Öğrencilerin Okulu Bırakma Durumunu ve Akademik Başarısını Tahmin Etmek

Osman Cihan Ergüden 23181616064

Emre Kalkan 24181617002

Özet:

Yükseköğretim kurumları, öğrencileri hakkında önemli miktarda veri kaydetmektedir. Bu veriler bilgi, izleme ve karar destek sistemleri üretme açısından önemli bir potansiyele sahiptir. Yükseköğretimde okul terki ve eğitimde başarısızlık; ekonomik büyüme, istihdam, rekabetçilik ve verimlilik açısından bir engel teşkil etmekte, öğrencilerin ve ailelerinin yaşamlarını, yükseköğretim kurumlarını ve tüm toplumu doğrudan etkilemektedir. Bu çalışmada tanıtılan veri kümesi, farklı ve bağımsız veri kaynaklarının birleştirilmesiyle oluşturulmuş olup; demografik, sosyoekonomik, makroekonomik verilerle kayıt ve akademik performans bilgilerini içermektedir. Bu veri kümesi, akademik başarı ve okul terki tahmini için makine öğrenimi modelleri geliştirmek amacıyla kullanılmıştır. Geliştirilen bu sistem, Portalegre Politeknik Enstitüsü bünyesindeki danışmanlık ekiplerine, öğrencilerin başarısızlık ve terk riski hakkında bilgi sağlayan bir öğrenme analitiği aracının parçasıdır. Bu veri kümesi, öğrenci performansıyla ilgili karşılaştırmalı çalışmalar yapmak isteyen araştırmacılar ve makine öğrenimi eğitimi için faydalı bir kaynak niteliğindedir

Anahtar Kelimeler: Öğrenci Devamsızlığı, Sınıflandırma, SMOTE, KNN, Karar Ağaçları, ANN, akademik performans; eğitimde makine öğrenmesi; dengesiz sınıflar; çok sınıflı sınıflandırma; eğitimsel veri madenciliği; öğrenme yönetim sistemi; tahmin

1.Giriş:

Öğrenci devamsızlığı ve başarısızlık, yükseköğretim kurumlarında önemli bir problemdir.

Bu çalışma, öğrencilerin ilk iki dönem sonunda okulda kalıp kalmayacağı veya mezun olup olamayacağına dair öngörüler sunmayı amaçlar. Problemi öğrenci durumu başlığı altında 3 sınıfta inceler ("Dropout", "Enrolled", "Graduate")

Materyal:

Öğrencilere ait birleştirilmiş kayıt sistemlerinden oluşturulmuş.

Öğrenci yaşı, kabul notu, birinci ve ikinci dönem sonu yıl sonu notları, cinsiyet, bursluluk, öğrenci milliyeti, öğrencinin yurtdışından gelip gelmediği, öğrencinin özel gereksiniminin olup olmadığı vb.

Hedef değişkenler: 0 = Dropout(Okulu bırakma), 1 = Enrolled(Kayıtlı öğrenci), 2 = Graduate(Mezun öğrenci)

Eksik veriler median ile doldurulmuş, kategorik değişkenler LabelEncoder ile kodlanmış.

Yöntem

Sayısal veriler StandardScaler ile ölçeklendirildi.

Dengesiz sınıf dağılımı SMOTE yöntemi ile dengelendi.

KNN: GridSearchCV ile en uygun "k" değeri belirlendi.

Karar Ağacı: Maksimum derinlik parametresi optimize edildi.

Yapay Sinir Ağı (ANN): Dropout ve EarlyStopping içeren, çok katmanlı bir model kullanıldı. Keras ile geliştirildi.

2. Veri Kümesinin Tanımı

Veri kümesi; demografik, sosyoekonomik, makroekonomik veriler ile öğrenci kayıt anındaki ve birinci ile ikinci dönem sonundaki akademik verileri içermektedir. Kullanılan veri kaynakları hem kurum içinden hem de dışından sağlanmıştır ve şunları kapsamaktadır:

Kurumun Akademik Yönetim Sistemi (AMS),

Kurumun öğretim faaliyetlerini desteklemek amacıyla geliştirilen iç sistem (PAE),

Yükseköğretime Ulusal Giriş Sınavı (CNAES) ile kabul edilen öğrencilerle ilgili Yükseköğretim Genel Müdürlüğü (DGES) verileri,

Portekiz'e ait çağdaş makroekonomik verileri içeren PORDATA veri tabanı.

