Research Article/Araştırma Makalesi

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Higher Education Planning and Decision Support System with Multi-Class and Imbalanced Educational Dataset: A Case Of Technology Faculty

Esra Yılmaz[*a], Zehra Aysun Altıkardeş[b], Hasan Erdal[c]

Submitted: 21.06.2022 Revised: 08.12.2022 Accepted: 21.01.2023 doi:10.30855/gmbd.0705053

ABSTRACT

Studies on academic performance prediction, a sub-branch of Educational Data Mining, have increased in recent years. Educational datasets in real environments often have class imbalanced and multi-class target variables. However, studies with these datasets are very few. In this context, in this study, with the ethical no decision of 23.05.2022-286783, using the data set of Marmara University (MU) Faculty of Technology (TF) students, a student graduation status estimation was made with the multiclass imbalanced educational dataset to identify the students at risk. 1394 samples and 11 features were obtained through data preprocessing and feature selection (FS) stages. 153 students belonging to 2016 were used for robustness control. 3 different datasets containing 11, 7 and 5 features obtained with 7 different FS were created. Using 9 different sampling methods and 16 different machine learning algorithms, 750 different models were created. Models were checked for robustness. F1 Score and Repeated Stratified 5*5 fold-CV were used as success criteria. Hyperparameter settings were made with GridSearchCV. As a result, although ROS+RF was the most successful algorithm with an F1 Score of 0.9935, the most successful and most consistent models were the 7-featured None+ET, None+MLP, None+Bagging_DT and None+RF models. With these models, the decision support system web application was developed and presented to MU TF faculty members.

Keywords: Educational data mining, ımbalanced dataset, multiclass dataset, prediction

[a],* Marmara University, Institute For Graduate Studies In Pure and Applied Sciences, Department of Computer Engineering 34730 - İstanbul, Türkiye
Orcid: 0000-0003-2411-4937
e mail: esra@marmara.edu.tr

[b] Marmara University, Vocational School Of Technical Sciences, Computer Technologies 34730 - İstanbul, Türkiye
Orcid: 0000-0003-3875-1793

[c] Marmara University, Faculty Of Technology, Electrical And Electronics Engineering 34730 - İstanbul, Türkiye
Orcid: 0000-0001-8296-0694

*Corresponding author: esra@marmara.edu.tr

Çok Sınıflı ve Dengesiz Eğitimsel Veri Kümesiyle Yükseköğretim Planlama ve Karar Destek Sistemi: Teknoloji Fakültesi Örneği

ÖZ

Eğitimsel Veri Madenciliğinin alt dalı olan akademik performans tahminiyle ilgili çalışmaların son yıllarda attığı görüldü. Gerçek ortamlarda eğitimsel veri kümeleri çoğunlukla sınıf dengesizliğine ve çok sınıflı hedef değişkene sahip olduğu ancak bu veri kümesi ile yapılan çalışmaların literatürde az olduğu görüldü. Bu bağlamda, bu çalışmada, 23.05.2022-286783 etik no kararı ile Marmara Üniversitesi (MÜ) Teknoloji Fakültesi (TF) öğrencilerine ait veri seti kullanılarak, çok sınıflı dengesiz eğitimsel veri kümesiyle, riskli öğrencileri tespit etmek için öğrenci mezuniyet durum tahmini yapıldı. Veri ön işleme ve özellik seçimi (FS) aşamalarıyla 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 örnekleme yöntemi ve 16 farklı makine öğrenmesi algoritması kullanılarak birbirinden farklı 750 model oluşturuldu. 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+ET, None+MLP, None+Bagging_DT ve None+RF modelleri oldu. Bu modellerle karar destek sistemi web uygulaması geliştirilerek MÜ TF öğretim üyelerine sunuldu

Anahtar Kelimeler: Eğitimsel Veri Madenciliği, Dengesiz Veri Kümesi, Çok Sınıflı Veri Kümesi, Tahmin

To cite this article: E.Yılmaz, Z. A. Altıkardeş and H. Erdal, "Higher Education Planning and Decision Support System with Multi-Class and Imbalanced Educational Dataset: A Case Of Technology Faculty" *Gazi Journal of Engineering Sciences*, vol. 9, no. 1, pp. 63-78, 2023. doi:10.30855/gmbd.0705053

1. Giriş (Introduction)

"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 başarısının artması yönünde destek vermek ve okuduğu bölüme geri kazandırmak yükseköğretimin amaçlarındandır. Yanlış bölüm seçiminden dolayı bölümü bırakma ihtimali olan öğrencilerin önceden tespit edilerek, doğru mesleğe henüz yükseköğretime başlamadan yönlendirilmesi ise hem öğrencinin hem yükseköğretimin hem de 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) yöntemi 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 yöntemi, eğitim verilerini analiz ederek, makine öğrenme yöntemleri ile öğrenme deneyimini ve kurumsal etkinliği geliştirmek için modeller geliştirir.

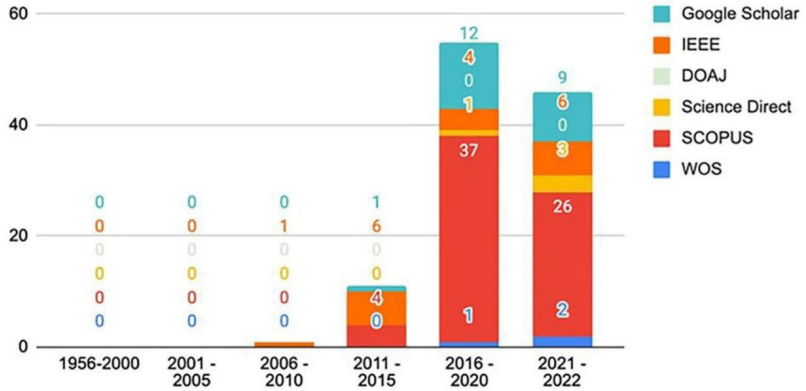
EDM kullanarak geliştirilen modelin başarısı, birçok kritere bağlıdır. Makine öğrenmesi algoritmalarını seçmeden önce kullanılacak verinin doğru analiz edilmesi büyük önem taşır. Bu nedenle eğitim verilerinin büyüklüğü, karmaşıklığı göz önüne alındığında, eğitim verisinin ön işlem aşamasında çok iyi 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 oluşturan One-vs-Rest ve One-vs-One şeklinde yaklaşım vardır [9].

Bu çalışmada, Marmara Üniversitesi (MÜ) Öğrenci İşleri Daire Başkanlığı (ÖİDB)'nın E-44174047-730.03.01-286783 etik kurul kararı ile MÜ Teknoloji Fakültesi (TF) mühendislik öğrencilerine ait 319 öznitelik ve 4023 örneklem içeren veri kümesi kullanıldı. Çok sınıflı dengesiz eğitimsel veri kümesi ile makine öğrenmesi yöntemleri kullanarak öğrenci mezuniyet tahmini yapan karar destek sistemi geliştirilmesi amaçlandı.

