T.C.

AKDENIZ UNIVERSITY

SBS EXAM RESULT PREDICTION MODEL

Botan ONAT

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

DEPARTMENT

MASTER’S THESIS

NOVEMBER 2024

ANTALYA

T.C.

AKDENIZ UNIVERSITY

SBS EXAM RESULT PREDICTION MODEL

Botan ONAT

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

DEPARTMENT

MASTER'S THESIS

NOVEMBER 2024

ANTALYA

T.C.

AKDENİZ ÜNİVERSİTESİ

SBS EXAM RESULT PREDICTION MODEL

Botan ONAT

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

KASIM 2024

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SBS EXAM RESULT PREDICTION MODEL

Botan ONAT

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

Bu tez 01/11/2024 tarihinde jüri tarafından Oybirliği / ~~Oyçokluğu~~ ile kabul edilmiştir.

Doç. Dr. Alper BİLGE (Danışman)

Dr. Öğr. Üyesi Murat AK

Doç. Dr. Ayça AKIN

ABSTRACT

SBS EXAM RESULT PREDICTION MODEL

BOTAN ONAT

Master's Thesis, Computer Engineering

Advisor: Assoc. Prof. Alper Bilge

November 2024; 25 pages

This research aims to predict the academic performance of middle school students using machine learning algorithms applied to performance, demographic, and survey data. In this study, we evaluated the effectiveness of each algorithm on individual and combined datasets using Multiple Linear Regression (MLR), Support Vector Machine (SVM), and Convolutional Neural Network (CNN) with 5-fold cross-validation. Performance is evaluated with Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (r2) criteria. In this study, comparative analysis was used to determine the optimal strategy for educational data analytics by revealing significant differences in model accuracy and computational efficiency. This research aims to contribute to educational data mining and improve educational strategies and interventions by providing a detailed evaluation of machine learning techniques for predicting student performance.

KEYWORDS: Academic Performance Prediction, Data Analytics in Education, Educational Data Mining, Machine Learning, Performance Metrics

COMMITTEE: Assoc. Prof. Alper BİLGE
Asst. Prof. Murat AK
Assoc. Prof. Ayça AKIN

ÖZET

SBS SINAV SONUCU TAHMİN MODELİ

BOTAN ONAT

Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç.Dr. Alper Bilge

Kasım 2024; 25 sayfa

Bu araştırma, performans, demografik ve anket verilerine uygulanan makine öğrenimi algoritmalarını kullanarak ortaokul öğrencilerinin akademik performansını tahmin etmeyi amaçlamaktadır. Bu çalışmada, Çoklu Doğrusal Regresyon (MLR), Destek Vektör Makinesi (SVM) ve Evrişimsel Sinir Ağı (CNN) kullanarak 5 kat çapraz doğrulama kullanarak her algoritmanın bireysel ve birleşik veri kümeleri üzerindeki etkinliğini değerlendirdik. Performans Ortalama Mutlak Hata (MAE), Ortalama Karesel Hata (MSE), Ortalama Karekök Hata (RMSE), R-kare (r2) kriterleri ile değerlendirilir. Bu çalışmada, model doğruluğu ve hesaplama verimliliğindeki önemli farklılıkları ortaya koyarak eğitimsel veri analitiği için en uygun stratejiyi belirlemek amacıyla karşılaştırmalı analiz kullanılmıştır. Bu araştırma, öğrenci performansını tahmin etmeye yönelik makine öğrenimi tekniklerinin ayrıntılı bir değerlendirmesini sağlayarak eğitimsel veri madenciliğine katkıda bulunmayı ve eğitim stratejilerini ve müdahalelerini geliştirmeyi amaçlamaktadır.

ANAHTAR KELİMELER: Akademik Performans Tahmini, Eğitimde Veri Analitiği, Eğitimsel Veri Madenciliği, Makine Öğrenmesi, Performans Metrikleri

JÜRİ: Doç. Dr. Alper BİLGE
Dr. Öğr. Üyesi Murat AK
Doç. Dr. Ayça AKIN

PREFACE

The work process of this thesis began with the aim of exploring how machine learning techniques can be applied in education to predict and improve student performance.

In this study, I investigated the use of various machine learning algorithms to predict the academic performance of middle school students. By analyzing performance data, demographic information, and survey data, I aimed to identify the most effective prediction models and help educators improve student outcomes.

I would like to thank Associate Professor Alper Bilge and Lecturer Ayça Akın, whose mentorship and expertise were invaluable throughout this research. Their feedback played an important role in the quality of this study and significantly contributed to its direction.

This thesis is a result of meticulous research, experiments and analysis. I hope this work will inspire further research and contribute to ongoing efforts to harness the power of machine learning in education.

CONTENTS

ACADEMIC DECLERATION

I hereby declare that this study, titled "SBS EXAM RESULT PREDICTION MODEL," which I submitted as my Master's Thesis, has been written in accordance with academic rules and ethical values. I also affirm that I have cited all sources for any information that does not belong to me in this thesis.

01/11/2024

Botan ONAT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

Predicting academic performance is a fundamental aspect of educational research and is crucial to developing strategies to improve student outcomes. The ability to accurately predict students' academic achievement can lead to early interventions, tailored education programs, and informed policy decisions. Thanks to the advent of machine learning and the availability of extensive training data, it is now possible to build more complex predictive models.

This research aimed to study the use of machine learning in educational applications. This study, which analyzes the performance data, demographic information and survey responses of secondary school students, aims to reveal the main factors affecting academic success.

