Proje Başlığı:

LMS (Learning Management System) Aktivite Verileri ile Öğrenci Başarısının Tahmini

Grup İsmi:
GradeSense

Grup Üyeleri

Ad-Soyad	Öğrenci No	E-Mail
Fatma Abazlı	22040101142	fatmaabazli@stu.topkapi.edu.tr
Hazal Al Mohammad Algharbi	22040101140	hazalalmohammadalgharbi@stu.topkapi.edu.tr
İmaduddin Hamo	22040101010	imaduddinhamo@stu.topkapi.edu.tr

Ders / Dönem:

FET445 – Veri Madenciliği / 2024–2025

2) Problem Tanımı

Bu proje, bir öğrencinin LMS platformundaki davranışsal aktivitelerine bakarak dersi geçip geçmeyeceğini (Pass/Fail) tahmin etmeyi amaçlar.

Veri seti;

- **sayfa görüntüleme oranları,**
- **ödev görüntülemeleri,**
- **quiz girişleri,**
- **kaynak erişimleri,**
- **zaman bazlı aktivite ölçümleri**

gibi yüksek boyutlu etkileşim verilerinden oluşmaktadır.

Görev Türü: Sınıflandırma (Classification)
Hedef Değişken: TARGET → PASS_FAIL (0 = Fail, 1 = Pass)

Başarı Kriterleri:

- **Accuracy ≥ 0.80**
- **F1-score ≥ 0.75**
- **ROC–AUC ≥ 0.80**

3) Proje Yönetimi

Zaman Çizelgesi

Hafta	Görev
1	Veri seti seçimi ve proje tanımı
2	Veri temizleme, eksik değer analizi, EDA
3	Temel (Base) modellerin kurulumu
4-5	Feature selection + dimension reduction + base model karşılaştırmaları
6	Performans analizi + hata analizi
7	Final raporu ve sunum

Roller ve Sorumluluklar

Üye	Farklılık Noktası
Fatma	Veri temizleme + Logistic Regression & Decision Tree + Variance Threshold FS
Hazal	Scaling + KNN & Naive Bayes + ANOVA (f_classif) FS
İmad	Chi-Square FS + PCA DR + SVM & LDA

Böylece her üyenin kullandığı modeller, FS yöntemleri ve dönüşümler farklıdır.

Çıktılar

- Final Proje Raporu (PDF)
- Jupyter Notebook dosyaları (her üye için ayrı)
- Temizlenmiş veri seti: *processed.csv*
- Proje sunum slaytları
- GitHub repository (TBD)

4) İlgili Çalışmalar (Mini Literatür)

Literatür Özeti:

1. LMS verileri ile öğrenci başarısı tahmini yapan çalışmalarda Logistic Regression ve Decision Tree gibi **base modellerin** yüksek doğruluk verdiği görülmüştür.
2. Quiz etkileşimi, ödev teslimi ve sayfa görüntüleme davranışları en etkili faktörlerdir.
3. Time-bucket yapılandırılmış veriler model performansını artırmaktadır.
4. Çoklu model aileleri kullanan çalışmaların daha güvenilir sonuç verdiği raporlanmıştır.

Boşluk:
Bu proje özellikle *25.000 satırlık yüksek boyutlu LMS aktivite verisi* üzerinde sistematik bir Pass/Fail tahmin modeli ortaya koymayı hedeflemektedir.

5) Veri Tanımı ve Yönetimi

Veri Seti:

input_50.csv
LMS etkinlik verilerinden oluşan, yaklaşık **50% sampling** uygulanmış yüksek boyutlu bir veri seti.