Çalışmanın diğer çalışmalardan en önemli farkı, veri kümesinin hem çoklu sınıf hem de sınıf dengesizliği problemlerine sahip olan bir eğitimsel veri kümesi olmasıdır. Literatür Araştırması bölümünde ayrıntılı bir şekilde açıklandığı ve Şekil 1.'de görüldüğü gibi çok sınıflı dengesiz eğitimsel veri kümesi ile yapılan çalışmaların 2010-2011 yıllarından itibaren arttığı görüldü. Bu bağlamda, veri

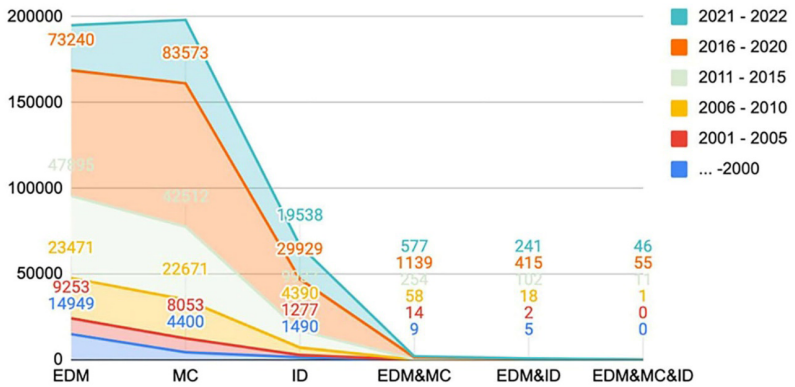
kümesinin özgün olması bu çalışmanın en önemli özgünlüğüdür. Ek olarak, MÜ TF Öğretim Üyelerinin kullanabileceği bir karar destek sistemi web uygulaması yapıldı. Yöntem olarak farklı makine öğrenmesi algoritmaları ve farklı örnekleme yöntemleriyle modelleme yapılarak, farklı modellere ait karşılaştırmalı sonuçların elde edilmesi, literatüre farklı çözümler sunulması açısından faydalı olacağı öngörüldü.



Şekil 1. Bilimsel veri kaynakları ve EDM, MC ve ID disiplinlerindeki çalışmaların yıllara göre dağılımı *(Scientific data sources and distribution of studies on disciplines EDM, MC and ID according to years)*

1.1. Literatür araştırması (Literature research)

Literatür araştırması sırasında, Şekil 2'den de anlaşıldığı gibi EDM, ID ve MC disiplinleri ile ayrı ayrı yapılan çalışmaların 2000'li yıllardan önce başladığı görüldü. Son yıllarda EDM, ID ve MC disiplinlerinin aynı veri kümesinde kullanıldığı çalışmaların sayısının ise arttığı görüldü. Bu çalışmaların içinden çok sınıflı ve dengesiz eğitimsel veri kümesi ile ilgili olanların özet bilgileri Tablo 1.'de gösterildi. Ayrıca yapılan bu çalışmaların çoğunlukla, üniversiteye kayıt yaptırmış ve okuduğu bölüme ait ders notu olan öğrencilerin performans tahmini üzerine olduğu görüldü. Diğer yandan henüz üniversite tercihinde bulunmamış ve üniversiteye yerleşmemiş aday öğrencilerin ders notu, yarıyıl dönemi gibi akademik bilgileri olmadığından bu şekilde öğrenci performans tahminini yapan çalışmanın yok denecek kadar az olduğu görüldü. Yükseköğretim verilerinde çok sınıflı ve dengesiz veri kümesi problemi tespit edildi. Mevcut çalışmalar, öğrencilerin öğrenci performans tahminini geliştirmeye odaklansa da, çok sınıflı dengesiz eğitimsel veri kümesi ile öğrenci performans tahmini üzerinde yapılan çalışmaların az olduğu görüldü. Literatürdeki bu boşluğu doldurmak üzere, bu çalışmada çok sınıflı ve dengesiz eğitimsel veri kümesi üzerinde öğrenci performansını makine öğrenmesi algoritmaları ile tahminlenmesi ve bu modeller kullanılarak geliştirilen bir karar destek sistemi web uygulaması yapıldı.



Şekil 2. EDM, MC ve ID disiplinlerinde yapılan çalışmaların yıllara göre dağılımı *(Distribution of studies on disciplines EDM, MC and ID according to years)*

Özetlemek gerekirse, Tablo 1.'de ve Şekil 1'de belirtilen literatür araştırması sonuçları neticesinde bu çalışmada, makine öğrenmesi algoritmaları ve örnekleme yöntemleri kullanılarak çok sınıflı dengesiz eğitimsel veri kümesi üzerinde mezuniyet tahmini yapan modeller oluşturuldu ve en tutarlı modeller kullanılarak pilot bir karar destek sistemi web uygulaması geliştirildi. Böylece, üniversiteye

başlamadan, bölüme kayıt yaptırdıkları takdirde, sınıfta kalma ihtimali, bölümü terk etme ihtimali olan riskli öğrencilerin önceden tahmin edilebilmesi ve bu öğrencilerle ilgili iyileştirme çalışmasının yapılmasına imkân sunuldu. Özgün çalışmada önerilen yöntemin literatürde EDM disiplinine yeni bir katkı sağlayacağı düşünüldü.

Tablo 1. Literatürde çok sınıflı ve dengesiz eğitimsel veri kümesi ile ilgili yapılan çalışmalar (Studies on multi-class and imbalanced educational dataset in the literature)

No	Veri Kümesi	Örneklem Sayısı	Özellik Sayısı	Hedef Sınıf Sayısı	Makine Öğrenmesi	Yeniden Örnekleme Yöntemleri	Cross Validation (CV)	Başarı Metrikleri	Sonuç
1 [11]	UCI Machine Learning	403	5	3	SVM	SMOTE	-	Accuracy, Sensitivity, Specificity, G-mean	Başarılı
2 [12]	A university	1282	10	5	RF, DT, SVM, NB, KNN, LR	SMOTE	10 fold CV	F Measure	RF + SMOTE en başarılı
3 [13]	Two universities	650 ve 394	19 ve 19	4 ve 4	XGBoost, RF, KNN, ANN, SVM, DT, LR, NB	Borderline SMOTE, ROS, SMOTE, SMOTE-ENN, SVMSMOTE, SMOTETomek	Shuffle 5-fold CV	Accuracy, Recall, Precision, F1 Score	RF + SVM SMOTE en başarılı
4 [14]	Kaggle Kalboard 360 LMS	480	16	3	Stacking, RF, LR, KNN, DT, CART, SVM	SMOTE, Borderline SMOTE, SMOTETomek	10-fold CV	Accuracy, G-Mean	SMOTE + RF %83 F1 Score en başarılı
5 [15]	A university	1334	43	7	RF, Boosting, Bagging, NB, SVM, NN, LR, KNN, DT C4.5		10-fold CV	Accuracy, ROC	RF
6 [16]	A university	101617	19	3	RF, Gradient Boost, AdaBoost, DT	SMOTE	Holdout	Accuracy	Smote + RF %85.03 accuracy
7 [17]	A university	497	18	?	RF, Adaboost, Stochastic GBM, xgbTree, C4.5	?	10 Fold CV	precision, specificity, recall, kappa metrics,balanced accuracy, F-score	RF ve Adaboost F-score 66.67%
8 [18]	Kaggle Kalboard 360 LMS	480	16	3	LDA, LR, RT, KNN, NB, SVM	SMOTE	5 ve 10 Fold CV	confusion matrix (CM), accuracy, precision, recall, Fl-Score	LR %86 accuracy
9 [19]	A university	6882	15	3	Gradient Boost, Adaboost, RF, MLP	SMOTE, Borderline SMOTE, SVMSMOTE ADASYN	?	?	MLP + Borderline SMOTE en başarılı
10 [20]	Kaggle Kalboard 360 LMS	480	16	3	Gradient Boost, SVM, KNN, CART, MLP, LDA	RUS ROS SMOTE	10 Fold CV	CM, Accuracy, AUC, Kappa, Precision, Recall	SMOTE + PCA SVM %94 Accuracy en başarılı

Bu bağlamda çalışmanın genel düzeni şu şekilde organize edildi: Materyal ve Yöntem bölümü altında ilk önce tasarlanan sistem mimarisi belirtildi. Veri setinin tanıtımı ve üzerinde yapılan ön işleme ve özellik seçimiyle ilgili bilgiler materyal yöntemi kısmında açıklandıktan sonra yöntem bölümünün altında kullanılan makine öğrenmesi algoritmaları ile tasarlanan sistem ve performans kriterleri ile

