GradeSense

LMS Aktivite Verileri ile Öğrenci Başarısının Tahmini

Ders Adı:

Veri Madenciliği

Proje Ekibi:

- Fatma Abazlı – 22040101142
- Hazal Al Mohammad Algharbi 22040101140
- İmaduddin Hamo – 22040101010

Danışman:

Dr. Yıldız Karadayı

GitHub Proje Linki: https://github.com/GradeSense210/GradeSense

PROJE AMACI

Öğrenci Başarısını Erken Aşamada Tahmin Etmek

LMS aktivitelerine bakarak öğrencinin dersi geçip geçmeyeceğini (PASS / FAIL) erken aşamada tahmin etmek, öğrencilere zamanında destek sağlamak ve akademik başarıyı artırmak için kritik öneme sahiptir.

Basit Modeller

Hızlı ve yorumlanabilir referans noktaları sunar.

Gelişmiş Modeller

Karmaşık öğrenci davranışlarını ve etkileşimlerini yakalar.

Derin Öğrenme

Zaman serisi verilerini, etkileşimleri ve gizli görüntüleri daha iyi öğrenir.

Kullanılan Veri Seti: Kaynak ve İçerik

Öğrenci başarısı tahmini için, Moodle tabanlı Öğrenme Yönetim Sistemi (LMS) üzerinden toplanan kapsamlı bir aktivite veri seti kullanılmıştır.

Veri Kaynağı ve Genel Yapı

Projede Moodle tabanlı bir LMS sisteminden toplanan öğrenci aktivite verileri kullanılmıştır. Her satır, bir öğrencinin LMS üzerindeki toplam aktivite özetini temsil eder.

- Açık kaynak veri seti: GitHub
- 25.260 satır (öğrenci-ders kayıtları)
- 69 özellik, 1 hedef değişken

Özellik Tipleri

Veri seti, öğrencilerin ders sürecindeki etkileşim davranışlarını temsil eden çeşitli aktivite ve zaman bazlı özelliklerden oluşur:

- **Course Activities:** Ders ve kaynak görüntülemeler
- **Assignment Activities:** Ödev görüntüleme, gönderme zamanları
- **Quiz Activities:** Quiz görüntüleme, deneme sayısı, tamamlama süreleri
- **Forum Activities:** Forum görüntüleme, tartışma katılımı
- **Time-based Features:** Platformda geçirilen süre, aktivite zaman ölçümleri

Hedef Değişken ve Ön İşleme

Modelin hedefi, erken uyarı sistemi oluşturmak için öğrencinin dersi geçip geçmeyeceğini tahmin etmektir.

- **TARGET:** Öğrencinin dönem sonu notu
- **PASS_FAIL:** 1 (Geçti) eğer TARGET ≥ 5, 0 (Kaldı) eğer TARGET < 5
- **Ön İşleme:** -1 değerleri NaN olarak, aktivite eksik değerleri 0 olarak ele alındı. Gereksiz sütunlar kaldırıldı ve dengeli sınıf dağılımı için Stratifiled Split uygulandı.

TARGET Dağılımı

Bu grafik, öğrencilerin dersteki **nihai başarı notunu temsil eden TARGET değişkeninin dağılımını** göstermektedir.

Grafikte **0 değerinde belirgin bir yoğunluk** bulunmaktadır. Bu durum, çok sayıda öğrencinin dersi tamamlamadığını veya başarısız olduğunu göstermektedir.

1 ile 10 arasındaki değerlerde dağılım daha dengeli olup, özellikle **6–10 aralığında** öğrenci sayısının arttığı görülmektedir.

