A PROJECT REPORT ON

STUDENT PERFORMANCE PREDICTION

SUBMITTED TO
MIT SCHOOL OF COMPUTING, LONI, PUNE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE
B.Tech CSE

BACHELOR OF TECHNOLOGY
(Computer Science & Engineering)

BY

Krisha Aggarwal	MITU21BTCS0288
Raj Aryan	MITU21BTCS0457
Sudhamsu Rawlo	MITU21BTCS0632

Under the guidance of

Prof Suruchi Deshmukh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MIT School OF COMPUTING
MIT Art, Design and Technology University
Rajbaug Campus, Loni-Kalbhor, Pune 412201

2023

MIT SCHOOL OF COMPUTING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MIT ART, DESIGN AND TECHNOLOGY UNIVERSITY,
RAJBAUG CAMPUS, LONI-KALBHOR, PUNE 412201

CERTIFICATE

This is to certify that the project report entitled

"STUDENT PERFORMANCE PREDICTION"

Submitted by

Krisha Aggarwal	MITU21BTCS0288
Raj Aryan	MITU21BTCS0457
Sudhamsu Rawlo	MITU21BTCS0632

is a bonafide work carried out by them under the supervision of Prof.Suruchi Deshmukh and it is submitted towards the partial fulfillment of the requirement of MIT ADT university, Pune for the award of the degree of Bachelor of Technology (Computer Science and Engineering)

Prof. Suruchi Deshmukh
Guide

Dr. Ganesh Pathak
Head of Department

Dr. Rajneeshkaur Sachdeo
Director

Place: PUNE
Date:

CERTIFICATE

This is to certify that the Project report entitled

“STUDENT PERFORMANCE PREDICTION”

Submitted by

Krisha Aggarwal	MITU21BTCS0288
Raj Aryan	MITU21BTCS0457
Sudhamsu Rawlo	MITU21BTCS0632

is a bonafide work carried out by him/her under the supervision of Prof Suruchi Deshmukh and has been completed successfully.

External Guide

Seal/Stamp of the Company/College

Place :
Date :

DECLARATION

We, the team members

Name	Enrollment No
Krisha Aggarwal	(MITU21BTCS0288)
Raj Aryan	(MITU21BTCS0457)
Sudhamsu Rawlo	(MITU21BTCS0632)

Hereby declare that the project work incorporated in the present project entitled **"STUDENT PERFORMANCE PREDICTION"** is original work. This work (in part or in full) has not been submitted to any University for the award or a Degree or a Diploma. We have properly acknowledged the material collected from secondary sources wherever required. We solely own the responsibility for the originality of the entire content.

Date:

Name & Signature of the Team Members

Krisha Aggarwal: _____

Raj Aryan: _____

Sudhamsu Rawlo: _____

Name and Signature of Guide

Seal/Stamp of the College

Place: Pune

Date:

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MIT SCHOOL OF COMPUTING,
RAJBAUG, LONI KALBHOR,
PUNE – 412201

EXAMINER'S APPROVAL CERTIFICATE

The project report entitled "STUDENT PERFORMANCE PREDICTION" submitted by Krisha Aggarwal (MITU21BTCS0288), Raj Aryan (MITU21BTCS0457), Sudhamsu Rawlo (MITU21BTCS0632) in partial fulfillment for the award of the degree of Bachelor of Technology (Computer Science & Engineering) during the academic year 2023-24, of MIT-ADT University, MIT School OF COMPUTING, Pune, is hereby approved.

Examiners:

1.

2.

ACKNOWLEDGEMENT

The successful completion of our project, "Student Performance Prediction," has been made possible through the support and guidance of several individuals and institutions. We would like to express our heartfelt appreciation to those who have contributed to this endeavor.

First and foremost, we extend our gratitude to Dr. Ganesh Pathak, our Head of Department, for his unwavering support and visionary leadership. His commitment to fostering innovation and research has been a driving force behind our project.