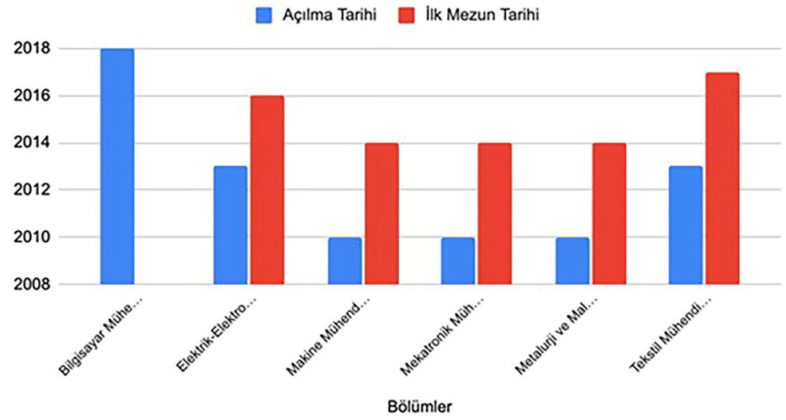
elde edilen bulgular paylaşıldıktan sonra Karşılaştırma bölümünde, literatürle ilgili karşılaştırma ve bu çalışmadaki en başarılı ve en tutarlı sonuçların karşılaştırması yaparak, çalışmanın güçlü ve zayıf yönleri ile geliştirilebilir yönleri belirtildi. "Web Uygulaması" alt başlığı altında geliştirilen web uygulaması karar destek sistemi anlatıldıktan sonra "Sonuç" bölümünde çalışmayla ilgili ulaşılan sonuçlar, öneriler ve gelecek çalışmalara ilişkin görüşler paylaşıldı.

2. Materyal ve Yöntem (Material and Method)

Bu çalışmada, veri analizine dayalı makine öğrenmesi ile tahminleme modeli geliştirildi. Bunun için gerekli olan süreçler verinin temin edilmesi, verinin işlenmesi ve makine öğrenmesi algoritması ile modellenmesi şeklinde yapıldı. Bu bağlamda, her sürecin verimini artıracak yazılımsal ve donanımsal materyaller kullanıldı. Çalışma boyunca, MacBook Pro: 2.3Ghz Dual-Core Intel Core i5 donanımı ve Python Jupyter ve WEKA yazılımları kullanıldı.

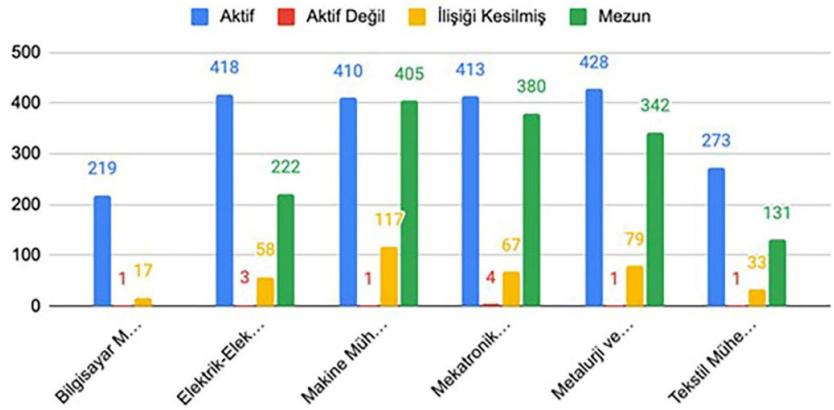
2.1. Veri kümesi (Dataset)

Bu çalışmada, MÜ ÖİDB'nın E-16110545-302.08.01-2100040487 etik no kararı ile MÜ TF mühendislik öğrencilerine ait veri kümesi kullanıldı. Elde edilen veri kümesinde 319 adet özellik ve 4023 adet örneklem tespit edildi. Veri kümesi üzerinde istatistiksel analizler yapıldığında TF'de Bilgisayar Mühendisliği (Müh.), Elektrik-Elektronik Müh., Makine Müh., Mekatronik Müh., Metalurji ve Malzeme Müh. ve Tekstil Müh. olmak üzere 6 bölüm olduğu görüldü. Şekil 3'te TF bölümlerinin ilk öğrenci kayıt yılı ve ilk öğrenci mezun yılları gösterildi.



Şekil 3. MÜ TF 2021 Şubat ayına kadar bölümlere ait açılma ve mezun yılları (MU TF opening and graduation years of departments until February 2021)

TF de bulunan 6 bölümdeki toplam 4023 öğrencinin mezuniyet durumları "Aktif", "Aktif Değil", "İlişiği Kesilmiş" ve "Mezun" şeklinde olduğu görüldü ve bölümlere göre öğrenci mezuniyet durum dağılımları Şekil 4.'teki gibi gösterildi.



Şekil 4. MÜ TF bölümlerine göre öğrenci ve öğrenci durum dağılımları (Student and student status distributions according to MU TF departments)

Şekil 4 incelendiğinde, "Mezun" durumda olan öğrenci sayısının 1480 iken, "İlişiği Kesilmiş" durumda olan öğrenci sayısının 371 ve "Aktif Değil" durumda olan öğrenci sayısının ise sadece 11 olduğu görüldü. Bu bağlamda toplam 4023 öğrenciden sadece %9.2 si "İlişiği Kesilmiş" durumda olan öğrenci olduğu tespit edildi. Bu durumda mezuniyet durum dağılımları arasındaki farkın çok olduğu görüldü. Bu nedenle bu veri setinin dengesiz bir veri seti olduğu belirlendi.

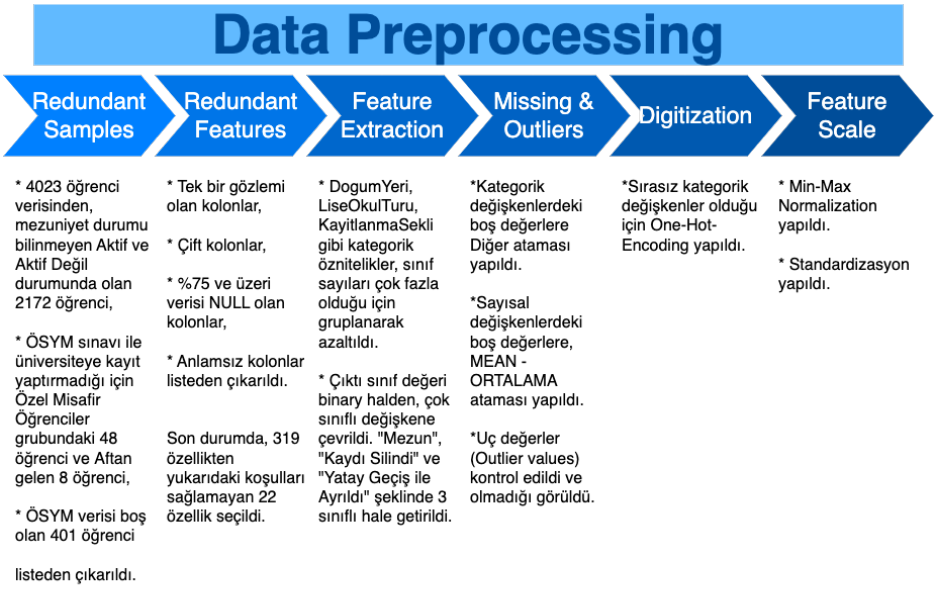
2.2. Veri ön işleme (Data preprocessing)

MÜ TF öğrenci veri kümesi Şekil 5'teki veri önişleme adımları sırasıyla uygulandı. Öncelikle, mezuniyet durumu bilinmeyen Aktif ve Aktif Değil durumundaki 2172 öğrenci, ÖSYM verisi boş olan 401 öğrenci ile üniversiteye kayıtlanma şekli ÖSYM olmayan 56 öğrenci, gereksiz örnekler olduğu için veri kümesinden çıkarıldı. Özellik mühendisliği yöntemleri ile gereksiz özellikler veri kümesinden çıkarıldı. Daha sonra özellik çıkarımı ile kategorik özellikler işlendi ve yeni özellikler elde edildi. Eksik aykırı değerler aşamasında boş olan değerler, ilgili özelliğin ortalaması ile güncellendi. Uç değerler kontrol edildi ve olmadığı görüldü. Kategorik değişkenlerden boş olanlar Diğer verisi ile güncellendi. One hot encoding işlemi ile veri kümesi sayısallaştırıldı. Son olarak özellik ölçeklendirme yapıldı ve 4023 öğrenciden, hedef değişkeni 1158 "Mezun", 110 "Kaydı Silindi" ve 126 "Yatay Geçiş / Program Değişikliği" mezuniyet durumlarını içeren toplam 1394 öğrenci elde edildi.