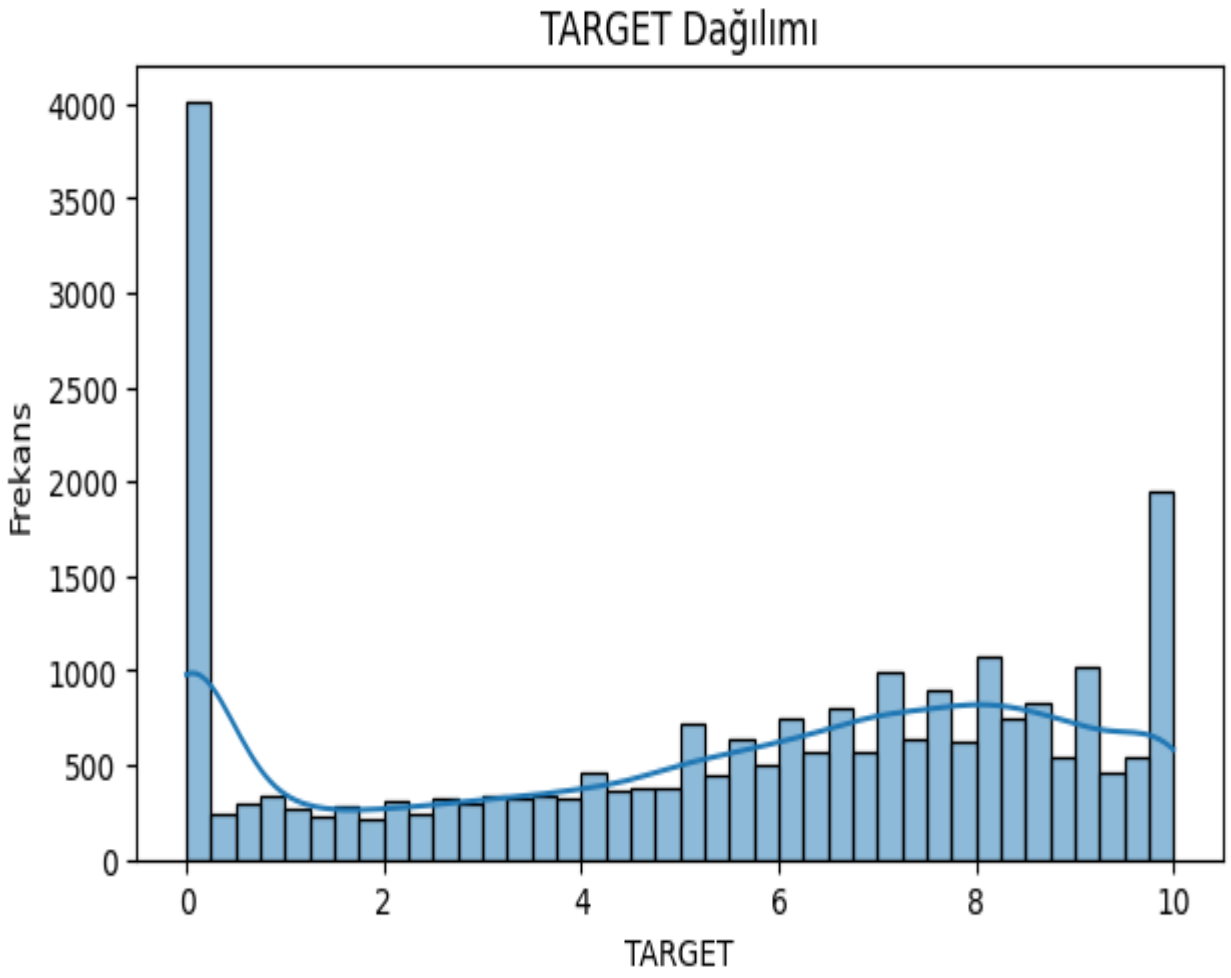
Bu dağılım, veri setinin **dengesiz (imbalanced)** olduğunu göstermektedir.

Bu nedenle TARGET değişkeni, analizlerde **ikili sınıflandırma** problemine dönüştürülmüştür:

0 = Fail (Başarısız)

1 = Pass (Başarılı)

Ayrıca sınıf dengesizliği nedeniyle, model performansı değerlendirilirken yalnızca **Accuracy** değil, **F1-score ve ROC-AUC** gibi metrikler de kullanılmıştır.



PASS_FAIL Dağılımı

Bu grafik, öğrencilerin **başarılı (Pass)** ve **başarısız (Fail)** olarak sınıflandırılmış dağılımını göstermektedir.

