Öğrenci Performanslarının Makine Öğrenmesi Teknikleri ile Sınıflandırılması

1. Ebrar Öztürk
Bilgisayar Mühendisliği
TOBB ETU
ebrar.ozturk@etu.edu.tr

Özet: Bu rapor, öğrenci başarı verilerini analiz etmek ve çeşitli öznitelikler ile öğrenci performansı arasındaki ilişkileri ortaya koymak amacıyla gerçekleştirilen bir projeyi sunmaktadır. Proje, student_success.csv veri setini kullanarak, öğrenci başarı düzeylerini sınıflandırmak için makine öğrenimi modelleri geliştirmeyi hedeflemiştir. Özellikle, veri işleme aşamalarının yanı sıra, veri dengesizliği ile başa çıkmak için örnekleme tekniklerinin başarı üzerinde büyük bir etkiye sahip olduğu tespit edilmiştir.

Anahtar Kelimeler — class imbalance, classification, ensamle methods, feature selection, outlier detection, histogram, boxplot, correlation, dimension reduction, scaling, sampling, F-score, ROC/AUC curve

I. GİRİŞ (*PROBLEM TANIMI)*

Öğrenci başarısını etkileyen birçok faktör bulunmaktadır. Bu faktörlerin analiz edilmesi, eğitim politikalarının geliştirilmesi ve öğrenci destek hizmetlerinin iyileştirilmesi açısından kritik öneme sahiptir. Bu projede, öğrenci verilerini kullanarak, başarı düzeylerinin belirlenmesi ve bu başarıyı etkileyen özniteliklerin incelenmesi daha sonra da eğitilen modeller arasında öğrencilerin başarılı olup olmayacağını tahmin edebilen en iyi modeli bulmak amaçlanmıştır.

Motivasyon: Eğitim sisteminde öğrencilerin başarısını artırmak için veri analizi ve makine öğrenimi yöntemlerinin entegrasyonu, akademik performansı artırmanın önemli bir yoludur.

Genel Metodoloji: Proje kapsamında, veri ön işleme, örnekleme, model geliştirme ve farklı modelleri ensemble metodları ile birleştirip tüm sonuçları değerlendirme aşamaları gerçekleştirilmiştir. Ayrıca cross validation ile tüm modellerin en iyi parametrelerinin seçilmesi sağlanmıştır. Örnekleme, özellikle dengesiz veri dağılımları ile başa çıkmada önemli bir rol oynamaktadır. Veriyi test, train setlerine ayırmada stratified sampling kullanılmıştır. Test ve veri setlerinde her sınıftan yaklaşık eşit miktarda bulunması eğitimin doğru gerçekleşmesi için önemlidir.

Amaç/Hedef: Bu projenin amacı, öğrenci başarı düzeylerini sınıflandırmak ve bu sınıflandırmayı etkileyen faktörleri belirlemektir. Daha sonra bu sınıflandırmaya dayanarak belli durumlar altında öğrencilerin ne kadar başarılı olacağını tahmin etmektir. Farklı makine öğrenimi modellerinin

performansını karşılaştırarak bu amaçlara ulaşmak hedeflenmiştir.

Başarım Metrikleri: Projede, model başarı oranı, doğruluk, F-skoru, ROC AUC gibi metrikler kullanılarak modellerin performansı değerlendirilmiştir.

II. LİTERATUR ARAŞTIRMASI

Eğitim alanında veri analizi ve makine öğrenimi kullanımı, son yıllarda artan bir ilgi görmektedir. Önceki çalışmalar, öğrenci başarı düzeylerini tahmin etmek için çeşitli makine öğrenimi algoritmalarının etkinliğini göstermektedir. Literatür araştırması yaparken kaggle'da kullanılan öğrenci başarısı veri setiyle daha önce yapılmış projeler incelendi. Ayrıca, veri dengesizliği ile başa çıkmak için kullanılan örnekleme tekniklerinin kullanımını ve ensamble metodları ile modellerin performans iyileştirmelerinin nasıl yapıldığı araştırıldı.

III. VERİ SETİ, VERİ ÖZELLİKLERİ, ÖZNİTELİKLER

Veri Kaynağı: Projede kullanılan veri seti student_success.csv dosyasıdır. Veri seti, öğrenci başarılarını ve başarıyı etkileyen çeşitli faktörleri içermektedir.

Veri Kümesi: Veri seti, 650 satırdan oluşmaktadır ve her satır, bir öğrenciyi temsil etmektedir. Öğrencilerin başarı notları (G1, G2, G3) ve demografik özellikleri gibi öznitelikler içermektedir. 17 kategorik, 16 numerik olmak üzere 33 öznitelikten oluşmaktadır.

Önişleme Aşamaları: Veri setinde eksik değerler kontrol edilmiş ve gerekirse doldurulmuştur. Outlier tespiti IQR metodu kullanılarak yapılmış ve outlierlar veri setinden silinmiştir. Başarının belirlenmesi için 3 farklı not yerine bu notların ortlamaları alınarak "ortalama notlar" adında yeni bir sutun eklenmiştir Ayrıca, dengesiz veri dağılımı ile başa çıkmak için örnekleme yöntemleri kullanılmıştır. Örnekleme işlemleri, dengesiz dağılımın başarıyı nasıl etkilediğini gösteren önemli bulgular sunmuştur.