2.3. Özellik seçimi (Feature Selection)

Ham veri setinde bu çalışmanın amacına uygun olarak kullanılacak verileri içeren Öğrenci Bilgileri, Öğrenci ÖSYM Bilgileri, ÖSYM Detay, OSYM Lise, OSYM Lise 2014, Öğrenci Bilgiler Kalanlar ve İl İlçe tabloları kullanıldığında, özellik sayısı, 319'den ilk olarak 81'e indirildi. Devamında, Şekil 5'te belirtilen "Gereksiz Özellikler" aşaması uygulandı ve özellik sayısı 81'den 20'e indirildi.

Şekil 5'teki "Gereksiz Özellikler" aşamasından sonra şu özellikler elde edildi: BirimAdi, Cinsiyet, DogumYerIl, DogumTarihi, KayitYili, Sinif, EgitimYariYili, Durum, MezuniyetTarihi, MezunGANO, OgrenciDurumKodID, OsymKayitYili, LiseNotu, osymOkulTuru, YerPuaniSirasi, YerBolumTercihSirasi, MtokOgrencisi, osymOkulBirincisi, KontenjanTuru, UniversiteYerlestirmeSirasi. Korelasyon haritası çıkarıldı ve yüksek korelasyona sahip özelliklerin 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 yüksek korelasyonlu diğer özellikler kaldırıldı. "Eksik&Aykırı Değerler" aşamasında, boş olan değerler güncellendi. Son durumda, veri ön işleme aşamaları tamamlandığında bu çalışmada kullanılan ve Tablo 2'de ayrıntıları verilen 4023 öğrenci ve 319 özellikten, toplam 1394 öğrenci ve 11 özellik elde edildi.



Şekil 5. Veri önişleme aşamaları ve yapılan işlemler (Data preprocessing stages and operations)

Tablo 2. Özellik değerleri ve açıklamaları (Feature values and descriptions)

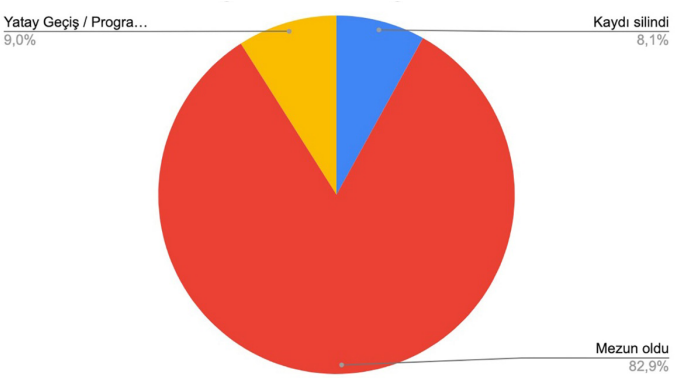
Özellik	Tür	Değerler	Toplam Öğrenci Sayısı	Açıklama
Birim Adı	Kategorik	Bilgisayar Müh Elektrik-Elektronik Müh Makine Müh Mekatronik Müh MetalurjiveMalzeme Müh Tekstil Müh	14 178 365 355 345 137	Öğrencilerin tercih ettiği bölümlerdir. Bilgisayar Mühendisliği mezun vermediği için listeden çıkarıldı.
Cinsiyet	İkili (Boolean)	Kadın: 0 Erkek: 1	316 1078	Cinsiyet. Öğrencilerin cinsiyetleri.
Doğum Yeri Bölge	Kategorik	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.
Lise Notu	Sayısal	142 <= L.N. <= 500		Puan aralığıdır. Normalize edilerek 0 – 1 aralığına indirildi.
Lise Okul Turu	Kategorik	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.
ÖSYM Yerleştirme Sırası	Sayısal	15141<=Y.P.S.<=413695		Puan sıralama. Normalize edilerek 0 – 1 aralığına indirildi.
Tercih Sırası	Sayısal	1 <= T.S. <=28		Tercih sırası. Normalize edilerek 0 – 1 aralığına indirildi.
Üniversite Yerleştirme Sırası	Sayısal	1 <= UYS <= 56		Üniversite yerleştirme sırası. Normalize edilerek 0- 1 aralığına indirildi.
Mtok Öğrencisi	İkili (Boolean)	Evet: 1 Hayır: 0	474 920	MTOK öğrencisi mi?
Okul Birincisi mi	İkili (Boolean)	Evet: 1 Hayır: 0	55 1339	Okul Birincisi mi? Öğrencilerin liseden birincilikle mi mezun olduğunu gösteren özellik
Mezuniyet Durum (Hedef Değişken)	Kategorik	Kaydı silindi Mezun oldu Yatay Geçiş / Program Değişikliği	110 1158 126	Öğrencilerin mezuniyet durumlarını gösteren heden değişken

2.4. Verinin eğitim ve sağlamlık veri setine bölünmesi (Splitting data into training and robustness dataset)

Temizlenmiş (Prepared) veri seti oluşturulduktan sonra, makine öğrenmesi ile oluşturulan modellerin sağlamlığını kontrol etmek için, 2016 yılında kayıt yapan öğrenciler kullanıldı. Bu nedenle, 2016 yılında kayıt yapan 153 öğrenci, temizlenmiş veri listesinden sağlamlık verisi olarak ayrıldı. 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 modeli eğitmek ve doğrulama yapmak üzere kullanıldı. Son durumda Şekil 6'de modeli eğitmek üzere kullanılan veri kümesinin sınıf dengesizliği ve çok sınıflı sınıflandırma problemlerine sahip olduğu görüldü.

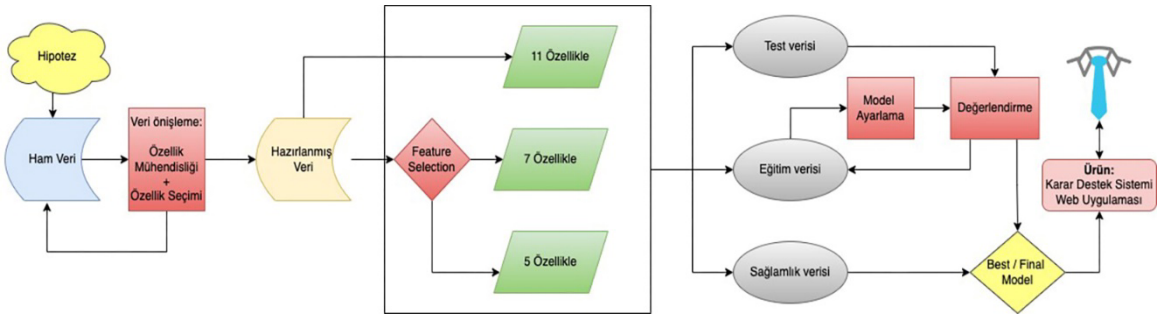
Şekil 7'de görüldüğü gibi veri önişleme ve özellik seçimi tamamlandıktan sonra, hazırlanmış veri sağlamlık verisi, eğitim verisi ve test verisi olarak ayrıldı ve makine öğrenme algoritmaları eğitilip, modeller oluşturuldu ve son modellerden pilot bir karar destek web uygulaması yapıldı. Önişleme ve özellik seçiminden sonra elde edilen hazırlanmış veri kümesinden 2016 yılına ait öğrenci bilgileri

sağlamlık testinde kullanılmak üzere ayrıldı. Böylece, sağlamlık testinde her sınıftan örnek olması sağlandı. Daha sonra, 2016 yılı hariç olan eğitim verisi, sınıf dengesizliği probleminde kullanımı önerilen Repeated Stratified 5*5 fold çapraz geçerleme yöntemi ile eğitildi. Böylece eğitim setinden elde edilen her örneklemde, her sınıftan örnek olması sağlandı. Döngü tamamlanana kadar eğitim verisine, sırasıyla makine öğrenmesi algoritmaları uygulandı ve modeller oluşturuldu.