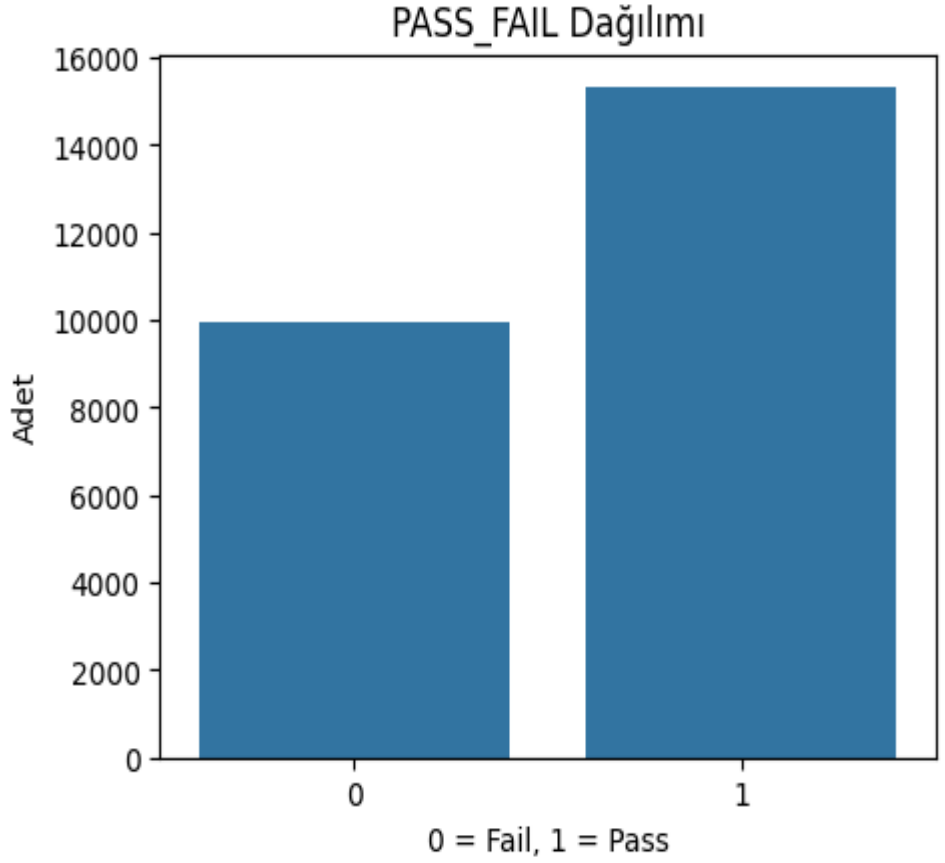
1 (Pass) sınıfının, **0 (Fail)** sınıfına göre daha fazla olduğu görülmektedir.

Buna rağmen sınıflar arasında **tam bir denge yoktur**, yani veri seti **kısmen dengesiz (imbalanced)** bir yapıdadır.

Bu durum, model performansını değerlendirirken yalnızca **Accuracy** metriğinin yeterli olmayacağını göstermektedir.

Bu nedenle çalışmada **F1-score, Precision, Recall ve ROC-AUC** gibi dengesiz veri setleri için daha anlamlı metrikler kullanılmıştır.

Ayrıca **Stratified Train/Test Split** uygulanarak eğitim ve test setlerinde sınıf oranlarının korunması sağlanmıştır.



Korelasyon Matrisi (Numerik Özellikler)

Bu korelasyon matrisi, LMS veri setindeki **sayısal değişkenler** arasındaki doğrusal ilişkileri göstermektedir.