Öznitelik Açıklamaları:

- **Hatalı Veriler:** Veri setindeki Outlier değerler IQR ile belirlenmiş ve uygun yöntemlerle temizlenmiştir. Sette eksik veriye rastlanmamıştır.
- **Veri Türleri:** Veri setindeki öznitelikler; binary, sayısal ordinal, nominal türleri içermektedir. Bu verilerin hepsi structured olarak verilmiştir.
- Binary veriler: school ('GP'veya 'MS'), sex ('F' - 'M'), address ('U' veya 'R'), famsize ('LE3' veya 'GT3'), Pstatus ('T' veya 'A'), schoolsup (evet veya hayır), famsup (evet veya hayır), paid (evet veya hayır), activities (evet veya hayır), nursery (evet veya hayır), higher (evet veya hayır), internet (evet veya hayır), romantic (evet veya hayır)
- Sayısal Ordinal veriler: age (15- 22), Medu (0- 4), Fedu (0- 4), traveltime (1- 4), studytime (1- 4), failures (1- 4), famrel (1- 5), freetime (1- 5), goout (1- 5), Dalc (1- 5), Walc (1- 5), health (1- 5), absences (0- 93), G1 (0- 20), G2 (0- 20), G3 (0- 20)
- Nominal Veriler: Mjob ('teacher', 'health', 'civil services', 'at_home', 'other'), Fjob ('teacher', 'health', 'civil services', 'at_home', 'other'), reason ('close to home', 'school reputation',

'course preference', 'other'), guardian ('mother', 'father', 'other')

- **Öznitelik Seçimi:** PCA (Principal Component Analysis) yöntemi kullanılarak veri setindeki boyut azaltma işlemleri gerçekleştirilmiştir. Başarı metriği olarak 3 notun ortalaması alınarak yeni bir öznitelik oluşturulmuştur.
- **Sınıf Veri Dağılımları:** Veri setindeki sınıf dağılımları dengesizdir ve bu durum, sınıflardaki veri sayılarını dengelemek için örnekleme işlemleriyle ele alınmıştır.
- **Veri Normalizasyonu:** StandardScale kullanarak, verilerin ortalamasını 0 ve standart sapmasını 1 haline getirildi. Not ortalamaları 1-5 arasında numerik değerlere maplendi. Kategorik veriler eğitimin kolaylığı için "get_dummies" metodu ile numerik verilere donüştürüldü. Bunlar da farklı ölçeklerdeki sayısal verilerin birbiriyle karşılaştırılabilir olmasını sağladı.
- **Öznitelikler arasındaki ilişkiler:** Farklı öznitelikler arasında nasıl bir korelasyon olduğu incelendi. Bu kısımda sınıflara eşit dağılım olmadığı için veriler arasında sıkı bir korelasyon bulunamadı. Ancak sampling metodu ile korelasyon matrisinde bazı ilişkiler gözlenebildi. Numerik verilerin birbiriyle olan korelasyonu matris şeklinde, kategorik verilerle not ortalaması arasındaki ilişki boxplot şeklinde appendix-1 ve appendix-2 de görselleştirilmiştir. Not ortalamaları da histogram ile appendix-3'teki gibi görselleştirilmiş ve normal dağılıma benzer bir dağılım elde edilmiştir.

IV. KULLANILAN METODOLOJİ

Proje kapsamında, veri analizi için çeşitli makine öğrenimi modelleri kullanılmıştır. Bu modeller arasında Logistic Regression, Random Forest ve Gradient Boosting yer almaktadır. Daha sonra bu modeller ensamble metodları ile performansı artırmak üzere kombine edilmiştir. Bu modellerin seçilmesinin nedeni, eğitim verileri üzerinde gösterdikleri yüksek başarı ve uygulama kolaylığıdır.

Logistic Regression: İkili sınıflandırma problemlerinde yaygın olarak kullanılan bir istatistiksel modeldir. Model, bağımsız değişkenlerle bağımlı değişken arasındaki ilişkiyi bir olasılık fonksiyonu aracılığıyla modelleyerek, belirli bir sınıfa ait olma olasılığını tahmin eder. Lojistik regresyon, açıklayıcı değişkenlerin doğrusal bir kombinasyonunu alır ve bu kombinasyonu sigmoid fonksiyonu aracılığıyla 0 ile 1 arasında bir değer üretir.

Random Forest: Birden fazla karar ağacının bir araya gelmesiyle oluşturulan bir topluluk (ensemble) yöntemidir. Her bir ağaç, eğitim veri kümesinin rastgele bir alt kümesi üzerinde eğitilir ve sonuçlar çoğunluk oyu ile belirlenir. Bu yaklaşım, aşırı öğrenmeyi (overfitting) azaltarak daha genel bir model oluşturur.

Gradient Boosting: Zayıf öğrenicileri (genellikle karar ağaçları) ardışık olarak birleştirerek güçlü bir tahmin modeli oluşturan bir yöntemdir. Her yeni model, önceki modellerin hatalarını düzeltmeye çalışarak öğrenir. Bu yaklaşım, yüksek doğruluk sağlama potansiyeline sahiptir ancak hiperparametre ayarlarına duyarlıdır.

