi.

3. yöntem olan Şekil 2.14'deki gibi, öncelikle tüm veri kümesine SMOTE, SMOTENC, SVMSMOTE, ADASYN, BSMOTE1, BSMOTE2, ROS, SMOTEENN ve SMOTETomek örnekleme yöntemleri uygulanarak, makine öğrenmesi algoritmaları yeniden çalıştırıldı. Her algoritmanın çapraz geçerleme sonuçlarının ortalamaları oluşturuldu. Tartışma bölümündeki tablolarda tüm veri kümesine uygulanan yeniden örnekleme yöntemleri ile elde edilen değerler de eklendi.

Son olarak, elde edilen modeller üzerinde sağlamlık ve tutarlılık kontrolü yapıldı.11 özellikle, 7 özellikle ve 5 özellikle yapılan 1., 2. ve 3. yöntemden elde edilen 250'şer model üzerinde yapılan sağlamlık ve tutarlılık kontrolüne ait sonuçların incelenmesi ile tartışmalar bölümündeki tablolar elde edildi.

Nihayet, en başarılı ve en tutarlı modellere karar verilerek, pilot bir karar destek sistemi web uygulaması yapılarak, MÜ TF öğretim üyelerinin kullanımına sunuldu.

3.3. Tartışma ve Karşılaştırma

3.3.1. 11 özellik ile oluşturulan modellere ait sonuçlar

Tablo 3.2'deki değerler incelendiğinde, RF ve Bagging_DT algoritmalarının yeniden örnekleme yapılmadığında, en başarılı sonucu verdiği görüldü. Sadece eğitim verisine yeniden örneklem yapıldığında ET, DT ve BernoulliNB hariç, diğer sonuçlarda artış olmadı ancak tüm veri kümesine yeniden örneklem uygulandığında, RF, GradientBoost, ET, Adaboost, Bagging_SVC, DT, SVM, KNN, BernoulliNB algoritmalarının sonuçlarında ciddi artış oldu. Bu sonuçlar sağlamlık testleriyle karşılaştırıldığında en tutarlı olan model, bu tez çalışmasında en başarılı model olarak seçildi. Ayrıca ROS ve SMOTEENN yönteminin başarı oranlarında diğer yeniden örnekleme yöntemlerine göre yüksek artış yaptı. SVMSMOTE en çok başarının artmasını sağlayan yöntem oldu.

Tablo 3.2. 11 özellik içeren veri kümesi modeli F1 Score değerleri

	1. Yöntem		**2. Yöntem: Sadece Eğitim Kümesine Yeniden Örnekleme**		**3. Yöntem: Bütün Veriye Yeniden Örnekleme**	
Model	**Yeniden Örnekleme**	**Test F1 Score**	**Yeniden Örnekleme**	**Test F1 Score**	**Yeniden Örnekleme**	**Test F1 Score**
RF	None	**0.8341**	ROS	0.8263	**ROS**	**0.9935**
Bagging_DT	None	**0.8302**	SVMSMOTE	0.8203	SVMSMOTE	0.7884
MLP	None	0.8255	SVMSMOTE	0.7982	SVMSMOTE	0.6951
GradientBoost	None	0.8250	SVMSMOTE	0.8194	**SVMSMOTE**	**0.8349**
LR	None	0.8188	SVMSMOTE	0.7451	SVMSMOTE	0.7193
ET	None	0.8187	**ROS**	**0.8196**	**ROS**	**0.9917**
BalancedBagging	None	0.8154	-	-	-	-
AdaBoost	None	0.7748	ROS	0.7734	**ROS**	**0.9542**
Bagging_SVC	None	0.7712	SVMSMOTE	0.7501	**SMOTEENN**	**0.9697**
DT	None	0.7698	**SMOTE2**	**0.7803**	**ROS**	**0.8790**
SVM	None	0.7676	SVMSMOTE	0.7568	**SMOTEENN**	**0.9862**
KNN	None	0.7577	ROS	0.7522	**SMOTEENN**	**0.9883**
GaussianNB	None	0.6976	SVMSMOTE	0.5776	SVMSMOTE	0.6324
BalancedRF	None	0.6410	-	-	-	-
EasyEnsemble	None	0.5908	-	-	-	-
BernoulliNB	None	0.5082	**SVMSMOTE**	**0.6521**	**SVMSMOTE**	**0.6824**

Tablo 3.3'de gösterilen sağlamlık (robustness) kontrolü F1 Score değerlerinde, 1. yöntem olan yeniden örneklem yapılmadığında en başarılı değerler MLP, ET, RF ve Bagging_DT algoritmaları oldu. 2. ve 3. yöntemlerde yapılan yeniden örnekleme sonuçlarına bakıldığında, yeniden örnekleme yapılmayan sonuçlara yakın değerler elde edildiği görüldü. Bu durumda, sağlamlık kontrolünde, yeniden örneklem yapılmadan base modellerin kullanılabileceği görüldü.

Tablo 3.3. 11 özellik içeren veri kümesi sağlamlık kontrolü F1 Score sonuçları

	1. Yöntem		**2. Yöntem: Sadece Eğitim Kümesine Yeniden Örneklem**		**3. Yöntem: Bütün Veriye Yeniden Örneklem**	
Model	**Yeniden örnekleme**	**Test F1 Score**	**Yeniden Örnekleme**	**Test F1 Score**	**Yeniden Örnekleme**	**Test F1 Score**
Bagging_DT	None	0.7719	SVMSMOTE	0.7531	SVMSMOTE	0.7468
MLP	None	0.7714	BSMOTE1	0.7412	**ROS**	**0.7728**
ET	None	0.7704	**ROS**	**0.7713**	**ROS**	**0.7713**
RF	None	0.7685	**ROS**	**0.7705**	**ROS**	**0.7709**
Bagging_SVC	None	0.7612	SVMSMOTE	0.7465	SMOTE	0.7388
BalancedBagging	None	0.7582	-	-	-	-
SVM	None	0.7508	**ADASYN**	**0.7524**	**SMOTE**	**0.7514**
GradientBoost	None	0.7408	SVMSMOTE	0.7179	SMOTENC	0.7149
LR	None	0.7282	SVMSMOTE	0.6420	SVMSMOTE	0.6617
KNN	None	0.7180	ROS	0.7147	ROS	0.7057
DT	None	**0.7089**	**BSMOTE2**	**0.7135**	**ADASYN**	**0.7266**
AdaBoost	None	0.6864	**ROS**	**0.7299**	**ROS**	**0.7101**
GaussianNB	None	0.6366	SVMSMOTE	0.5133	SVMSMOTE	0.4832
BalancedRF	None	0.4943	-	-	-	-
EasyEnsemble	None	0.4468	-	-	-	-
BernoulliNB	None	0.3989	**BSMOTE2**	**0.7017**	**BSMOTE2**	**0.6735**

3.3.1.1. 1. yönteme ait modellerin tutarlılık değerleri

Tablo 3.4.'de oranları verilen tutarlılık değeri, bir algoritma en başarılı sonucu verse de, daha önce hiç görmediği veri kümesinde test sonuçlarına yakın değer vermesi, modelin tutarlılığını belirler.

Tablo 3.2. incelendiğinde, en başarılı algoritmanın %83.41 oranında RF algoritması olduğu görülmekle birlikte sağlamlık kontrolünde Tablo 3.4'te görüldüğü gibi %7.9 oranında bir farkla tutarlılık değeri bulunmaktadır. RF Algoritmasına en yakın %83.02 F1 başarı oranı ile Bagging_DT algoritmasıdır ve %7.02 oranında bir farkla tutarlılık değeri bulunmaktadır. Bu sonuçlara bakıldığında, Bagging_DT, RF algoritmasından daha tutarlı sonuç vermiştir. %75 değeri üzerindeki F1 Score larına sahip algoritmalar arasında ise en tutarlı değeri %1.3 oranında bir farkla Bagging_SVC modeli vermiştir ancak bu modelin F1 Score değeri %77 dir.

Tablo 3.4. 11 özellik ve 1. yöntem ile oluşturulan modellerin tutarlılık oranları

Model	Yeniden Örnekleme	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası
Bagging_SVC	None	0.7712	0.7612	-0,01296680498	-1,296680498	1
SVM	None	0.7676	0.7508	-0,02188639917	-2,188639917	2
KNN	None	0.7577	0.7180	-0,05239540715	-5,239540715	3
ET	None	0.8187	0.7704	-0,05899596922	-5,899596922	4
MLP	None	0.8255	0.7714	-0,06553603876	-6,553603876	5
BalancedBagging	None	0.8154	0.7582	-0,07014961982	-7,014961982	6
Bagging_DT	None	0.8302	0.7719	-0,0702240424	-7,02240424	7
RF	None	0.8341	0.7685	-0,07864764417	-7,864764417	8
DT	None	0.7698	0.7089	-0,07911145752	-7,911145752	9
GaussianNB	None	0.6976	0.6366	-0,08744266055	-8,744266055	10
GradientBoost	None	0.8250	0.7408	-0,1020606061	-10,20606061	11
LR	None	0.8188	0.7282	-0,1106497313	-11,06497313	12
AdaBoost	None	0.7748	0.6864	-0,1140939597	-11,40939597	13
BernoulliNB	None	0.5082	0.3989	-0,215072806	-21,5072806	14
BalancedRF	None	0.6410	0.4943	-0,2288611544	-22,88611544	15
EasyEnsemble	None	0.5908	0.4468	-0,2437373053	-24,37373053	16

3.3.1.2. 2. yönteme ait modellerin tutarlılık değerleri

2. yöntem olan, sadece eğitim veri kümesine yeniden örneklem yöntemleri uygulanarak elde edilen modellerin tutarlılık oranları Tablo 3.5'te gösterildi.

Tablo 3.5. 11 özellik ve 2. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden Örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
None	RF	8.341	7.685	-0,079	-7,9	63	7	1
None	Bagging_DT	8.302	7.719	-0,070	-7	60	1	2
ROS	RF	8.263	7.705	-0,068	-6,8	56	4	3
None	MLP	8.255	7.714	-0,066	-6,6	55	2	4
None	GradientBoost	8.250	7.408	-0,102	-10,2	80	35	5
SVMSMOTE	Bagging_DT	8.203	7.531	-0,082	-8,2	66	23	6
SVMSMOTE	RF	8.202	7.648	-0,068	-6,8	57	8	7
ROS	ET	8.196	7.713	-0,059	-5,9	50	3	8
SVMSMOTE	GradientBoost	8.194	7.179	-0,124	-12,4	96	52	9
None	LR	8.188	7.282	-0,111	-11,1	86	48	10
None	ET	8.187	7.704	-0,059	-5,9	51	5	11
BSMOTE2	RF	8.173	7.597	-0,070	-7	61	19	12
BSMOTE2	GradientBoost	8.170	6.947	-0,150	-15	110	69	13
None	BalancedBagging	8.154	7.582	-0,070	-7	59	21	14

(Tablo 3.5.'in devamı sonraki sayfadadır.)

Tablo 3.5 (devam). 11 özellik ve 2. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden Örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
SMOTENC	GradientBoost	8.152	6.836	-0,161	-16,1	112	77	15
BSMOTE1	RF	8.143	7.629	-0,063	-6,3	54	13	16
SMOTETomek	RF	8.136	7.577	-0,069	-6,9	58	22	17
SMOTENC	RF	8.132	7.402	-0,090	-9	71	37	18
SMOTE	GradientBoost	8.127	6.959	-0,144	-14,4	108	67	19
SMOTETomek	GradientBoost	8.126	6.964	-0,143	-14,3	107	66	20
BSMOTE2	Bagging_DT	8.116	7.305	-0,100	-10	77	45	21
ADASYN	GradientBoost	8.113	7.047	-0,131	-13,1	99	60	22
SMOTE	RF	8.107	7.604	-0,062	-6,2	53	18	23
BSMOTE1	GradientBoost	8.104	6.967	-0,140	-14	106	65	24
ADASYN	RF	8.098	7.597	-0,062	-6,2	52	20	25
BSMOTE2	ET	8.042	7.610	-0,054	-5,4	47	17	26
SVMSMOTE	ET	8.040	7.628	-0,051	-5,1	45	14	27
SMOTE	ET	7.992	7.636	-0,045	-4,5	42	12	28
ADASYN	ET	7.984	7.643	-0,043	-4,3	40	9	29
SVMSMOTE	MLP	7.982	7.702	-0,035	-3,5	37	6	30
BSMOTE1	ET	7.976	7.640	-0,042	-4,2	39	10	31
SMOTENC	ET	7.975	7.408	-0,071	-7,1	62	36	32
ROS	MLP	7.974	7.612	-0,045	-4,5	43	15	33
SMOTETomek	ET	7.955	7.638	-0,040	-4	38	11	34
BSMOTE1	Bagging_DT	7.839	7.019	-0,105	-10,5	81	61	35
BSMOTE2	DT	7.803	7.135	-0,086	-8,6	68	55	36
SVMSMOTE	DT	7.758	7.123	-0,082	-8,2	65	56	37
None	AdaBoost	7.748	6.864	-0,114	-11,4	90	74	38
ROS	AdaBoost	7.734	7.299	-0,056	-5,6	49	47	39
None	Bagging_SVC	7.712	7.612	-0,013	-1,3	32	16	40
None	DT	7.698	7.089	-0,079	-7,9	64	57	41
ADASYN	Bagging_DT	7.695	6.792	-0,117	-11,7	93	80	42
SMOTETomek	DT	7.692	6.858	-0,108	-10,8	83	75	43
BSMOTE1	DT	7.684	6.955	-0,095	-9,5	73	68	44
None	SVM	7.676	7.508	-0,022	-2,2	36	27	45
ADASYN	DT	7.662	6.830	-0,109	-10,9	84	78	46
SMOTETomek	Bagging_DT	7.660	6.934	-0,095	-9,5	72	70	47
SMOTENC	DT	7.656	6.300	-0,177	-17,7	123	103	48
BSMOTE2	AdaBoost	7.653	6.649	-0,131	-13,1	98	87	49
SMOTENC	AdaBoost	7.646	6.327	-0,173	-17,3	120	102	50
SMOTETomek	AdaBoost	7.635	6.707	-0,122	-12,2	95	84	51
SVMSMOTE	AdaBoost	7.629	6.987	-0,084	-8,4	67	64	52

(Tablo 3.5.'in devamı sonraki sayfadadır.)

Tablo 3.5 (devam). 11 özellik ve 2. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden Örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
ADASYN	AdaBoost	7.611	6.865	-0,098	-9,8	75	73	53
BSMOTE1	AdaBoost	7.611	6.857	-0,099	-9,9	76	76	54
SMOTE	Bagging_DT	7.582	6.905	-0,089	-8,9	70	71	55
None	KNN	7.577	7.180	-0,052	-5,2	46	51	56
SMOTE	AdaBoost	7.571	6.802	-0,102	-10,2	78	79	57
SVMSMOTE	SVM	7.568	7.501	-0,009	-0,9	29	28	58
BSMOTE2	SVM	7.567	7.496	-0,009	-0,9	30	29	59
BSMOTE1	SVM	7.560	7.460	-0,013	-1,3	33	32	60
SMOTENC	Bagging_DT	7.549	6.011	-0,204	-20,4	127	108	61
SMOTEENN	GradientBoost	7.543	6.254	-0,171	-17,1	119	104	62
SMOTE	DT	7.540	6.492	-0,139	-13,9	103	94	63
SMOTENC	SVM	7.535	7.492	-0,006	-0,6	26	30	64
SMOTE	SVM	7.530	7.521	-0,001	-0,1	23	25	65
ROS	KNN	7.522	7.147	-0,050	-5	44	54	66
ADASYN	SVM	7.520	7.524	0,001	0,1	19	24	67
SMOTETomek	SVM	7.520	7.521	0,000	0	22	26	68
SVMSMOTE	Bagging_SVC	7.501	7.465	-0,005	-0,5	24	31	69

Tablo 3.5'de görüldüğü üzere, en başarılı F1 Score değerlerine sahip algoritmalara, Test sırası, en başarılı Sağlamlık F1 Score değerine Sağlamlık sırası ve Tutarlılık oranlarına göre Tutarlılık sırası eklendi. İlk olarak %75 ve üzeri sonuçları olan 69 model alındı. Diğerleri başarısız kabul edildi. Daha sonra sağlamlık testinden %75 ve üzeri sonuçları olan 28 model alındı. ET, RF, SVM, MLP ve Bagging_DT algoritmaları kaldı. Bunlardan None, SVMSMOTE, ADASYN, SMOTE, SMOTETomek yeniden örnekleme yöntemleri kaldı. Daha sonra Robustness sırasında en başarılı ilk 10 tane model alındı ve diğerleri elendi. Bagging_DT, MLP, ET, RF algoritmaları kalmış oldu. İlk 7 tanesinin Robustness başarı oranının %77 olduğu görüldü, bu nedenle ilk 7 alındı. Geliştirilen Web uygulamada kullanılmasına karar verilecek en tutarlı modeller Tablo 3.6'da gösterildi.