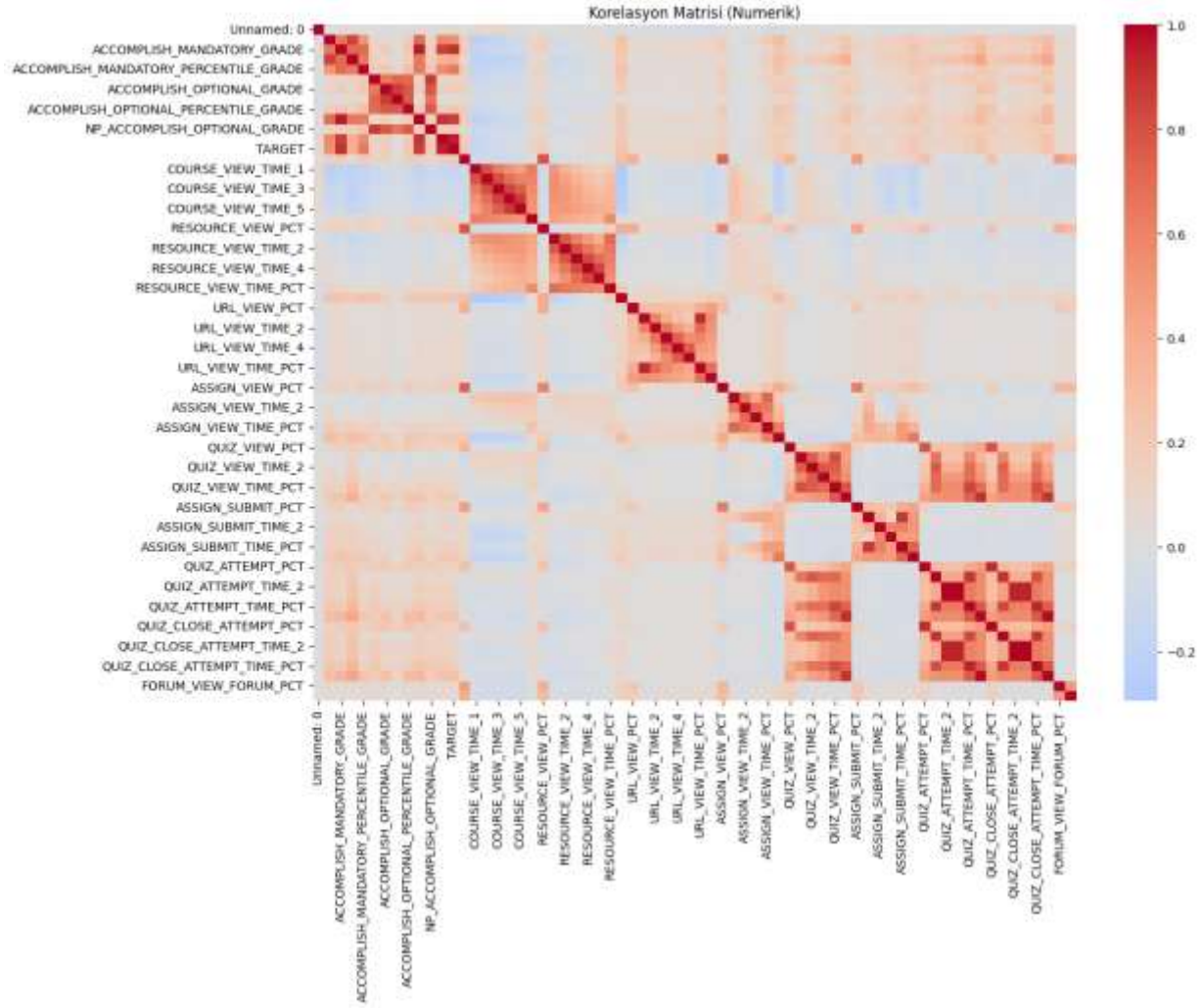
Kırmızı tonlar **pozitif korelasyonu**, mavi tonlar ise **negatif korelasyonu** ifade etmektedir.

Aynı türdeki aktivitelerin (örneğin *quiz attempt*, *quiz close attempt*, *assign submit* gibi) kendi aralarında **yüksek korelasyona** sahip olduğu görülmektedir.

Bu durum, öğrencinin platformdaki belirli bir davranış türünde aktif olmasının, benzer aktivitelerde de aktif olma ihtimalini artırdığını göstermektedir.

TARGET değişkeni ile bazı aktivite ve not temelli değişkenler arasında **orta düzeyde pozitif korelasyon** bulunmaktadır. Bu, LMS etkileşimlerinin öğrenci başarısı ile ilişkili olduğunu desteklemektedir.

Yüksek korelasyonlu özellik grupları, veri setinde **çoklu bağlantı (multicollinearity)** riski oluşturabilir. Bu nedenle çalışmada **Feature Selection (Chi-Square / ANOVA)** ve **PCA** gibi boyut indirgeme teknikleri uygulanmıştır