Blending ve Majority Voting: Farklı modellerin tahminlerini birleştirerek, her bir modelin güçlü yönlerinden yararlanmayı

amaçlar. Blending, birkaç modelin sonuçlarının bir başka model (genellikle basit bir lojistik regresyon) ile birleştirilmesi ile gerçekleştirilirken, çoğunluk oyu, en çok oy alan sınıfın seçilmesi ile yapılır.

Neden Bu Modeller Seçildi?

1-)Çeşitlilik ve Performans: Farklı makine öğrenimi yöntemleri, veri kümesinin özelliklerine bağlı olarak farklı performans sergileyebilir. Lojistik regresyon, basit ve hızlı bir model sunarken, rastgele orman ve gradyan artırma, daha karmaşık ilişkileri öğrenme yeteneğine sahiptir. Bu çeşitlilik, en iyi sonucu elde etmek için önemli bir stratejidir.**,**

2-) Overfitting Kontrolü: Random Forest ve Gradient Boosting, ağaç tabanlı yöntemlerdir ve genellikle aşırı öğrenmeyi kontrol etmede daha etkilidir. Bu, özellikle sınıf dengesizliği gibi durumlarda modelin genelleme yeteneğini artırır.

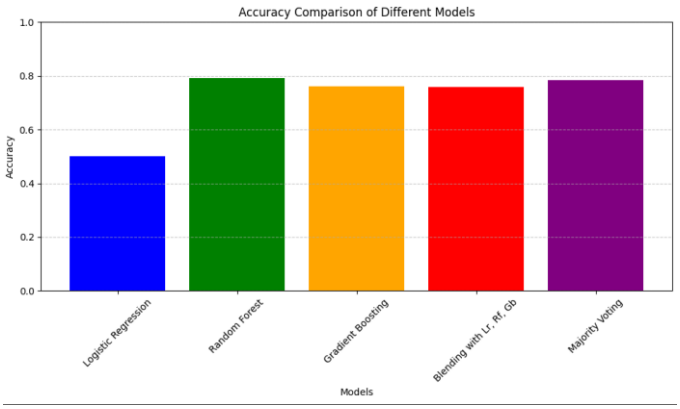
3-) Hedef Değişkenin Doğası: Öğrenci başarısı gibi çok sınıflı bir hedef değişken, bu tür topluluk yöntemleri ile daha iyi sınıflandırılabilir. Aksi takdirde, daha basit yöntemler karmaşık sınıf yapılarını yakalamakta zorlanabilir.

Bu metodoloji, öğrenci başarısını etkileyen çeşitli faktörlerin etkisini daha iyi anlamak ve bu faktörlerin başarı üzerinde nasıl bir rol oynadığını incelemek için tasarlanmıştır. Her modelin çıktıları, eğitim politikalarının ve stratejilerinin geliştirilmesine katkıda bulunmak için değerlendirilecektir.

V. TEST SONUÇLARI

Elde edilen sonuçlar aşağıdaki gibidir:

- **Logistic Regression:** %50 accuracy
- **Random Forest**: %79 accuracy
- **Gradient Boosting**: %76 accuracy
- **Blending**: %76 accuracy
- **Majority Voting**: %78 accuracy



Logistic regression, sınıflandırma doğruluğunda belirgin bir şekilde düşük performans göstermiştir. Diğer üç model ise daha yüksek doğruluk oranları ile dikkat çekmektedir. Random Forest, en yüksek doğruluğu sağlarken, Gradient Boosting ve blending benzer sonuçlar elde etmiştir. Majority Voting yöntemi ise blending'e göre biraz daha iyi bir performans göstermektedir.

Random forest modelinin daha iyi performans göstermesinin birkaç nedeni bulunmaktadır:

1. **Ağaç Tabanlı Yapı**: Rastgele orman, birden fazla karar ağacının bir araya gelmesiyle oluştuğu için karmaşık verilerde daha iyi genelleme yapabilir. Bu, aşırı öğrenmeyi önlemeye yardımcı olur.
2. **Özelliklerin Rastgele Seçilmesi**: Model, her ağaç için rastgele özellikler kullanarak eğitildiği için, farklı özelliklerin etkilerini öğrenme kabiliyeti artar.
3. **Daha İyi Sınıf Dengelemesi**: Özellikle dengesiz veri kümelerinde, rastgele ormanın çoğunluk sınıflarını daha iyi tahmin etme yeteneği vardır. Logistic regression gibi daha basit

modeller, bu tür durumlarda performans kaybı yaşayabilir.

Sonuç olarak, model seçiminde genel performans göz önünde bulundurulduğunda, random forest, bu çalışmada en iyi seçenek olarak öne çıkmaktadır. Bu modelin diğerlerinden daha iyi olma sebepleri, parametre aralığının daha iyi seçilmesi, minority sınıflar için sampling yapılması gibi işlemler de olabilir. Modellerin detaylı performansları (ROC-AUC curves, f1 scores) Appendix-4'te gösterilmiştir.