Veri Kaynağı:

Açık veri paylaşım platformu / Araştırma amaçlı kullanılan LMS logları
(Lisans durumu: TBD – eğitim kullanımına uygundur)

Veri Şeması:

- **Sayısal Değişkenler:** course_view_time, resource_view_time, quiz_attempt_time, assign_submit_time, url_view_pct, forum_view_pct vb.
- **Kategorik:** TARGET, BIN_TARGET
- **Beklenen Aralıklar:** 0–1 normalleştirilmiş yüzdelikler, bazı eksik değerler (-1)

Boyut:

- Yaklaşık **25.000 satır**
- **70+ sütun**
- Sınıf dağılımı: Pass / Fail (dengesizlik analiz edilecek)

Etik & Gizlilik:

- Kişisel bilgiler içermemektedir
- Öğrenci ID anonimdir (UID)
- Tüm analizler eğitim amaçlı yapılmaktadır

6) Keşifsel Veri Analizi (EDA)

- Eksik değerlerin belirlenmesi (`-1` → NaN)
- Aykırı değer tespiti (IQR, boxplot)
- Sızıntı kontrolü (TARGET ile doğrudan ilişkili kolonlar)
- Hedef sınıf dengesinin incelenmesi
- Korelasyon analizleri
- Başarılı vs başarısız öğrenci grupları arasındaki davranış farklarının karşılaştırılması

7) Veri Hazırlama Planı

- Eksik değerler için imputasyon: mean/median veya 0 (aktivite yok)
- Normalizasyon / Standartlaştırma
- One-hot encoding (varsa kategoriler)
- Feature scaling (MinMax veya StandardScaler)
- Zaman dilimlerine göre yeni özellikler türetme
- PCA ile boyut indirgeme (isteğe bağlı)

8) Modelleme Planı

Baseline

- Majority baseline
- Logistic Regression (baseline)

Aday Modeller

- Logistic Regression
- Decision Tree
- KNN
- Naive Bayes

Hiper-Parametre Ayarlama

- Grid Search
- Randomized Search
- Stratified K-Fold

Dengesizlik Yönetimi

- Class weights
- SMOTE (gerekiyorsa)

9) Değerlendirme Tasarımı

Metrikler

- Accuracy
- F1-score
- ROC–AUC
- Precision, Recall
- Confusion Matrix

Validation

- Train/Test → 80/20
- Stratified K-Fold
- Veri sızıntısı kontrolü

10) Riskler ve Azaltma Yöntemleri

Risk	Çözüm
Eksik değerler	**İmputasyon**
Sınıf dengesizliği	**class weights/SMOTE**
Overfirring	**CV+Regularization**
Yüksek boyut	**PCA/Feature selection**

11) Kullanılan Araçlar

- **Python:** 3.11
- **Kütüphaneler:** pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost
- **Environment:** Jupyter Notebook

12) Beklenen Sonuçlar ve Görselleştirme

- Confusion matrix
- ROC/AUC eğrileri
- Feature importance (Decision Tree)
- LR katsayı analizi
- Model karşılaştırma tablosu

(Fatma)

Model	F1	Accuracy
LogReg(Orignal)	**0.963984**	**0.956057**
DecisionTree (Original)	**0.973578**	**0.967736**
LogReg (Chi-Square)	**0.964378**	**0.956453**
DecisionTree (Chi-Square)	**0.975467**	**0.970111**
LogReg (PCA)	**0.945701**	**0.933492**
DecisionTree (PCA)	**0.904247**	**0.882621**

(Hazal)

Model	Accuracy	F1 Score
KNN (Baseline)	0.8856	0.9087
KNN (Tuned)	0.8939	0.9162
KNN (After ANOVA FS)	0.9416	0.9529
Naive Bayes (Baseline)	0.8504	0.8772
Naive Bayes (After ANOVA FS)	0.8935	0.9135

(İmad)

Model	Accuracy	F1 Score
SVM (Original)	0.94+	0.95+
LDA (Original)	0.91 – 0.92	0.92 civarı
SVM (Chi-Square)	0.95+	0.96+
LDA (Chi-Square)	0.92 – 0.93	0.93
SVM (PCA)	0.96+	0.97+
LDA (PCA)	0.94 civarı	0.95

13) Referanslar

[1] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 6, pp. 601–618, 2010.

[2] A. Mueen, B. Zafar, and U. Manzoor, “Modeling and predicting students’ academic performance using data mining techniques,” *International Journal of Modern Education and Computer Science*, vol. 7, no. 11, pp. 36–42, 2015.

[3] R. J. Marwani, “Predicting student performance using LMS log data and machine learning,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, pp. 124–131, 2021.

[4] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[5] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[6] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338–345.

[7] J. W. Tukey, “Exploratory data analysis,” *Addison-Wesley*, 1977.