Tablo 3.6. 11 özellik içeren 1. ve 2. yöntem ile elde edilen modellerden web uygulamasında kullanılabilecek en tutarlı ve en başarılı algoritmalar

Yeniden Örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Tutarlılık Sırası
SVMSMOTE	MLP	0.7982	0.7702	-0,035	37
ROS	ET	0.8196	0.7713	-0,059	50
None	ET	0.8187	0.7704	-0,059	51
None	MLP	0.8255	0.7714	-0,066	55
ROS	RF	0.8263	0.7705	-0,068	56
None	Bagging_DT	0.8302	0.7719	-0,070	60
None	RF	0.8341	0.7685	-0,079	63

3.3.1.3. 3. yönteme ait modellerin tutarlılık değerleri

3. yöntem olan, sağlamlık verisi hariç tüm veri kümesine yeniden örneklem yöntemleri uygulanarak elde edilen modellerin tutarlılık oranları Tablo 3.7'te gösterildi.

Tablo 3.7. 11 özellik ve 3. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
ROS	RF	9.935	7.709	-0,224	-22,4	102	7	1
ROS	ET	9.917	7.713	-0,222	-22,2	98	6	2
SMOTEENN	KNN	9.883	5.293	-0,464	-46,4	133	117	3
SMOTEENN	SVM	9.862	7.216	-0,268	-26,8	118	52	4
SMOTEENN	ET	9.812	6.914	-0,295	-29,5	123	65	5
SMOTEENN	Bagging_SVC	9.697	6.866	-0,292	-29,2	120	69	6
SMOTEENN	RF	9.607	6.863	-0,286	-28,6	119	71	7
ROS	AdaBoost	9.542	7.101	-0,256	-25,6	115	57	8
ROS	KNN	9.534	7.057	-0,260	-26	116	59	9
SMOTETomek	ET	9.525	7.634	-0,199	-19,9	85	14	10
ADASYN	ET	9.519	7.650	-0,196	-19,6	79	12	11
BSMOTE1	ET	9.507	7.631	-0,197	-19,7	83	16	12
SMOTE	ET	9.499	7.632	-0,197	-19,7	81	15	13
ROS	SVM	9.496	7.190	-0,243	-24,3	110	54	14
SMOTENC	ET	9.483	7.361	-0,224	-22,4	100	43	15
BSMOTE1	RF	9.436	7.579	-0,197	-19,7	82	22	16
SMOTETomek	SVM	9.423	7.448	-0,210	-21	93	32	17
SMOTE	SVM	9.415	7.514	-0,202	-20,2	88	28	18
ADASYN	SVM	9.406	7.485	-0,204	-20,4	89	30	19

(Tablo 3.7.'in devamı sonraki sayfadadır.)

Tablo 3.7 (devam). 11 özellik ve 3. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
SMOTETomek	RF	9.373	7.555	-0,194	-19,4	75	24	20
SVMSMOTE	ET	9.370	7.659	-0,183	-18,3	67	11	21
BSMOTE2	ET	9.363	7.604	-0,188	-18,8	70	20	22
SMOTENC	SVM	9.355	7.423	-0,207	-20,7	90	34	23
BSMOTE1	SVM	9.347	7.390	-0,209	-20,9	92	36	24
SMOTE	RF	9.324	7.544	-0,191	-19,1	72	25	25
ADASYN	RF	9.322	7.620	-0,183	-18,3	66	17	26
BSMOTE2	SVM	9.313	7.438	-0,201	-20,1	87	33	27
SMOTENC	RF	9.313	7.308	-0,215	-21,5	96	47	28
SMOTETomek	KNN	9.278	6.408	-0,309	-30,9	130	91	29
BSMOTE1	KNN	9.262	6.440	-0,305	-30,5	126	90	30
ADASYN	KNN	9.251	6.400	-0,308	-30,8	128	92	31
SVMSMOTE	SVM	9.231	7.387	-0,200	-20	86	40	32
SVMSMOTE	RF	9.230	7.643	-0,172	-17,2	60	13	33
BSMOTE2	RF	9.225	7.569	-0,180	-18	62	23	34
SMOTENC	KNN	9.211	6.195	-0,327	-32,7	132	101	35
SMOTE	KNN	9.208	6.509	-0,293	-29,3	121	89	36
BSMOTE1	Bagging_SVC	9.148	7.350	-0,197	-19,7	80	44	37
ROS	Bagging_SVC	9.139	7.012	-0,233	-23,3	106	61	38
SVMSMOTE	KNN	9.133	6.700	-0,266	-26,6	117	81	39
BSMOTE2	Bagging_SVC	9.093	7.374	-0,189	-18,9	71	42	40
BSMOTE2	KNN	9.091	6.306	-0,306	-30,6	127	95	41
SMOTETomek	Bagging_SVC	9.088	7.345	-0,192	-19,2	73	45	42
ADASYN	Bagging_SVC	9.085	7.317	-0,195	-19,5	76	46	43
SMOTENC	Bagging_SVC	9.065	7.275	-0,197	-19,7	84	50	44
SMOTE	Bagging_SVC	9.061	7.388	-0,185	-18,5	69	38	45
SVMSMOTE	Bagging_SVC	9.040	7.386	-0,183	-18,3	68	41	46
SMOTEENN	AdaBoost	8.845	5.992	-0,323	-32,3	131	107	47
ROS	DT	8.790	6.836	-0,222	-22,2	99	73	48
BSMOTE1	AdaBoost	8.535	6.867	-0,195	-19,5	77	68	49
SVMSMOTE	AdaBoost	8.429	6.893	-0,182	-18,2	65	66	50
SVMSMOTE	GradientBoost	8.349	7.031	-0,158	-15,8	53	60	51
SMOTEENN	DT	8.342	6.270	-0,248	-24,8	113	97	52

(Tablo 3.7.'in devamı sonraki sayfadadır.)

Tablo 3.7 (devam). 11 özellik ve 3. yöntem %75 üzeri F1 Score ve tutarlılık oranları

Yeniden örnekleme	Model	Test F1 Score	Sağlamlık F1 Score	Oran	Yüzde	Tutarlılık Sırası	Sağlamlık Sırası	Test Sırası
None	RF	8.341	7.685	-0,079	-7,9	34	10	53
SMOTENC	AdaBoost	8.333	5.860	-0,297	-29,7	124	109	54
SMOTETomek	AdaBoost	8.328	6.596	-0,208	-20,8	91	88	55
SMOTE	AdaBoost	8.314	6.718	-0,192	-19,2	74	79	56
None	Bagging_DT	8.302	7.719	-0,070	-7	32	4	57
ADASYN	AdaBoost	8.280	6.883	-0,169	-16,9	58	67	58
BSMOTE2	AdaBoost	8.276	6.654	-0,196	-19,6	78	84	59
None	MLP	8.255	7.714	-0,066	-6,6	29	5	60
BSMOTE1	DT	8.253	6.933	-0,160	-16	56	64	61
None	GradientBoost	8.250	7.408	-0,102	-10,2	42	35	62
SVMSMOTE	DT	8.240	6.962	-0,155	-15,5	51	63	63
None	LR	8.188	7.282	-0,111	-11,1	43	48	64
None	ET	8.187	7.704	-0,059	-5,9	28	8	65
BSMOTE1	GradientBoost	8.187	6.846	-0,164	-16,4	57	72	66
None	BalancedBagging	8.154	7.582	-0,070	-7	30	21	67
BSMOTE2	DT	8.129	6.662	-0,180	-18	63	83	68
SMOTETomek	DT	8.083	6.718	-0,169	-16,9	59	80	69
SMOTEENN	GradientBoost	8.075	6.609	-0,182	-18,2	64	87	70
BSMOTE2	GradientBoost	8.050	6.774	-0,159	-15,9	55	75	71
SMOTENC	DT	8.048	6.094	-0,243	-24,3	109	105	72
ADASYN	DT	8.039	7.266	-0,096	-9,6	41	51	73
SMOTETomek	GradientBoost	7.947	7.009	-0,118	-11,8	45	62	74
SVMSMOTE	Bagging_DTe	7.884	7.468	-0,053	-5,3	27	31	75
SMOTE	DT	7.840	6.183	-0,211	-21,1	94	103	76
SMOTENC	GradientBoost	7.817	7.149	-0,085	-8,5	38	56	77
SMOTE	GradientBoost	7.812	6.748	-0,136	-13,6	50	77	78
None	AdaBoost	7.748	6.864	-0,114	-11,4	44	70	79
None	Bagging_SVC	7.712	7.612	-0,013	-1,3	19	18	80
None	DT	7.698	7.089	-0,079	-7,9	35	58	81
None	SVM	7.676	7.508	-0,022	-2,2	20	29	82
ADASYN	GradientBoost	7.670	6.674	-0,130	-13	49	82	83
None	KNN	7.577	7.180	-0,052	-5,2	26	55	84
BSMOTE1	Bagging_DT	7.507	6.613	-0,119	-11,9	46	86	85

Tablo 3.7'de görüldüğü üzere, en başarılı F1 Score değerlerine sahip algoritmalara, Test sırası, en başarılı Sağlamlık F1 Score değerine Sağlamlık sırası ve Tutuarlılık oranlarına göre Tutarlılık sırası eklendi. İlk olarak %75 ve üzeri sonuçları olan 85 model alındı. Diğerleri başarısız kabul edildi. Daha sonra sağlamlık testinden %75 ve üzeri sonuçları olan 22 model alındı. RF, ET, SVM, Bagging_DT, MLP, Bagging_SVC ve BalancedBagging algoritmaları kaldı. None, ROS, BSMOTE2, SMOTETomek, ADASYN, BSMOTE1, SMOTE, SVMSMOTE yeniden örnekleme yöntemleri kaldı. Daha sonra Robustness sırasında en başarılı ilk 10 tane model alındı ve diğerleri elendi. Bagging_DT, MLP, ET, RF algoritmaları kaldı. İlk 8 tanesinin Robustness başarı oranının %77 olduğu görüldü, bu nedenle ilk 8 tanesi alındı. Geliştirilen Web uygulamasında kullanılmasına karar verilecek en tutarlı modeller Tablo 3.8'de gösterildi.

Tablo 3.8. 11 özellik içeren 1. ve 3. yöntem ile elde edilen modellerden web uygulamasında kullanılabilecek en tutarlı ve en başarılı algoritmalar

	Model	**Test F1 Score**	**Sağlamlık F1 Score**	**Oran**	**Tutarlılık Sırası**
None	**ET**	**0.8187**	**0.7704**	-0,059	28
None	**MLP**	**0.8255**	**0.7714**	-0,066	29
None	**Bagging_DT**	**0.8302**	**0.7719**	-0,070	32
None	**RF**	**0.8341**	**0.7685**	-0,079	34
SVMSMOTE	ET	0.9370	0.7659	-0,183	67
ADASYN	ET	0.9519	0.7650	-0,196	79
ROS	ET	0.9917	0.7713	-0,222	98
ROS	RF	0.9935	0.7709	-0,224	102

3.3.1.4. Sağlamlık kontrolü sonrası modele karar verilmesi

Tablo 3.6 ve Tablo 3.8'de görüldüğü üzere, yeniden örnekleme yönteminin tüm veri kümesine uygulandığı ve sadece eğitim verisine uygulandığı iki farklı senaryoda, Sağlamlık F1 Score bilgilerinin benzer algoritmalarda benzer sonuçlar verdiği ve en tutarlı olan MLP, ET, Bagging_DT ve RF algoritmalarında ise %77 değerinde olduğu görüldü. Yeniden örnekleme yöntemi tüm veri kümesine uygulandığında, test F1-Score değerlerinde ciddi artış olmasına rağmen, sağlamlık ve tutarlılık olarak, yeniden örnekleme yöntemi sadece eğitim verisine uygulandığında daha tutarlı olduğu görüldü. Bu durum, orjinal veri kümesinin, sentetik veri oluşturularak elde edilen veri kümesinden daha doğru sonuçlar verdiğini ve daha tutarlı olduğunu gösterir. Veri gürültüsünün de göstergesidir. Ayrıca, hem CPU, hem dosya sistem gereksinimi olarak, maliyet tüm veriye yeniden örneklem yöntemi uygulandığında arttı.

Bundan dolayı, tutarlılık, maliyet ve zaman açısından tablo Tablo 3.8'deki 11 özellikle oluşturulan modellerden yeniden örneklem yapılmamış ve en başarılı ve en tutarlı ilk 4 modelin web uygulamasında kullanılmasına karar verildi.

3.3.2. Özellik seçimi ve oluşturulan modellerin sonuçları

Tablo 3.9 ve Tablo 3.10'da görüldüğü üzere, Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF ve Anova F testleri ile özellik sayısı 7 iken ve 5 iken özellikler önem sırasına göre karşılaştırıldı. Değerler incelendiğinde, Cinsiyet, DoğumYeriBolge gibi demografik bilgilerin en az etki eden özellikler olduğu görüldü. Ayrıca Osym Okul Birincisi ve Lise Okul Türü gibi lise akademik bilgileri de en az etkili olan özellikler olduğu görüldü. Bu değerlere göre, Şekil 2.12.'deki gibi 1., 2. ve 3. yöntemler 7 özellikli ve 5 özellikli olarak çalıştırılarak hem test sonuçları hem de sağlamlık sonuçları elde edildi ve karşılaştırma tabloları oluşturuldu.

Tablo 3.9. Özellik sayısı 7 olduğunda karşılaştırma tablosu

	Özellik	Pearson	Chi-2	Mutual - Info	RFE	Logistics	RF	Anova F Test	Toplam
1	YerBolumTercihSirasi	Hayır	Evet	Evet	Evet	Evet	Evet	Evet	6
2	MtokOgrencisi	Evet	Evet	Evet	Evet	Evet	Hayır	Evet	6
3	BirimAdi	Evet	Evet	Evet	Evet	Evet	Hayır	Evet	6
4	YerPuaniSirasi	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	5
5	MarmaraUniYerlestirmeSirasi	Evet	Evet	Hayır	Evet	Hayır	Evet	Evet	5
6	LiseNotu	Evet	Hayır	Evet	Evet	Hayır	Evet	Evet	5
7	KayitYili	Evet	Hayır	Evet	Hayır	Hayır	Evet	Evet	4
8	LiseOkulTuru	Evet	Evet	Evet	Hayır	Hayır	Hayır	Hayır	3
9	osymOkulBirincisi	Hayır	Evet	Hayır	Evet	Hayır	Hayır	Hayır	2
10	Cinsiyet	Hayır	Hayır	Hayır	Evet	Hayır	Hayır	Hayır	1
11	DogumYeriBolge	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	0

Tablo 3.9 ve Tablo 3.10 oluşturulurken, Python diliyle yazılan sklearn.feature_selection kütüphanesi kullanıldı. Özellikleri önem sırasına göre belirlemek için karşılaştırmalı kullanılan 7 model oluşturulup, egitildikten sonra, get_support() fonksiyonları kullanılarak, hangi özelliklerin özellik seçimi metodu için desteklendiğinin bilgisi elde edildi ve tablolar oluşturuldu.

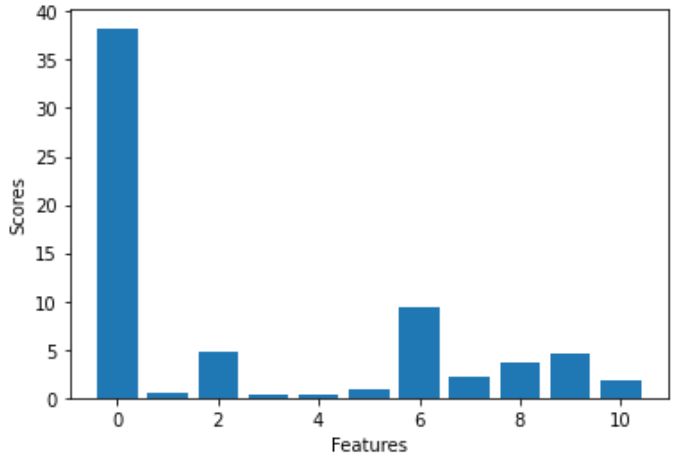
Tablo 3.10. Özellik sayısı 5 olduğunda karşılaştırma tablosu

	Özellik	Pearson	Chi-2	Mutual - Info	RFE	Logistics	RF	Anova F Test	Toplam
1	MtokOgrencisi	Evet	Evet	Evet	Evet	Evet	Hayır	Evet	6
2	YerBolumTercihSirasi	Hayır	Evet	Hayır	Evet	Evet	Evet	Evet	5
3	LiseNotu	Evet	Hayır	Evet	Evet	Hayır	Evet	Evet	5
4	BirimAdi	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	5
5	YerPuaniSirasi	Evet	Hayır	Evet	Hayır	Hayır	Evet	Evet	4
6	KayitYili	Evet	Hayır	Evet	Hayır	Hayır	Evet	Evet	4
7	MarmaraUniYerlestirme Sirasi	Hayır	Hayır	Hayır	Evet	Hayır	Evet	Hayır	2
8	osymOkulBirincisi	Hayır	Evet	Hayır	Hayır	Hayır	Hayır	Hayır	1
9	LiseOkulTuru	Hayır	Evet	Hayır	Hayır	Hayır	Hayır	Hayır	1
10	DogumYeriBolge	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	0
11	Cinsiyet	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	0

Chi-2 özellik seçim modeline göre özelliklerin değerleri Tablo 3.11'deki gibi ve görsel hâli Şekil 3.1.'deki gibidir.