VI. SONUÇ

Bu çalışmada öğrencilerin başarılının tahmin edilebilmesinde etkili olan temel faktörler belirlenip farklı modeller eğiterek ve bu modellerin kombinasyonları ile farklı sonuçlar elde ettik. Veri hazırlığı için dengesiz olan sınıflar dengelenmeye çalışıldı, boyut kısıtlamasına ve yeni öznitelikler çıkarılamsına başvuruldu. Her modelin değerlendirilmesi için farklı metodlar kullanıldı.

Çalışma sonucunda, farklı makine öğrenimi modelleri ile farklı özniteliklerin öğrenci başarıları üzerindeki etkilerini karşılaştırdık ve Random Forest modelinin diğerlerine göre en yüksek doğruluğu sağladığını gördük. Bu, model seçiminde ve sınıflandırma görevlerinde hangi yöntemlerin daha etkili olduğunu, veri analizine göre sampling, boyut azaltma gibi yöntemleri nasıl yapacağımızı anlamamıza yardımcı oldu. Ayrıca, istatistiksel testlerin önemini öğrenerek, elde ettiğimiz sonuçların güvenilirliğini değerlendirmenin yollarını keşfettik. Bu çalışma, eğitim alanında veri analizi ve makine öğrenimi uygulamalarına katkı sağladı.

Çalışmada, bazı veri ön işleme adımlarını (örneğin, daha fazla veri temizliği veya daha karmaşık feature engineering teknikleri) uygulama fırsatımız olmadı. Bu, zaman kısıtlamaları ve veri setinin belirli özellikleriyle sınırlı kalmamızdan kaynaklandı. Ayrıca projenin başlangıçtaki amacı yalnızca korelasyonları incelemekti, fakat veri dağılımlarının yeterince düzgün olmaması sebebiyle güçlü korelasyonlar gözlemlenemedi ve bu yüzden projeye sınıflandırma adı altında devam edildi.

Gelecek çalışmalar, daha büyük, çeşitli ve dengeli dağılmış veri setleri üzerinde çalışarak model performansını artırmayı amaçlayabilir. Farklı makine öğrenimi tekniklerini (örneğin, derin öğrenme yöntemleri) denemek ve elde edilen sonuçları daha kapsamlı bir analiz ile karşılaştırmak da bir hedef olabilir. Bunun yanı sıra, elde edilen sonuçların eğitim politikalarına ve stratejilerine nasıl entegre edileceği üzerine çalışmak, uygulama alanında önemli bir katkı sağlayabilir.

KAYNAKLAR

[1] https://www.kaggle.com/datasets/larsen0966/student-performance-data-set/data

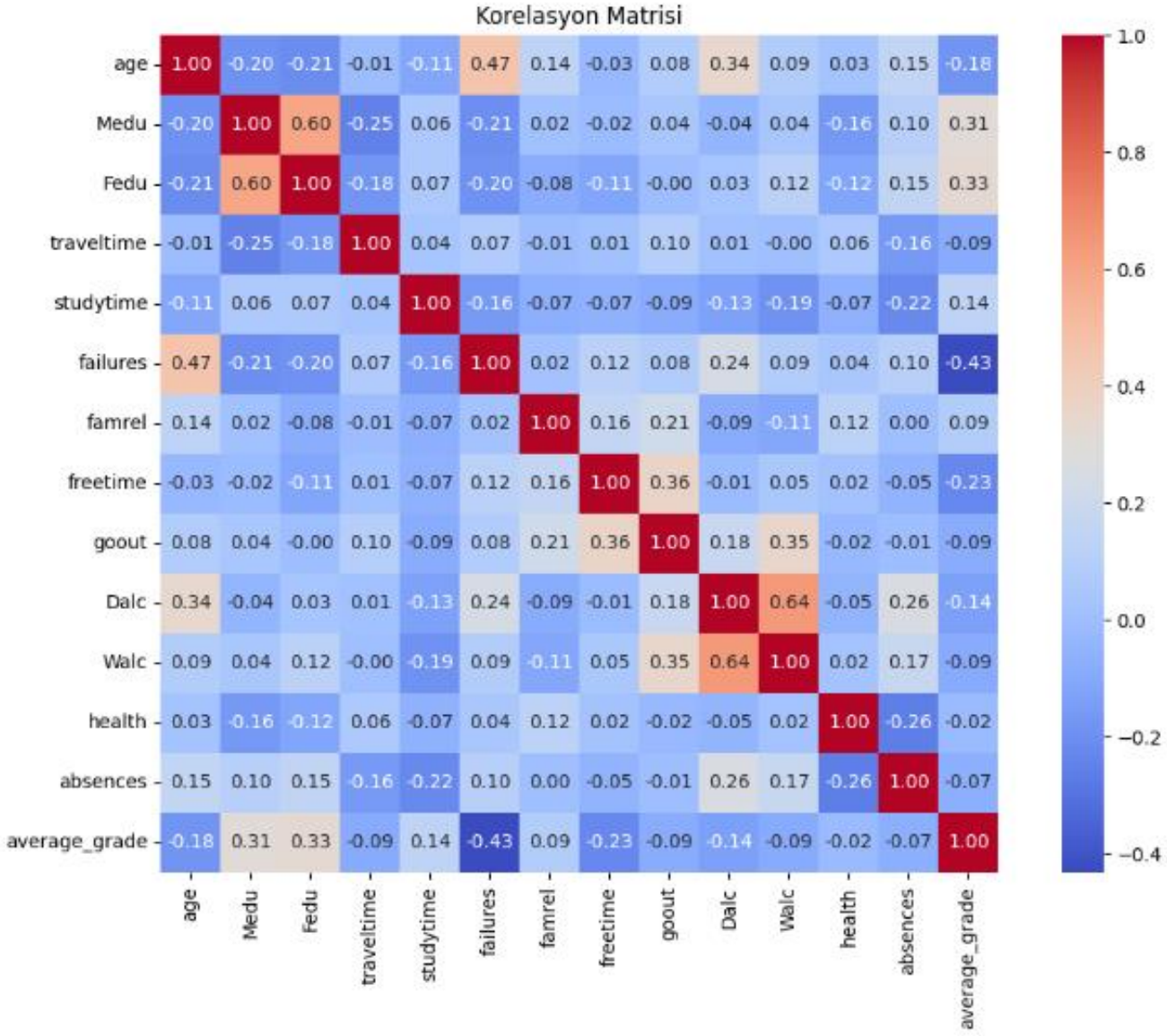
[2] https://www.qualtrics.com/experience-management/research/sampling-methods/