Şekil 6. Model oluşturmak için kullanılan 1241 öğrencinin hedef özelliği dağılımları (Target feature distributions of 1241 students used to create the model)

2.5. Yöntem (Method)



Şekil 7. Genel akış şeması(General flow chart)

3 ayrı yöntem ile eğitim yaparak modeller oluşturuldu:

1. yöntemde, makine öğrenmesi algoritmaları yeniden örnekleme(resampling) yapılmadan ve hiper parametreler her algoritmaya özel belirlenerek modeller oluşturuldu.

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

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

Her 3 yöntemde de her döngüde oluşan modelin F1 Score, Recall, Accuracy, ROC Area, confusion matrix gibi performans ölçüm değerleri elde edildi. Ayrıca sağlamlık testinde kullanmak için, eğitim sonucunda oluşan her model, .pkl dosya formatında, dosya sistemine kayıt edildi. Döngüler tamamlandıktan sonra tüm değerlerin ortalama sonuçları alındı.

Ek olarak, Şekil 5.'teki Özellik Seçimi aşamasında elde edilen 1241 satır ve 11 özellik içeren hazırlanmış veri seti, Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF, Anova F Test şeklindeki özellik seçimi yöntemleri ile tekrar ele alındı. Elde edilen sonuçlara göre, 1., 2. ve 3. yöntemler, 7 tane en çok kullanılan özellik ve 5 tane en çok kullanılan özellik ile ayrıca eğitildi. Böylece Tablo 3'te gösterildiği gibi, 3 farklı yöntem ve 3 farklı özellik seçimi ile toplam 9 çeşit model türü içeren 750 model elde edildi. Tüm oluşan modeller için sağlamlık kontrolü yapıldı ve eğitilen modellerin Test F1 score değerleri ile karşılaştırılarak, tutarlılık oranları elde edildi. En başarılı ve en tutarlı ilk 4 model seçildi ve Django-Python frameworku kullanılarak, pilot bir karar destek web uygulaması yapıldı.

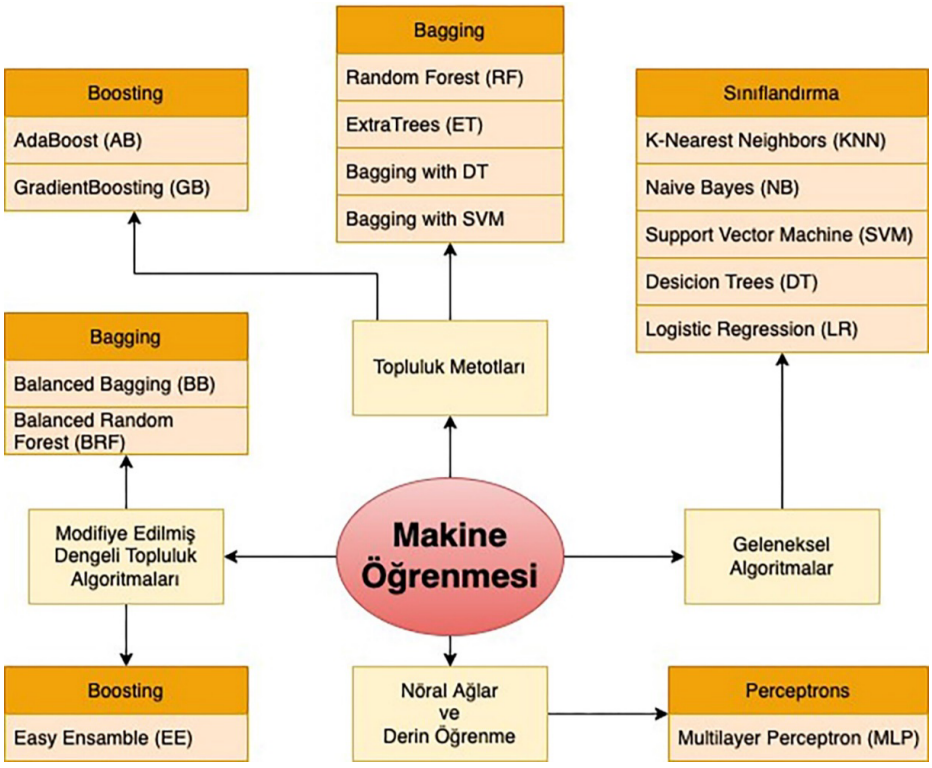
Tablo 3. Yöntemler sonucunda oluşan model sayıları (Number of models resulting from the methods)

	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**

Bu çalışmada kullanılan sınıflandırma algoritmaları, Şekil 8'de gösterildi. Ayrca kullanılan makine öğrenmesi algoritmalarının parametreleri, GridSearchCV yöntemi kullanılarak optimize edildi.

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 olduğu literatürde bulundu[7]. Bu çalışmada, veri kümesine uygun olan örnekleme yöntemlerinden, SMOTE, SMOTENC, Borderline SMOTE, SVMSMOTE, ADASYN, Random Over-sampling ve hibrit yöntemlerden SMOTE - ENN, SMOTE - Tomek kullanıldı.

Dengesiz veri kümelerinde, hold out cross validation, k-fold cross validation, Repeated random subsampling validation her sınıftan örneklem almayabileceğinden ötürü kullanımının uygun olmadığı görüldü. StratifiedKFold, katmanlı kıvrımları döndüren bir k-fold çeşidi olduğu ve her iterasyonda, her set, tam set olarak her bir hedef sınıftan yaklaşık olarak aynı yüzdede örnek içerdiği görüldü. Böylece, özellikle dengesiz veri kümelerinde, doğrulama veya eğitim verilerinde belirli bir sınıfın fazla bulunmasını engellediği tespit edildi. Böylece, az örneklem içeren sınıfların da eğitilmesi sağlandı. Bu nedenle dengesiz veri setinde kullanımının uygun olduğu görüldü. Bu çalışmada da kullanılan veri kümesinde sınıf dengesizliği problemi olduğu için, Repeated Stratified 5*5-fold cross-validation yöntemi kullanıldı.



Şekil 8. Modellemede kullanılan makine öğrenmesi algoritmaları. (Machine learning algorithms used in modeling)

3. Bulgular (Findings)

Sınıf dengesizliği problemi olan veri kümelerinde, performans ölçümü olarak Doğruluk/Accuracy değerinin kullanılmasının yanıltıcı başarı oranı sağlayabildiği görüldü. Bu nedenle Doğruluk/Accuracy yerine, dengesiz veri setlerinde kullanımı güvenilir olan F1 Puanı/F1 Score, Dengeli Doğruluk/Balanced Accuracy değerleri kullanıldı. Doğruluk/Accuracy, Geri Çağırma/Recall ve Özgünlük/Specificity değerleri ayrıca incelendi. Ayrıca, bu çalışmada dengesiz sınıf dağılımı olduğu için, Doğruluk/Accuracy değerinin başarı oranında Dengeli Doğruluk/Balanced Accuracy ve F1 Puanı/F1 Score yükseldiğinde düşüş olması ihtimali olduğu belirtildi.