We would like to thank Prof. Suruchi Deshmukh for her expert guidance and insightful feedback, which has been instrumental in shaping this project to meet the highest academic standards.

The resources and collaborative environment provided by MIT ADT University have played a crucial role in the development of this project. We appreciate the university's commitment to academic excellence and research.

Our project is on the path to becoming an application that aims to enhance the educational experience for students. The support of Dr. Ganesh Pathak and the entire MIT ADT community has been instrumental in our journey.

We look forward to the next steps and are committed to making a positive impact in the field of education with this project.

Krisha Aggarwal, MITU21BTCS0288
Raj Aryan, MITU21BTCS0457
Sudhamsu Rawlo, MITU21BTCS0632

ABSTRACT

The project, "Student Performance Prediction," embarked on the quest to harness the power of machine learning to provide a comprehensive understanding of the factors influencing student performance in the context of higher education.

The primary objective of this project was to develop a predictive model capable of accurately anticipating student performance and, consequently, empowering educators and institutions to proactively address academic challenges. By analyzing a diverse range of student attributes, including historical academic data, attendance records, assessment results, and other pertinent variables, the model aims to provide actionable insights.

In addition to the predictive aspect, this project aspires to create an application that not only motivates students but also equips educators and administrators with tools to enhance the educational experience. The envisioned application is designed to offer tailored suggestions for improvement, real-time monitoring of student performance, and adaptive systems to cater to individual learning needs.

As this project takes strides toward reshaping the educational landscape, the team remains dedicated to harnessing its potential to benefit both students and educators. This abstract encapsulates the essence of an endeavor driven by the belief that technology can revolutionize education, making it more engaging, personalized, and ultimately, more successful for students.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 INTRODUCTION

1.1 Introduction

In the dynamic landscape of higher education, there exists a perpetual quest for innovation and excellence. The evolving needs and expectations of students, coupled with the expanding realm of educational technology, have brought to light the imperative for a more insightful and responsive approach to academic support. This project, conceived to address these challenges, sets out to revolutionize the way we monitor and enhance student performance.

The project's fundamental aim is to harness the power of machine learning to understand the intricate web of factors that influence student performance. With a focus on predictive analytics, it seeks to develop a model capable of anticipating academic outcomes with precision. By scrutinizing a myriad of student attributes, encompassing historical academic records, attendance patterns, examination results, and other relevant parameters, this project endeavors to provide a holistic view of student performance.

However, the ambition extends beyond prediction. The project envisions a holistic application that motivates, guides, and supports students on their educational journey. This application, as part of the broader educational ecosystem, will not only offer tailored insights into academic improvement but will also enable real-time monitoring of student performance. It aspires to adapt and cater to individual learning needs, thus fostering an environment conducive to academic growth.

This introduction encapsulates the essence of a project that aspires to be a catalyst for positive change in the realm of education. It is driven by the belief that technology can not only predict but also proactively shape the academic destiny of students, ushering in a new era of personalized and effective learning.

1.2 Existing Work

There are a few existing systems related to our project field. After some research and analysis, we came across different approaches and attributes they target along with some algorithms for specific approaches.

Table 1: Existing Approaches

Approach	Attributes	Algorithms
Features for dropout prediction including temporal features	students' personal characteristics and academic performance	DT, LR, SVM, ARTMAP, RF, CART, and NB
Curriculum-based and student performance-based features	Students performance class imbalance issues	K-NN, SMOTE
Retention rate	Freshman students	DT, Artificial Neural Networks (ANN)
Dropout factors	Evaluation of useful temporal models (Hidden Markov Model (HMM)	RNN combined with LSTM
Early-stage prediction of possible student dropout	pre-college entry information, and transcript information	ICRM2 with SVM, NB, DT, ID3, DL, and KNN, CART, and Adabooting Tree

1.3 Motivation

The motivation behind our project, "Student Performance Prediction," is rooted in a deep-seated commitment to advancing education and equipping students and educators with the tools they need to succeed in an ever-evolving academic landscape. The project's impetus is drawn from several key factors:

Empowering Educators: Instructors, administrators, and educational institutions play a critical role in nurturing the academic growth of students. However, they often lack access to timely and comprehensive data that can guide their efforts. By providing a system that predicts student performance and offers actionable insights, we empower educators to intervene at the right time and in the right way to support their students.