Activity Features Korelasyon Isı Haritası

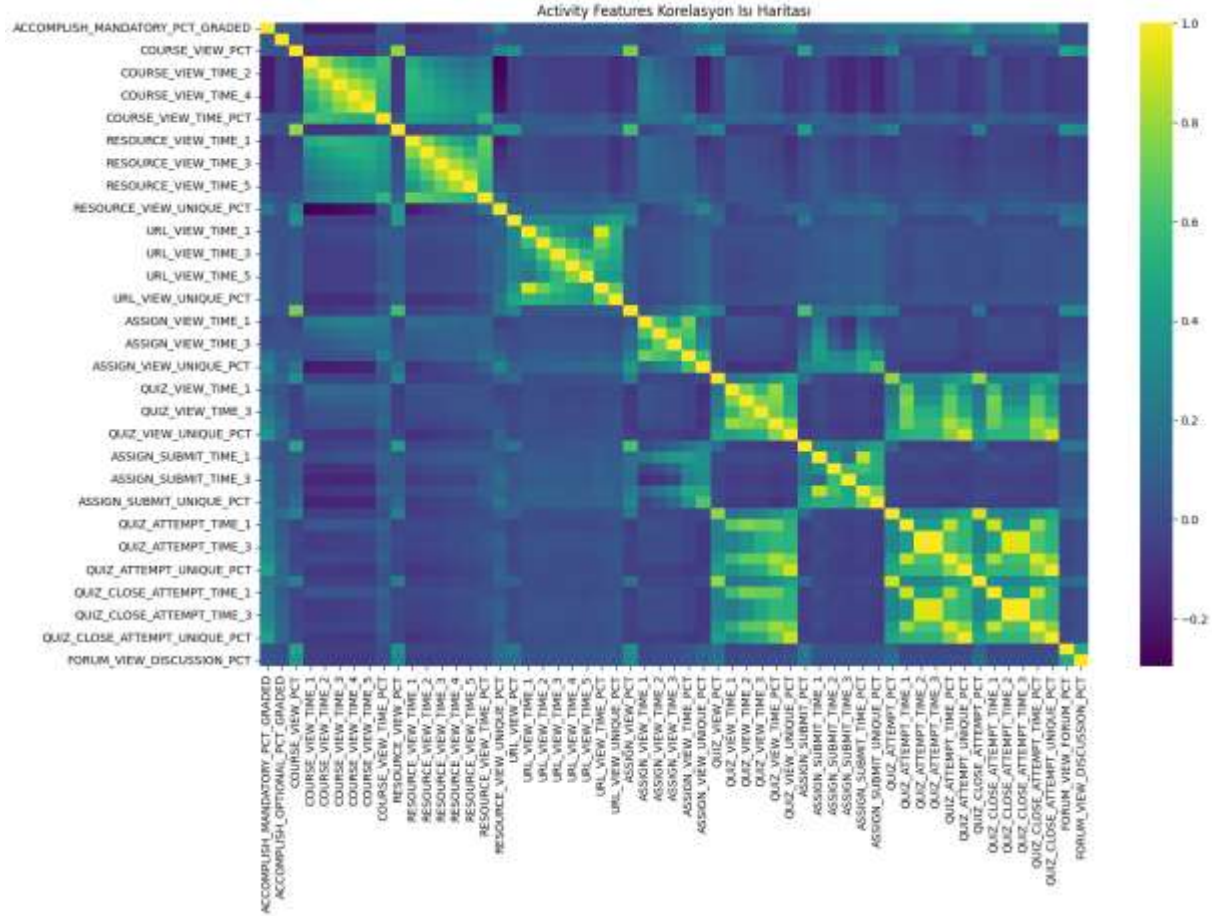
Bu ısı haritası, **LMS üzerindeki aktivite tabanlı özellikler** (view, time, attempt, submit vb.) arasındaki korelasyonları göstermektedir.

Aynı aktivite türüne ait zaman ve yüzde değişkenleri (ör. *COURSE_VIEW_TIME*, *COURSE_VIEW_PCT*) arasında **yüksek pozitif korelasyon** bulunmaktadır.

Quiz attempt, **quiz close attempt** ve **assignment submit** gibi etkileşimler kendi grupları içinde güçlü ilişkiler sergilemektedir. Bu durum, öğrencinin bir aktiviteyi yapma eğiliminin benzer aktivitelerde de devam ettiğini göstermektedir.

Farklı aktivite türleri arasında ise genellikle **düşük veya zayıf korelasyon** görülmektedir. Bu, LMS davranışlarının çok boyutlu ve birbirinden kısmen bağımsız olduğunu göstermektedir.

Yüksek korelasyonlu özellik kümeleri, veri setinde **çoklu bağlantı (multicollinearity)** oluşturabileceğinden, modelleme aşamasında **Feature Selection** ve **PCA** gibi yöntemlerin uygulanması gerekmiştir.



Bu projede label olarak PASS_FAIL değişkenini kullandık. Bu değişken, öğrencinin dersten geçip geçmediğini göstermektedir. Feature olarak ise LMS üzerindeki course, resource, quiz, assignment ve forum aktivitelerine ait zaman ve oran tabanlı tüm sayısal özellikleri kullandık.

Kullanılan Modeller: Kapsamlı Bir Bakış

Öğrenci başarısı tahmini için hem geleneksel makine öğrenimi algoritmaları hem de derin öğrenme tabanlı yaklaşımlar kullanılmıştır.