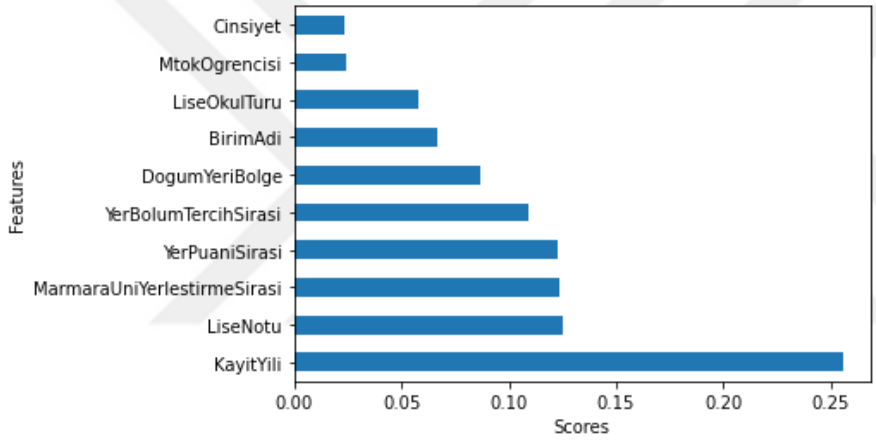
Tablo 3.11. chi2 ile yapılan özellik seçimi değerleri

	Specs	Score
0	MtokOgrencisi	38.276968
6	LiseOkulTuru	9.415682
2	BirimAdi	4.858737
9	osymOkulBirincisi	4.496603
8	YerBolumTercihSirasi	3.621819
7	YerPuaniSirasi	2.145171
10	MarUniYerlestirmeSirasi	1.817568
5	LiseNotu	0.800060
1	KayitYili	0.440568
4	DogumYeriBolge	0.374759
3	Cinsiyet	0.285156



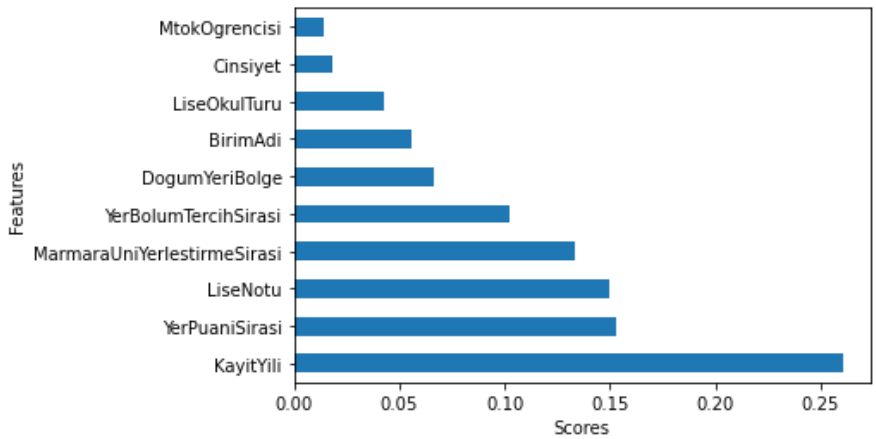
Şekil 3.1. chi-2 ile yapılan özellik seçimi değerleri

ET feature importance bilgileri Şekil 3.2.'deki gibidir.



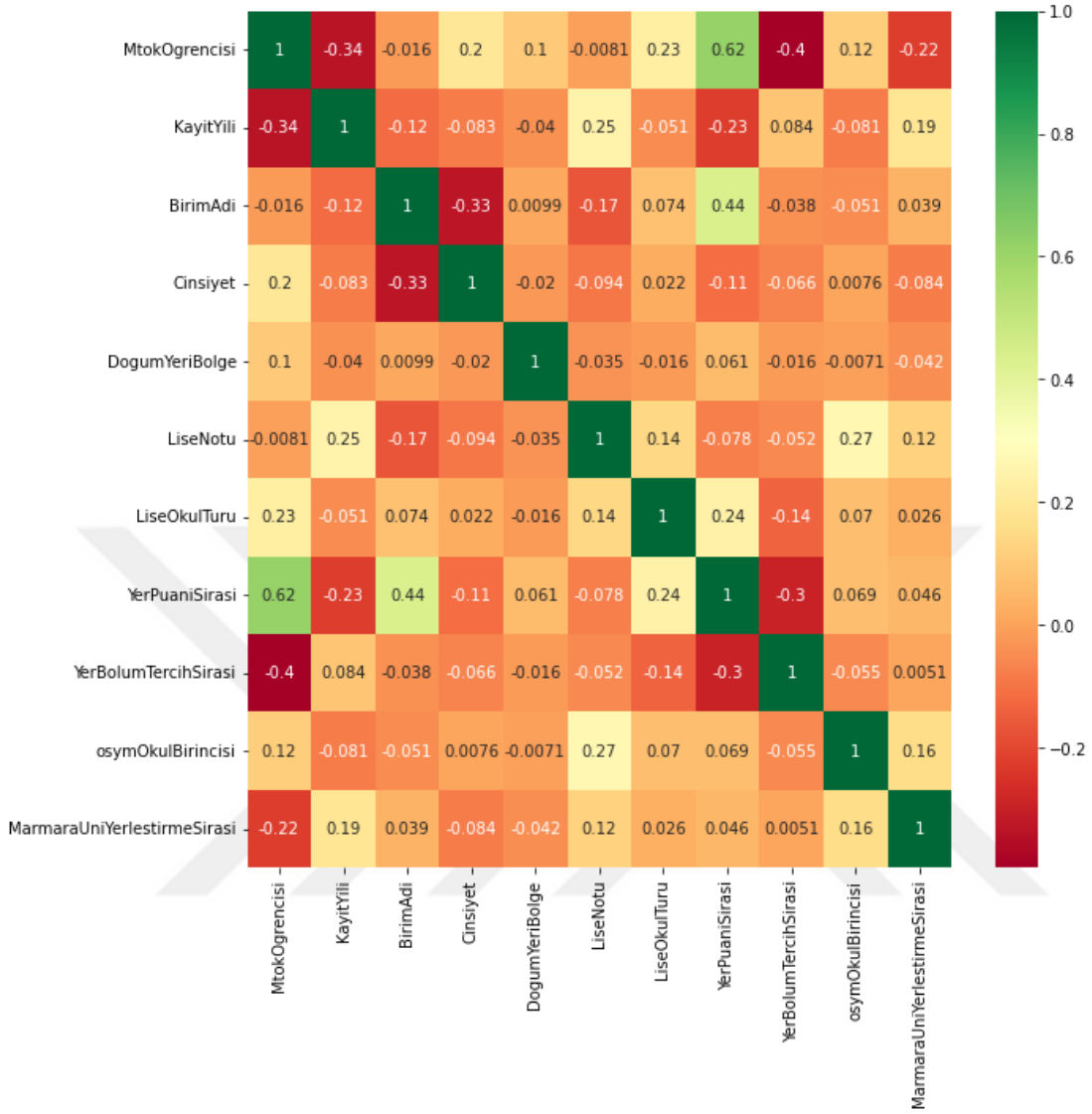
Şekil 3.2. ET ile yapılan özellik seçimi değerleri

RF feature importance bilgileri grafiği Şekil 3.3.'teki gibidir.



Şekil 3.3. RF ile yapılan özellik seçimi değerleri

Ek olarak özelliklerin korelasyon bilgileri Şekil 3.4.'teki gibidir.



Şekil 3.4. 11 özelliğin korelasyon haritası

Korelasyon haritasına bakıldığında aşağıdaki düşünceler elde edilebilir:

- MtokÖğrenci'lerinin kayıt kontenjanları farklı olmaktadır. Ek puan aldıkları için ek kontenjan ile giriş yapabilmektedirler.
- Marmara üniversitesi yerleştirme sırası, kişinin kendisiyle ilgilidir, başarısıyla ilgili değildir. Bu üniversiteyi, evi yakın olduğu için gibi özel sebeplerle birinci sırada tercih etmiş olabilir. Bu da bu öğrencinin başarısıyla doğrudan ilişkilidir diyemeyiz.
- Yerleştirme bölüm tercih sırası ve yerleştirme puan sırası arasında ters ilişki görüldü. Bu da ÖSYM'de yüksek puan alan öğrencilerin, ilk olarak başka üniversitelerdeki başka bölümleri tercih ettikleri, başarı sıralaması düştükçe ise TF bölümlerinin tercih edildiği söylenebilir.

- BirimAdı ve Cinsiyet arasında da ters bir ilişki olduğu görüldü. Bu da bazı bölümlerde erkeklerin bazı bölümlerde de kadınların yoğunlaştığını söyleyebiliriz.
- MtokÖğrenci ve bölüm tercih sırası arasında da ters ilişki olduğu görüldü. Bu da, MTOK öğrencilerinin TF bölümlerini öncelikli olarak tercih ettiğini söyleyebiliriz.
- Yerleştirme Puan sırası ile MtokÖğrenci arasında pozitif ilişki görüldü. Başarı puanı düşük olan MTOK öğrencilerinin, TF bölümlerini tercih ettiğini söyleyebiliriz.
- Lise Notu ve okul birincisi arasında pozitif ilişki olduğu görüldü. Lise notu arttıkça başarının artması beklenen bir durumdur.
- MtokÖğrencisi ile KayıtYılı arasında ters ilişki olduğu görüldü.

FS işleminde, Tablo 3.9 ve Tablo 3.10'daki değerlere göre DogumYeriBolge, Cinsiyet, LiseOkulTuru ve osymOkulBirincisi alanları çıkarılarak, 7 özellikle 1., 2. ve 3. yöntem çalıştırıldı. Bu alanlara ek olarak MarmaraUniYerlestirmeSirasi ve KayıtYılı alanları da çıkarılarak 5 özellikle 1., 2. ve 3. yöntem çalıştırıldı.

3.3.2.1. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre model sonuçları

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 1. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 3.12 gibidir. Bu tabloya göre 7 özellik ile elde edilen değerler, 11 özellik ile elde edilen değerlerle oldukça benzerlik göstermektedir.

Tablo 3.12. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre sonuçları

	1. Yöntem		11 özellik		7 özellik		5 özellik	
	Model	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**	**Test F1 Score**	**Sağlamlık F1 Score**	**Test F1 Score**	**Sağlamlık F1 Score**
1	AdaBoost	None	0.7748	0.6864	0.8126	0.7501	0.7663	0.7200
2	**Bagging_DT**	None	**0.8302**	**0.7719**	**0.8199**	**0.7696**	0.7634	0.6740
3	Bagging_SVC	None	0.7712	0.7612	0.7374	0.6086	0.6369	0.2530
4	BalancedBagging	None	**0.8154**	0.7582	0.8050	0.7581	0.7280	0.4853
5	Balanced-RF	None	0.6410	0.4943	0.6372	0.5224	0.5594	0.3955
6	BernoulliNB	None	0.5082	0.3989	0.5250	0.2893	0.5579	0.2644
7	DT	None	0.7698	0.7089	0.7062	0.7083	0.6782	0.5301
8	EasyEnsemble	None	0.5908	0.4468	0.5995	0.4578	0.5185	0.4484
9	**ET**	None	**0.8187**	**0.7704**	**0.8222**	**0.7713**	0.7819	0.6757
10	GaussianNB	None	0.6976	0.6366	0	**0.7714**	0	**0.7618**
11	GradientBoost	None	**0.8250**	0.7408	**0.8264**	0.7487	0.7716	0.7322
12	KNN	None	0.7577	0.7180	0.7729	0.7185	0.7327	0.6118
13	LR	None	**0.8188**	0.7282	0	**0.7714**	0	0.7597
14	**MLP**	None	**0.8255**	**0.7714**	**0.8285**	**0.7714**	0.7719	0.7049
15	**RF**	None	**0.8341**	**0.7685**	**0.8292**	**0.7666**	**0.7891**	0.7013
16	SVM	None	0.7676	0.7508	**0.8248**	0.6076	0.7657	0.6699
	En yüksek değerler:		**0.82-0.83**	**0.77**	**0.82-0.83**	**0.77**	**0.77-0.78**	**0.70-0.76**

Tablo 3.12'de görüldüğü gibi Adaboost ve SVM algoritma başarılarında ciddi yükseliş vardır. Adaboost sağlamlık değerinde de ciddi yükseliş olmuştur. Bununla beraber, RF, MLP, GradientBoost, ET, BalancedBagging ve Bagging_DT algoritmaları neredeyse aynı sonucu vermiştir. Bu tabloya bakıldığında, model olarak 7 özellikli kullanılması maliyet ve zaman açısından oldukça mantıklı görünmektedir.

3.3.2.2. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre model sonuçları

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 2. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 3.13 gibidir. Tablo 3.13 ve Tablo 3.12 deki değerler karşılaştırıldığında, sadece eğitim setine yapılan yeniden örnekleme neticesinde ve sağlamlık sonuçlarına göre, 1. yöntem ile oluşturulan modellerin değerlerini kullanmak, maliyet ve zaman açısından daha verimli olduğu görüldü. Çünkü en yüksek değerler 1. ve 2. yöntem sonuçlarında hemen hemen aynıdır.

Tablo 3.13. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre sonuçları

2. Yöntem		**11 özellik**		**7 özellik**			**5 özellik**		
Model	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**
AdaBoost	ROS	0.773	0.7299	SVMSMOTE	0.7761	0.7248	BSMOTE2	0.6879	0.3171
Bagging_DT	SVMSMOTE	**0.820**	0.7531	SVMSMOTE	0.7973	0.7474	ROS	**0.7287**	0.5884
Bagging_SVC	SVMSMOTE	0.750	0.7465	BSMOTE2	0.7344	0.5372	BSMOTE2	0.6916	0.4053
BernoulliNB	SVMSMOTE	0.652	0.6524	SVMSMOTE	0.6010	0.5313	SVMSMOTE	0.5961	0.3898
DT	BSMOTE2	0.780	0.7135	SMOTENC	0.7621	0.6228	ROS	0.6968	0.6856
ET	ROS	**0.820**	**0.7713**	ROS	**0.8221**	**0.7714**	SMOTENC	**0.7171**	0.4017
GaussianNB	SVMSMOTE	0.578	0.5133	ROS	0.5996	0.4696	SVMSMOTE	0.5985	0.3724
GradientBoost	SVMSMOTE	**0.819**	0.7179	SVMSMOTE	0.8138	0.7276	SVMSMOTE	**0.7189**	0.4564
KNN	ROS	0.752	0.7147	ROS	0.7729	0.7185	ROS	**0.7327**	0.6118
LR	SVMSMOTE	0.745	0.6420	SVMSMOTE	0.7317	0.6999	SVMSMOTE	0.6934	0.3784
MLP	SVMSMOTE	0.798	**0.7702**	ROS	**0.8195**	**0.7726**	SVMSMOTE	**0.7223**	0.3300
RF	ROS	0.826	**0.7705**	ROS	**0.8174**	**0.7714**	BSMOTE1	**0.7258**	0.5851
SVM	SVMSMOTE	0.757	0.7501	SVMSMOTE	0.7503	0.6544	SVMSMOTE	0.7130	0.3201
En yüksek değerler:		**0.82 - 0.83**	**0.77**		**0.82**	**0.77**		**0.72 - 0.73**	0.68

3.3.2.3. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre model sonuçları

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 3. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 3.14 gibidir. Tablo 3.14, Tablo 3.12 ve Tablo 3.13'deki değerler karşılaştırıldığında, tüm veri kümesine yapılan yeniden örnekleme neticesinde ve sağlamlık sonuçlarına göre, 1. yöntem ile oluşturulan modellerin değerlerini kullanmak, maliyet ve zaman açısından daha verimli olduğu görüldü. Çünkü en yüksek değerler 1. ve 3. yöntem sonuçlarında hemen hemen aynıdır.