[3] *https://www.javatpoint.com/majority-voting-algorithm-in-machine-learning#:~:text=In%20Majority%20Voting%2C%20a%20group,determined%20by%20the%20majority's%20decision*.

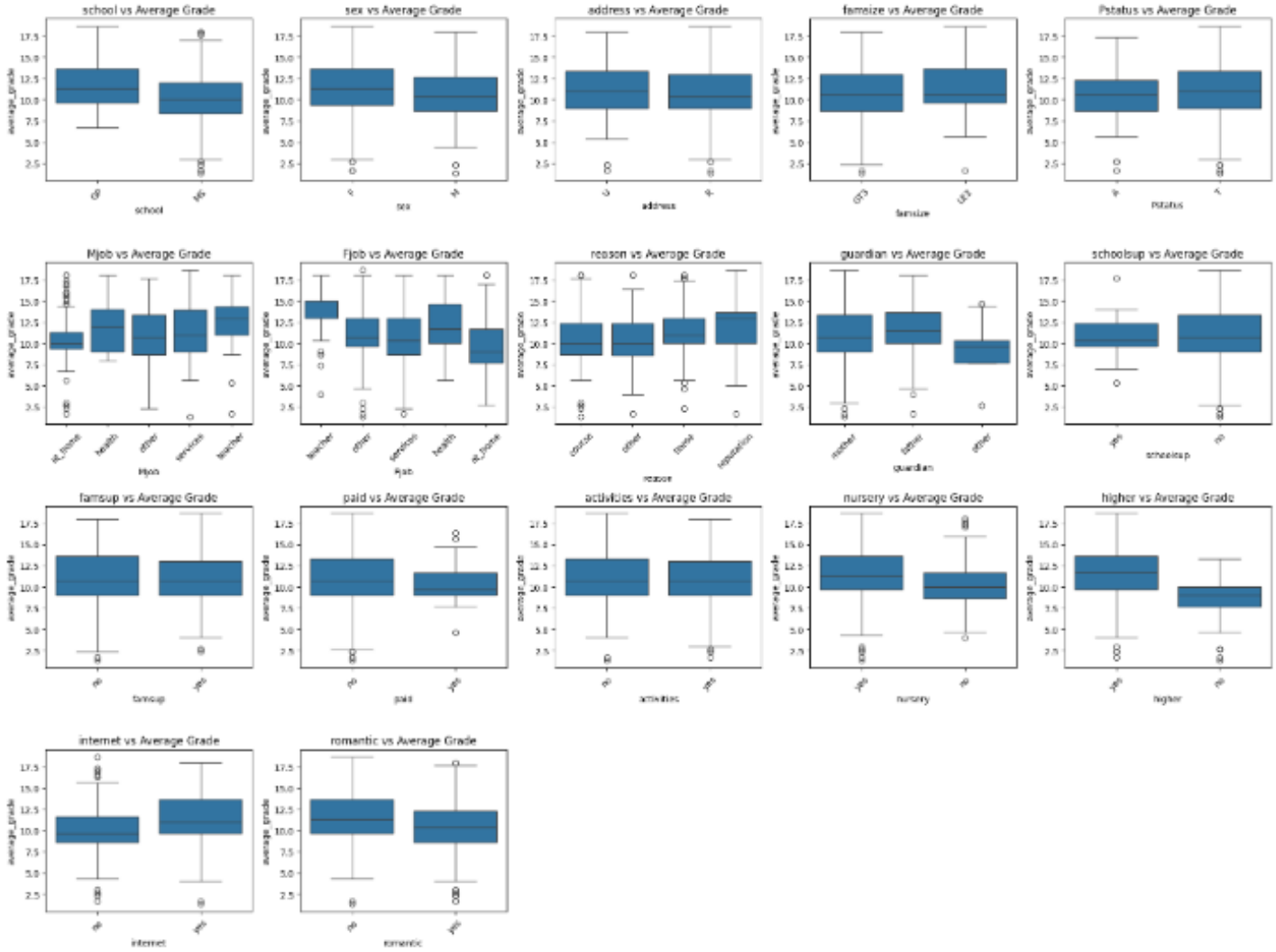
[4] https://chatgpt.com/

[5] https://www.geeksforgeeks.org/metrics-for-machine-learning-model/

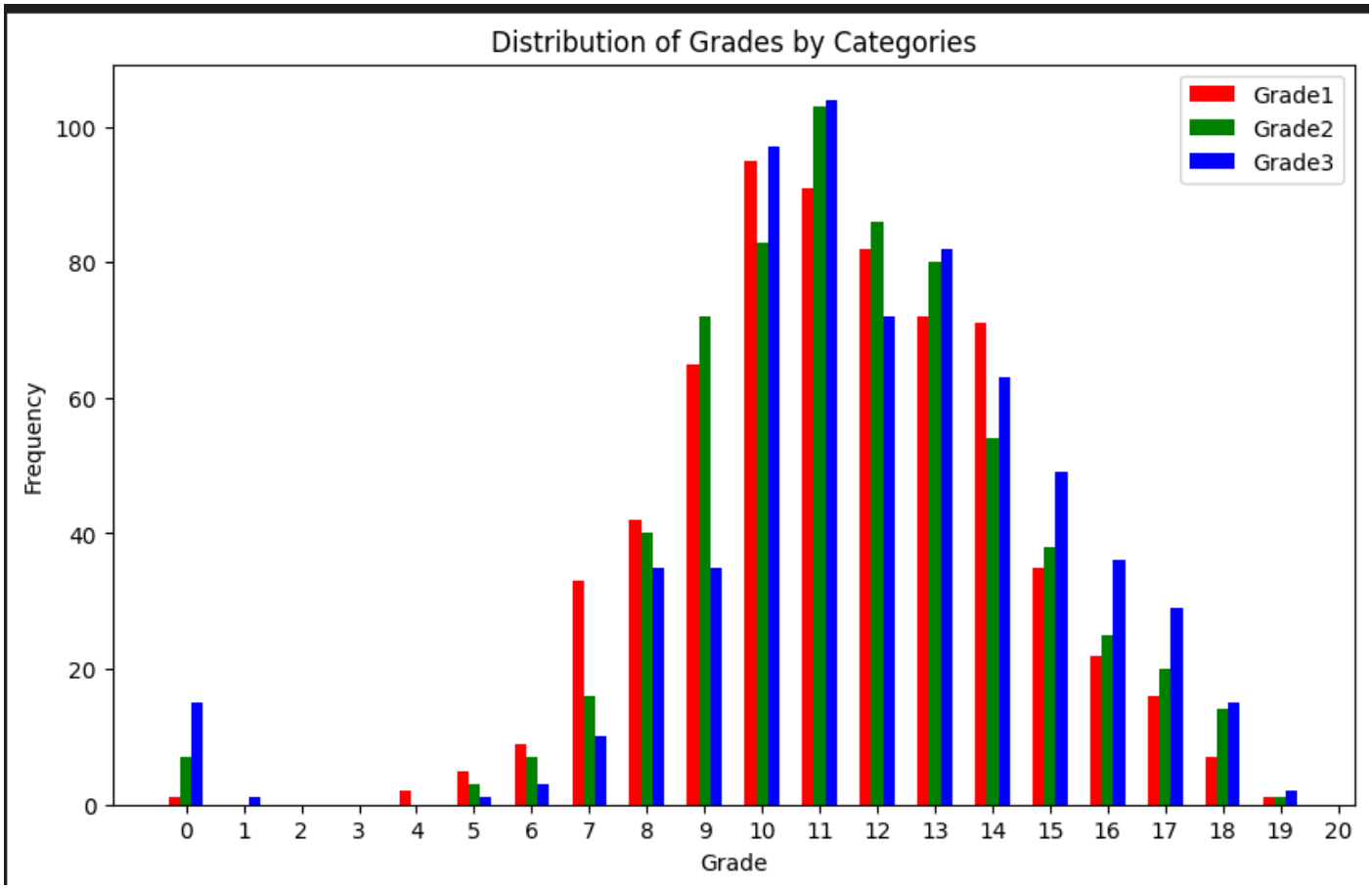
APPENDIX-1


Korelasyon Matrisi
age 1.00 -0.20 -0.21 -0.01 -0.11 0.47 0.14 -0.03 0.08 0.34 0.09 0.03 0.15 -0.18
Medu -0.20 1.00 0.60 -0.25 0.06 -0.21 0.02 -0.02 0.04 -0.04 0.04 -0.16 0.10 0.31
Fedu -0.21 0.60 1.00 -0.18 0.07 -0.20 -0.08 -0.11 -0.00 0.03 0.12 -0.12 0.15 0.33
traveltime -0.01 -0.25 -0.18 1.00 0.04 0.07 -0.01 0.01 0.10 0.01 -0.00 0.06 -0.16 -0.09
studytime -0.11 0.06 0.07 0.04 1.00 -0.16 -0.07 -0.07 -0.09 -0.13 -0.19 -0.07 -0.22 0.14