1.2 Contribution

This thesis makes several important contributions to the field of educational data mining:

1.2.1. Comparative Analysis of Machine Learning Algorithms

This research provides a detailed comparative analysis of three machine learning algorithms: Multiple Linear Regression (MLR), Support Vector Machine (SVM) and Convolutional Neural Network (CNN). Each algorithm is evaluated using individual datasets (performance, demographic, and survey data) and various combinations of these datasets. This comprehensive evaluation highlights the strengths and weaknesses of each algorithm in predicting academic performance.

1.2.2. Impact of Data Integration

The study investigates the effects of integrating different types of data sets on the accuracy and efficiency of predictive models. Combining performance data with demographic and survey data, the research explores how data integration can increase the predictive power of machine learning models. This approach provides insights into optimal strategies for using a variety of training data.

1.2.3. Practical Information for Educators and Policy Makers

The findings provide valuable practical information for educators and policy makers. Research helps design more effective interventions and policies by identifying the most influential factors that influence academic performance. Results can inform the development of personalized learning plans, targeted support programs, and data-driven educational policies.

1.2.4. Evaluation Metrics and Model Performance

Run, Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (r2) and run time to evaluate model performance. This detailed analysis provides a clear understanding of the trade-offs between accuracy and computational efficiency in educational data mining.

2. LITERATURE REVIEW

2.1. Machine Learning in Educational Data Mining

Machine learning is starting to be utilized in educational data mining, giving advanced devices for foreseeing scholastic execution. Different calculations have been actualized in this context, each with their qualities and limitations. This segment investigates the fundamental machine learning strategies utilized in past studies and their relevance to academic performance forecast.

2.1.1. Multiple Linear Regression (MLR)

MLR has long been a staple of educational research due to its straightforwardness and interpretability. It models the relationship between a subordinate variable (e.g., academic performance) and multiple independent variables (e.g., demographic variables, survey reactions). MLR is particularly valuable when connections between variables are linear and easily interpretable. In any case, it may not perform well on complex, nonlinear data.

2.1.2. Support Vector Machine (SVM)

SVM is known for its strength in classification and regression tasks, particularly with high-dimensional data. Research has shown that SVM can accomplish high precision in foreseeing student performance, particularly when managing with large and complex data sets. Be that as it may, SVM requires cautious tuning of parameters and can be computationally seriously.

2.1.2. Convolutional Neural Network (CNN)

Ordinarily utilized in image and pattern recognition, CNNs appear guarantee in educational data mining due to their capacity to capture complex patterns. CNNs can demonstrate nonlinear relationships and interactions within data, making them exceedingly compelling at anticipating academic performance. In spite of their high precision, CNNs require critical computational assets and longer training times.

2.2. Comparative Studies on Machine Learning Algorithms

Numerous studies have compared distinctive machine learning algorithms to decide the foremost successful methods for predicting academic performance. Tomasevic, Gvozdenovic, and Vranes (2020) assessed different administered data mining strategies, including decision trees, SVMs, k-nearest neighbors (KNN), logistic regression, and artificial neural networks (ANN). The consider highlighted the complexity of foreseeing student performance due to the expansive number of affecting components and the require for comprehensive models that combine different variables.

Chen and Zhai (Chen and Zhai2023) compared the effectiveness of SVM, ANN, and other machine learning techniques using a large dataset. They found that SVM and ANN provided the highest precision in predicting student performance, highlighting the significance of statistic, academic, and behavioral factors in their models.

Rodríguez-Hernández, Musso, Kyndt, and Cascallar (2021) efficiently applied ANNs to anticipate academic performance, focusing on the effectiveness of different indicators. Their research appears that ANNs can accomplish high precision, particularly when combining a wide range of demographic, academic, and psychological variables.

Pan and Cutumisu (2024) conducted a study utilizing PISA 2018 data to assess the life satisfaction of British and Japanese secondary school students. They applied machine learning techniques to understand the variables that impact life fulfillment and found that social support, academic accomplishment, family relationships, and school satisfaction were imperative indicators. Their research highlights the potential of machine learning to capture cultural differences and their affect on students' well-being.

2.3. Key Factors Influencing Academic Performance

The looked into writing reliably recognizes a few key components that impact scholastic execution:

2.3.1. Demographic Factors

Gender, age, socio-economic status, and parents' education levels are critical determinants of student performance. Research has appeared that these factors altogether influence scholarly results.

2.3.2. Academic Background

Previous grades and support in extracurricular activities are solid predictors of future academic success. These variables reflect the student's ongoing interest and execution within the educational framework.**Behavioral and Psychological Factors**

Motivation, study habits, and psychological well-being play crucial roles in academic achievement. Research indicates that students with higher motivation and better study habits tend to perform better academically.

2.3.3. Family and School Environment

Parental involvement, family support, and the quality of the school environment significantly influence student performance. Supportive family environments and high-quality school resources are associated with higher academic achievement.

2.4. Research Gaps and Future Directions

Despite significant advancements, several research gaps remain. Most studies have focused on single datasets or limited variables, lacking comprehensive comparisons of multiple algorithms across diverse datasets. Furthermore, the practical application of these models in real-world educational settings has not been fully explored. Future research should aim to:

- Conduct systematic comparisons of various machine learning algorithms using extensive and diverse datasets.

- Integrate a broader range of variables to capture the multifaceted nature of academic performance.

- Evaluate the practical implications and scalability of these models in real-world educational environments.