Tablo 3.14. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre sonuçları

3. Yöntem		**11 özellikle**		**7 özellikle**			**5 özellikle**		
Model	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**	**Yöntem**	**Test F1 Score**	**Sağlamlık F1 Score**
AdaBoost	ROS	0.9542	0.7101	SMOTENC	0.6546	0.3978	SMOTEENN	0.6387	0.3043
Bagging_DT	SVMSMOTE	0.7884	0.7468	ROS	0.9599	0.7547	ROS	0.9342	0.5390
Bagging_SVC	SMOTEENN	0.9697	0.6866	SMOTEENN	0.8488	0.4840	SMOTEENN	0.7581	0.2027
BernoulliNB	SVMSMOTE	0.6824	0.6314	SVMSMOTE	0.5735	0.4105	SVMSMOTE	0.5806	0.3635
DT	ROS	0.8790	0.6836	BSMOTE2	0.7742	0.7401	SMOTEENN	0.8814	0.3287
ET	ROS	**0.9917**	**0.7713**	ROS	**0.9929**	**0.7713**	ROS	0.9842	0.6649
GaussianNB	SVMSMOTE	0.6324	0.4832	SVMSMOTE	0.5687	0.4138	SMOTEENN	0.6038	0.2043
GradientBoost	SVMSMOTE	0.8349	0.7031	SVMSMOTE	0.7765	0.7125	SMOTEENN	0.7441	0.2947
KNN	SMOTEENN	**0.9883**	0.5293	SMOTEENN	**0.9866**	0.5284	SMOTEENN	0.9826	0.3352
LR	SVMSMOTE	0.7193	0.6617	SMOTEENN	0.5967	0.4355	SMOTEENN	0.6306	0.1805
MLP	SVMSMOTE	0.6951	**0.7721**	SVMSMOTE	0.6080	**0.7714**	SMOTEENN	0.5678	0.1458
RF	ROS	**0.9935**	**0.7709**	ROS	**0.9870**	**0.7706**	ROS	0.9719	0.6574
SVM	SMOTEENN	**0.9862**	0.7216	SMOTEENN	0.8718	0.5364	SMOTEENN	0.7723	0.1813
En yüksek değerler:		**0.99**	0.77		**0.99**	**0.77**		**0.97 - 0.98**	**0.65 - 0.66**

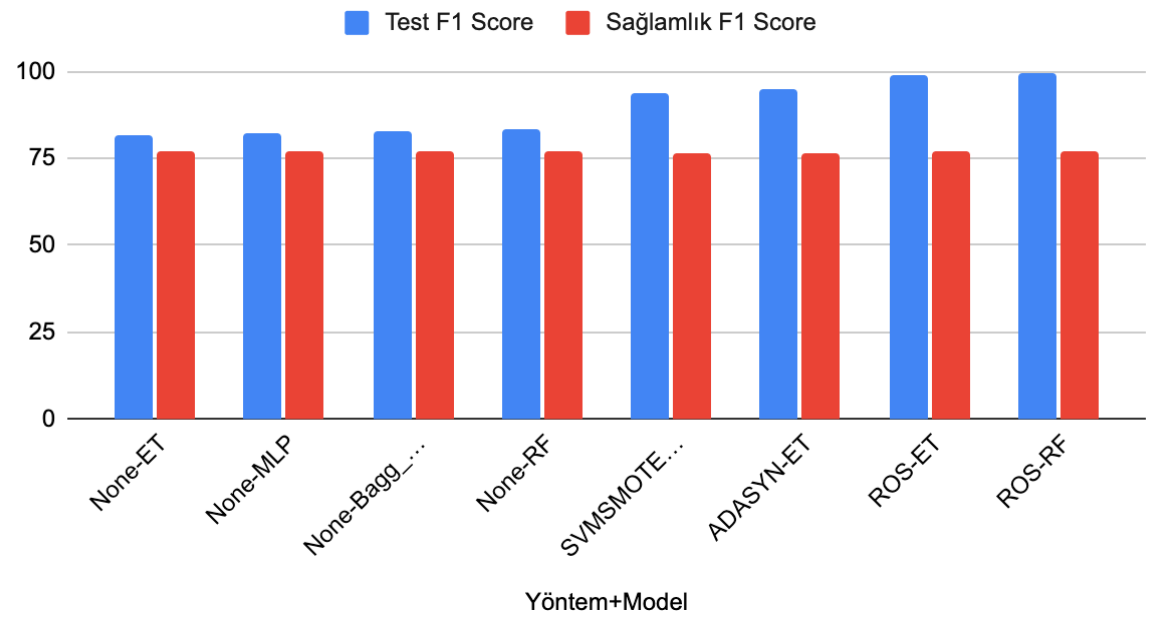
Ek olarak, tablolar incelendiğinde, 5 özellik ile model geliştirildiğinde, hem test F1 Score değerinde, hem de sağlamlık testlerinde ciddi düşüşler oldu. Ancak 7 özellik ile yapılan sonuçların, 11 özellik kullanarak yapılan sonuçlarla benzer olduğu görüldü.

3.3.3. Karşılaştırma

Sonuç olarak Şekil 3.5'te görüldüğü üzere Tablo 3.8.'de belirtilen ilk 4 modelin, Tablo 3.12 deki 7 özellikle oluşturulan modelin değerleri arasındaki F1 Score Başarı Oranı ve tutarlılık değeri yakın olduğu için, bu modellere ait 7 özellikli modellerin karar destek sistemi web uygulamasında kullanılmasına karar verildi. Tablo 3.12.'deki 2, 9, 14, 15 numaralı ET, MLP, Bagging_DT ve RF algoritmalarıyla 1. yöntem olan ve yeniden örneklem yapılmadan oluşturulan 7 özellikli modeller Tablo 3.15'te görüldüğü gibi karar destek sistemi web uygulamasında kullanıldı.

Tablo 3.15. KDS Web Uygulamasında kullanılan en tutarlı ve en başarılı algoritmalar

		11 özellikle		7 özellikle	
Yöntem	**MÖ Algoritması**	**Test F1 Score**	**Sağlamlık F1 Score**	**Test F1 Score**	**Sağlamlık F1 Score**
None	**ET**	0.8187	0.7704	0.8222	0.7713
None	**MLP**	0.8255	0.7714	0.8285	0.7714
None	**Bagging_DT**	0.8302	0.7719	0.8199	0.7696
None	**RF**	0.8341	0.7685	0.8292	0.7666



Şekil 3.5. En başarılı ve en tutarlı modellerin test ve sağlamlık F1 Score yüzdeleri

3.4. Çalışmanın Güçlü ve Zayıf Yönleri

Güçlü Yönler;

- Tezin veri kümesiyle alakalı, çok sınıflı dengesiz eğitimsel veri kümesi içeren makalelerle ilgili sistematik literatür incelemesinin yapılmış olması.
- Literatürde daha önce kullanılmayan yeni bir veri kümesi ile çalışmanın yapılmış olması.
- Marmara üniversitesinde alanında yapılan ilk çalışma olması.
- Türkiye'de yapılan tezlerde çok sınıflı ve dengesiz eğitimsel veri kümesiyle daha önce çalışmanın yapılmamış olması.
- Öğrenciler henüz kayıt olmadan, erken uyarı sistemi olarak çalışmanın yapılmış olması, bu şekilde de literatürde çalışmanın yok denecek kadar az olması.
- Topluluk metotlarıyla, hem çok sınıflı hem dengesiz hem de eğitimsel veri kümesinde başarılı sonuçların elde edilmiş olması.
- Karar destek sistemi web uygulamasının yapılıp, Teknoloji Fakültesi öğretim üyelerinin kullanımına sunulan pilot uygulamanın yapılmış olması.

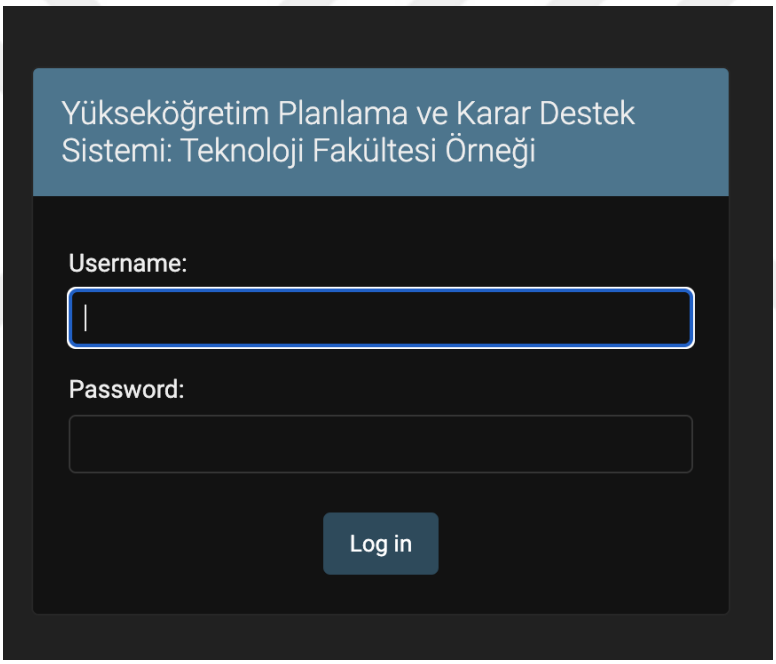
Zayıf Yönler;

- Çalışmada kullanılan toplam öğrenci sayısı 1394'tür ve veri kümesi sınıf dengesizliğine sahiptir. Daha çok öğrenci bilgisi içeren bir veri kümesi ile çalışmak daha başarılı sonuçlar verebilir.
- Bazı algoritmaların, çok sınıflı ve dengesiz veri kümesinde başarılı sonuç vermiyor olması.
- Bölümlerin aynı yılda açılıp, aynı yılda mezun vermeye başlamamış olması.
- Af, mühendislik tamamlama gibi farklı türdeki öğrencilerin de veri kümesinde yer alması ve bunlar çıkarıldığı için veri kümesinin azalması.
- Teknoloji Fakültesinin henüz yeni açılmış bir fakülte olması ve henüz çok fazla mezun vermemiş olması.

3.5. Karar Destek Sistemi Web Uygulaması

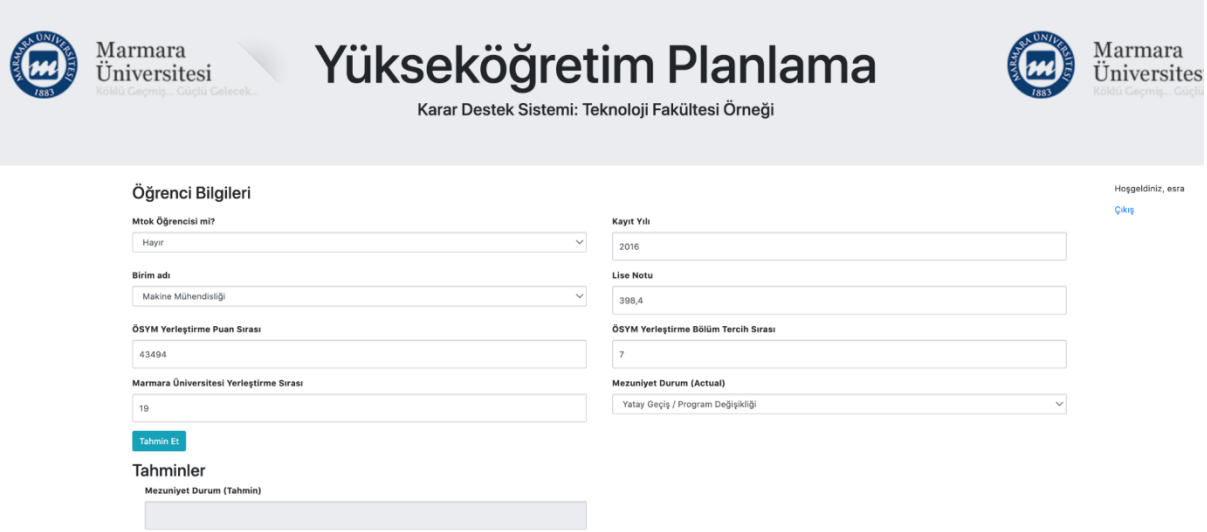
Tartışmalar bölümünde yapılan incelemeler ve karşılaştırma sonucunda, oluşturulan karar destek sistemi web uygulaması için seçilen modeller, yeniden örnekleme yapılmadan, 7 özellikle oluşturulan ve Şekil 3.5'te F1 Score değerleri gösterilen None-RF, None-MLP, None-Bagging_DT ve None-ET algoritmalarına ait modeller oldu.

Karar destek sistemi web uygulaması için, Django Python frameworku seçildi. Başlangıçta, Şekil 3.6.'daki gibi yetkilendirme ile sayfaya giriş yapılması sağlandı. Böylece, web uygulamasındaki güvenlik üst düzeye çıkarıldı. Şekil 3.7.'deki gibi, öğrenciye ait 7 özellik bilgisinin ekrandan girilmesi sağlandı ve tahmin et butonu ile girilen özellik bilgilerine göre, 4 model ayrı ayrı tahmin yaparak, sonuçları hem grafik şeklinde, hem de yazı şeklinde Şekil 3.8.'deki gibi gösterilmesi sağlandı.

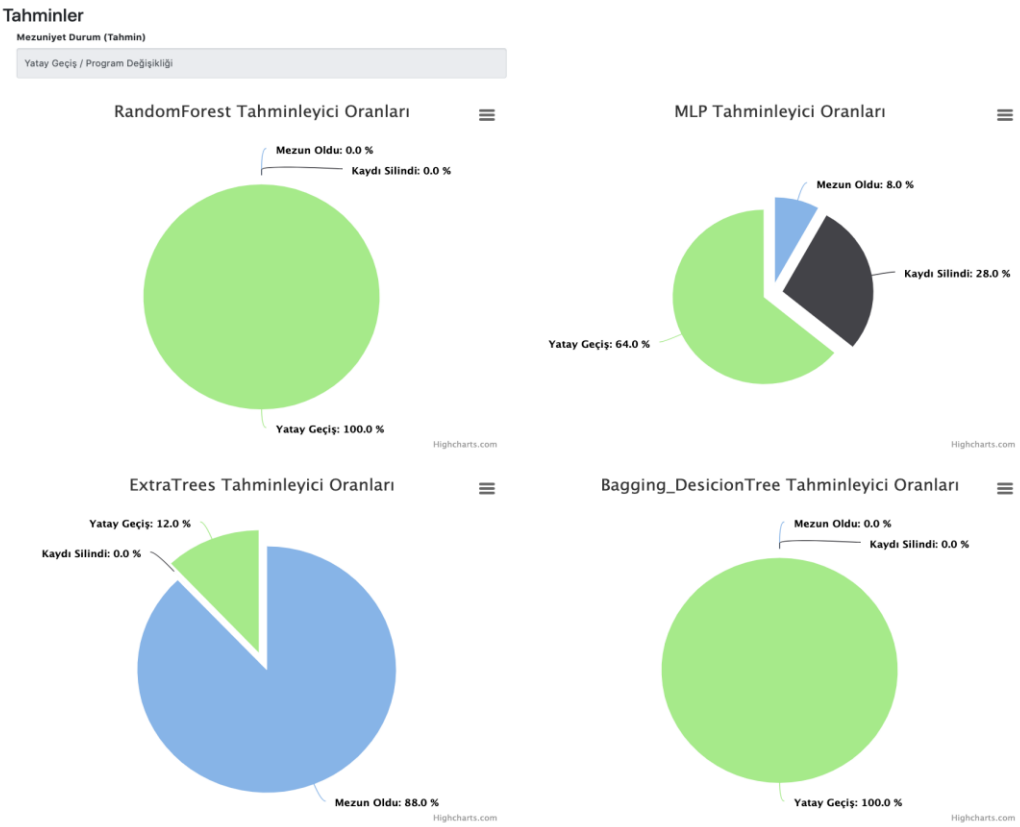


Şekil 3.6. Karar Destek Sistemi giriş paneli

Bu pilot çalışma, Marmara Üniversitesi, Teknoloji Fakültesi öğretim üyelerinin, kullanabileceği bir uygulama oldu. Riskli öğrencileri tahmin edecek bir karar destek sistemi ile öğrenciler hakkında ön görüleri ile ilgili düzenlemeler yapılabilir.