failures 0.47 -0.21 -0.20 0.07 -0.16 1.00 0.02 0.12 0.08 0.24 0.09 0.04 0.10 -0.43
famrel 0.14 0.02 -0.08 -0.01 -0.07 0.02 1.00 0.16 0.21 -0.09 -0.11 0.12 0.00 0.09
freetime -0.03 -0.02 -0.11 0.01 -0.07 0.12 0.16 1.00 0.36 -0.01 0.05 0.02 -0.05 -0.23
goout 0.08 0.04 -0.00 0.10 -0.09 0.08 0.21 0.36 1.00 0.18 0.35 -0.02 -0.01 -0.09
Dalc 0.34 -0.04 0.03 0.01 -0.13 0.24 -0.09 -0.01 0.18 1.00 0.64 -0.05 0.26 -0.14
Walc 0.09 0.04 0.12 -0.00 -0.19 0.09 -0.11 0.05 0.35 0.64 1.00 0.02 0.17 -0.09
health 0.03 -0.16 -0.12 0.06 -0.07 0.04 0.12 0.02 -0.02 -0.05 0.02 1.00 -0.26 -0.02
absences 0.15 0.10 0.15 -0.16 -0.22 0.10 0.00 -0.05 -0.01 0.26 0.17 -0.26 1.00 -0.07
average_grade -0.18 0.31 0.33 -0.09 0.14 -0.43 0.09 -0.23 -0.09 -0.14 -0.09 -0.02 -0.07 1.00