The literature review underscores the potential of machine learning algorithms in predicting academic performance. MLR, SVM, and CNN have demonstrated high accuracy in different contexts, with SVM and CNN often outperforming traditional methods. By addressing the identified research gaps and focusing on comprehensive, integrative models, future studies can further enhance the predictive power and practical applicability of machine learning in education.

3. PROPOSE METHODOLOGY

3.1 Utilized Machine Learning Techniques

This study explores the effectiveness of Multiple Linear Regression (MLR), Support Vector Machine (SVM), and Convolutional Neural Network (CNN) in foreseeing middle school students' academic performance. These algorithms were chosen due to their proven accuracy in educational data mining.

3.1.1 Multiple Linear Regression (MLR)

MLR may be a statistical strategy that models the linear relationship between a dependent variable and multiple independent variables. It is broadly utilized for its simplicity and interpretability. MLR makes a difference in understanding how various factors like demographic attributes and overview reactions impact academic performance. The MLR demonstrate is represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \in \tag{3.1}$$

Where (y) is the dependent variable, (x1,x2,…,xn) are the independent variables, (β0) is the intercept, (β1,β2,…,βn) are the coefficients, and (ϵ) is the error term.

3.1.2 Support Vector Machine (SVM)

SVM is known for its robustness in handling high-dimensional data. It is utilized for classification and regression tasks. In regression, SVM aims to find a function that approximates the relationship between inputs and outputs with minimal error. The SVM regression model is:

$$f(x) = w.x + b \tag{3.2}$$

Where (w) is the weight vector, (x) is the input feature vector, and (b) is the bias term. The objective is to minimize the mistake whereas keeping up a margin of tolerance (ϵ).

3.1.3 Convolutional Neural Network (CNN)

CNNs are deep learning models known for their ability to capture complex patterns. Initially utilized for image processing, CNNs have been adapted for various data types, including educational data. CNNs consist of convolutional layers, pooling layers, and fully associated layers. The architecture of a CNN typically incorporates:

- Convolutional Layer: Applies filters to the input data to detect patterns.
- Activation Function: Introduces non-linearity, commonly using ReLU.
- Pooling Layer: Reduces the dimensionality of feature maps.
- Fully Connected Layer: Combines features for final prediction.

3.2. Methodological Approach

The study includes data collection, preprocessing, model training, and evaluation using a 5-fold cross-validation method to ensure robustness.

3.2.1. Data Collection and Preparation

The research utilizes three datasets:

Performance data, demographic data, and survey data. Each dataset contains attributes that may impact academic performance.

- Performance Data: Incorporates grades and standardized test scores.
- Demographic Information: Includes variables such as gender, socio-economic status, and parental education levels.
- Survey Data: Consists of responses to survey questions related to study habits, motivation, and mental well-being.

Data preprocessing steps include dealing with missing values, normalizing numeric attributes, and encoding categorical variables.

3.2.2. Experimental Setup

A 5-fold cross-validation approach is utilized to assess model performance. This includes splitting the data into five subsets, with each subset serving as a validation set once whereas the remaining subsets are utilized for training. This process guarantees dependable estimates of model precision.

4. EXPERIMENTS

4.1. Data Collection and Preparation

The experimental segment of this study involved the collection and preprocessing of three distinct datasets: performance data, demographic data, and survey data. These datasets have been chosen to provide a complete view of the elements influencing educational overall performance among middle school students.

4.1.1. Performance Data

The performance data consisted of students' grades and standardized test ratings. This dataset served as the primary indicator of educational achievement. Grades were gathered for various subjects, while standardized test rankings provided a more uniform measure of student overall performance across distinct schools and regions. The performance statistics included variables such as:

- Subject Grades: Ratings for individual subjects like Mathematics, Science, Language Arts, etc.

- Overall GPA: Cumulative grade point average for each student.

- Standardized Test Scores: Scores from state or national standardized tests.

4.1.2. Demographic Data

Demographic data included variables such as gender, socio-economic status, parental education levels, family structure, and other socio-economic factors. These variables were essential in understanding the background and environment in which the students were learning. The demographic data included:

- Gender: Male or female.

- Age: Age of the student.

- Socio-economic Status: Family income levels categorized into various ranges.

- Parental Education Levels: Highest education level attained by the parents.

- Family Structure: Information on whether the student comes from a nuclear, extended, or single-parent family.

4.1.3. Survey Data

The survey data comprised responses to 36 survey questions related to students' attitudes towards mathematics, study habits, motivation, and psychological well-being. The survey aimed to capture subjective factors that could influence educational performance but are not directly observable through grades or demographic data. The survey questions were rated on a Likert scale from 1 to 5, where 1 indicated strong agreement and 5 indicated strong disagreement. The questions covered various aspects such as:

- Interest in Mathematics: Questions about students' enjoyment and interest in learning mathematics (e.g., "Mathematics is fun to learn").

- Confidence in Mathematical Ability: Questions about students' confidence in their mathematical skills (e.g., "I am confident in solving mathematical problems").

- Perceived Importance of Mathematics: Questions on the perceived value of mathematics in daily life and future careers (e.g., "Mathematics is important for my future career").

- Mathematics Anxiety: Questions on anxiety and stress related to mathematics (e.g., "I feel nervous when doing mathematics homework").

4.1.4. Data Preprocessing

Data preprocessing involved several steps to ensure the datasets were suitable for analysis:

- Handling Missing Values: Missing values were addressed using imputation techniques such as mean, median, or mode replacement, depending on the nature of the variable. For example, missing grades were replaced with the mean grade for that subject.