Şekil 3.7. Karar Destek Sistemi web uygulaması öğrenci bilgilerinin girildiği sayfa



Şekil 3.8. Karar Destek Sistemi web uygulaması tahmin sonuçları

MÜ TF öğretim üyeleri, bir öğrencinin mezuniyet ihtimali hakkında yorum yaparken karar destek sisteminden faydalanabilir. Böylece, objektif bir sisteme de referans verilmiş ve subjektif sebeplerden ötürü gözden kaçabilecek olası ihtimal, gündeme alınarak, öğrencinin bölüme geri kazandırılması hususunda çalışmalar yapılabilir.

BÖLÜM 4 - SONUÇLAR

Eğitimsel Veri Madenciliği(EDM) disiplininin bir alt dalı olan öğrenci performans tahmini ile ilgili yapılan çalışmalar son yıllarda arttı. Yapılan literatür taramasında, bu alanda çok sınıflı ve dengesiz eğitimsel veri kümesi ile ilgili yapılan çalışmaların hem ulusal hem uluslararası yayınlarda oldukça az olduğu görüldü. Ayrıca bu şekilde veri kümesiyle henüz bölüme kayıt olmayan öğrencilerin mezuniyet tahminini yapan erken uyarı sistemi şeklinde bir çalışma yok denecek kadar azdır. Literatürdeki bu iki boşluğun doldurulması amacıyla, bu tez çalışmasında, çok sınıflı ve dengesiz eğitimsel veri kümesi kullanarak, henüz üniversiteye başlamamış öğrencilerin mezuniyet tahminini yaparak, riskli öğrencileri önceden tahmin edecek bir karar destek sistemi yapıldı.

Bu çalışmada, 23.05.2022-286783 etik no kararı ile Marmara Üniversitesi Teknoloji Fakültesi öğrenci verileri kullanıldı. Marmara Üniversitesi Teknoloji Fakültesi 2010 - 2021 yıllarında öğrenim gören, 6 bölüme ait 4023 öğrenci içerisinden, ön işlem aşamasından sonra kalan 1394 öğrenci ve 12 özellik üzerinde çalışıldı. Bu verinin 2016 yılında kayıt yapan 153 öğrencisi sağlamlık verisi olarak ayrıldı ve eğitime dahil edilmedi.

Kullanılan veri kümesi çok sınıflı ve sınıf dengesizliği olan eğitimsel bir veri kümesidir. Çok sınıflı, dengesiz bir EDM veri kümesi üzerinde yapılan araştırmada, başarı tahminine yönelik önerilen model, birkaç yöntem kullanarak geliştirilen modeller arasında en başarılı sonuç veren yöntem üzerine inşa edilmiştir. Bu yöntemler şunlardır;

- Topluluk metotları kullanarak model üretmektir.
- Yeniden örnekleme yöntemleri ile beraber kullanılan geleneksel sınıflandırma algoritmaları ile model üretmektir.
- Dengesiz veri kümelerinde kullanmak üzere modifiye edilmiş topluluk metotlarıyla model üretmektir.

Bu tez çalışmasında her üç yöntem de kullanıldı. Kullanılan metotlar ve yöntemler şunlardır;

- Yukarı örnekleme yöntemi olarak SMOTE, SMOTENC, BSMOTE1, BSMOTE2, SVMSMOTE, ADASYN, ROS ve hibrit yöntemlerden SMOTE - ENN, SMOTE - Tomek olmak üzere toplam 9 tane yeniden örnekleme yöntemi,
- BalancedBagging, BalancedRF, EasyEnsemble olmak üzere 3 tane modifiye edilmiş topluluk algoritması,
- DT, SVM, LR, KNN, GaussianNB, BernoulliNB, MLP olmak üzere 7 tane geleneksel sınıflandırma algoritması,

- RF, ET, GradientBoost, AdaBoost, Bagging_SVC, Bagging_DT, 6 tane topluluk algoritması olmak üzere 16 farklı algoritma ile çalışıldı ve karşılaştırmalar yapıldı.

Kategorik veriler üzerinde One-Hot-Encoding ile özellik mühendisliği yapıldı ve veriler [0, 1] aralığına normalizasyon ile getirildi. Daha sonra eğitim aşamasında algoritmalar için GridSearchCV kullanarak hiper-parametre ayarları yapıldı.

Dengesiz veri kümeleri kullanımında uygun olduğu için repeated stratified 5*5-fold-CV kullanıldı. Bu aşamada oluşturulan tüm değerlerin ortalaması alınarak, son model başarı değerleri elde edildi ve modeller sağlamlık testinde kullanılmak üzere kayıt edildi. Sağlamlık testi ile modellerin tutarlılığı kontrol edildi. Oluşan modellerin yöntemlere göre kaç tane oldukları Tablo 2.4'de gösterildi.

Elde edilen sonuçlar ile çok sınıflı ve dengesiz eğitimsel veri kümesinde, topluluk algoritmalarının, yeniden örnekleme yöntemiyle beraber kullanılan geleneksel sınıflandırma algoritmalarından daha başarılı olduğu görüldü. Modellerin başarılı olmasıyla beraber, tutarlı olması da bu tez çalışmasında incelendi. Daha sonra Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF ve Anova F testleri olmak üzere 7 farklı özellik seçimi algoritması kullanılarak, FS karşılaştırması yapıldı. Her 3 yöntem de, 7 özellik ve 5 özellik için tekrar çalıştırıldı ve sağlamlık testleri yapıldı. Birbiriyle karşılaştırıldığında, 11 özelliğin kullanıldığı model başarı oranları ile 7 özellik kullanılarak elde edilen modellerin başarı oranları ve tutarlılık değerlerinin neredeyse aynı olduğu ancak 5 özellik kullanılarak model geliştirildiğinde başarı oranlarında düşüş olduğu görüldü.

Sonuç olarak, ROS+RF F1 Score 0.9935 değeriyle en başarılı algoritma olmasına rağmen, sırasıyla en tutarlı olan None+ET F1 Score 0.8187, None+MLP F1 Score 0.8255, None+Bagging_DT F1 Score 0.8302 ve None+RF F1 score 0.8341 değeriyle en tutarlı 4 model oldu. Bu modellerin 11 özellikli ve 7 özellikli değerleri birbirine çok yakın olduğu görüldü. Bu nedenle maliyet, zaman açısından düşünüldüğünde, 7 özellikle oluşturulan modellerin kullanılmasına karar verildi.

Daha sonra, en başarılı ve en tutarlı 4 model ile Marmara Üniversitesi Teknoloji Fakültesi Öğretim Üyelerinin kullanabileceği pilot bir "Karar Destek Sistemi Web Uygulaması" yapıldı. Bu uygulamada, yetkiye bağlı olarak, öğretim üyeleri, öğrenci bilgilerini girerek, 4 tahminleyici üzerinden öğrenci mezuniyet tahmini bilgisini almaktadır.

Gelecekte, önerilen çözümün genelliğini değerlendirmek için Marmara Üniversitesinin diğer bölümlerine ve başka üniversitelere de uygulanarak, birçok öğrenciden elde edilen veriler

üzerinde daha fazla araştırma yapılıp, oluşturulan sistemin genelleştirilip genelleştirilemeyeceği araştırılabilir. Ayrıca, performansın iyileştirilebilmesi için dengesiz veri kümesi durumunun daha fazla dikkate alınması gerekir.

Son olarak, çok sınıflı ve dengesiz eğitimsel veri kümesi içeren makalelerle ilgili yapılan sistematik literatür taraması uluslararası alanda; çok sınıflı ve dengesiz eğitimsel veri kümesiyle ilgili yapılan bu tez çalışması da ulusal alanda alanında tek çalışma olması dolayısıyla, gelecekteki çalışmalar için bu çalışmanın ilham kaynağı olduğuna inanılmaktadır.

KAYNAKLAR

1. Yükseköğretim Kanunu (1981) Resmi Gazete (06.11.1981, 17506). https://www.mevzuat.gov.tr/MevzuatMetin/1.5.2547.pdf, 10 Nisan 2022.

2. Karademirci, A. H. (2011) Eğitsel İletişim ve Teknoloji Derneği'ne (AECT) göre öğretim teknolojileri nedir? https://nenedir.com.tr/egitsel-iletisim-ve-teknoloji-dernegine-aect-gore-ogretim-teknolojileri-nedir/, 13 Nisan 2022.

3. Educational Data Mining (2021) EDM Tanımı. https://educationaldatamining.org/, 11 Mart 2022.

4. Max, K., Kjell, J. (2013) Uygulamalı Tahmine Dayalı Modelleme, Springer, New York.

5. *[Dengesiz verilerden öğrenme - Açık zorluklar ve geleceğe yönelik talimatlar , 2016]*

6. *[Sayfa vii, Dengesiz Veri Kümelerinden Öğrenme , 2018.]*

7. Dutt, A., Ismail, M. A., Herawan, T. (2017) A systematic review on educational data mining. IEEE Access, 5, 15991-16005.

8. Tatlıdil, M. (2020) Veri Türleri ve İstatistiğe Giriş. https://mervetatlidil.medium.com/veri-t%C3%BCrleri-ve-i%CC%87statisti%C4%9Fe-giri%C5%9F-2959f509f768, 13 Mart 2022.

9. Brownlee, J. (2021) One-vs-Rest and One-vs-One for Multi-Class Classification. https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/, 20 Nisan 2022.

10. Yılmaz, E., Altıkardeş, Z. A., Erdal, H. (2022) Systematic Literature Review on Multiclass Imbalanced Educational Data Mining. Educational Research Review.

11. Khan, A., Ghosh, S. K. (2021) Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and Information Technologies, 26(1), 205-240.

12. [https://tez.yok.gov.tr/UlusalTezMerkezi]

13. Pristyanto, Y., Pratama, I., Nugraha, A. F. (2018) Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification. 2018

International Conference on Information and Communications Technology (ICOIACT), IEEE.

14. Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., Ghani, N. A. M. (2021) Multiclass prediction model for student grade prediction using machine learning. IEEE Access, 9, 95608-95621.

15. Ghorbani, R., Ghousi, R. (2020) Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access, 8, 67899-67911.

16. Pratama, I., Pristyanto, Y., & Prasetyaningrum, P. T. (2021) Imbalanced Class handling and Classification on Educational Dataset. 2021 4th International Conference on Information and Communications Technology (ICOIACT), IEEE.

17. Chau, V. T. N., Phung, N. H. (2013) Imbalanced Educational Data Classification: An Effective Approach With Resampling and Random Forest. The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF), IEEE.

18. Purwoningsih, T., Santoso, H. B., Puspitasari, K. A., Hasibuan, Z. A. (2021) Early prediction of students' academic achievement: Categorical data from fully online learning on machine-learning classification algorithms. Journal of Hunan University Natural Sciences, 48(9), 131-141.

19. Lenin, T., & Chandrasekaran, N. (2021) Learning from Imbalanced Educational Data Using Ensemble Machine Learning Algorithms. Webology, 18(SI01), 183-195.

20. Buraimoh, E., Ajoodha, R., Padayachee, K. (2021) Application of machine learning techniques to the prediction of student success. 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE.

21. Docker (2022) The Magic of Docker Desktop is Now Available on Linux. https://www.docker.com/, 10 Mart 2022.

22. PostgreSQL (2022) The World's Most Advanced Open Source Relational Database. https://www.postgresql.org/, 20 Şubat 2022.

23. JetBrains (2022) DataGrip. https://www.jetbrains.com/datagrip/, 12 Mart 2022.

24. Waikato. Weka Wiki. https://waikato.github.io/weka-wiki/, 21 Mart 2022.

25. Anaconda (2022) Data Science Technology for a Competitive Edge a Better World. https://www.anaconda.com/, 8 Nisan 2022.

26. Han, E. (2022) https://github.com/esrahan, 11 Nisan 2022.

27. Sourcetree. https://www.sourcetreeapp.com/, 10 Mart 2022.

28. ÖSYM (2021) Okul Türleri. https://www.osym.gov.tr/TR,2666/tablo-7-okul-turleri.html, 13 Nisan 2022.

29. Django. Meet Django. https://www.djangoproject.com/, 12 Nisan 2022.

30. Rocca, J. (2019) Ensemble Methods: Bagging, Boosting and Stacking. https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205, 18 Nisan 2022.

31. Lutins, E. (2017) Ensemble Methods in Machine Learning: What are They and Why Use Them? https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f, 10 Mart 2022.

32. Ay, Ş. (2019) Ensemble Learning — Bagging ve Boosting. https://medium.com/deep-learning-turkiye/ensemble-learning-bagging-ve-boosting-50643428b22b, 9 Şubat 2022.

33. Imbalanced Learn. Ensemble of Samplers. https://imbalanced-learn.org/stable/ensemble.html 12 Mart 2022.

34. IBM (2020) Neural Networks. https://www.ibm.com/cloud/learn/neural-networks, 23 Nisan 2022.

35. Vikipedi (2005) Yapay Sinir Ağları. https://tr.wikipedia.org/wiki/Yapay_sinir_a%C4%9Flar%C4%B1 21 Nisan 2022.

36. Vikipedi (2021) Perceptron. https://tr.wikipedia.org/wiki/Perceptron, 10 Nisan 2022.

37. Hacıooğlu, M. (2022) Web Tabanlı Bir Dinamik Karar Ağacı Uygulaması. Yüksek Lisans Tezi, Trakya Üniversitesi Fen Bilimleri Enstitüsü, Edirne.

38. Akca, M. F. (2020) Karar Ağaçları. https://medium.com/deep-learning-turkiye/karar-a%C4%9Fa%C3%A7lar%C4%B1-makine-%C3%B6%C4%9Frenmesi-serisi-3-a03f3ff00ba5, 12 Mart 2022.

39. Alpaydin, E. (2021) Machine Learning, MIT Press.

40. Ukil, A. (2007) Support Vector Machine, Springer, Berlin, Heidelberg.

41. Gandhi, R. (2018) Support Vector Machine — Introduction to Machine Learning Algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, 12 Mart 2022.

42. Akca, M. F. (2020) Nedir Bu Destek Vektör Makineleri. https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e, 10 Nisan 2022.

43. Uğur, B. (2020) Destek Vektör Makinesi. https://burakugur.medium.com/destek-vekt%C3%B6r-makinesi-svm-f07f57f0a641, 12 Mart 2022.

44. Ukil, A. (2007) Support Vector Machine, Springer, Berlin, Heidelberg.

45. Akca, M. F. (2021) Lojistik Regresyon Nedir? Nasıl Çalışır? https://mfakca.medium.com/lojistik-regresyon-nedir-nas%C4%B1l-%C3%A7al%C4%B1%C5%9F%C4%B1r-4e1d2951c5c1, 23 Mart 2022.

46. Arc (2018) Derivative of the Sigmoid Function. https://towardsdatascience.com/derivative-of-the-sigmoid-function-536880cf918e, 8 Mart 2022.

47. Terzi, Y. Lojistik Regresyon Analizi.https://avys.omu.edu.tr/storage/app/public/yukselt/118305/Lojistik%20regresyon%20analizi.pdf, 23 Mart 2022.

48. Sharma, P. (2020) 4 Types of Distance metrics. Https://Www.Analyticsvidhya.Com/Blog/2020/02/4-Types-Of-Distance-Metrics-İn-Machine-Learning/,10 Nisan 2022.

49. Alpaydin, E. (2021) Machine Learning, MIT Press.

50. Öztürk, M. (2021) Python ile Sınıflandırma Analizleri. https://miracozturk.com/python-ile-siniflandirma-analizleri-knn-k-nearest-neighbours-k-en-yakin-komsu-algoritmasi/#Avantaj_Dezavantaj, 11 Mart 2022.

51. Dilber, B. (2020) AlgoRithm. https://www.datasciencearth.com/algorithm-naive-bayes-classifier/, 13 Mart 2022.

52. Kdag, I. K. (2021). Naive Bayes Algorithm. https://kdagiit.medium.com/naive-bayes-algorithm-4b8b990c7319, 12 Nisan 2022.

53. Theastrologypage (2022) Çok Katmanlı Algılayıcı (mlp) nedir? https://tr.theastrologypage.com/multilayer-perceptron, 21 Nisan 2022.

54. Akkaya, B. (2019) Multiplayer Perceptron. https://www.researchgate.net/figure/Multilayer-Perceptron-Advantages-and-Disadvantages_tbl4_338950098, 14 Nisan 2022.

55. Mbaabu, O. (2020) ntroduction to Random Forest in Machine Learning, https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/, 5 Nisan 2022.

56. Nazlı, R. (2021) R ile Random Forest. https://medium.com/@rmys.nzl/r-ile-random-forest-3909b0706d86, 1 Nisan 2022.

57. Freud, Y., Schapire, R. E. (1997). A Decision-theoretic Generalization of on-line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55, 119-139.