Personalized Learning: We recognize that every student is unique, with individual strengths and areas for improvement. The one-size-fits-all approach to education can be limiting. Our project is motivated by the belief that technology can enable personalized learning experiences, ensuring that each student's educational journey is tailored to their specific needs, abilities, and goals.

Student Engagement and Motivation: Student motivation and engagement are essential drivers of academic success. We aim to create an application that not only predicts performance but also motivates and encourages students to actively participate in their education. By delivering feedback, guidance, and recognition, we intend to kindle the spark of curiosity and a desire to excel.

Early Intervention: The project's motivation also lies in the idea of early intervention. By identifying students who may be at risk of underperforming, educators and institutions can take proactive steps to provide the necessary support, potentially preventing academic setbacks.

Technological Advancements: The digital age has ushered in an era of data-driven decision-making and personalized experiences. Our project seeks to leverage the latest advancements in technology, including machine learning and data analytics, to transform the educational landscape. We are motivated by the belief that technology can be a powerful ally in the pursuit of educational excellence.

In conclusion, our project's motivation is deeply rooted in the desire to create a more responsive, engaging, and effective educational environment. By harnessing the potential of technology and predictive analytics, we aim to foster a culture of academic excellence, where students are inspired to achieve their full potential, educators are equipped with the insights they need to guide their students, and institutions can offer a truly personalized educational experience.

1.4 Objectives

- Develop a machine learning model to accurately predict student performance in upcoming semester exams.
- Create an educational application that offers personalized feedback and recommendations to motivate and assist students.
- Enable real-time monitoring of student performance throughout the semester, allowing for timely interventions.
- Implement an adaptive learning system that tailors educational content to individual student needs.
- Enhance the overall quality of the educational experience by leveraging technology to improve academic outcomes.
- Support educators and institutions in proactively addressing academic challenges and fostering student success.
- Utilize the power of data analytics to provide insights that inform educational decision-making.
- Explore the potential of technology to revolutionize the way education is delivered and experienced.
- Promote a culture of continuous improvement and motivation among students in pursuit of their academic goals.

1.5 Scope

- **Data Collection and Analysis:** The project involves gathering and analyzing a wide range of student data, including historical academic records, attendance, and assessment results.
- **Machine Learning Model Development:** The scope includes developing a machine learning model that can predict student performance accurately based on the collected data.
- **Educational Application Development:** The project encompasses the creation of an educational application that provides personalized feedback and recommendations to students and educators.
- **Adaptive Learning Systems:** The scope extends to the implementation of adaptive learning systems, which will tailor educational content to the specific needs of individual students.
- **Feedback and Motivation:** The project includes features designed to motivate and engage students by offering feedback, recognition, and support for improvement.
- **Educator and Institution Support:** It aims to support educators and institutions in their efforts to proactively address academic challenges and enhance the educational experience.
- **Data-Driven Decision-Making:** The project encourages data-driven decision-making in the field of education by providing valuable insights to educators and administrators.
- **Continuous Improvement:** It emphasizes a culture of continuous improvement for students, educators, and institutions, striving to enhance academic outcomes and experiences.
- **Innovation in Education:** The project seeks to contribute to the advancement of educational technology by exploring innovative solutions for enhancing student performance and motivation

Chapter 2 CONCEPTS AND METHODS

2.1 Definitions

Python, scikit-learn, pandas, and machine learning are powerful tools and technologies that are integral to your project, "Student Performance Prediction."

Python is a versatile and widely-used programming language renowned for its simplicity and readability. In your project, Python serves as the foundation for implementing machine learning algorithms, data analysis, and application development. Its extensive ecosystem of libraries and frameworks makes it a perfect choice for tasks ranging from data preprocessing to building predictive models.