age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc health absences average_grade
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

APPENDIX-2

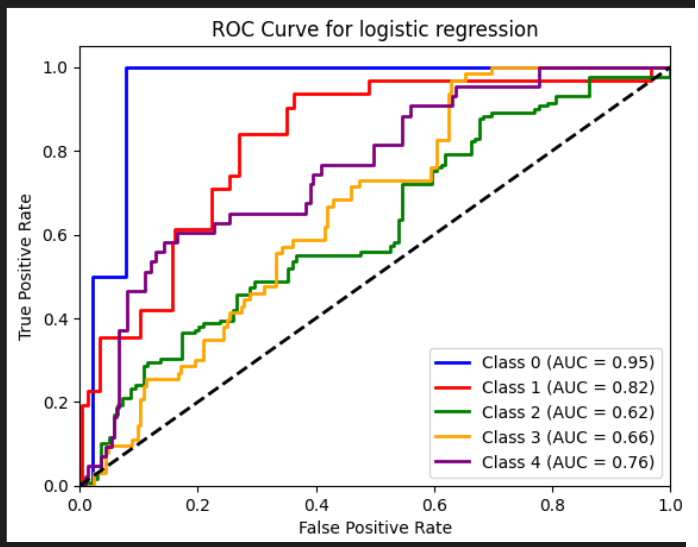

school vs Average Grade
sex vs Average Grade
address vs Average Grade
famsize vs Average Grade
Pstatus vs Average Grade
Mjob vs Average Grade
Fjob vs Average Grade
reason vs Average Grade
guardian vs Average Grade
schoolsup vs Average Grade
famsup vs Average Grade
paid vs Average Grade
activities vs Average Grade
nursery vs Average Grade
higher vs Average Grade
internet vs Average Grade
romantic vs Average Grade
average_grade

APPENDIX-3

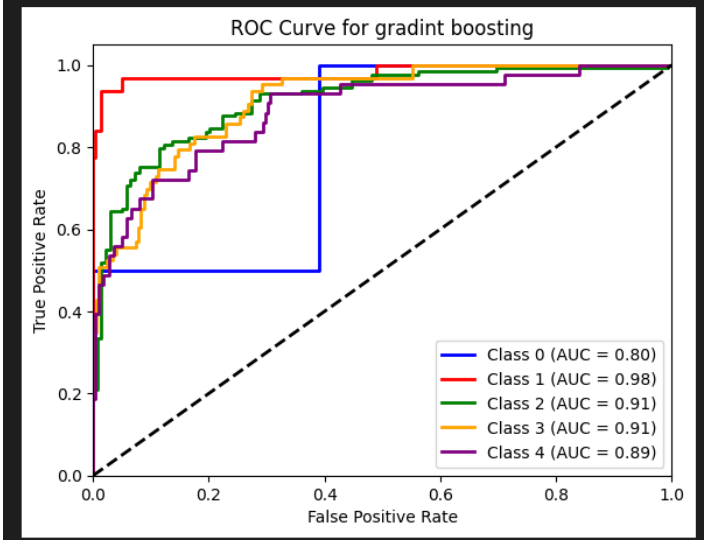

Distribution of Grades by Categories
Grade1
Grade2
Grade3
Frequency
Grade
0
20
40
60
80
100
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