58. Veranyurt, Ü., Deveci, A., Esen, M. F., Veranyurt, O. (2020) Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neighbour ve Adaboost Algoritmaları Uygulaması. Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi, 6(2), 275-286.

59. Bulut, F. (2016) AdaBoost ile Kalp Krizi Risk Tespiti. CBÜ Fen Bil. Dergi, 12(3), 459-472.

60. Wang, Z., Zhang, J., Verma, N. (2015) Realizing Low-Energy Classification Systems by Implementing Matrix Multiplication Directly Within an ADC.

https://www.researchgate.net/figure/Illustration-of-AdaBoost-algorithm-for-creating-a-strong-classifier-based-on-multiple_fig9_288699540, 11 Nisan 2022.

61. Heo, J., Yang, J. Y. (2014) AdaBoost Based Bankruptcy Forecasting of Korean Construction Companies. https://reader.elsevier.com/reader/sd/pii/S1568494614003743?token=1DCA597D672F4FB3EA8844CF25AC6E28BF22926B581F9D2F3E769504B5829B2CEB9377C273800CAD95E2CAD9FA798310&originRegion=eu-west-1&originCreation=20220228112058, 15 Nisan 2022.

62. Blogathon (2022) Mutlak Yeni Başlayanlar İçin Adaboosta Giriş. https://zephyrnet.com/tr/mutlak-yeni-ba%C5%9Flayanlar-i%C3%A7in-adaboost%27a-giri%C5%9F/, 17 Mart 2022.

63. Muratlar, E. M. (2020) Gradient Boosted Regresyon Ağaçları. https://www.veribilimiokulu.com/gradient-boosted-regresyon-agaclari/ 21 Mart 2022.

64. Muratlar, E. M. (2020) Gradient Boosted Regresyon Ağaçları. https://www.veribilimiokulu.com/gradient-boosted-regresyon-agaclari/ 21 Mart 2022.

65. Ay, Ş. (2019) Ensemble Learning — Bagging ve Boosting. https://medium.com/deep-learning-turkiye/ensemble-learning-bagging-ve-boosting-50643428b22b, 9 Şubat 2022.

66. Deng, H., Zhou, Y., Wang, L., Zhang, C. (2021) Ensemble Learning for the Early Prediction of Neonatal Jaundice with Genetic Features. BMC Medical Informatics and Decision Making, 21(1), 1-11.

67. Liu, Y., Gu, Y., Nguyen, J. C., Li, H., Zhang, J., Gao, Y., Huang, Y. (2017) Symptom Severity Classification with Gradient Tree Boosting. Journal of Biomedical Informatics, 75, 105-111.

68. Erarslan, Y. (2020) Araştırma – Büyük Veri Setlerinin Belleğe Sığdırılmasına Yardımcı Yöntemlerin İncelenmesi. https://www.yasinerarslan.com/arastirma/arastirma-buyuk-veri-setlerinin-bellege-sigdirilmasina-yardimci-yontemlerin-incelenmesi/, 9 Şubat 2022.

69. Yazğılı, E., Baykara, M. Türkçe Metinlerde Makine Öğrenmesi Yöntemleri ile Siber Zorbalık Tespiti. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12(2), 443-453.

70. Morde, V. (2019) XGBoost Algorithm: Long May She Reign! https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d, 20 Mart 2022.

71. Wang, Z., Zhang, J., Verma, N. (2015) Realizing Low-Energy Classification Systems by Implementing Matrix Multiplication Directly Within an ADC. https://www.researchgate.net/figure/Illustration-of-AdaBoost-algorithm-for-creating-a-strong-classifier-based-on-multiple_fig9_288699540, 11 Nisan 2022.

72. Kuzubaşılı, A. (2019) Neden İlk Adresiniz XGBoost olmalı? https://ahmetkuzubasli.medium.com/neden-ilk-adresiniz-xgboost-75010416a583, 17 Mart 2022.

73. Geurts, P., Ernst, D., Wehenkel, L. (2006) Extremely Randomized Trees. Machine Learning, 63(1), 3-42.

74. Berrouachedi, A., Jaziri, R., Bernard, G. (2019) Deep Cascade of Extra Trees. *Pacific*-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham.

75. MLJAR. Random Forest. https://mljar.com/machine-learning/extra-trees-vs-random-forest/, 21 Nisan 2022.

76. Imbalanced Learn. Ensemble of Samplers. https://imbalanced-learn.org/stable/ensemble.html 12 Mart 2022.

77. Barros, T. M., Souza Neto, P. A., Silva, I., Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. Education Sciences, 9(4), 275.

78. Imbalanced Learn. Ensemble of Samplers. https://imbalanced-learn.org/stable/ensemble.html 12 Mart 2022.

79. Agusta, Z. P. (2019) Modified Balanced Random Forest for Improving Imbalanced Data Prediction. International Journal of Advances in Intelligent Informatics, 5(1), 58-65.

80. Imbalanced Learn. Ensemble of Samplers. https://imbalanced-learn.org/stable/ensemble.html 12 Mart 2022.

81. Liu, X. Y., Wu, J. Zhou, Z. H. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.

82. Aghalarova, S., Keser, S. B. (2021). Önerilen Yapay Sinir Ağı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini. Veri Bilimi, 4(2), 19-32.

83. Imbalanced Learn. Ensemble of Samplers. https://imbalanced-learn.org/stable/ensemble.html 12 Mart 2022.

84. MDS (2022) What is Undersampling? https://www.mastersindatascience.org/learning/statistics-data-science/undersampling, 15 Nisan 2022.

85. Erek, S. (2020) Dengesiz Veri Setiyle Nasıl Başa Çıkılır? https://simgeerek.medium.com/dengesiz-veri-setiyle-nas%C4%B1l-ba%C5%9Fa-%C3%A7%C4%B1k%C4%B1l%C4%B1r-dd1f8b5ff99d, 10 Şubat 2022.

86. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence research, 16, 321-357.

87. Müller, A. C. (2020) Data Splitting Strategies. https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html, 13 Mart 2022.

88. Veribilimcisi (2017) Karışıklık Matrisi. https://veribilimcisi.com/2017/07/18/karisiklik-matrisi-nedir/, 21 Mart 2022.

89. Olugbenga, M. (2022) Balanced Accuracy: When Should You Use It? https://neptune.ai/blog/balanced-accuracy, 13 Nisan 2022.

90. Öğündür, G. (2020) ROC ve AUC. https://medium.com/@gulcanogundur/roc-ve-auc-1fefcfc71a14, 17 Mart 2022.

91. Aljarah, İ. (2016) About Dataset, https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data, 21 Mart 2022.

EKLER

EK 1 - Excel sayfa içerikleri

Tablo 1. Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
1	Öğrenci Durum Log	33165	OgrenciTarihceID OgrenciID Durum OgrenciDurumKodID DurumAciklama OgrenciDurumAciklama DurumTarihi AsamaAciklama TarihceDali BaslamaYil BaslamaDonem BitisYil BitisDonem BitisTarihi Sinif EgitimYariYili Aciklama	Öğrencilerin yıl dönem bazlı, durum değişiklik loglarının tutulduğu sayfa. Eğitim yarıyıl, sınıf gibi değişiklikler de tutulmaktadır.
2	Kayıt Dondurma	3555	KayitDondurmaID OgrenciID Tarih IslemeAl Aciklama OgretimYili OgretimDonemi Sebep	Öğrencilerin Kovid Den itibaren online kayıt dondurma taleplerinin ve durumlarının tutulduğu sayfa.
3	Öğrenci Bilgileri	4023	OgrenciID CinsiyetId Cinsiyet DogumTarihi Uyruk_Ulke DogumYeri Engelli EngellilikOrani KayitYili Sinif EgitimYariYili ProgramMufredatID OrganizasyonProgramID OrganizasyonID UstBirimAdi BirimAdi AltBirimAdi Program ProgramTuru DurumId Durum KayitlanmaSekliId KayitlanmaSekli MezuniyetTarihi HazirlikSayisi MezunGANO UniversiteYerlestirmeSirasi	Öğrenciye ait demografik bilgiler tutulur, kayıtlanma şekilleri ve mezun ise mezuniyet GANOsu da tutulur.

(Tablo 1.'in devamı sonraki sayfadadır.)

Tablo 1 (devam). Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
4	Öğrenci ÖSYM Bilgileri	3502	OgrenciID OSYMOgrenciID OsymKayitYili LiseTuru LiseDerece LiseKolu LiseNotu osymOkulKodu osymOkulTuru osymOkulKolu osymOkulBirincisi YerPuani YerPuaniSirasi YerPuaniYuzdesi YerPuanTuru YerBolumTercihSirasi EkPuan MesLiseMezun MSGBasariPuani KontenjanTuru GelisTuruId GelisTuru	ÖSYM den alınan öğrenciye ait yerleştirme puan bilgileri.
5	ÖSYM Detay	54190	OsymOgrenciId PuanTuru PuanDeger	ÖSYM puanlarına ait tüm puanlar
6	Program Müfredat	19	OrganizasyonId OrganizasyonProgramId MufredatTur MufredatAdi OrgProgAciklama MinDersAdeti MaxDersAdeti MinZorunluKredi MinSecmeliKredi ToplamKredi AktifOgretimYili	Fakülteye ait bölümler ve bölümlere ait açılan müfredat bilgileri
7	Müfredat	690	ProgramMufredatID OrganizasyonId OrganizasyonProgramId MufredatTur MufredatAdi Donem Sinif MufredatId DersId DersKodu ZorunluSecmeli ZorunluMu	Bölümlere ait açılan müfredatlardaki yıl ve dönem bazlı ders isimleri ve ders türleri seçmeli ve zorunlu
8	Öğretim Üyeleri	134	UstBirimAdi BirimAdi AltBirimAdi AkademikUnvan OgretimUyesiID	Bölümlere ait öğretim üyeleri ve akademik unvanları

(Tablo 1.'in devamı sonraki sayfadadır.)

Tablo 1 (devam). Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
9	Ders Açılan Öğretmen	18125	NotSorumlu ProgramMufredatID OrganizasyonID OrganizasyonProgramId OgretimUyesiID AkademikUnvan DersAcilanID OgretimYili OgretimDonemi DersId DersKodu DersAdi	Açılan derslerdeki öğretim üyeleri
10	Öğrenci Bazlı Ders Bilgileri	977	OrganizasyonID DersID DersKodu DersAdi Kredi ZorunluSecmeli ZorunluSecmeli YillikDonemlik EctsKredi LisanKodu DersSaati UzemDersi	Bölümde açılan ve öğrencilerin aldığı ders bilgileri. Yine de bu tablodaki bilgiler tam anlaşılmadı, belki tekrar çekilir sistemden.
11	Bölüm Ders Bilgileri	927	UstBirimAdi BirimAdi AltBirimAdi OrganizasyonID DersID DersKodu DersAdi Kredi ZorunluSecmeli ZorunluSecmeli YillikDonemlik EctsKredi LisanKodu DersSaati UzemDersi	Teknoloji fakültesine ait olan bütün dersler
12	Öğrenci Ders Açılan Bilgileri	6250	DersAcilanID DersKodu OgretimYili OgretimDonemi DersiAlanOgrenciSayisi DersiBirakanOgrenciSayisi	Açılan dersler de dersi alan öğrenci ve dersi bırakan öğrencilerin yıl ve dönem bazlı derslere göre listesi
13	Öğrenci Ders < 2015	59367	OgrenciID OgretimYili OgretimDonemi DersKodu GecmeNotu HarfNotID DevamKodu ButGirildi DevamsizlikSaati BuDersiKacKereAldi YonetmelikHarf YerineAldigiOgrenciDersId YerineAldigiDers	Yıl dönem bazlı öğrenci ders bilgileri 2015 ve öncesi yıllara ait

(Tablo 1.'in devamı sonraki sayfadadır.)

Tablo 1 (devam). Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
14	Bölüm Ders Açılan Bilgileri	11319	UstBirimAdi BirimAdi AltBirimAdi DersAcilanID OgretimYili OgretimDonemi OrganizasyonID OrganizasyonProgramID DersID GrupNo LabNo Sinif LisanKodu DegerlendirmeTipi SinifOrtalamasi OgretimTuru Kapali HazirlikToplamDersSaati DersiAlanOgrenciSayisi UzemDersi DersKodu ZorunluSecmeli ZorunluSecmeli YillikDonemlik LisanKodu DersiAlanOgrenciSayisi DersiBirakanOgrenciSayisi	Bölümlere göre yıl ve dönem bazlı ders bilgileri.
15	Öğrenci Ders 2016-2017	47397	OgrenciID OgretimYili OgretimDonemi DersKodu GecmeNotu HarfNotID DevamKodu ButGirildi DevamsizlikSaati BuDersiKacKereAldi YonetmelikHarf YerineAldigiOgrenciDersId YerineAldigiDers	2016 ve 2017 yıllarına ait yıl dönem bazlı öğrenci ders bilgileri
16	Öğrenci Ders 2018-2019	59475	OgrenciID OgretimYili OgretimDonemi DersKodu GecmeNotu HarfNotID DevamKodu ButGirildi DevamsizlikSaati BuDersiKacKereAldi YonetmelikHarf YerineAldigiOgrenciDersId YerineAldigiDers	2018 ve 2019 yıllarına ait yıl dönem bazlı öğrenci ders bilgileri
17	Kodlar	1818	KodID KodGrup KodNo KodDuzey Aciklama	Excel sayfalarında id yerlere karşılık gelen kod bilgileri

(Tablo 1.'in devamı sonraki sayfadadır.)

Tablo 1 (devam). Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
19	Transkript	65189	OgrenciID Yil Donem Ortalama OrtalamaGano OrtalamaYuzlu TamamlananKredi ToplamKredi EkKredi Türü	Öğrencilerin yıl dönem bazlı mezuniyet ve detaylı Gano transkript bilgileri
20	Öğrenci Durum	209	OgrenciDurumKodID Aciklama Durum	Öğrenci bilgileri tablosundaki öğrenci durum id alanına karşılık gelen kod bilgileri
21	Öğrenci Bilgiler Kalanlar	4023	OgrenciID OgrenciDurumKodID UyrukID KayitliIl KayitliIlce	Öğrenci bilgilerinden Nu
22	OSYM Lise	12222	LiseID OsymKodu Tur Adi Il	OSYM lise isimleri kod olarak tutuluyor.
23	OSYM Lise 2014	16467	id ilKod ilad OsymKod Okulad	OSYM lise isimleri kod olarak tutuluyor.
24	**SecmeliDerslerTumFakulte**	1882	MufredatID MüfredatAdi ProgramMufredatID Sinif Donem HavuzAdi SecmeliAnaDersKodu SecmeliAnaDersAdi SecmeliHavuzID ZorunluSecmeli OrganizasyonID SecmeliHavuzDersId SecmeliDersKodu SecmeliDersAdi	Teknoloji Fakültesine ait tüm müfredatlara ait seçmeli derslere ait bilgiler
25	SecmeliDersler6Mufredat	399	MufredatID MüfredatAdi ProgramMufredatID Sinif Donem HavuzAdi SecmeliAnaDersKod SecmeliAnaDersAdi SecmeliHavuzID ZorunluSecmeli OrganizasyonID SecmeliHavuzDersId SecmeliDersKodu SecmeliDersAdi	Teknoloji Fakültesine ait sadece aktif ana müfredatlara ait seçmeli derslere ait bilgiler

(Tablo 1.'in devamı sonraki sayfadadır.)