3.1. Özellik seçimi (Feature Selection)

Tablo 4'te görüldüğü üzere, Pearson, Chi-2, Mutual-Info, RFE, Logistics, RF ve Anova F testleri ile özellik sayısı 5 olduğunda iken özellikler karşılaştırıldı. Değerler incelendiğinde, Cinsiyet, DoğumYeriBolge gibi demografik bilgilerin ve Osym Okul Birincisi ve Lise Okul Türü olan lise akademik bilgilerin de en az etkili olan özellikler olduğu görüldü. Bu değerlere göre, 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 4. Özelliklerin feature=5 olduğunda önem sıralamalarına göre karşılaştırmalı tablosu (Comparative table of features in order of importance when feature=5)

Feature	Pearson	Chi-2	Mutual-Info	RFE	Logistics	RF	Anova F Test	Total
MtokOgrencisi	Evet	Evet	Evet	Evet	Evet	Hayır	Evet	6
YerBolumTercihSirasi	Hayır	Evet	Hayır	Evet	Evet	Evet	Evet	5
LiseNotu	Evet	Hayır	Evet	Evet	Hayır	Evet	Evet	5
BirimAdi	Evet	Evet	Evet	Evet	Evet	Hayır	Hayır	5
YerPuaniSirasi	Evet	Hayır	Evet	Hayır	Hayır	Evet	Evet	4
KayitYili	Evet	Hayır	Evet	Hayır	Hayır	Evet	Evet	4
MarmaraUniYerlestirmeSirasi	Hayır	Hayır	Hayır	Evet	Hayır	Evet	Hayır	2
OsymOkulBirincisi	Hayır	Evet	Hayır	Hayır	Hayır	Hayır	Hayır	1
LiseOkulTuru	Hayır	Evet	Hayır	Hayır	Hayır	Hayır	Hayır	1
DogumYeriBolge	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	0
Cinsiyet	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	0

3.2. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre model sonuçları (Model results of data containing 11, 7 and 5 features according to method 1)

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 1. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 5.'te gösterildi. Bu tabloya göre 7 özellik ile elde edilen değerlerin, 11 özellik ile elde edilen değerlerle oldukça benzerlik gösterdiği görüldü. Adaboost ve SVM algoritmalarında ciddi yükseliş olduğu görüldü. Adaboost sağlamlık değerinde de ciddi yükseliş oldu. Bununla beraber, RF, MLP, GradientBoost, ET, BalancedBagging ve Bagging_DT algoritmaları neredeyse aynı sonucu verdi. Bu tabloya bakıldığında, model olarak 7 özellikli kullanılması maliyet ve zaman açısından oldukça mantıklı görüldü.

3.3. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre model sonuçları (Model results of the data containing 11, 7 and 5 features according to the 2nd method)

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 2. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 6.'da gösterildi. 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 değerlerin 1. ve 2. yöntem sonuçlarında hemen hemen aynı olduğu görüldü.

Tablo 5. 11, 7 ve 5 özellik içeren verinin 1. yönteme göre modellendiğinde elde edilen sonuçlar (The results obtained when the data containing 11, 7 and 5 features are modeled according to the 1st method)

1. Yöntem			**11 özellik**		**7 özellik**		**5 özellik**	
	Model	**Yöntem**	**Test F1 Score**	**Robustness F1 Score**	**Test F1 Score**	**Robustness F1 Score**	**Test F1 Score**	**Robustness 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	Easy Ensemble	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	nan	0.7618
11	Gradient Boost	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	nan	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 6. 11, 7 ve 5 özellik içeren verinin 2. yönteme göre modellendiğinde elde edilen sonuçlar (The results obtained when the data containing 11, 7 and 5 features are modeled according to the 2nd method)

2. Yöntem	**11 özellik ile**			**7 özellik ile**			**5 özellik ile**	
Model	**Yöntem**	**Test F1 Score**	**Robustness F1 Score**	**Yöntem**	**Test F1 Score**	**Robustness F1 Score**	**Yöntem**	**Test F1 Score**	**Robustness F1 Score**
AdaBoost	ROS	0.7734	0.7299	SVM SMOTE	0.7761	0.7248	Borderline SMOTE2	0.6879	0.3171
Bagging_DT	SVM SMOTE	0.8203	0.7531	SVM SMOTE	0.7973	0.7474	ROS	0.7287	0.5884
Bagging_SVC	SVM SMOTE	0.7501	0.7465	Borderline SMOTE2	0.7344	0.5372	Borderline SMOTE2	0.6916	0.4053
BernoulliNB	SVM SMOTE	0.6521	0.6524	SVM SMOTE	0.6010	0.5313	SVM SMOTE	0.5961	0.3898
Decision Tree	BorderlineSMOTE2	0.7803	0.7135	SMOTENC	0.7621	0.6228	ROS	0.6968	0.6856
Extra Trees	ROS	0.8196	0.7713	ROS	0.8221	0.7714	SMOTENC	0.7171	0.4017
GaussianNB	SVM SMOTE	0.5776	0.5133	ROS	0.5996	0.4696	SVM SMOTE	0.5985	0.3724
Gradient Boost	SVM SMOTE	0.8194	0.7179	SVM SMOTE	0.8138	0.7276	SVM SMOTE	0.7189	0.4564
KNN	ROS	0.7522	0.7147	ROS	0.7729	0.7185	ROS	0.7327	0.6118
LR	SVM SMOTE	0.7451	0.6420	SVM SMOTE	0.7317	0.6999	SVM SMOTE	0.6934	0.3784
MLP	SVM SMOTE	0.7982	0.7702	ROS	0.8195	0.7726	SVM SMOTE	0.7223	0.3300
RF	ROS	0.8263	0.7705	ROS	0.8174	0.7714	Borderline SMOTE1	0.7258	0.5851
SVM	SVM SMOTE	0.7568	0.7501	SVM SMOTE	0.7503	0.6544	SVM SMOTE	0.7130	0.3201
En yüksek değerler:		0.82 - 0.83	0.77		0.82	0.77		0.72 - 0.73	0.68

3.4. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre model sonuçları (Model results of data containing 11, 7 and 5 features according to the 3rd method)

7 ve 5 özellik kullanılarak yapılan modellerin sonuçlarının 11 özellik içeren 3. yöntem modelinin F1 Score başarı ve sağlamlık sonuçlarıyla kıyas tablosu Tablo 7'de gösterildi.

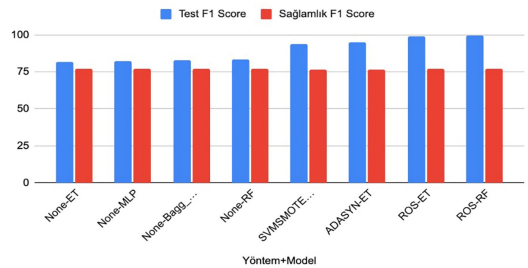
Tablo 7. 11, 7 ve 5 özellik içeren verinin 3. yönteme göre modellendiğinde elde edilen sonuçlar (The results obtained when the data containing 11, 7 and 5 features are modeled according to the 3rd method)

3. Yöntem	**11 özellik ile**			**7 özellik ile**			**5 özellik ile**		
Model	**Yöntem**	**Test F1 Score**	**Robust-ness F1 Socre**		**Test F1 Score**	**Robust-ness F1 Socre**		**Test F1 Score**	**Robust-ness F1 Socre**
AdaBoost	ROS	0.9542	0.7101	SMOTENC	0.6546	0.3978	SMOTEENN	0.6387	0.3043
Bagging_DT	SVM SMOTE	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
Bernoulli NB	SVM SMOTE	0.6824	0.6314	SVM SMOTE	0.5735	0.4105	SVM SMOTE	0.5806	0.3635
Decision Tree	ROS	0.8790	0.6836	BorderlineSMOTE2	0.7742	0.7401	SMOTEENN	0.8814	0.3287
ExtraTrees	ROS	0.9917	0.7713	ROS	0.9929	0.7713	ROS	0.9842	0.6649
Gaussian NB	SVM SMOTE	0.6324	0.4832	SVM SMOTE	0.5687	0.4138	SMOTEENN	0.6038	0.2043
Gradient Boost	SVM SMOTE	0.8349	0.7031	SVM SMOTE	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	SVM SMOTE	0.7193	0.6617	SMOTEENN	0.5967	0.4355	SMOTEENN	0.6306	0.1805
MLP	SVM SMOTE	0.6951	0.7721	SVM SMOTE	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

Tablo 3.14, Tablo 6 ve Tablo 5'teki değerler karşılaştırıldığında, tüm veri 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 değerlerin 1. ve 3. yöntem sonuçlarında hemen hemen aynı olduğu görüldü.