Veriler, Avrupa'da Bologna Süreci'nin uygulanmasından sonraki yılları kapsayan 2008/2009 ila 2018/2019 akademik yılları arasında kayıt yaptırmış öğrencilere aittir. Veri kümesi; tarım, tasarım, eğitim, hemşirelik, gazetecilik, işletme, sosyal hizmet ve teknolojiler gibi farklı disiplinlerdeki 17 lisans programına kayıtlı öğrencileri içermektedir.

Nihai veri kümesi UTF-8 ile kodlanmış virgülle ayrılmış değerler (CSV) biçimindedir ve 35 öznitelikten oluşan toplam 4424 öğrenci kaydını içermektedir. Veride eksik bilgi bulunmamaktadır.

Veri kümesindeki öznitelikler, aşağıdaki başlıklar altında gruplanmıştır:

Demografik Veriler: Medeni durum, uyruk, yer değiştirme durumu, cinsiyet, kayıt yaşı, uluslararası öğrenci olup olmama.

Sosyoekonomik Veriler: Anne-baba eğitim durumu, mesleği, özel eğitim ihtiyacı, borçlu olma durumu, harç ödemeleri, burs durumu.

Makroekonomik Veriler: İşsizlik oranı, enflasyon oranı, gayrisafi yurt içi hasıla (GSYH).

Kayıt Anı Akademik Veriler: Başvuru şekli, sıralama, bölüm, gündüz/gece öğretimi, önceki eğitim durumu.

1. Dönem Sonu Akademik Veriler: Kredi aktarımı, alınan ders sayısı, değerlendirilen dersler, geçilen dersler, ortalama not, değerlendirilmeyen dersler.

2. Dönem Sonu Akademik Veriler: Yukarıdakilerin ikinci dönem için karşılıkları.

Hedef Değişken (Target): Öğrencinin mezun, kayıtlı ya da terk durumu.

Veri kümesi hem araştırma hem de makine öğrenimi algoritmalarının eğitimi açısından kullanılabilecek niteliktedir.

3. Materyal ve Yöntem

Bu bölümde veri kümesinin oluşturulma süreci ve örnek bir keşifsel veri analizi açıklanmaktadır. Bu analiz, verinin dengesiz doğası, öznitelikler arasındaki çoklu bağlantılar (multicollinearity) ve literatürde benzer problemlerde en çok kullanılan algoritmalara göre öznitelik önemini ortaya koymaktadır.

3.1. Veri Ön İşleme (Data Preprocessing)

Veriler üç farklı formatta toplanmıştır:

AMS verileri: Virgülle ayrılmış CSV dosyaları şeklinde,

PORDATA verileri: Makroekonomik göstergeleri içeren manuel olarak toplanmış bilgiler.

AMS'den alınan öğrenci kayıtları, 13.992 satır ve 398 sütun içermekteydi. Ancak birçok satır ve sütun tekrarlı ya da çalışmayla ilgisiz olduğundan temizlenmiştir. Eski programlara kayıtlı öğrenciler ve Erasmus gibi istisnai yollarla gelenler dışlanmıştır. Seçilen öznitelikler adlandırılmış, tekrarlar kaldırılmış ve veriler "Demografik" ve "Sosyoekonomik" veri gruplarına ayrılmıştır.

Öğrenci değerlendirme verileri de ayrı bir CSV dosyasından alınmış ve önceki adımdan gelen her öğrenciye karşılık gelecek şekilde işlenmiştir. Burada, birinci ve ikinci dönem sonu akademik veriler hesaplanmıştır.

Son adımda, tüm veriler birleştirilmiş ve "Makroekonomik Veriler" eklenmiştir. Ardından anormallikler, açıklanamayan aykırı değerler ve eksik veriler temizlenmiştir. Her öğrenci, program süresi sonunda (çoğu için 3 yıl, hemşirelik için 4 yıl) mezun, kayıtlı ya da bırakmış şeklinde sınıflandırılmıştır

