Proje Raporu

Abstract

Bu çalışmada, öğrencilerin akademik performansını (sınav skoru) günlüksel yaşam ve çalışma alışkanlıklarına dayalı olarak tahmin etmek amacıyla yapay sinir ağları (ANN) modelleri geliştirilmiştir. Kaggle'da yer alan Student Habits vs Academic Performance veri seti kullanılarak, eksik veri yönetimi, kategorik dönüşümler, özellik mühendisliği, ölçeklendirme ve veri bölme işlemleri uygulanmıştır. Hem temel hem de optimize edilmiş modellerin performansları MSE ve MAE metrikleriyle karşılaştırılmıştır. Sonuçlar, özellik mühendisliği sonrası manuel olarak optimize edilmiş modelin en iyi genelleme performansını sunduğunu göstermektedir.

1. Introduction

Akademik başarı, öğrencilerin yaşam tarzı ve alışkanlıklarıyla yakından alakalıdır. Bu proje, bu ilişkiyi niceliksel olarak modelleyerek öğrencilerin sınav skorlarını öngörmeyi amaçlar. Elde edilen tahminler, eğitim planlaması ve erken müdahale süreçlerinde kullanılabilir.

2. Data Description

2.1 Dataset Overview

- **Kaynak:** Kaggle – Student Habits vs Academic Performance
- **Boyut:** 1000 gözlem, 16 değişken (1 hedef, 15 özellik)
- **Özellikler:** age, gender, study_hours_per_day, social_media_hours, netflix_hours, part_time_job, attendance_percentage, sleep_hours, diet_quality, exercise_frequency, parental_education_level, internet_quality, mental_health_rating, extracurricular_participation, exam_score

2.2 Data Quality Issues

- Eksik veri: parental_education_level (%9.1)
- Aykırı değer: study_hours_per_day ve social_media_hours sütunlarında gözlemlenen uç değerler

3. Methodology

3.1 Data Preprocessing

- **Missing Value Imputation:** parental_education_level → "Unknown"
- **Ordinal Encoding:** diet_quality, internet_quality, parental_education_level
- **One-Hot Encoding:** gender, part_time_job, extracurricular_participation

3.2 Feature Engineering

- Interaction: study_hours_per_day × attendance_percentage
- Aggregation: social_media_hours + netflix_hours
- Ratios: total_social_hours / (study_hours_per_day + 1), sleep_hours / (study_hours_per_day + 1)

3.3 Scaling and Splitting

- StandardScaler ile sayısal özelliklerin normalize edilmesi
- Veri bölme: %70 eğitim, %15 validasyon, %15 test (random_state=42)

3.4 Model Architectures

Model	Architecture	Hyperparameters
Baseline ANN	[64, 32]	lr=0.001, batch=32
Manual Tuned ANN	[128, 64, 32] + dropout(0.1)	lr=5e-4, batch=16
KerasTuner RandomSearch	2–4 katman, units=[32–256], dropout=[0–0.5]	max_trials=20, epochs=50
KerasTuner Hyperband	2–4 katman, factor=3, max_epochs=30	
FE + Tuned ANN	Manual Tuned üzerine Feature Engineered	lr=5e-4, batch=16

3.5 Training Strategy

- Optimizör: Adam
- Loss: Mean Squared Error (MSE)
- Metric: Mean Absolute Error (MAE)
- EarlyStopping: patience=10–15 (val_loss)

4. Results

Model performansları test seti üzerinde aşağıdaki tabloda özetlenmiştir:

Model	Test MSE	Test MAE
Baseline ANN	58.42	6.19
Manual Tuned ANN	39.83	5.18
RandomSearch Tuning	49.15	5.66
Hyperband Tuning	43.93	5.29
FE + Baseline	135.70	9.23
FE + Tuned ANN	37.68	4.75
5-Fold CV FE+Tuned	36.52±4.58	4.85±0.31

5. Discussion

- Özellik mühendisliği, model performansını %20–30 oranında iyileştirmiştir.
- Manuel tuning, otomatik tuning yöntemlerinden daha iyi sonuç vermiştir.
- 5-fold CV sonuçları, modelin genelleme yeteneğinin istikrarlı olduğunu göstermektedir.

6. Conclusion and Future Work

Bu çalışmada, ANN tabanlı modellerle öğrencilerin sınav skorlarının tahmini başarıyla gerçekleştirilmiştir. Gelecekte:

- Derin öğrenme (CNN/RNN) veya ensemble yöntemleri test edilebilir.
- Gerçek zamanlı değerlendirme için bir API servisi oluşturulabilir.
- Veri setine yeni çevresel veya psikometrik değişkenler eklenerek model zenginleştirilebilir.

Hazırlayan: Çağatay Dişli