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.

Sonuç olarak Tablo 5'de belirtilen en başarılı ve en tutarlı ilk 4 modelin, 7 özellikle oluşturulan modelin değerleri arasındaki F1 Score Başarı Oranı ve tutarlılık değeri yakın olduğu görüldü. Şekil 9'da görüldüğü üzere bu modellere ait 7 özellik içeren modellerin web uygulamasında kullanılmasına karar verildi. ET, MLP, Bagging_DT ve RF algoritmalarıyla 1. yöntem olan ve yeniden örneklem yapılmadan oluşturulan 7 özellikli modeller, karar destek sistemi web uygulamasında kullanıldı.



Şekil 9. En başarılı ve en tutarlı modellerin test ve sağlamlık F1 Score yüzdeleri (Testing and robustness F1 Score percentages of the most successful and consistent models)

4. Çalışmanın Güçlü ve Zayıf Yönleri (Strengths and Aspects of the Study)

Güçlü Yönler;

- Bu çalışmaya başlamadan önce sistematik literatür incelemesi yapıldı.
- Literatürde daha önce kullanılmayan yeni bir veri seti ile çalışma yapıldı.
- MÜ'nde alanında yapılan ilk çalışma oldu.
- Türkiyedeki tezlerde çok sınıflı ve dengesiz eğitimsel veri kümesi kullanılarak daha önce tez çalışması yapılmadı.
- Öğrenciler henüz kayıt olmadan, erken uyarı sistemi olarak çalışma yapıldı ve bu şekilde de literatürde çalışmanın yok denecek kadar az olduğu görüldü.
- Topluluk algoritmalarla, hem çok sınıflı hem dengesiz veri hem de eğitimsel verisinde kabul edilebilir, başarılı sonuçlar elde edildi.
- En başarılı ve en tutarlı modellerle pilot bir karar destek sistemi web uygulaması yapıldı ve MÜ TF öğretim üyelerinin kullanımına sunuldu.

Zayıf Yönler;

- MÜ TF yeni açılmış bir fakültedir ve henüz çok fazla mezun vermedi.
- Bölümler aynı yılda açılıp, aynı yılda mezun vermeye başlamadı.
- Af, mühendislik tamamlama gibi farklı türdeki öğrenciler de veri kümesinde yer aldı.
- Çalışmada kullanılan toplam öğrenci sayısı 1394 oldu ve veri kümesi sınıf dengesizliğine sahipti. Daha çok öğrenci bilgisi içeren bir veri seti ile çalışmak daha başarılı sonuçlar verebilir diye düşünüldü.
- Bazı algoritmalar, çok sınıflı ve dengesiz veri setinde başarılı sonuç vermedi.

5. Web Uygulaması (Web Application)

Tartışmalar bölümünde yapılan incelemeler ve karşılaştırmalar sonucunda, oluşturulan karar destek sistemi web uygulaması için seçilen modeller, yeniden örnekleme yapılmadan, 7 özellikle oluşturulan ve Şekil 9'da F1 Score değerleri gösterilen None-RF, None-MLP, None-Bagging_DT ve None-ET algoritmalarına ait modeller oldu.

Web uygulaması için, Django Python frameworku kullanıldı, Docker üzerinde çalıştırıldı. Başlangıçta, yetkilendirme ile sayfaya giriş yapılması sağlandı. Böylece, web uygulamasındaki güvenlik üst düzeye çıkarıldı. Şekil 10.'daki 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 grafik şeklinde Şekil 11.'deki gibi gösterildi. Bu pilot çalışma, MÜ TF öğretim üyelerinin, kullanabileceği bir uygulama olarak geliştirildi.

Şekil 10. Karar Destek Sistemi web uygulamasında öğrenci bilgilerinin girildiği sayfa (The page where student information is entered in the Decision Support System web application)

Şekil 11. Karar Destek Sistemi web uygulamasında yapılan tahmin sonuçları (Prediction results made in Decision Support System web application)

6. Sonuçlar ve Karşılaştırma (Results and Comparison)

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 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 az olduğu görüldü. Literatürdeki bu boşluğun doldurulması amacıyla, bu tez çalışmasında, çok sınıflı ve dengesiz eğitimsel veri kümesi kullanarak, öğrenciler henüz üniversiteye başlamadan riskli öğrencilerin mezuniyet tahminini yaparak, bu öğrencilerin risklerini engelleyebilecek bir karar destek öneri sistemi çalışması yapıldı. Bu çalışmada, 23.05.2022-286783 etik no kararı ile MÜ TF öğrenci verileri kullanıldı. MÜ TF 2010-2021 yıllarında öğrenim gören, 6 bölüme ait 4023 öğrenci içerisinden, ön işlem aşamasından sonra kalan 1394 öğrenci ve 11 özellik üzerinde çalışıldı.

Kullanılan veri kümesi çok sınıflı ve sınıf dengesizliği olan eğitimsel bir veri kümesi oldu. 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 edildi. Bu yöntemlerden birincisi, topluluk yöntemleri kullanarak başarı tahminini yapmak oldu. Bir diğeri, yeniden örnekleme yöntemleriyle beraber kullanılan geleneksel sınıflandırma algoritmaları ile model üretmek oldu. Ve sonuncusu ise, dengesiz veri setlerinde kullanmak üzere modifiye edilmiş topluluk metotlarıyla model üretmek oldu.

Toplamda, 9 tane yeniden örnekleme yöntemi kullanıldı. 3 tane modifiye edilmiş topluluk algoritması, 7 tane geleneksel sınıflandırma algoritması, 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 kullanılarak hiper-parametre ayarlaması yapıldı. 2016 yılında kayıt olan 153 öğrenci sağlamlık verisi olarak ayrıldı. Dengesiz veri setleri kullanımında uygun olduğu için repeated stratified 5*5-fold cross-validation kullanıldı. Bu aşamada oluşturulan tüm modeller, sağlamlık kontrolü için kayıt edildi. Ve tüm değerlerin ortalaması alınarak, son model başarı değerleri belirlendi.

Elde edilen sonuçlar ile çok sınıflı 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.

Sonuç olarak RandomOverSampler+RandomForest F1 Score 0.9935 değeriyle en başarılı algoritma olmasına rağmen, sırasıyla en tutarlı olan 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 en tutarlı 4 model olduğu görüldü. Bu modellerin 11 özellikli ve 7 özellikli değerleri birbirine çok yakın olduğu tespit edildi. Bu nedenle maliyet, zaman açısından düşünüldüğünde, 7 özellikle oluşturulan modellerin kullanılmasına karar verildi.

Son olarak, en başarılı ve en tutarlı 4 model ile MÜ TF 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 aldı.

