ABSTRACT

The primary aim of this thesis is to model and predict student academic performance using machine learning techniques. The focus is on 'G_avg', a composite measure derived from the average of three periodic academic grades (G1, G2, G3), providing a holistic view of a student's performance over time. The research aims to uncover patterns and factors significantly affecting academic success by predicting 'G_avg'. The insights gained from this study aim to inform educators and policymakers in developing strategies to enhance student performance and identify students who may need additional support. The thesis also aims to compare various machine learning algorithms to determine the most effective approach for such predictive analysis in an educational context.

As time progresses and machine learning algorithms and technological proficiency continue to advance, our analytical capabilities have become more refined, enabling us to develop deeper insights into complex datasets. With over twenty years of experience in teaching mathematics in private and international schools, my inclination towards education has inevitably evolved to overlap with these technological advancements. I believe the analytical prowess of machine learning algorithms can provide a more comprehensive understanding of student data, thereby illuminating their academic achievements in greater detail.

The dataset used in this research was obtained from Kaggle, a well-established online platform providing authentic datasets for a wide range of analytical endeavors. Despite being sourced from a school in Portugal, its inclusion of universal aspects of education makes it meaningful and beneficial for this study. Its extensive features provided a robust foundation for thorough analysis.

The dataset's evaluations consisted of three different exam results (G1, G2, and G3), with the final exam (target value) determined as G3. Contrary to the expected cumulative structure of

G3, it is designed to function as an independent assessment, similar to G1 and G2. This observation led to the development of a new feature, 'G_avg', calculated as the arithmetic mean of these three grades, to provide a more comprehensive target variable for the analysis. A range of regression models, including *Ridge* Regression, *Lasso* Regression, Elastic Net, and Logistic Regression, were incorporated into my methodology. These models were selected based on their ability to effectively capture the complex, non-linear relationships in the data. In the classification part, three success categories were applied to 'G_avg': 'Unsatisfactory', 'Satisfactory', and 'Successful'. In this context, various classification models were applied. The chosen algorithms are widely recognized for their high efficiency and accuracy in classification tasks: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Neural Networks (MLP Classifier).
The purpose of this research is to provide predictive understandings that can guide educators, thereby enhancing students' educational experiences.

Keywords: Machine Learning, Educational Data Analysis, Academic Performance, Regression Models, Classification Algorithms, Predictive Modeling, Data-Driven Education, Student Success Metrics, Elastic Net, *Ridge* Regression, *Lasso* Regression, Logistic Regression, Decision Trees, Random Forest, SVM, Gradient Boosting, Neural Networks, MLP Classifier, Pedagogical Strategies, Kaggle Dataset, Performance Prediction, Success Interval Classification

ÖZET

Bu tezin temel amacı, makine öğrenmesi tekniklerini kullanarak öğrenci akademik performansını modellemek ve tahmin etmektir. Odak noktası, öğrencinin zaman içindeki performansına ilişkin bütünsel bir görünüm sağlayan, üç periyodik akademik notun (G1, G2, G3) ortalamasından türetilen bileşik bir ölçüm olan ' G_avg ' değişkenidir. Araştırma, ' G_avg ' tahminini yaparak akademik başarıyı önemli ölçüde etkileyen kalıpları ve faktörleri ortaya çıkarmayı amaçlıyor. Bu çalışmadan elde edilen bilgiler, eğitimcileri ve politika yapıcıları öğrenci performansını artırmaya ve ek desteğe ihtiyaç duyabilecek öğrencileri belirlemeye yönelik stratejiler geliştirme konusunda bilgilendirmeyi amaçlamaktadır. Ayrıca tez, eğitim bağlamında bu tür tahmine dayalı analiz için en etkili yaklaşımı belirlemek amacıyla çeşitli makine öğrenimi algoritmalarını karşılaştırmayı amaçlamaktadır.

Zaman ilerledikçe ve makine öğrenimi algoritmaları ve teknolojik beceri gelişmeye devam ettikçe, analitik kapasitelerimiz daha da gelişti ve karmaşık veri kümelerine ilişkin derin içgörüler geliştirmemizi sağladı. Özel ve uluslararası okullarda matematik eğitimi verme konusunda yirmi yıldan fazla deneyime sahip biri olarak, eğitime olan eğilimim kaçınılmaz olarak bu teknolojik ilerlemelerle örtüşecek şekilde gelişti. Makine öğrenimi algoritmalarının analitik yeteneklerinin, öğrenci verilerinin daha kapsamlı anlaşılmasını sağlayabileceğine ve böylece akademik başarılarını daha ayrıntılı bir şekilde aydınlatabileceğine düşünüyorum.

Bu araştırmada kullanılan veri seti, çok çeşitli analitik çabalar için özgün veri setleri sağlayan köklü bir çevrimiçi platform olan Kaggle'dan alındı. Veriler Portekiz'deki bir okuldan alınmış olmasına rağmen, eğitimin evrensel yönlerini bünyesinde barındırması nedeniyle bu çalışma için anlamlı ve faydalıdır. Sahip olduğu detaylı öznitelikler kapsamlı bir inceleme için güçlü bir temel oluşturdu.

Bu veri setindeki değerlendirmeler üç farklı sınav sonucundan (G1, G2 ve G3) oluşuyordu ve sonuç değeri (*target value*) G3 olarak belirlenmişti. G3 ise beklenen kümülatif yapısının aksine, yapısı G1 ve G2'ninkine benzeyen ve bağımsız bir değerlendirme işlevini görecek şekilde tasarlanmıştı. Bu gözlem, analize daha kapsamlı bir hedef değişken sağlamak amacıyla bu üç notun aritmetik ortalaması olarak hesaplanan 'G_avg' adlı yeni bir özniteliğin geliştirilmesine yol açtı.

Ridge Regresyonu, *Lasso Regresyonu, Elastik Net* ve Lojistik regresyon gibi bir dizi regresyon modeli metodolojime dahil edildi. Bu modellerin seçimi, verilerde mevcut olan karmaşık, doğrusal olmayan ilişkileri daha etkili bir şekilde yakalama yeteneklerine dayanıyordu.

Sınıflandırma kısmında ise 'G_avg'e üç başarı kategorisi uygulandı: 'Yetersiz', 'Tatmin Edici' ve 'Başarılı'. Bu bağlamda çeşitli sınıflandırma modelleri uygulandı. Seçilen algoritmalar, sınıflandırma görevlerindeki yüksek verimlilikleri ve doğruluklarıyla geniş çapta bilinen algoritmalardır; Lojistik Regresyon, Karar Ağaçları, Rassal Orman, Destek Vektör Makineleri (SVM), Gradyan Arttırma ve Sinir Ağları (MLP Sınıflandırıcı).

Bu araştırmanın amacı, eğitimcilere rehberlik edebilecek, dolayısıyla öğrencilerin eğitim deneyimlerini geliştirebilecek öngörülü anlayışlar sunmaktır.

Anahtar Kelimeler: Makine Öğrenimi, Eğitimsel Veri Analizi, Akademik Performans, Regresyon Modelleri, Sınıflandırma Algoritmaları, Tahmine Dayalı Modelleme, Veriye Dayalı Eğitim, Öğrenci Başarı Metrikleri, Elastik Net, *Ridge* Regresyonu, *Lasso* Regresyonu, Lojistik Regresyon, Karar Ağaçları, Rassal Orman, SVM, Gradyan Artırma, Sinir Ağları, MLP Sınıflandırıcı, Pedagojik Stratejiler, Kaggle Veri Kümesi, Performans Tahmini, Başarı Aralığı Sınıflandırması