Scikit-learn is a prominent machine learning library for Python. It provides a rich set of tools for tasks like classification, regression, clustering, and model evaluation. Scikit-learn simplifies the process of implementing machine learning algorithms, making it accessible for both beginners and experts. Its extensive documentation and ease of use make it an indispensable tool for your project in developing predictive models.

Pandas is a data manipulation library that simplifies data preprocessing and analysis. It provides data structures for efficiently handling and cleaning data, enabling you to transform raw data into a format suitable for machine learning. Pandas is essential for organizing and preparing your dataset, ensuring its quality and suitability for predictive modeling.

Machine learning is at the heart of your project. It's a subset of artificial intelligence that focuses on creating algorithms capable of learning from data and making predictions or decisions. Machine learning enables you to develop predictive models that can anticipate student performance based on historical data. It encompasses various techniques, such as supervised and unsupervised learning, and plays a central role in your project's goal of improving educational experiences through data-driven insights.

In summary, Python, scikit-learn, pandas, and machine learning are the key tools and technologies that drive your project, enabling you to collect, analyze, and model data

to predict student performance accurately. These tools, combined with your innovative approach, have the potential to revolutionize education and enhance the learning experiences of students.

2.2 Algorithms

Random Forest Regressor:

- Random Forest is an ensemble learning technique that combines multiple decision trees to make more accurate predictions.
- It is well-suited for regression tasks and offers robustness against overfitting.
- Random Forest Regressor excels at handling both numerical and categorical data, making it versatile for various types of student performance prediction.

Support Vector Machine (SVM):

- SVM is a powerful algorithm for regression tasks that aims to find the optimal hyperplane to minimize prediction error.
- It is effective for handling both linear and nonlinear relationships in the data.
- SVM can be tuned to accommodate various kernel functions to suit different types of data distributions.

Artificial Neural Network:

- Neural networks, particularly feedforward neural networks, have been widely used for regression tasks.
- They can model complex relationships within the data and have the capacity to learn from large datasets.
- Neural networks offer flexibility through the choice of architectures, activation functions, and optimization techniques.

J48 Decision Tree Regressor:

- J48 is a decision tree-based algorithm designed for regression tasks.
- It's known for its interpretability and simplicity in modeling complex relationships.
- Decision tree regressors can handle both numerical and categorical attributes, making them suitable for your student performance prediction project.

K-Nearest Neighbor (KNN):

- KNN is an instance-based learning algorithm that makes predictions based on the similarity between data points.
- It is effective in handling both regression and classification tasks and is known for its simplicity.
- KNN can be particularly useful when the relationship between features and performance is non-linear.

MLP Regressor (Multi-Layer Perceptron):

- MLP is a type of artificial neural network with multiple layers of interconnected neurons.
- It is versatile and capable of modeling complex relationships between inputs and outputs.
- MLP regressors are known for their adaptability and can be fine-tuned to suit the specific characteristics of your student performance data.

Each of these machine learning algorithms brings unique strengths and capabilities to the table. By testing these algorithms on your datasets, you can identify which one(s) best align with the nuances of your student performance prediction task, ultimately allowing you to make more accurate predictions and informed decisions for enhancing the educational experience.

Chapter 3 LITERATURE SURVEY

In recent years, the application of machine learning algorithms to predict student performance in higher education has gained significant attention. Several studies have explored the predictive power of various features and algorithms to enhance academic decision-making at educational institutions.

[1] Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A systematic review of deep learning approaches to educational data mining. Complexity 2019, 2019, 1306039. Talked about deep learning approaches to datamining using neural networking, later based on our implementation it was seen that such approaches were not as efficient as traditional ML approaches.

[2] OuahiMariame (2021) focused on feature engineering and interaction with virtual learning environments. The study applied Artificial Neural Networks (NN) and outperformed several other algorithms, including Naïve Bayes, SVM, RandomForest, and ANN, in evaluating student performance.