Bu bağlamda, bu çalışma ve literatür çalışmaları arasında bazı benzerliklerin ve farklılıkların olduğu görüldü. Öncelikle kullanılan veri kümeleri farklı olduğu için sonuçlarla ilgili karşılaştırma yapılmadı ancak literatürdeki çalışmaların başarı değerlerine benzer ve yüksek başarı değerleri elde edildi. Çok sınıflı dengesiz veri kümesi ile ilgili kullanılan yöntemlerin benzer olduğu görüldü. Yani, hem tüm veri setine örnekleme yapılması literatürdeki çalışmalarda da kullanıldı hem de bu çalışmada kullanılan makine öğrenmesi algoritmaları ve örneklem yöntemleri farklı farklı literatürdeki çalışmalarda kullanıldı. Ancak literatürden farklı olarak daha tutarlı olduğu gözlemlenen, sadece eğitim veri kümesine örneklem yöntemlerinin uygulanması, yöntem olarak bu çalışmada kullanıldı. Ayrıca literatürdeki çalışmalardan farklı olarak bu çalışmada sağlamlık kontrolü yapıldı ve en başarılı algoritmaların tutarlı olup olmadığı kontrol edildi. Ayrıca literatürdeki çalışmalardan farklı olarak dengesiz veri kümesinde kullanımı daha uygun olan Repeated Stratified 5*5-fold cross-validation yöntemi bu çalışmada kullanıldı. Ayrıca literatürdeki çalışmalardan farklı olarak, en tutarlı modellerin kullanıldığı karar destek sistemi web uygulaması yapıldı. Yine 7 farklı özellik seçim yöntemleri kullanarak elde edilen veri uzunluğu aynı ancak özellik sayısı farklı toplam 3 farklı veri kümesi ile tüm yöntemler denenerek sonuçların karşılaştırılması literatürden farklı olarak bu çalışmada yapıldı.

Gelecekte, çalışmanın geliştirilebilir yönleri olarak öncelikle çalışmanın zayıf yönleri belirlendi. Ayrıca dengesiz veri kümesine karşı daha dayanıklı algoritmaları çalışmaya dahil ederek başarının artıp artmadığı test edilebilir. Yine çalışmaya dahil edilmeyen ama literatürde güncel olan örneklem yöntemleri sisteme eklenip, başarının artıp artmadığı test edilebilir. MÜ'nden başarı durumunu daha belirgin tahmin etmemizi sağlayacak yeni özellikler içeren veri listesi alınarak veri ön işleme ve özellik seçimi kısmı çalışmada geliştirilebilir. Anomali tespiti gibi farklı kapsamlardaki çalışmalara ait yöntemler incelenerek çok sınıflı dengesiz eğitimsel veri kümesinde denenerek başarının artıp artmadığı test edilebilir. Literatürdeki çalışmalarda kullanılan öncelikle açık kaynak veri kümeleri ile bu çalışmanın yöntem bölümü test edilerek, bu çalışmanın yönteminin genelleştirilebilirliği tartışılabilir. Ayrıca önerilen çözümün genelliğini değerlendirmek için bu çalışma MÜ'nin diğer bölümlerine ve başka üniversitelere de uygulanarak, birçok öğrenciden elde edilen veriler üzerinde daha fazla araştırma yapılıp, oluşturulan sistemin genelleştirilip genelleştirilemeyeceği araştırılabilir. Sadece öğretim üyelerinin kullanımına sunulan karar destek sistemi pilot web uygulaması, öğrencilerin de kullanabileceği şekilde geliştirilebilir.

Son olarak, çok sınıflı dengesiz eğitimsel veri kümesi kullanarak yapılan bu çalışma ulusal alanda alanında tek olması dolayısıyla, gelecekteki çalışmalar için bu çalışmanın ilham kaynağı olduğuna inanılmaktadır.

Teşekkür (Acknowledgment)

Bu çalışmada, 23.05.2022-286783 etik no kararı ile Marmara Üniversitesi Teknoloji Fakültesi öğrenci verilerinin kullanılmasına müsaade eden Marmara Üniversitesi rektörü Sayın Prof. Dr. Mustafa Kurt Hocamıza teşekkür ederiz. Ayrıca bu çalışmada maddi bir destek alınmamıştır.

Çıkar Çatışması Beyanı (Conflict of Interest Statement)

Yazarlar tarafından herhangi bir çıkar çatışması bildirilmemiştir.

Kaynaklar (Resources)

[1] Mevzuat, "Yükseköğretim Kanunu," 2022. [Çevrimiçi]. Erişilebilir: https://www.mevzuat.gov.tr/MevzuatMetin/1.5.2547.pdf, [Erişim tarihi: 25.04.2022].

[2] A. Hancı Karademirci, "Öğretim teknolojileri: tanımı ve tarihsel gelişimine yeniden bakmak," *in Akademik Bilişim'10 – XII. Akademik Bilişim Konferansı Bildirileri*, Muğla Üniversitesi, Şubat 10-12, 2010, ss. 397-403.

[3] Datamining, "EDM Tanımı" 2022. [Çevrimiçi]. Erişilebilir: https://educationaldatamining.org/. [Erişim tarihi: 25.04.2022].

[4] M. Kuhn ve K. Johnson, *Applied predictive modeling*, Vol. 26. New York: Springer, 2013. doi:10.1007/978-1-4614-6849-3

[5] D. Çelik, "11. sınıf öğrencilerinin düşünme stilleri, öğrenme stratejileri ve düşünme stilleri ile öğrenme stratejileri arasındaki ilişki," Yüksek Lisans tezi, Pamukkale Üniversitesi Eğitim Bilimleri Enstitüsü, Denizli, 2016.

[6] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk ve F. Herrera. Learning from Imbalanced Data Sets. vol. 10. Springer International Publishing, 2018.

[7] A. Dutt, M. A. Ismail, T. Herawan, "A systematic review on educational data mining," *IEEE Access*, Vol. 5, pp. 15991-16005, 2017. doi: 10.1109/ACCESS.2017.2654247.

[8] M. Tatlıdil, "Veri Türleri ve İstatistiğe Giriş", Mayıs 15, 2020. [Çevrimiçi]. Erişilebilir: https://mervetatlidil.medium.com/veri-t%C3%BCrleri-ve-i%CC%87statisti%C4%9Fe-giri%C5%9F-2959f509f768. [Erişim tarihi: 25.04.2022].

[9] J. Brownlee, "One-vs-Rest and One-vs-One for Multi-Class Classification", Nisan 13, 2020. [Çevrimiçi]. Erişilebilir: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification./. [Erişim tarihi: 25.04.2022].

[10] A. Khan ve S.K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ Inf Technol,* vol. 26, pp. 205–240, 2021. doi: 10.1007/s10639-020-10230-3.

[11] Y. Pristyanto, I. Pratama ve A. F. Nugraha, "Data level approach for imbalanced class handling on educational data mining multiclass classification," *in 2018 International Conference on Information and Communications Technology (ICOIACT)*, March, 2018. pp. 310-314.

[12] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, ve N. A. M. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608-95621, 2021.

[13] R. Ghorbani ve R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899-67911, 2020.

[14] I. Pratama, Y. Pristyanto, P. T. Prasetyaningrum, "Imbalanced Class handling and Classification on Educational Dataset," *in 4th International Conference on Information and Communications Technology (ICOIACT)*, August, 2021. pp. 180-185.

[15] V. T. N Chau ve N. H. Phung, "Imbalanced educational data classification: An effective approach with resampling and random forest," *in The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*, November, 2013, pp. 135-140.

[16] T. Purwoningsih, H. B. Santoso, K. A. Puspitasari and Z. A. Hasibuan "Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine-Learning Classification Algorithm," *Journal of Hunan University Natural Sciences*, vol. 48, no. 9, 2021.

[17] T. Lenin, N. Chandrasekaran, "Learning from imbalanced educational data using ensemble machine learning algorithms," *Webology*, vol. 18, no. SI01, pp. 183-195, 2021.

[18] E. Buraimoh, R. Ajoodha and K. Padayachee, "Application of machine learning techniques to the prediction of student success," *in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1-6. doi:10.1109/IEMTRONICS52119.2021.9422545.

[19] N. Rachburee, W. Punlumjeak, "Oversampling technique in student performance classification from engineering course," *International Journal of Electrical and Computer Engineering; Yogyakarta,* vol. 11, No. 4, pp. 3567-3574, August 2021.

[20] E. Buraimoh, R. Ajoodha, K. Padayachee, "Importance of data re-sampling and dimensionality reduction in predicting students' success," *in 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE),* June, 2021, pp. 1-6.