4.Bulgular ve Tartışma

Algoritma	Doğruluk
KNN	%83.71
Karar Ağacı	%74.81
Yapay Sinir Ağı (ANN)	%79.64

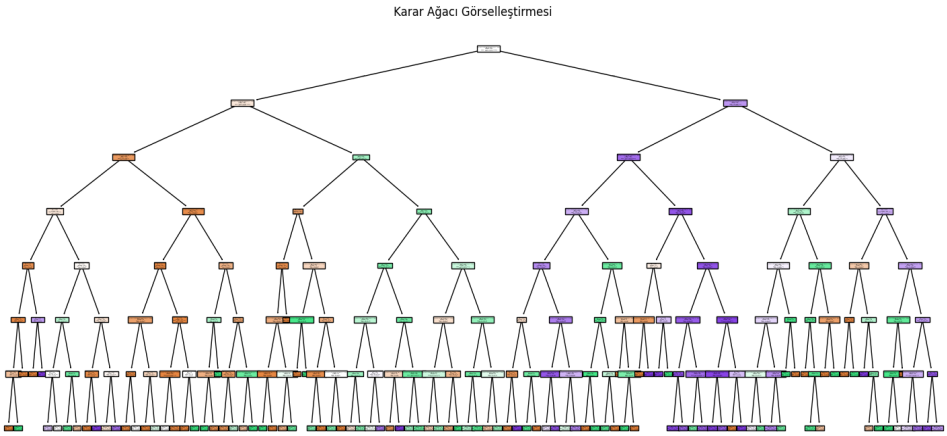
En iyi sonuç KNN ve ANN ile alınmıştır.

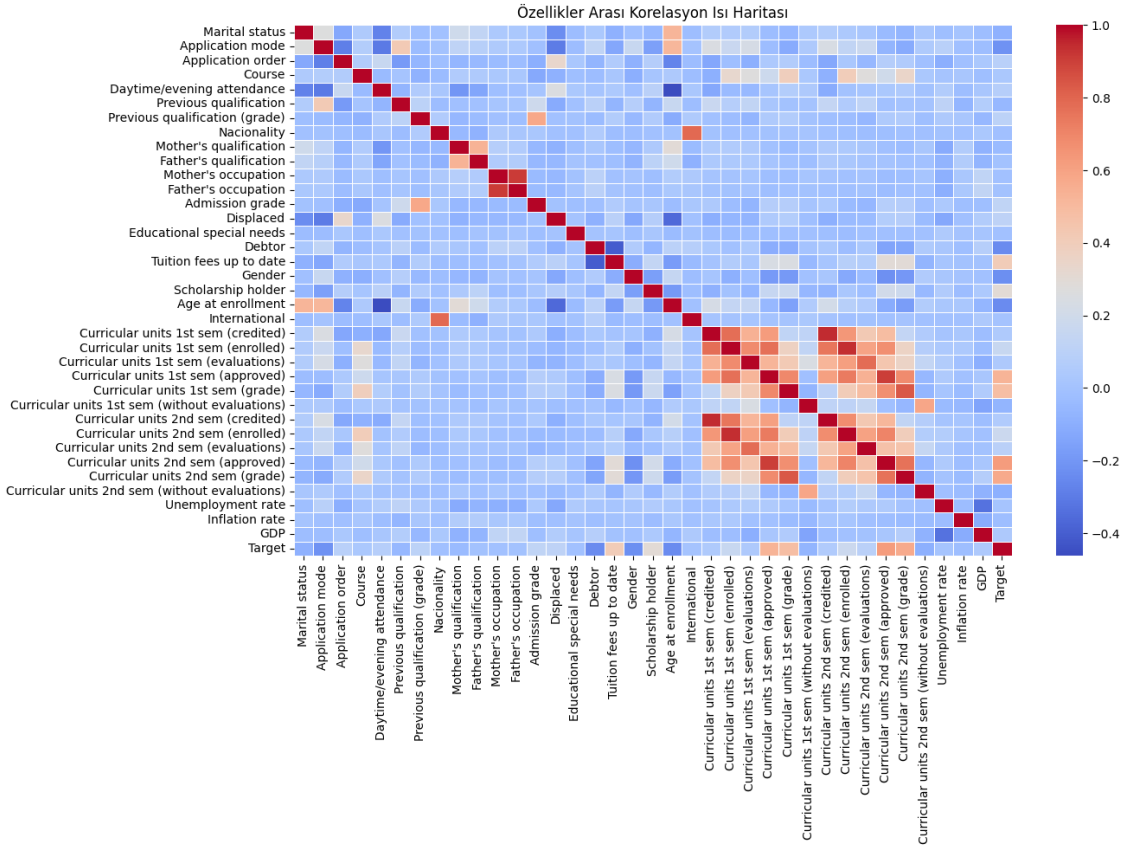
Karar Ağacı modeli, görselleştirilmiş ve yorumlanabilir yapısı ile avantaj sağlar.