Tablo 1 (devam). Marmara Üniversitesinden alınan excel listesine ait genel bilgiler

	Sayfa Adı	Satır Sayısı	Kolon İsimleri	Sayfa Geçen Açıklama
26	İl İlçe	86	IlIlceID IlKodu IlceKodu Adi AskerlikDairesiIlID AskerlikDairesIlceID AskerlikDairesiPostaKodu BolgeKodu AskerlikSubesiIlID Latitude Longitude IlHaritaKodu SonIslemTarihi SonIslemYapanID	Şehir bilgileri
27	Harf Not	250	YonetmelikHarfNotID YonetmelikID HarfNotDegerlendirmeID BasariNotu HarfNot Katsayi BasariDegerlemesi Durum Etkisiz TekrarEdilebilir Aciklama IsaretTuru BasKrediEtki TamKrediEtki YanoEtki GanoyaEtki AlabilecegiHBN OrtalamaylaGecer SonIslemTarihi SonIslemYapanID	Öğrencilerin aldığı notların harf not karşılıkları

EK 2 - 4023 öğrenciye ait öğrenci durumları

Tablo 2. Öğrenci mezuniyet durumlarına ait alt durum bilgileri

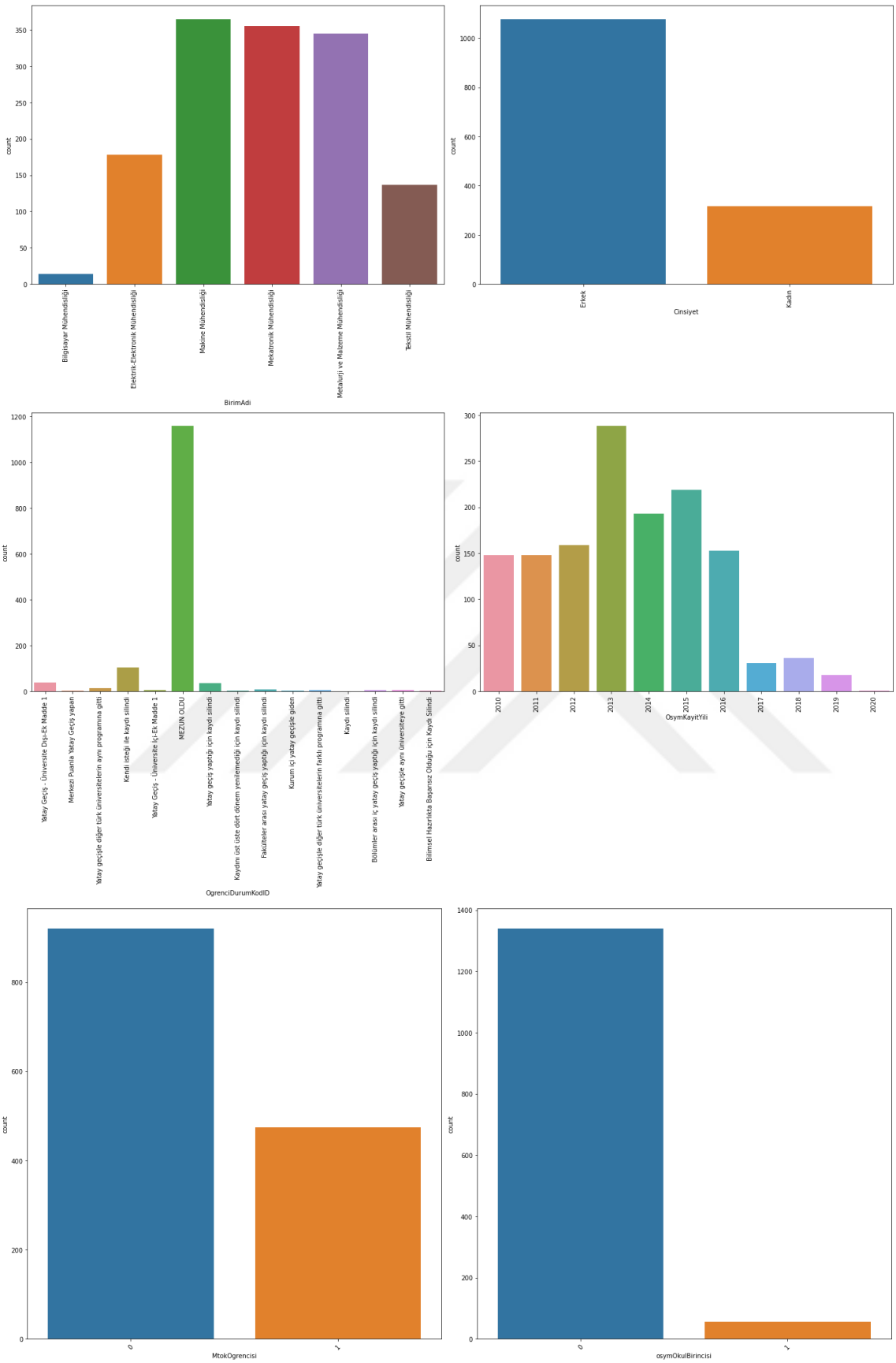
Durum	Alt Durum	Toplam Öğrenci
Aktif	Aftan yararlanıp kaydı yenilendi	1
Aktif	Hazırlık sınıfından başarısız olup türkçe programa alınan	1
Aktif	İZİN/KAYIT DONDURMADAN DÖNÜŞ YAPTI	14
Aktif	KAYIT YENİLEME	2109
Aktif	ÖSYMyi kazanıp kaydını yaptırdı	20
Aktif	Türkiye Burslusu ile gelen	9
Aktif	Yatay geçişle gelen	4
Aktif	Yök Kararı İle Aktarılan	1
Aktif	YÖS ile eğitime alınan	2
Aktif Değil	Askere Gitmek İstediğinden	2
Aktif Değil	KAYIT DONDURMA	9
İlişiği Kesilmiş	Bilimsel Hazırlıkta Başarısız Olduğu için Kaydı Silindi	2
İlişiği Kesilmiş	Bölümler arası iç yatay geçiş yaptığı için kaydı silindi	5
İlişiği Kesilmiş	Diğer nedenlerle kaydı silindi	1
İlişiği Kesilmiş	Fakülteler arası yatay geçiş yaptığı için kaydı silindi	9
İlişiği Kesilmiş	Kaydını üst üste dört dönem yenilemediği için kaydı silindi	6
İlişiği Kesilmiş	Kaydı silindi	3
İlişiği Kesilmiş	Kendi isteği ile kaydı silindi	169
İlişiği Kesilmiş	Kurum içi yatay geçişle giden	3
İlişiği Kesilmiş	Merkezi Puanla Yatay Geçiş yapan	4
İlişiği Kesilmiş	Misafir öğrenciliği sona eren	1
İlişiği Kesilmiş	Özel Öğrenciliği Sona Eren	47
İlişiği Kesilmiş	Program değişikliği nedeniyle kaydı silindi	1
İlişiği Kesilmiş	Yatay geçişle aynı üniversiteye gitti	11
İlişiği Kesilmiş	Yatay geçişle diğer türk üniversitelerin aynı programına gitti	16
İlişiği Kesilmiş	Yatay geçişle diğer türk üniversitelerin farklı programına gitti	5
İlişiği Kesilmiş	Yatay geçişten kendi isteği ile vazgeçme	2
İlişiği Kesilmiş	Yatay Geçiş - Üniversite Dışı-Ek Madde 1	44
İlişiği Kesilmiş	Yatay Geçiş - Üniversite İçi-Ek Madde 1	6
İlişiği Kesilmiş	Yatay geçiş yaptığı için kaydı silindi	36
Mezun	MEZUN OLDU	1480