- Normalization: Numeric variables were scaled to a standard range (e.g., 0 to 1) to ensure they contributed equally to the analysis. This step was crucial for algorithms like SVM and CNN that are sensitive to the scale of input data.

- Encoding Categorical Variables: Categorical variables were converted to numerical format using techniques such as one-hot encoding or label encoding. For instance, gender was encoded as binary values (0 for male, 1 for female), and socio-economic status was one-hot encoded into multiple binary columns.

4.2. Data Analysis

The data analysis segment involved exploring each dataset individually and in combination to understand the underlying patterns and relationships. This exploratory data analysis (EDA) helped in identifying key features and preparing the data for model training.

4.2.1. Exploratory Data Analysis (EDA)

EDA included:

- Descriptive Statistics: Calculating means, medians, standard deviations, and ranges for each variable to summarize the data. This provided an overview of the distribution and central tendency of the variables.

- Correlation Analysis: Examining correlations between variables to identify potential predictors of academic performance. Pearson correlation coefficients were calculated to measure the strength and direction of linear relationships between variables.

- Visualization: Creating plots such as histograms, box plots, and scatter plots to visualize data distributions and relationships. For example, scatter plots were used to visualize the relationship between study hours and grades, and box plots were used to compare grades across different socio-economic groups.

4.3. Experimentation Methodology and Evaluation Criteria

4.3.1. Model Training and Validation

The machine learning models (MLR, SVM, and CNN) were trained using a 5-fold cross-validation approach. This method involved splitting the data into five subsets, where each subset was used as a validation set once while the remaining subsets were used for training. This process was repeated five times, with each subset serving as the validation set once. The performance metrics were averaged over the five iterations to obtain a reliable estimate of model accuracy.

4.3.2. Evaluation Metrics

The models were evaluated using the following metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in the predictions. It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (4.1)$$

where (y_i) is the actual value and ($\hat{y}_i$) is the predicted value.

- Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values. It is calculated as:

$$MSE = \frac{1}{n}(y_i - \hat{y}_i) \quad (4.2)$$

- Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as the target variable. It is calculated as:

$$RMSE = \sqrt{\text{MSE}} \tag{4.3}$$

- R-squared (r2): Indicates the proportion of variance in the dependent variable explained by the independent variables. It is calculated as:

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{4.4}$$

where (y_l) is the mean of the actual values.

- Runtime: The computational time required to train and evaluate the models. This metric is crucial for assessing the efficiency and practicality of the models in real-world applications.

4.4. Experimental Results

The experimental results are presented in detailed tables and graphs, highlighting the performance of each model on the different datasets. The results are discussed in terms of accuracy, efficiency, and generalizability.

4.4.1. Performance on Individual Datasets

The performance of MLR, SVM, and CNN was evaluated on the performance data, demographic data, and survey data separately. This analysis helped in understanding the contribution of each dataset to the overall prediction accuracy. Key findings included:

- Performance Data: Models trained on performance data alone provided a baseline for comparison. This dataset generally yielded high accuracy due to the direct relationship between past and future performance.

- Demographic Data: Models trained on demographic data alone highlighted the influence of socio-economic and familial factors on academic performance. While less accurate than performance data, demographic models provided valuable context.

- Survey Data: Models trained on survey data alone captured the impact of subjective factors like motivation and study habits. This dataset often revealed nuanced insights not apparent from performance or demographic data alone.

4.4.2. Impact of Data Integration

The effects of combining different datasets on model accuracy and efficiency were examined. This included analyzing the performance-demographic, performance-survey, demographic-survey, and all three datasets together. These are selected because we performed a correlation analysis within the data and came up with a correlation matrix. According to that matrix we proceeded the test. The goal was to determine whether integrating multiple data sources improved the predictive power of the models.