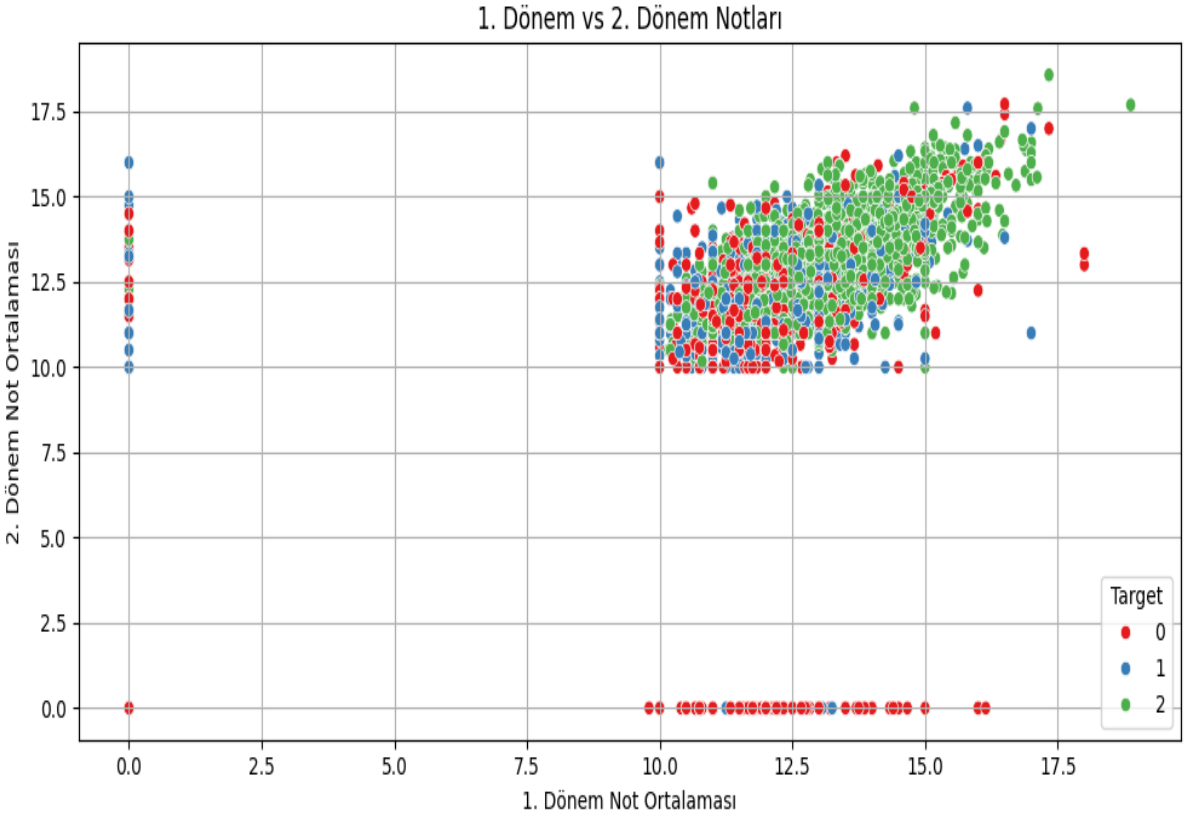
KNN performansı düşüktür, çünkü özellik sayısı fazla ve sınıflar karmaşıktır.