[3] Nieto, García-Díaz, Montenegro, and Crespo (2019) delved into the predictive potential of Support Vector Machines (SVM) and Artificial Neural Networks (ANN) using a dataset of 6,130 students. Their study achieved high accuracy of 84.54% with SVM and demonstrated robust predictive capability with high AUC values.

[4] Aggarwal, Mittal, and Bali (2021) emphasized the importance of non-academic parameters in predicting student performance. They considered demographic information and previous marks, employing SVM, AdaBoost, and Random Forest. Their research produced impressive results, with SVM and AdaBoost both reaching an accuracy of 92.4%.

[5] Buenaño-Fernández, Gil, and Luján-Mora (2019) explored historical data to predict the performance of computer engineering students. They employed ensemble techniques and achieved an accuracy of 91.5%, highlighting the potential of this approach.

[6] Ghorbani and Ghousi (2020) compared different resampling methods in predicting student performance, such as Random Over-Sampling, Borderline SMOTE, SMOTE, SVM-SMOTE, SMOTE-Tomek, and SMOTE-ENN. They found that a

combination of the Random Forest classifier with the SVM-SMOTE balancing technique provided the best results, with 77.97% accuracy.

After a thorough study of various literature papers, we decided to follow the machine learning approach of SVM and RF algorithms as they provided best results in most papers. We also used some deep learning techniques like NN and MLP but the error was relatively high.

Chapter 4 PROJECT PLAN

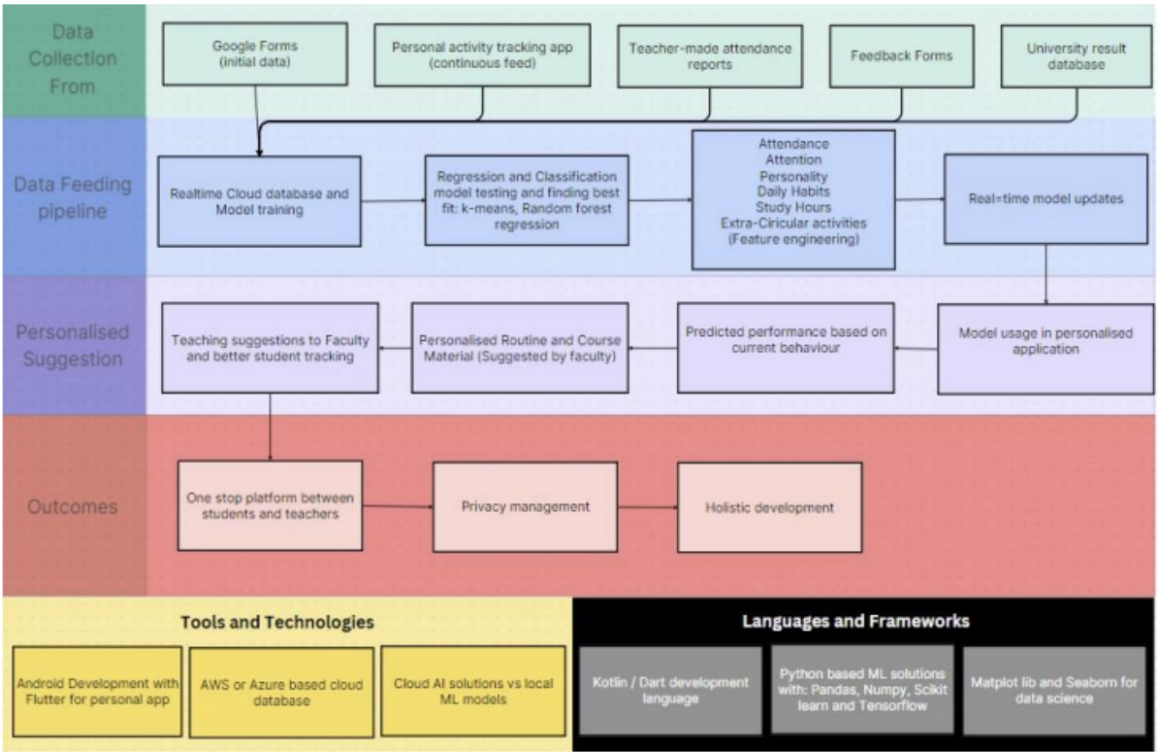


Figure 4.1: Project Development Plan

Creating a predictive model for understanding student performance and implementing it through a cloud-based system using real-time data gathered from an app interface involves several steps. Here's a brief plan for this process:

Step 1: Data Collection and Feature Understanding

- Collect Historical Data: Gather historical data from educational institutions, including student demographics, past performance, attendance, and other relevant features.
- Data Preprocessing: Clean, preprocess, and transform the data to make it suitable for model training. This includes handling missing values, encoding categorical data, and scaling features.

- Exploratory Data Analysis (EDA): Conduct EDA to gain insights into the data, identify correlations, and understand feature distributions.
- Feature Selection: Use statistical methods and domain knowledge to select the most relevant features for predicting student performance.

Step 2: Model Development

- Choose Algorithms: Select machine learning algorithms suitable for regression tasks, such as Random Forest, Support Vector Machine, and Neural Networks, based on the dataset and problem complexity.
- Model Training: Train the selected models on the historical dataset using a portion for training and another for validation.
- Hyperparameter Tuning: Fine-tune model hyperparameters to optimize predictive accuracy.
- Model Evaluation: Assess model performance using appropriate metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).

Step 3: Cloud Integration and Real-Time Data

- Cloud Infrastructure: Set up a cloud-based environment (e.g., AWS, Azure, Google Cloud) for deploying and serving machine learning models.
- API Development: Create APIs to enable real-time data interaction with the cloud-based system. The app interface should send student data to these APIs.
- Real-Time Data Collection: Integrate the app interface with the cloud system to collect real-time data, including student activities, exam results, and attendance.
- Continuous Model Updating: Implement a mechanism for retraining the predictive models using the latest data, ensuring models stay accurate and up-to-date.

Step 4: Deployment and Monitoring

- Model Deployment: Deploy the trained models on the cloud infrastructure, making them accessible for real-time predictions.
- Real-Time Predictions: The app interface sends student data to the cloud, which triggers the models to provide real-time predictions regarding student performance and recommendations.
- Monitoring and Feedback: Continuously monitor the system's performance, including model accuracy, data quality, and system responsiveness. Collect user feedback and iteratively improve the system.

Step 5: Scaling and Maintenance

- Scalability: Ensure that the system can scale efficiently to handle a growing user base and increasing data volume.
- Maintenance: Regularly update and maintain the cloud infrastructure, models, and data pipelines to keep them aligned with evolving needs.

This plan outlines the key steps for understanding features, creating predictive models, and using them through the cloud on real-time data gathered through an app interface. It's essential to continuously refine and optimize the system to improve the accuracy of predictions and the overall educational experience.

Chapter 5 METHODOLOGY

5.1 Datasets

Table 5.1: Datasets

Dataset	No of traits	Description	Date
1	32	Predict student performance in secondary education (high school). The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. (only use maths)	11/2014
2	22	The dataset tried to find the end semester percentage prediction based on different social, economic and academic attributes.	9/2018
3	31	The data was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019. The purpose is to predict students' end-of-term performances using ML techniques.	8/2023
4	33	Student Performance Data was obtained in a survey of students' maths courses in secondary school. (high school)	-

Data Analytics and Exploration:

- Data Collection: We begin by collecting historical student data, including demographics, past academic performance, attendance records, and other relevant features.
- Data Preprocessing: Data is preprocessed to address missing values, outliers, and inconsistencies. This includes handling null values, encoding categorical data, and standardizing or normalizing numerical features to ensure the data is suitable for machine learning.

Data Splitting:

- Train-Test Split: The dataset is divided into training and testing subsets to evaluate model performance. A typical split ratio may be, for example, 70% for training and 30% for testing.