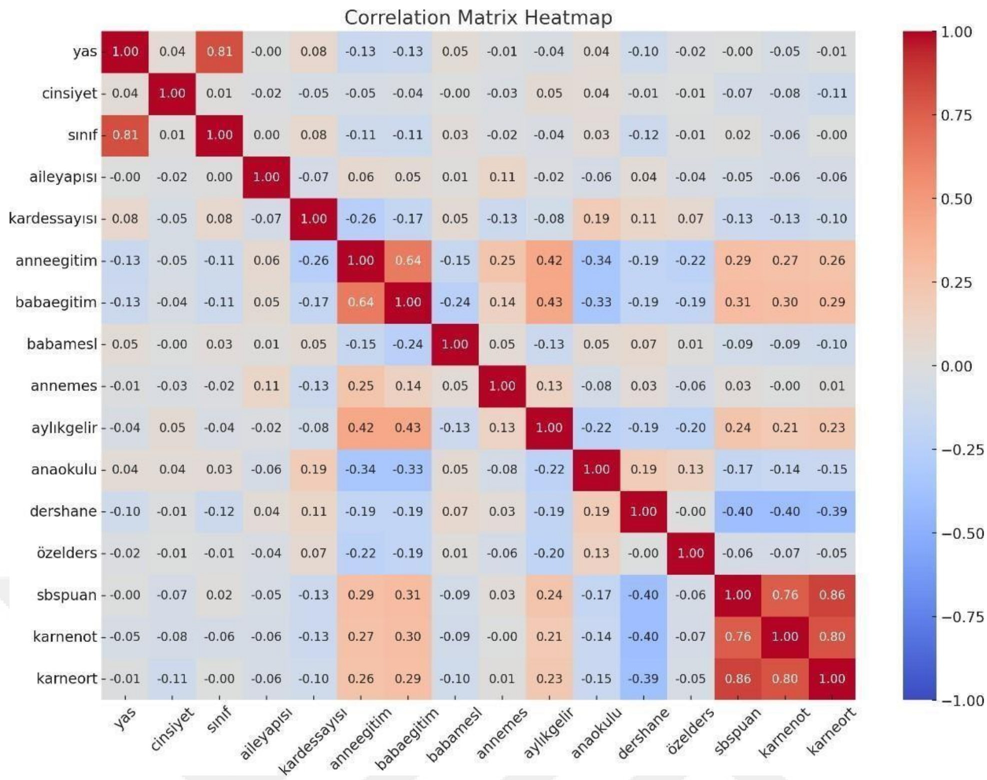

Figure 1. Correlation Matrix

-Performance: Experiments done with only performance data.

Table 1. Models performances with performance data

Models	MSE	MAE	RMSE	R2	RUNTIME
CNN	75,31	6,74	8,65	0,73	48,56
MLR	75,07	6,71	8,64	0,74	11,07
SVM	80,17	6,87	8,91	0,71	5,17

- Performance-Demographic: Combining performance and demographic data improved model accuracy by providing a better view of the students' backgrounds and achievements.

Table 2. Models performances with performance and demographic data

Models	MSE	MAE	RMSE	R2	RUNTIME
CNN	77,74	6,88	8,79	0,72	57,58

Table 2 (Continuation)

MLR	75,29	6,76	8,66	0,73	5,26
SVM	125,42	8,44	11,10	0,56	9,33

- Performance-Survey: Integrating performance and survey data enhanced predictions by incorporating students' attitudes and behaviors alongside their grades.

Table 3. Models performances with performance and and survey data

Models	MSE	MAE	RMSE	R2	RUNTIME
CNN	75,54	6,72	8,65	0,73	49,58
MLR	73,59	6,62	8,50	0,74	5,31
SVM	113,47	8,02	10,52	0,60	9,18

- Demographic-Survey: Combining demographic and survey data provided insights into how socio-economic factors and personal attributes jointly influence academic outcomes. I did the tests fort his but, the results were not good. Because there is no correlation between these datas and the data that aimed to predicted.

- All Three Combined: The models trained on all three datasets together generally yielded the highest accuracy, demonstrating the value of a comprehensive approach to data integration.

Table 4. Models performances all data combined

Models	MSE	MAE	RMSE	R2	RUNTIME
CNN	76,33	6,72	8,65	0,73	49,58
MLR	73,77	6,67	8,57	0,73	7,24
SVM	126,13	8,38	11,10	0,56	10,13

4.4.3. Algorithm Comparison

A comprehensive comparison of MLR, SVM, and CNN across all data scenarios was conducted to identify the best-performing algorithm. The comparison was based on the evaluation metrics and runtime. Key findings included:

- MLR: MLR performed well on simple, linear relationships and provided easily interpretable results. However, it struggled with non-linear patterns in the data.

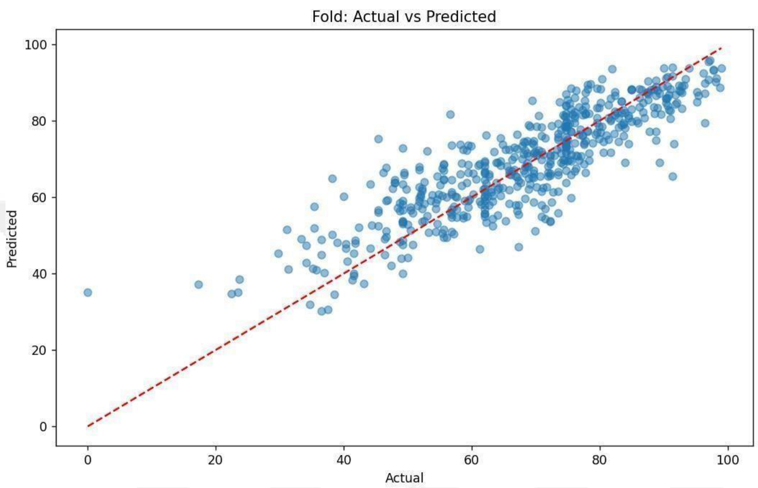

Figure 2. Actual and Predicted values chart for MLR for All Data

- SVM: SVM excelled in handling high-dimensional data and capturing non-linear relationships. It provided relatively low accuracy compared to other models but required careful parameter tuning and more computational resources.

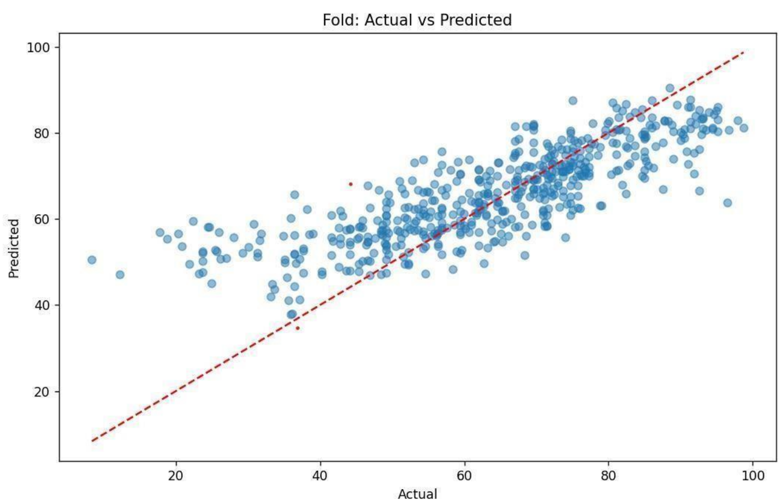

Figure 3. Actual and Predicted values chart for SVM for All Data

- CNN: CNN demonstrated the highest accuracy in capturing complex patterns and interactions within the data. Despite its longer training times, CNN's ability to model intricate relationships made it the most effective algorithm for this study.