EK 3 - Özellik Seçimi Sonrası Kalan 21 Özelliğin İstatistiksel Bilgileri

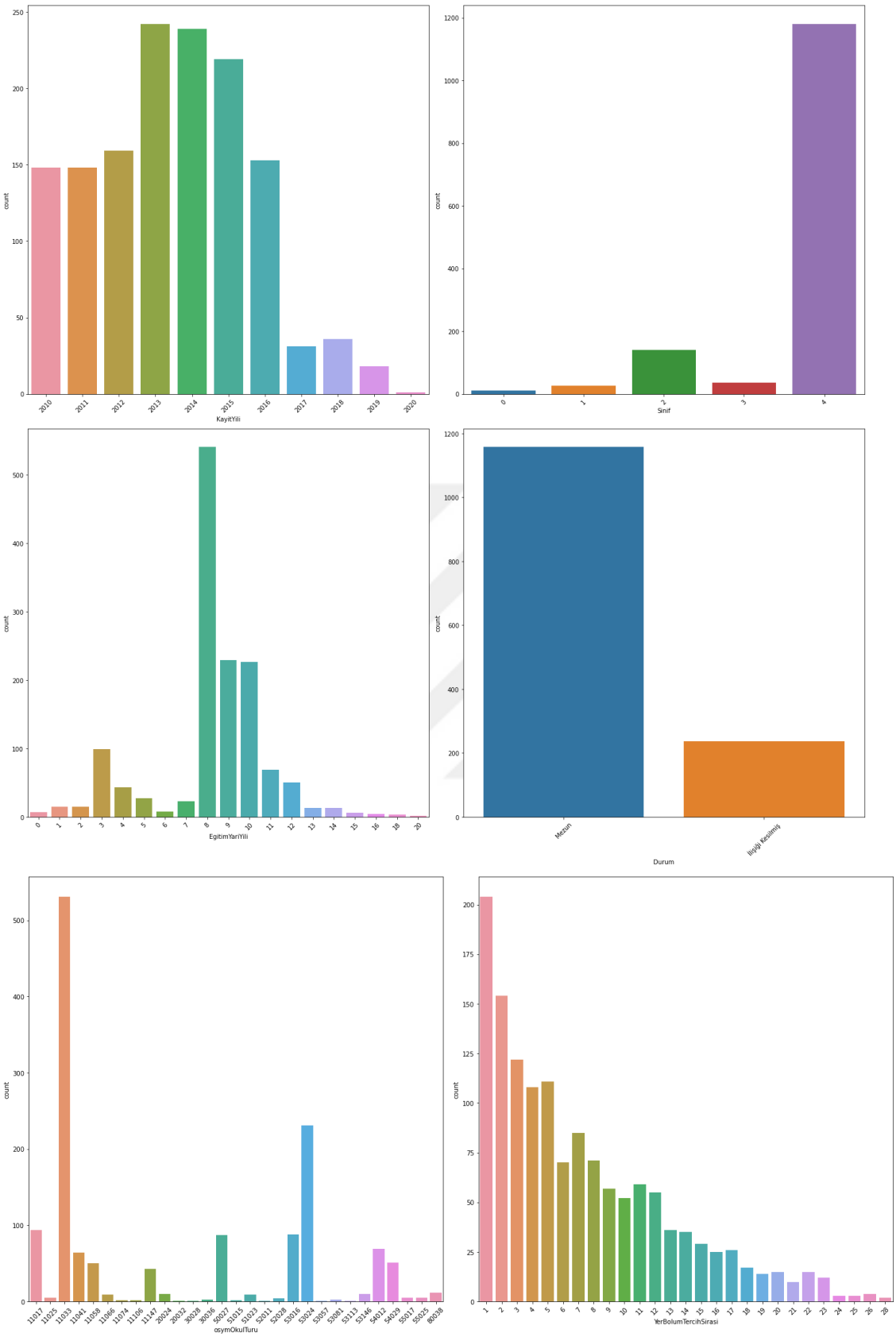
Özellikler: BirimAdi, Cinsiyet, DogumYeriIl, DogumTarihi, KayitYili, Sinif, EgitimYariYili, Durum, MezuniyetTarihi, MezunGANO, OgrenciDurumKodID, OsymKayitYili, LiseNotu, osymOkulTuru, YerPuaniSirasi, YerBolumTercihSirasi, MtokOgrencisi, osymOkulBirincisi, KontenjanTuru, UniversiteYerlestirmeSirasi.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1394 entries, 0 to 1393
Data columns (total 20 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   BirimAdi                      1394 non-null   object
 1   Cinsiyet                      1394 non-null   object
 2   DogumYeri                     1394 non-null   object
 3   DogumTarihi                   1394 non-null   int64
 4   KayitYili                     1394 non-null   int64
 5   Sinif                         1394 non-null   int64
 6   EgitimYariYili                1394 non-null   int64
 7   Durum                         1394 non-null   object
 8   MezuniyetTarihi               1158 non-null   float64
 9   MezuniyetGano                 1295 non-null   float64
 10  OgrenciDurumKodID             1394 non-null   object
 11  OsymKayitYili                 1394 non-null   int64
 12  LiseNotu                      1394 non-null   float64
 13  osymOkulTuru                  1394 non-null   int64
 14  YerPuaniSirasi                1394 non-null   int64
 15  YerBolumTercihSirasi          1394 non-null   int64
 16  MtokOgrencisi                 1394 non-null   int64
 17  osymOkulBirincisi             1394 non-null   int64
 18  KontenjanTuru                 1394 non-null   object
 19  UniversiteYerlestirmeSirasi   1394 non-null   int64
dtypes: float64(3), int64(11), object(6)
memory usage: 217.9+ KB
```

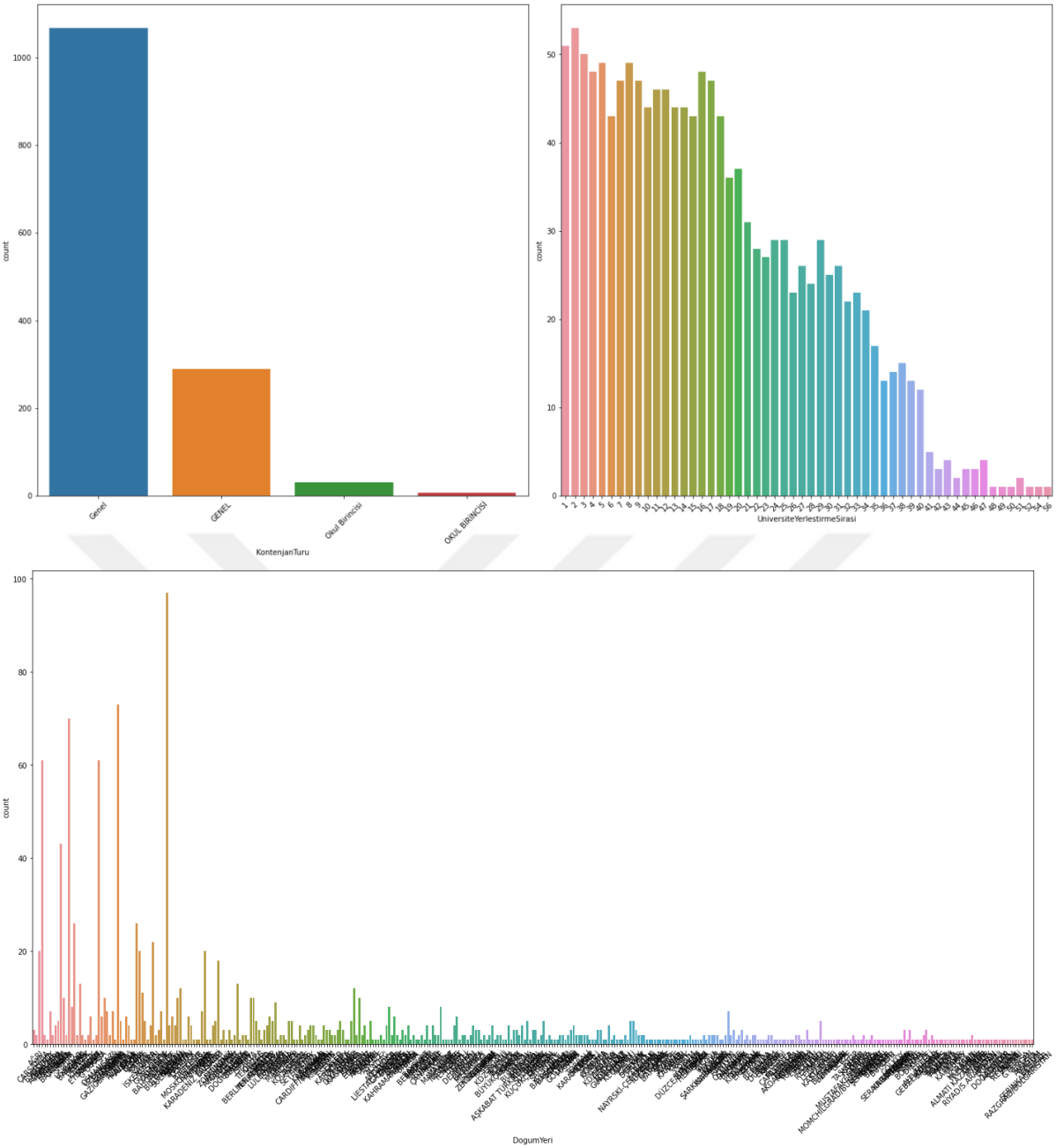
Şekil 1. Özellik bilgileri



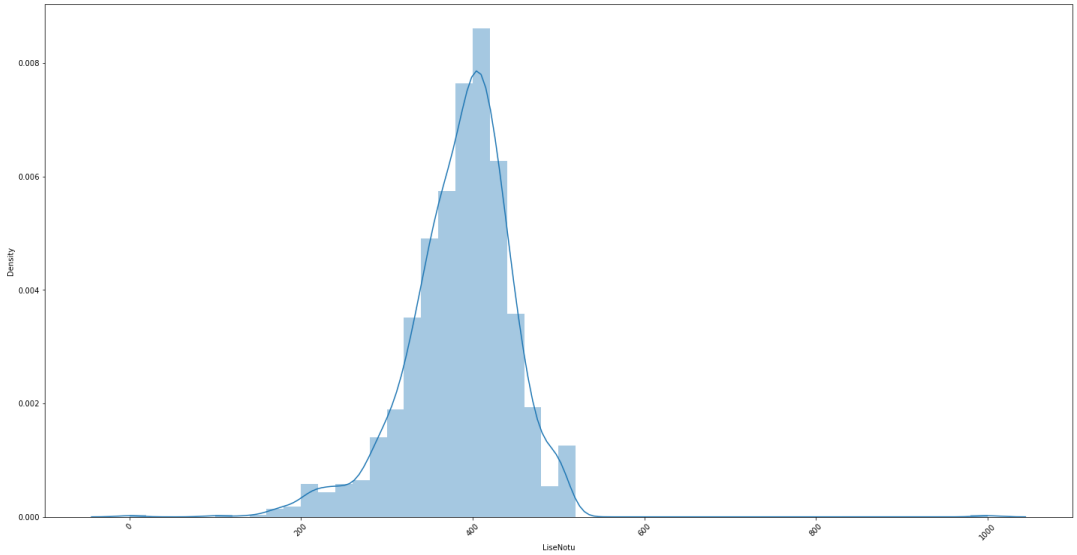
Şekil 2. Histogram bilgileri

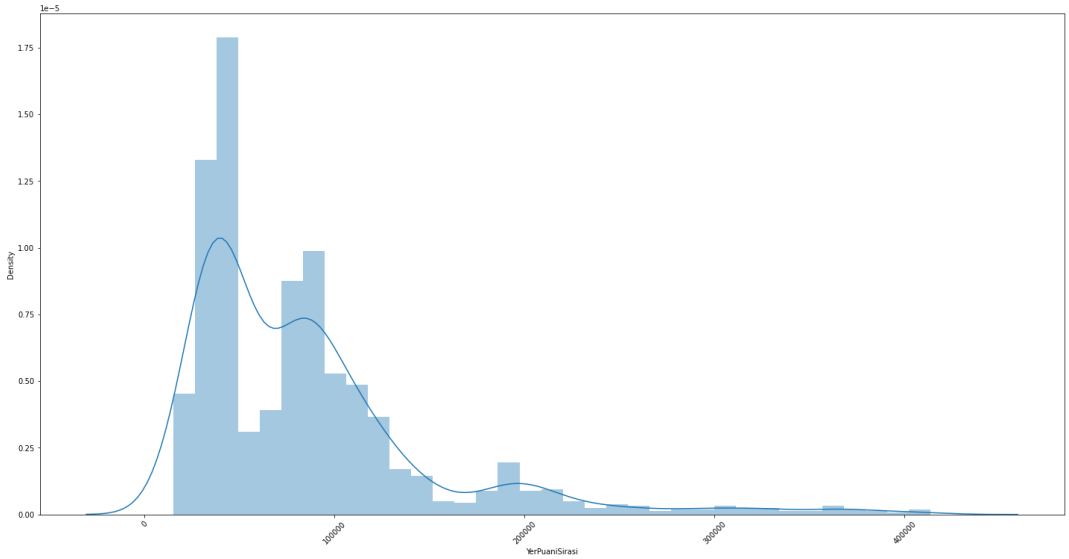


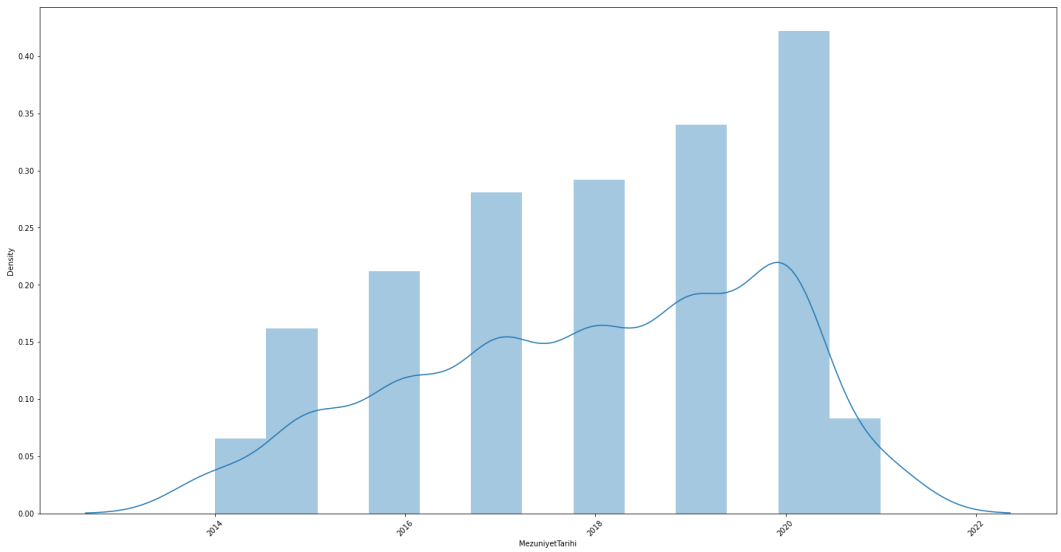
Şekil 2 (devam). Histogram bilgileri



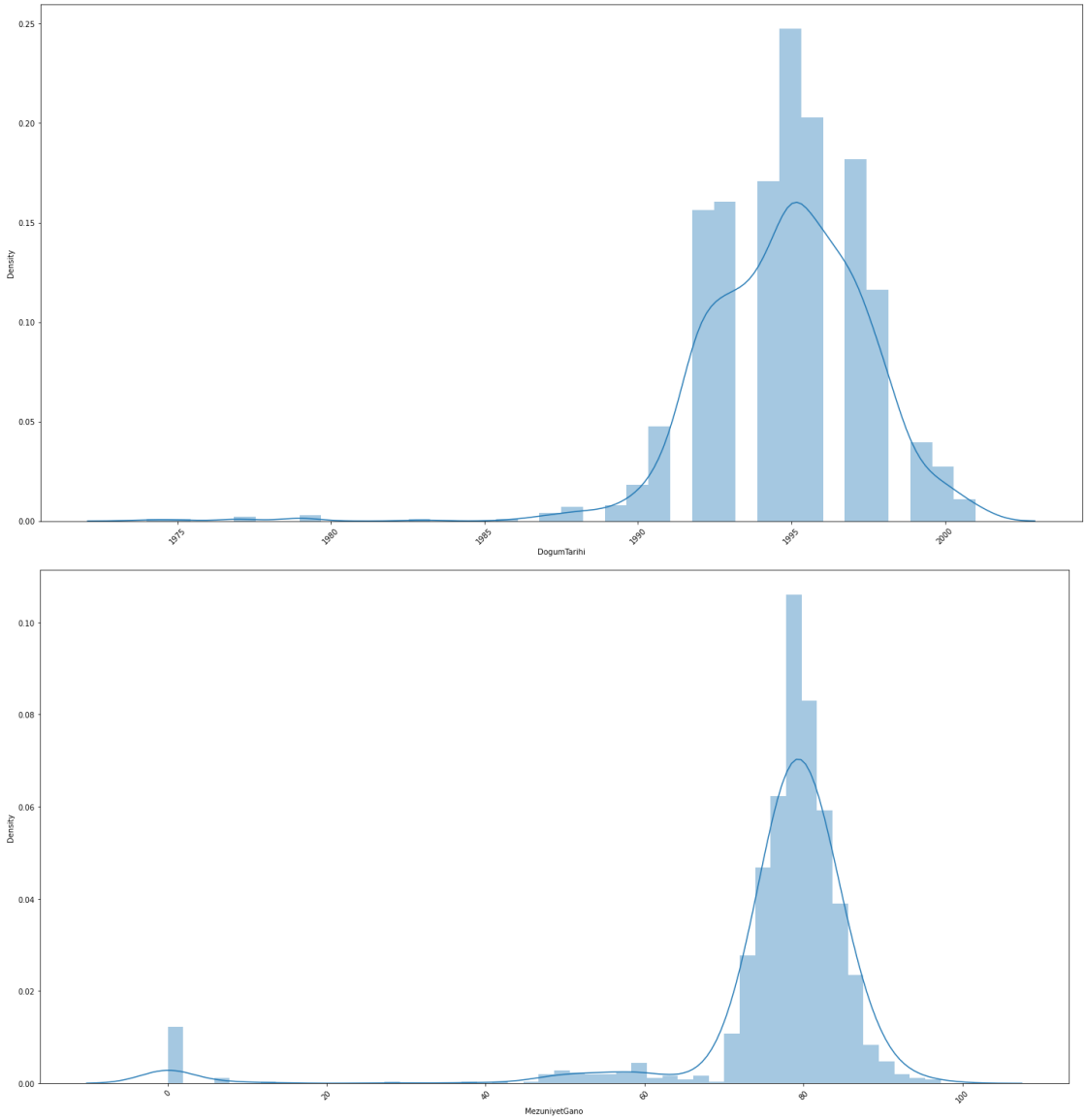
Şekil 2 (devam). Histogram bilgileri







Şekil 2 (devam). Histogram bilgileri



Şekil 2 (devam). Histogram bilgileri

	DogumTarihi	KayitYili	Sinif	EgitimYariYili	MezuniyetTarihi	MezuniyetGano	OsymKayitYili	LiseNotu	osymOkulTuru	YerPuaniSirasi	YerBolumTercihSirasi	MtokOgrencisi	osymOkulBirincisi	UniversiteYerlestirmeSirasi
DogumTarihi	1.000000	0.829563	-0.173389	-0.209281	0.723173	-0.192100	0.828240	0.260875	-0.317477	-0.269664	0.084510	-0.374984	-0.019516	0.213463
KayitYili	0.829563	1.000000	-0.300359	-0.297466	0.879455	-0.288730	0.996501	0.250943	-0.312047	-0.231277	0.055264	-0.368720	-0.075709	0.213970
Sinif	-0.173389	-0.300359	1.000000	0.810963	nan	0.631834	-0.301529	-0.099059	0.117186	0.030591	-0.106129	0.135267	0.052955	-0.079406
EgitimYariYili	-0.209281	-0.297466	0.810963	1.000000	0.340758	0.350267	-0.297993	-0.165391	0.111391	0.042271	-0.106056	0.120713	0.020020	-0.076136
MezuniyetTarihi	0.723173	0.879455	nan	0.340758	1.000000	-0.078735	0.869002	0.118164	-0.069511	-0.022345	-0.122749	-0.101785	-0.054882	0.146085
MezuniyetGano	-0.192100	-0.288730	0.631834	0.350267	-0.078735	1.000000	-0.290682	-0.037772	0.094236	0.013265	-0.066629	0.082976	0.055398	-0.083921
OsymKayitYili	0.828240	0.996501	-0.301529	-0.297993	0.869002	-0.290682	1.000000	0.255028	-0.330223	-0.237098	0.062138	-0.390374	-0.077779	0.225047
LiseNotu	0.260875	0.250943	-0.099059	-0.165391	0.118164	-0.037772	0.255028	1.000000	0.007286	-0.087555	-0.067087	-0.027664	0.244248	0.110808
osymOkulTuru	-0.317477	-0.312047	0.117186	0.111391	-0.069511	0.094236	-0.330223	0.007286	1.000000	0.504006	-0.342006	0.858519	0.131997	-0.180132
YerPuaniSirasi	-0.269664	-0.231277	0.030591	0.042271	-0.022345	0.013265	-0.237098	-0.087555	0.504006	1.000000	-0.281466	0.603376	0.069011	0.034309
YerBolumTercihSirasi	0.084510	0.055264	-0.106129	-0.106056	-0.122749	-0.066629	0.062138	-0.067087	-0.342006	-0.281466	1.000000	-0.365098	-0.044953	0.001890
MtokOgrencisi	-0.374984	-0.368720	0.135267	0.120713	-0.101785	0.082976	-0.390374	-0.027664	0.858519	0.603376	-0.365098	1.000000	0.111223	-0.253250
osymOkulBirincisi	-0.019516	-0.075709	0.052955	0.020020	-0.054882	0.055398	-0.077779	0.244248	0.131997	0.069011	-0.044953	0.111223	1.000000	0.164638
UniversiteYerlestirmeSirasi	0.213463	0.213970	-0.079406	-0.076136	0.146085	-0.083921	0.225047	0.110808	-0.180132	0.034309	0.001890	-0.253250	0.164638	1.000000

Şekil 3. Pearson metodu kullanılarak elde edilen korelasyon bilgileri

	count	mean	std	min	25%	50%	75%	max
DogumTarihi	1394.0	1994.826399	2.729750	1974.0	1993.0000	1995.000	1997.000	2001.000
KayitYili	1394.0	2013.470588	2.136628	2010.0	2012.0000	2013.500	2015.000	2020.000
Sinif	1394.0	3.684362	0.790794	0.0	4.0000	4.000	4.000	4.000
EgitimYariYili	1394.0	8.231707	2.624934	0.0	8.0000	8.000	10.000	20.000
MezuniyetTarihi	1158.0	2017.989637	1.859330	2014.0	2017.0000	2018.000	2020.000	2021.000
MezuniyetGano	1295.0	76.440039	14.286514	0.0	76.1500	79.100	82.000	97.250
OsymKayitYili	1394.0	2013.437590	2.135918	2010.0	2012.0000	2013.000	2015.000	2020.000
LiseNotu	1394.0	385.272937	62.458102	0.0	354.5125	393.495	422.875	999.999
osymOkulTuru	1394.0	28730.243902	20947.424899	11017.0	11033.0000	11058.000	53024.000	80038.000
YerPuaniSirasi	1394.0	85353.256098	64859.077891	15141.0	39796.2500	73909.500	103660.000	413695.000
YerBolumTercihSirasi	1394.0	7.187948	5.767744	1.0	2.0000	5.000	11.000	28.000
MtokOgrencisi	1394.0	0.340029	0.473888	0.0	0.0000	0.000	1.000	1.000
osymOkulBirincisi	1394.0	0.039455	0.194744	0.0	0.0000	0.000	0.000	1.000
UniversiteYerlestirmeSirasi	1394.0	17.000000	11.329307	1.0	8.0000	15.000	25.000	56.000

Şekil 4: İstatistiksel bilgiler

EK 4 - Etik Kurul Kararı

Evrak Tarih ve Sayısı: 23.05.2022-286783

T.C.
MARMARA ÜNİVERSİTESİ REKTÖRLÜĞÜ
Fen Bilimleri Enstitüsü Müdürlüğü

Sayı : E-44174047-730.03.01-286783 23.05.2022
Konu : Esra YILMAZ' ın (523618009) Etik Kurul Kararı

İLGİLİ MAKAMA

Enstitümüz 523618009 numaralı Yüksek Lisans öğrencisi Esra YILMAZ' ın "Yüksek Öğretim Planlama ve Karar Destek Sistemi: Teknoloji Fakültesi Örneği" başlıklı tez önerisinin Etik olarak incelenmesi yapılmış olup, Etik Kurulumuz tarafından uygun bulunmuştur.

Bilgilerinize arz ve rica ederim.

Prof. Dr. Bülent EKİCİ
Enstitü Müdürü

Bu belge, güvenli elektronik imza ile imzalanmıştır.
Doğrulama Kodu :BSAKH025E5 Pin Kodu :87032 Belge Takip Adresi : https://turkiye.gov.tr/ebd?eK=5709&eD=BSAKH025E5&eS=286783
Adres:Marmara Üniversitesi Göztepe Yerleşkesi Fen Bilimleri Enst. Enstitüler Binası 34722 Kadıköy / İSTANBUL
Telefon:0216 777 1860 Faks:0216 777 1861
e-Posta:fbe@marmara.edu.tr Elektronik Ağ:http://fbe.marmara.edu.tr
Kep Adresi:marmarauniversitesi@hs01.kep.tr
Bilgi için: ERSİN ÇELİK
Unvanı: Şef
Tel No: (216) 777 18 67

ÖZGEÇMİŞ

Adı Soyadı : Esra YILMAZ

Öğrenim Durumu

Derece	Bölüm/Program	Lise/Üniversite	Mezuniyet Tarihi
Lise	Yabancı Dil Ağırlıklı Lise	Bağcılar Süper Lisesi	2006 Haziran
Lisans	Bilgisayar Mühendisliği	Marmara Üniversitesi / Mühendislik Fakültesi	2011 Haziran

İş Deneyimi

Görevi	Firma / Kurum	Link	Başlangıç - Bitiş Tarihleri
Yazılım Geliştirme Uzmanı	Bilgi Sistemleri	www.bilgisistemleri.com	2011 Haziran - 2013 Ocak
Yazılım Geliştirme Uzmanı	Marmara Üniversitesi / Bilgi İşlem Daire Başkanlığı	bys.marmara.edu.tr basvuru.marmara.edu.tr harc.marmara.edu.tr	2013 Ocak - 2022 Ocak

Uzmanlık Alanları

Diller	C#, PYTHON
Veri Tabanları	MSSQL, POSTGRESQL