APPENDIX-4

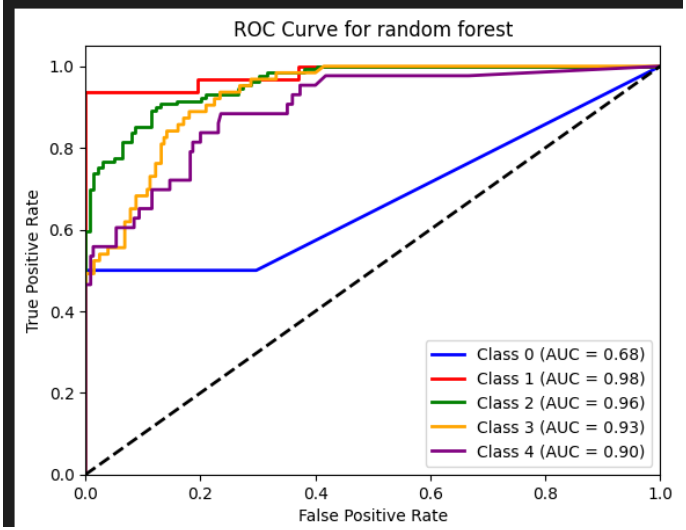
```
logistic regression Evaluation:
Accuracy: 0.50
Confusion Matrix:
[[  0   0   2   0   0]
 [  0   0  30   0   1]
 [  0   1 115   6   7]
 [  0   0  52   3   8]
 [  0   0  23   4  16]]
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00        31
           3       0.52      0.89      0.66       129
           4       0.23      0.05      0.08        63
           5       0.50      0.37      0.43        43

    accuracy                           0.50       268
   macro avg       0.25      0.26      0.23       268
weighted avg       0.38      0.50      0.40       268
```



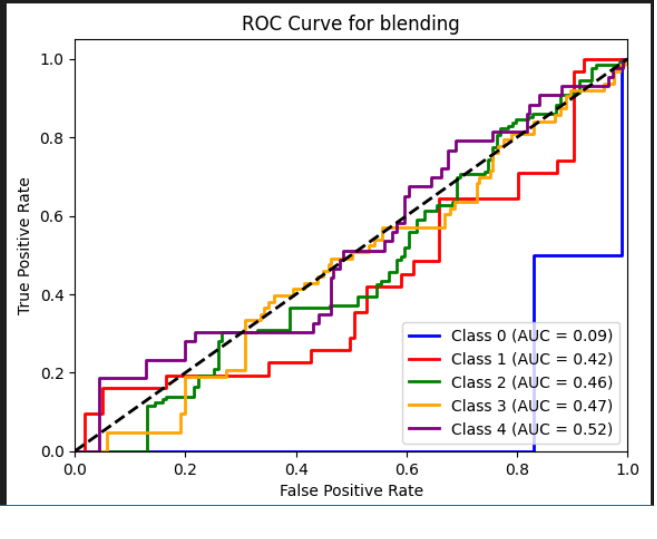

```
gradint boosting Evaluation:
Accuracy: 0.76
Confusion Matrix:
[[  1   0   1   0   0]
 [  0  26   4   0   1]
 [  1   1 117   6   4]
 [  0   0  21  35   7]
 [  0   2  13   3  25]]
Classification Report:
              precision    recall  f1-score   support

           1       0.50      0.50      0.50         2
           2       0.90      0.84      0.87        31
           3       0.75      0.91      0.82       129
           4       0.80      0.56      0.65        63
           5       0.68      0.58      0.63        43

    accuracy                           0.76       268
   macro avg       0.72      0.68      0.69       268
weighted avg       0.76      0.76      0.75       268
```




```
random forest Evaluation:
Accuracy: 0.79
Confusion Matrix:
[[  1   0   1   0   0]
 [  0  29   1   0   1]
 [  0   0 118   6   5]
 [  0   1  14  38  10]
 [  0   2   7   8  26]]
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.50      0.67         2
           2       0.91      0.94      0.92        31
           3       0.84      0.91      0.87       129
           4       0.73      0.60      0.66        63
           5       0.62      0.60      0.61        43

    accuracy                           0.79       268
   macro avg       0.82      0.71      0.75       268
weighted avg       0.79      0.79      0.79       268
```




```
blending Evaluation:
Accuracy: 0.76
Confusion Matrix:
[[  0   1   1   0   0]
 [  0  29   1   0   1]
 [  0   0 118   7   4]
 [  0   1  19  37   6]
 [  0   1  13  10  19]]
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         2
           2       0.91      0.94      0.92        31
           3       0.78      0.91      0.84       129
           4       0.69      0.59      0.63        63
           5       0.63      0.44      0.52        43

    accuracy                           0.76       268
   macro avg       0.60      0.58      0.58       268
weighted avg       0.74      0.76      0.74       268
```



```
majority voting Evaluation:
Accuracy: 0.78
Confusion Matrix:
[[  1   0   1   0   0]
 [  0  29   1   0   1]
 [  0   0 119   6   4]
 [  0   0  20  35   8]
 [  0   1  12   4  26]]
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.50      0.67         2
           2       0.97      0.94      0.95        31
           3       0.78      0.92      0.84       129
           4       0.78      0.56      0.65        63
           5       0.67      0.60      0.63        43

    accuracy                           0.78       268
   macro avg       0.84      0.70      0.75       268
weighted avg       0.78      0.78      0.78       268
```