Geleneksel Modeller (Baseline)

- Logistic Regression
- Decision Tree
- KNN (K-Nearest Neighbors)
- Naive Bayes
- SVM (Support Vector Machine)
- LDA (Linear Discriminant Analysis)

Gelişmiş & Derin Öğrenme Modelleri

- MLP (Multi-Layer Perceptron)
- 1D-CNN (1 Boyutlu Evrişimsel Sinir Ağı)
- GRU (Gated Recurrent Unit)
- TabNet (Tabular Veri için Özel Ağ)
- Autoencoder + Sınıflandırıcı
- Wide & Deep (Hibrit Yaklaşım)

Her Model Ne Problemi Çözüyor?

Her bir modelin amacı ve öğrenci davranışlarını analiz etmedeki rolü aşağıda açıklanmıştır.

Logistic Regression	max_iter=500 (yakınsama için iterasyon sayısı)
Decision Tree	max_depth=8 (ağacın maksimum derinliği, aşırı uydurmayı engeller)
KNN	n_neighbors=5 (sınıflandırma için komşu sayısı)
Naive Bayes	default (varsayılan ayarlar)
SVM	kernel=rbf (Radial Basis Function çekirdeği, doğrusal olmayan ayırım için)
LDA	default (varsayılan ayarlar)
MLP	hidden=(128,64), dropout=0.25 (iki gizli katman ve aşırı uydurmayı önleyici dropout oranı)
1D-CNN	conv(16,32), kernel=3 (16 ve 32 filtreli, 3 boyutlu evrişim katmanları)
GRU	hidden=128, layers=2 (128 birimli iki GRU katmanı)
TabNet	n_d=32, n_steps=5 (karar adımları ve boyut ayarları)
Autoencoder	latent_dim=32 (gizli temsilin boyutu)
Wide & Deep	deep=(256,128), dropout=0.2 (derin bileşen için katmanlar ve dropout)

Detaylı Model Performans Karşılaştırması

Tabloda, öğrenci başarısı tahmini için kullanılan tüm modellerin test seti üzerindeki detaylı performans karşılaştırmasını sunmaktadır. Farklı özellik mühendisliği yaklaşımları (Orijinal, Chi-Square, PCA) ve derin öğrenme çerçeveleri (PyTorch) altındaki performansları da dahil edilmiştir.

Model	Accuracy	F1 Score	ROC-AUC
Logistic Regression (Original)	0.897	0.914	0.963
Logistic Regression (Chi-Square)	0.897	0.914	0.963
Logistic Regression (PCA)	0.893	0.910	0.962
Decision Tree (Original)	0.910	0.925	0.962
Decision Tree (Chi-Square)	0.911	0.926	0.967
Decision Tree (PCA)	0.823	0.853	0.900
KNN (Baseline)	0.739	0.548	0.785
Naive Bayes (Baseline)	0.756	0.577	0.817
MLP (PyTorch)	0.911	0.926	0.971
1D-CNN (PyTorch)	0.909	0.923	0.973
SVM (PCA)	0.943	0.954	0.981
LDA (PCA)	0.936	0.949	0.975
SVM (Chi-Square)	0.961	0.968	0.987
LDA (Chi-Square)	0.953	0.962	0.980
AE + Classifier	0.958	0.965	0.989
GRU (PyTorch)	0.962	0.969	0.989
TabNet (PyTorch)	0.962	0.969	0.986
Wide & Deep (Best)	0.964	0.971	0.990

En İyi Model: Wide & Deep

Kapsamlı karşılaştırmalar sonucunda Wide & Deep modeli, öğrenci başarısı tahmininde en üstün performansı sergilemiştir.

Yüksek Accuracy

Genel doğru tahmin oranı en yüksek modeldir.

Üstün F1-Score

Dengeli hassasiyet ve geri çağırma (recall) ile sınıflandırma başarısı.

Mükemmel ROC-AUC

Pozitif ve negatif sınıfları ayırmadaki üstün yeteneği.

Hibrit Öğrenme Yeteneği

Hem basit lineer hem de karmaşık derin ilişkileri öğrenir.

> LMS öğrenci aktiviteleri, akademik başarıyı tahmin etmek için güçlü bir göstergedir ve Wide & Deep gibi gelişmiş modellerle bu potansiyel en üst düzeyde değerlendirilebilir.