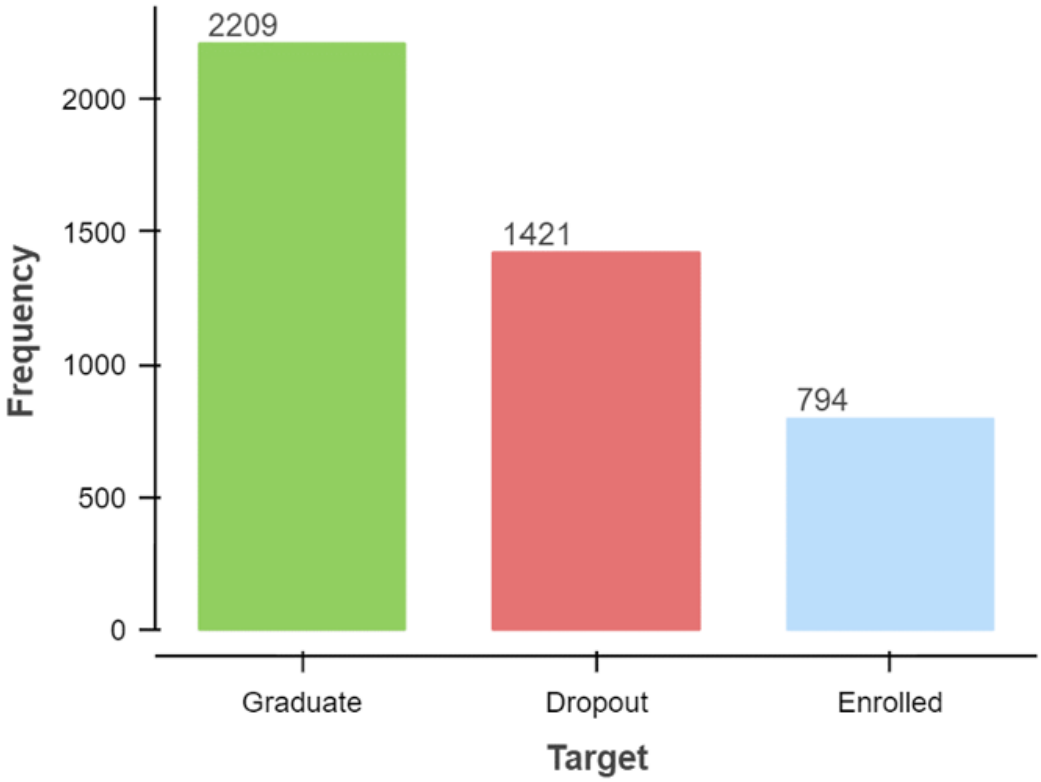
Görsel analizlerde dönem notları, geçilen ders sayısı ve yaş-kabul notu ilişkisi grafiklerle ortaya konmuştur.

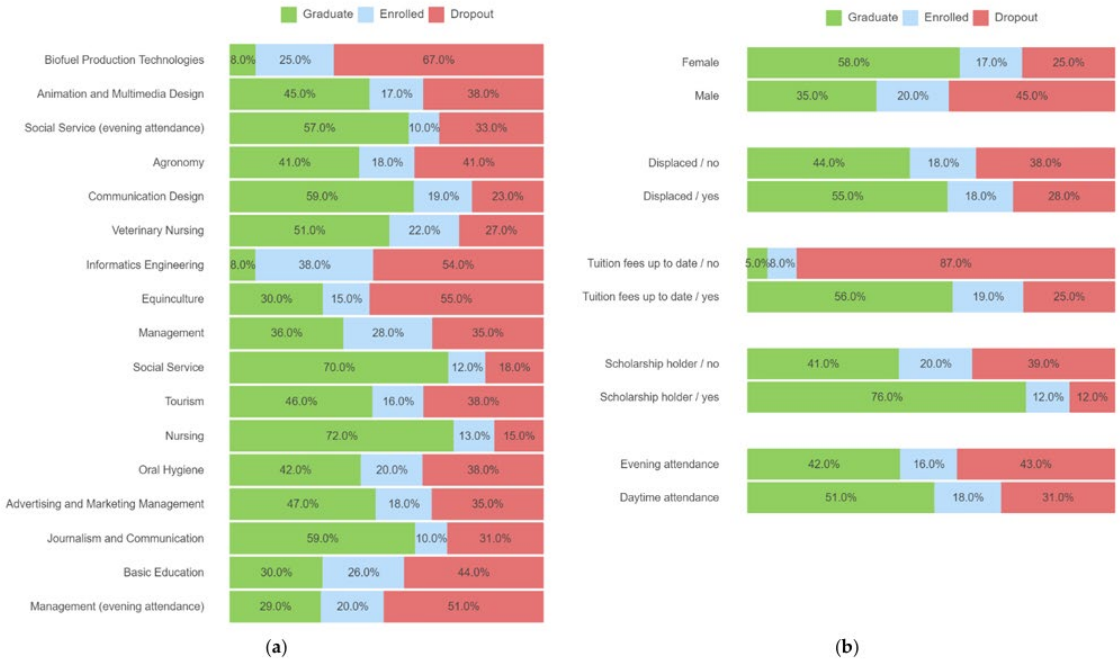
5.Ekran Görüntüleri:






1. Dönem vs 2. Dönem Notları
2. Dönem Not Ortalaması
1. Dönem Not Ortalaması
Target
0
1
2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


2209
1421
794
Frequency
0
500
1000
1500
2000
Graduate
Dropout
Enrolled
Target



6.Sonuç ve Öneriler:

Öğrencilerin akademik başarısı, kayıt anında bilinen verilerle büyük oranda öngörülebilir.