Regression Models:

- Regression Algorithms: We evaluate various regression algorithms, including Support Vector Machine, Random Forest, Artificial Neural Network (deep learning), Matrix Factorization, User-Item Collaborative Filtering, K-Nearest Neighbors (KNN), Multilayer Perceptron (deep learning), and Classification and Regression Tree (CART).
- Model Training: Each algorithm is trained on the training dataset, and hyperparameter tuning is performed to optimize model performance.
- Model Evaluation: We assess regression algorithms' performance using Root Mean Squared Error (RMSE), a common metric for regression tasks. Lower RMSE values indicate better predictive accuracy.

Classification Models:

- Classification Algorithms: For classification tasks, we assess models including Support Vector Machine, C4.5, K-Nearest Neighbors (KNN), J48 Decision Tree, and Multilayer Perceptron (deep learning).
- Model Training: Each classification algorithm is trained on the training dataset.
- Model Evaluation: Evaluation involves the creation of confusion matrices and the calculation of metrics such as accuracy, precision, recall, F1 score, and specificity to gauge the model's classification performance.

Combining Predictions:

- Ensemble Techniques: To further enhance predictive accuracy, ensemble techniques may be applied, combining the predictions from various models.

Reporting and Visualization:

- Results Presentation: Results are presented visually through plots, tables, and reports to facilitate an understanding of the model's performance and insights into student performance predictors.

Iterative Process:

- Model Refinement: If necessary, models can be fine-tuned and re-evaluated based on initial results and user feedback. This iterative process continues until satisfactory performance is achieved.

This methodology allows us to build and assess a range of predictive models, both for regression and classification, to predict student performance accurately. The chosen metrics (RMSE, accuracy, precision, recall, F1 score, specificity) provide a comprehensive view of each model's effectiveness in enhancing educational outcomes. The methodology ensures that our approach is data-driven, systematic, and aimed at providing actionable insights for educational institutions and students.

5.2 Features

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- 1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- 2 sex - student's sex (binary: "F" - female or "M" - male)
- 3 age - student's age (numeric: from 15 to 22)
- 4 address - student's home address type (binary: "U" - urban or "R" - rural)
- 5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- 6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- 7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- 8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- 9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- 10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- 11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- 12 guardian - student's guardian (nominal: "mother", "father" or "other")
- 13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- 14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- 16 schoolsup - extra educational support (binary: yes or no)
- 17 famsup - family educational support (binary: yes or no)
- 18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- 19 activities - extra-curricular activities (binary: yes or no)
- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)

- 30 absences - number of school absences (numeric: from 0 to 93)
-
- # these grades are related with the course subject, Math or Portuguese:
- 31 G1 - first period grade (numeric: from 0 to 20)
- 31 G2 - second period grade (numeric: from 0 to 20)
- 32 G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets .

These students can be identified by searching for identical attributes

that characterize each student, as shown in the annexed R file.