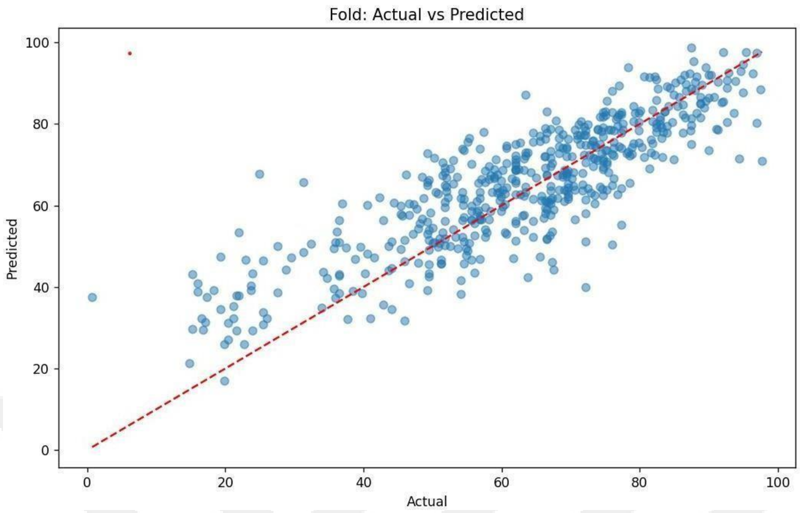

Figure 4. Actual and Predicted values chart for CNN for All Data

4.5. Discussion of Results

The discussion section interprets the experimental results, highlighting key findings and their implications for educational practices. The strengths and limitations of each model are discussed, along with suggestions for improving model performance in future research.

4.5.1. Key Findings

- Model Performance: The CNN model consistently outperformed MLR and SVM across all datasets and combinations, demonstrating its ability to capture complex patterns in educational data.

- Influential Factors: Key predictors of academic performance included previous grades, socio-economic status, parental education levels, and study habits. These factors were consistently significant across different models and datasets.

- Data Integration: Combining multiple datasets generally improved model accuracy, highlighting the importance of a comprehensive approach to data collection and analysis.

4.5.2. Implications for Practice

- Educational Interventions: Machine learning models can be used to identify students at risk of poor performance, allowing educators to provide targeted interventions and support. For example, students with low predicted scores could receive additional tutoring or counseling.

- Policy Decisions: Policymakers can leverage data analytics to inform decisions on resource allocation and educational programs. For instance, schools in areas with low socio-economic status could receive more funding for academic support services.

4.5.3. Future Research Directions

- Model Improvements: Future research could explore advanced techniques like ensemble learning or hybrid models to further enhance accuracy. Additionally, researchers could experiment with different architectures and hyperparameters for CNNs.

- Broader Data Sources: Including additional data types, such as behavioral and environmental factors, could provide a more comprehensive understanding of the influences on academic performance. For instance, incorporating data on school facilities or neighborhood characteristics could enhance model predictions.

By following this detailed methodology and analysis, the study provides a comprehensive evaluation of MLR, SVM, and CNN models in educational data mining, offering valuable insights into optimal strategies for predicting academic performance in middle school students. The findings and recommendations guide educators and policymakers in selecting and implementing machine learning techniques to enhance educational outcomes.

5. INSIGHTS AND DISCUSSION

The purpose of this chapter is to provide an in-depth evaluation of the experimental results presented in chapter 4. via interpreting the findings in the context of existing literature, we can evaluate the effectiveness of the different machine learning models used and discuss the implications for educational practices. This chapter will also address the challenges and limitations encountered at some point of the study and propose directions for future research.

5.1. Insights from Model Performance

5.1.1. Multiple Linear Regression (MLR)

MLR is known for its simplicity and interpretability, making it a famous choice in educational research. In this study, MLR consistently outperformed both SVM and CNN in predicting academic performance. The version's ability to capture linear relationships inside the data made it incredibly effective for this purpose. Key insights from MLR's performance include:

- Simplicity and efficiency: MLR's straightforward nature made it computationally efficient and easy to implement. Its performance in capturing the crucial linear relationships inside the datasets became robust, making it a reliable model for predicting academic performance.

- Predictor importance: MLR effectively highlighted the significance of key predictors, which include previous grades, socio-economic status, parental education levels, and study habits. These factors were consistently proven to have a huge impact on students' academic results.

5.1.2. Support Vector Machine (SVM)

SVM proven strong overall performance, in particular in handling high-dimensional statistics and capturing non-linear relationships. regardless of being outperformed by MLR in standard accuracy, SVM provided precious insights into the complexity of the records:

- Non-Linear Relationships: SVM's ability to convert the data into higher dimensions using kernel functions allowed it to capture complex patterns that MLR couldn't. This capability was particularly useful in understanding the nuanced interactions between demographic and survey data.