ANN gibi derin öğrenme modelleri yüksek performans sağlamaktadır.

Gelecekte, öğrencilerin psikolojik durumları veya dönem içi performansları da modele dahil edilerek tahmin gücü artırılabilir.

7.Kaynakça

1.https://chatgpt.com/

2.https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

3.Behr, Motives for Dropping out from Higher Education—An Analysis of Bachelor's Degree Students in Germany, Eur. J. Educ., № 56, c. 325 DOI: 10.1111/ejed.12433

4.Kehm, Student Dropout from Universities in Europe: A Review of Empirical Literature, Hungarian Educ. Res. J., № 9, c. 147 DOI: 10.1556/063.9.2019.1.18

5.Atchley, Comparison of Course Completion and Student Performance through Online and Traditional Courses, Int. Rev. Res. Open Distance Learn., № 14, c. 104 DOI: 10.19173/irrodl.v14i4.1461

6.Quinn, J. (2013). Dropout and Completion in Higher Education in Europe among Students from Under-Represented Groups.

7.Namoun, A., and Alshanqiti, A. (2020). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Appl. Sci., 11. DOI: 10.3390/app11010237

8.Saa, Mining Student Information System Records to Predict Students' Academic Performance, Adv. Intell. Syst. Comput., № 921, c. 229 DOI: 10.1007/978-3-030-14118-9_23

9.Altun, Using Learning Analytics to Develop Early-Warning System for at-Risk Students, Int. J. Educ. Technol. High. Educ., № 16, c. 40 DOI: 10.1186/s41239-019-0172-z

9.Daud, A., Lytras, M.D., Aljohani, N.R., Abbas, F., Abbasi, R.A., and Alowibdi, J.S. (2017, January 3–7). Predicting Student Performance Using Advanced Learning Analytics. Proceedings of the 26th International World Wide Web Conference 2017, WWW 2017 Companion, Perth, Australia. DOI: 10.1145/3041021.3054164

10.Martins, Early Prediction of Student's Performance in Higher Education: A Case Study, Adv. Intell. Syst. Comput., № 1365, c. 166 DOI: 10.1007/978-3-030-72657-7_16

11.Chawla, SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res., № 16, c. 321 DOI: 10.1613/jair.953

12.He, H., Bai, Y., Garcia, E.A., and Li, S. (2008, January 1–8). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the International Joint Conference on Neural Networks, Hong Kong, China.

13.Chen, Using Random Forest to Learn Imbalanced Data, Univ. Calif. Berkeley, № 110, c. 1

14.Liu, Exploratory Undersampling for Class-Imbalance Learning, IEEE Trans. Syst. Man Cybern. Part B Cybern., № 39, c. 539 DOI: 10.1109/TSMCB.2008.2007853

15.Maclin, R., and Opitz, D. An Empirical Evaluation of Bagging and Boosting. Proceedings of the National Conference on Artificial Intelligence, Providence, RI, USA.

16.Hido, Roughly Balanced Bagging for Imbalanced Data, Stat. Anal. Data Min., № 2, c. 412 DOI: 10.1002/sam.10061

17.Wang, S., and Yao, X. (April, January 30). Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA. DOI: 10.1109/CIDM.2009.4938667

18.Saarela, Comparison of Feature Importance Measures as Explanations for Classification Models, SN Appl. Sci., № 3, c. 272 DOI: 10.1007/s42452-021-04148-9

19.Spelmen, V.S., and Porkodi, R. (2018, January 1–3). A Review on Handling Imbalanced Data. Proceedings of the 2018 International Conference on Current Trends

towards Converging Technologies (ICCTCT), Coimbatore, India. DOI: 10.1109/ICCTCT.2018.8551020

20.Ali, Imbalance Class Problems in Data Mining: A Review, Indones. J. Electr. Eng. Comput. Sci., № 14, c. 1552

21.Ho, T.K. (1995, January 14–16). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada.

22.Chen, T., and Guestrin, C. (2016, January 13–17). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference, San Francisco, CA, USA. DOI: 10.1145/2939672.2939785

23.Ke, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Adv. Neural Inf. Process. Syst., № 30, c. 3147

24.Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2017). CatBoost: Unbiased Boosting with Categorical Features. arXiv.

25.Wilkinson, The FAIR Guiding Principles for Scientific Data Management and Stewardship, Sci. Data, № 3, c. 160018 DOI: 10.1038/sdata.2016.18