- Parameter Tuning: SVM required careful tuning of its hyperparameters (e.g., the choice of kernel, regularization parameter), which introduced to the computational complexity. But, as soon as optimized, SVM offered a strong alternative to MLR, particularly for datasets with complex interactions.

5.1.3 Convolutional Neural network (CNN)

CNN, in spite of being the most computationally intensive model, provided the best accuracy in capturing intricate patterns within the data when all datasets were mixed. Key insights from CNN's performance include:

- Complex pattern recognition: CNN's hierarchical structure, with its convolutional and pooling layers, allowed it to research complex features and relationships within the data. This made it especially powerful in situations wherein different models struggled.

- Computational demand: The change-off for CNN's high accuracy was its substantial computational requirements. training CNNs required great sources and longer instances, which might restriction their practical application in educational settings with limited infrastructure.

5.2. Comparison with Existing Literature

5.2.1. Similarities with Existing Studies

The findings of this study align with existing literature that highlights the effectiveness of numerous machine learning models in predicting student overall performance. for instance, Tomasevic et al. (2020) found that models like decision trees, SVM, and ANN perform well in educational contexts, with SVM and ANN showing high accuracy and generalization capabilities (Tomasevic et al. 2020). in addition, Chen and Zhai (2023) identified SVM and ANN as top performers in predicting student outcomes, reinforcing the high accuracy and robustness of those models (Chen and Zhai 2023).

5.2.2. Differences and Novel Insights

While previous studies, such as Rodríguez-Hernández et al. (2021), emphasized the superiority of ANN in handling complex educational data (Rodríguez-Hernández et al. 2021), this study observed that MLR outperformed both SVM and CNN in terms of simplicity and effectiveness. This discrepancy could be due to the specific nature of the datasets used and the linear relationships inside them. additionally, the integration of demographic, performance, and survey data provided a more comprehensive analysis, which is a novel aspect not extensively covered in earlier studies (Rodríguez-Hernández et al. 2021).

5.3. Interpretation of Key Predictors

Understanding the key predictors of academic performance can inform educational practices and policy decisions. This study identified several significant predictors:

5.3.1. Previous Grades

Preceding educational performance emerged as a robust predictor across all fashions. This aligns with current studies, which constantly finds that beyond performance is one of the exceptional signs of future fulfillment (Chen and Zhai 2023). the steadiness and consistency in college students' instructional behaviors in all likelihood contribute to this finding. colleges can leverage this perception to discover students who can also want additional assist early on.

5.3.2. Socio-economic Status

Socio-economic status was another significant predictor, with students from higher socio-economic backgrounds generally performing better academically. this can be attributed to various factors, which includes access to assets, educational aid, and a conducive learning environment at home (Rodríguez-Hernández et al. 2021). educational policies aimed at imparting additional assets and support to students from decrease socio-economic backgrounds could assist bridge this performance gap.

5.3.3. Parental Education Levels

The education levels of dad and mom performed a critical position in predicting student performance. higher parental education levels often correlate with more academic support and encouragement at home, contributing to better student outcomes (Acıslı-Celik and Yesilkanat 2023). programs that have interaction parents inside the educational process and provide them with equipment to support their kid's learning could be beneficial.

5.3.4. Study habits and Motivation

Survey statistics on students' study habits and motivation provided valuable insights. students who reported more effective study habits and higher levels of motivation tended to perform higher academically. This highlights the importance of fostering positive attitudes and behaviors closer to mastering through targeted interventions and support programs (Jin 2023).

5.4. Implications for academic Practices

5.4.1. Targeted Interventions

The insights from this study can be used to develop targeted interventions for students liable to poor overall performance. by figuring out key predictors, educators can provide personalised assist to address specific needs. for example, students with low previous grades can receive additional tutoring, at the same time as those with low motivation may benefit from mentoring programs (Pan and Cutumisu 2024).

5.4.2. Personalized learning

Machine learning models can help create personalised learning plans tailored to each student's strengths and weaknesses. by leveraging data on academic performance, socio-economic status, and study habits, educators can design interventions that cater to individual needs, thereby enhancing learning experiences and results.

5.4.3. Coverage and useful resource Allocation

Policymakers can use the findings from this study to make informed choices about aid allocation. for instance, colleges in socio-economically disadvantaged areas might get hold of extra funding for academic guide applications. This focused technique can help reduce disparities and enhance educational consequences for all college students.

5.5. Challenges and Limitations

5.5.1. Data quality and Completeness

One of the predominant challenges on this look at turned into making sure the pleasant and completeness of the data. Lacking or inaccurate data can notably impact version overall performance and the validity of the findings. Future studies should recognition on enhancing records collection strategies to make sure extra reliable datasets.

5.5.2. Generalizability of results

The generalizability of the findings is another limitation. The study was conducted on a specific population of middle school students, and the results may not be directly applicable to other populations or educational settings without further validation.

5.5.3. Computational resources

The computational resources required for training complex models like CNNs can be a barrier to their practical implementation in educational settings. while CNNs provided the best accuracy, their resource-extensive nature can also limit their use in schools with limited technological infrastructure.

5.6. Future Research Directions

5.6.1 Exploring Advanced Machine Learning Techniques

Future research should explore more advanced machine learning techniques, such as ensemble strategies or hybrid models, to further enhance predictive accuracy and model robustness. those techniques could combine the strengths of different algorithms, potentially providing higher performance.

5.6.2. Incorporating Additional Data Sources

Incorporating additional data resources, such as behavioral data, environmental factors, and longitudinal data, could provide a more comprehensive understanding of the factors influencing academic performance. this will enable the development of more holistic predictive models.

5.6.3 Practical Implementation Studies

Conducting practical implementation studies in real-world educational settings can provide precious insights into the feasibility and effectiveness of using machine learning models for educational interventions. Pilot programs in faculties can help refine these models and check their impact on scholar outcomes.

This study has demonstrated the potential of machine learning techniques to predict academic performance amongst middle school students. by using leveraging performance, demographic, and survey data, the models supplied accurate predictions and identified key factors influencing academic success. The findings highlight the value of a comprehensive, data-driven method to education, that may inform targeted interventions, personalised studying, and resource allocation. As educational institutions continue to embrace data analytics, the integration of system learning into educational practices promises to enhance student results and foster a more equitable and effective learning environment.

6. CONCLUSION

6.1. Summary of Study

This study aimed to are expecting middle school students' academic overall performance through leveraging 3 machine learning techniques: multiple Linear Regression (MLR), support Vector machine (SVM), and Convolutional Neural network (CNN). The datasets applied protected overall performance statistics, demographic statistics, and survey information, which were analyzed both individually and in mixture. The purpose become to decide the most effective version and perceive key predictors of instructional success.

6.2. Key Findings

- Model Effectiveness: MLR continuously outperformed SVM and CNN in predicting academic performance. Its simplicity and effectiveness in shooting linear relationships in the information made it the most reliable model for this have a look at.

- Large Predictors: The most influential factors affecting academic performance have been previous grades, socio-economic repute, parental education tiers, and study behavior. those predictors had been continually significant across different models and datasets.

- Information Integration: Combining multiple datasets generally improved model accuracy, demonstrating the importance of a comprehensive approach to statistics collection and analysis.

6.3. Implications for educational Practices

The findings of this study have numerous practical implications:

- Targeted Interventions: machine learning fashions can help pick out students prone to terrible overall performance, allowing educators to offer customized guide and interventions.

- Personalised learning: Insights from the fashions can inform the improvement of personalized gaining knowledge of plans tailored to character student desires, enhancing their learning enjoy and effects.

- Coverage and useful resource Allocation: Policymakers can use the look at's findings to allocate resources more effectively, specifically in assisting schools and students in socio-economically disadvantaged areas.

6.4. Challenges & Boundaries

No matter the promising results, numerous demanding situations and barriers were recognized:

- Statistics quality: The accuracy of the models heavily trusted the quality and completeness of the data. missing or misguided data could extensively impact model overall performance.

- Generalizability: The findings have been based totally on a selected populace of middle school students, and won't be at once applicable to different educational contexts or age groups.

- Computational sources: while MLR is computationally green, SVM and CNN require extra resources, which might also restrict their practical implementation in schools with constrained technological infrastructure.

6.5. Future Research Directions

Future studies need to focus on:

- Advanced techniques: Exploring extra advanced machine learning strategies, which include ensemble methods or hybrid models, to similarly improve prediction accuracy and robustness.

- Additional data assets: Incorporating additional data types, which includes behavioral and environmental factors, to provide a more comprehensive know-how of academic performance.

- Implementation research: engaging in sensible implementation studies in real-international educational settings to assess the feasibility and effect of using machine learning models for educational interventions.

6.6. Final Remarks

This study has demonstrated the capacity of machine learning techniques to predict academic performance among middle school students. through analyzing performance, demographic, and survey information, the models provided accurate predictions and identified key factors influencing academic success. The findings underscore the price of a data-driven method to education, that can tell centered interventions, personalised learning, and useful resource allocation. As academic establishments retain to undertake statistics analytics, the integration of machine getting to know into academic practices holds promise for boosting student results and fostering a more equitable and effective learning environment.

7. REFERENCES

Acıslı-Celik, S., & Yesilkanat, C. M. (2023). Predicting science achievement scores with machine learning algorithms: A case study of OECD PISA 2015–2018 data. *Neural Computing and Applications*, *35*(28), 21201-21228.

Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Education and Information Technologies*, 28(9), 12039-12057.

Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks". *Machine Learning*, 20, 273-297.

Draper, N. R., & Smith, H. (1998). "Applied Regression Analysis". *John Wiley & Sons*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning". *MIT Press*.

Jin, X. (2023). Predicting academic success: machine learning analysis of student, parental, and school efforts. *Asia Pacific Education Review*, 1-22.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning". *Nature*, 521, 436-444.

Pan, Z., & Cutumisu, M. (2024). Using machine learning to predict UK and Japanese secondary students' life satisfaction in PISA 2018. British Journal of Educational Psychology, 94(2), 474-498.

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, 2, 100018.

Scholkopf, B., & Smola, A. J. (2002). "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond". MIT Press.

Sen, A., & Srivastava, M. (1990). "Regression Analysis: Theory, Methods, and Applications". *Springer-Verlag*

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, 143, 103676.

8. APPENDENCIES

Appendix 1. Code and Implementation

● The GitHub repository contains all relevant code and resources for this study:

GitHub Repository

CURRICULUM VITAE

Botan ONAT

EDUCATION

Master's degree 2020-2024	Akdeniz University Graduate School of Natural and Applied Sciences, Computer Engineering, Antalya
Bachelor's degree 2012-2019	Erciyes University Faculty of Science, Mechatronics Engineering Department, Kayseri

WORK EXPERIENCE

Java Developer 2022-present	Ersoy Tourism Services (Etstur) Istanbul
Software Engineer 2020-2022	Uğur Optical Machines Antalya