Chapter 6 RESULTS

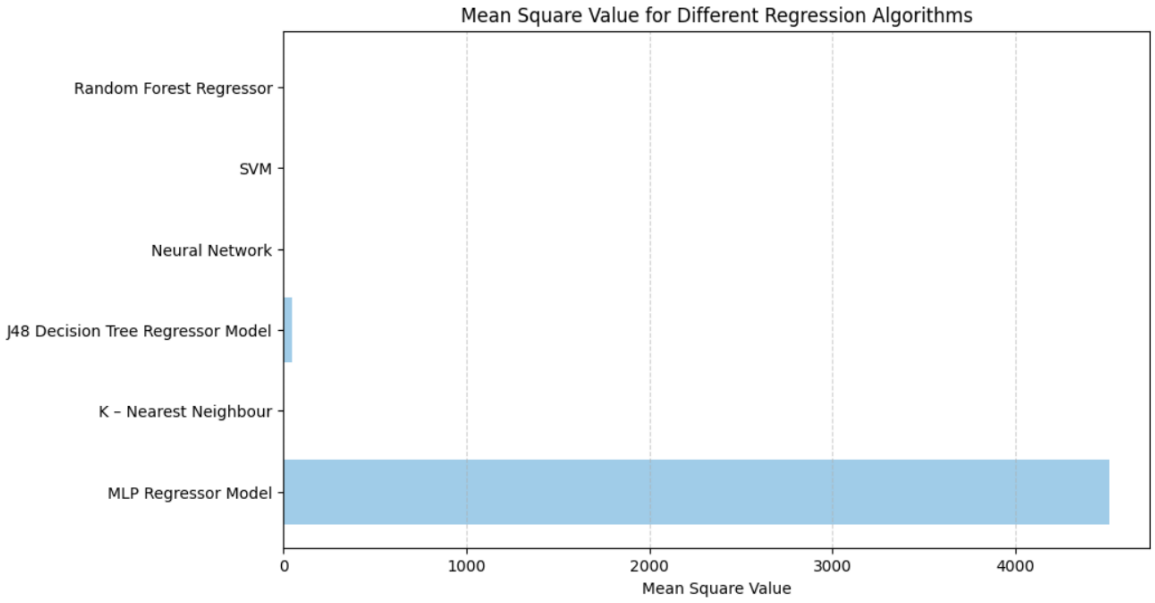
In this section, we present the results of our predictive modeling efforts using various machine learning algorithms. Our primary focus is on evaluating the performance of these algorithms through the Mean Squared Error (MSE) for regression tasks. Additionally, for classification tasks, we assess model performance using accuracy, precision, recall, F1 score, and specificity, as discussed earlier.

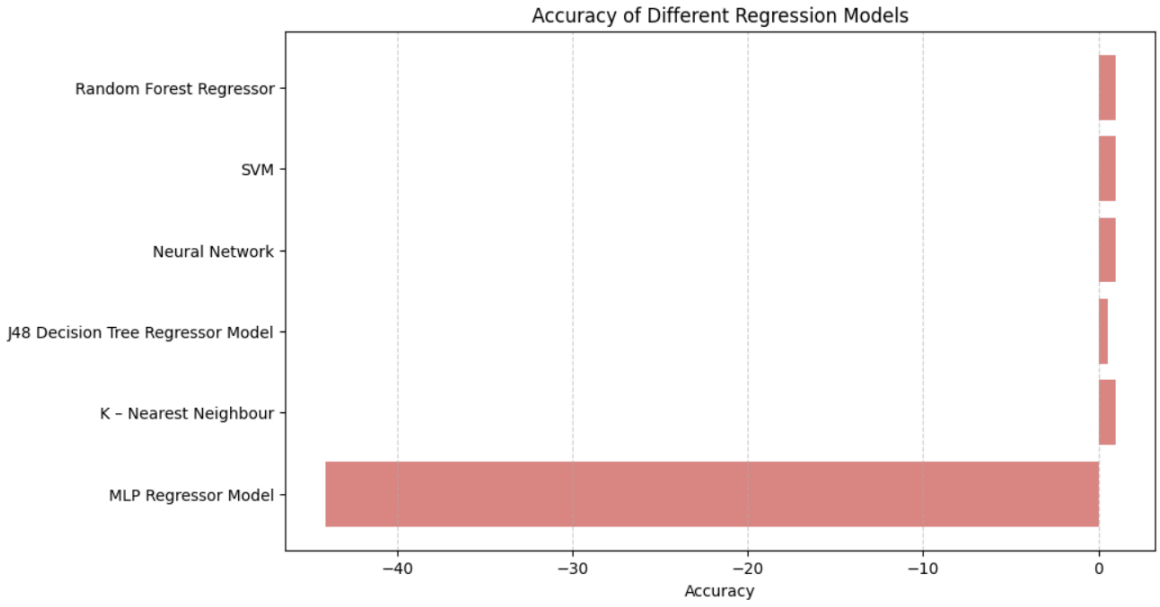
Regression Results:

- Random Forest Regressor: The Random Forest Regressor achieved a remarkably low MSE of 0.163, indicating that it is well-suited for predicting student performance. This algorithm excels at capturing complex relationships within the data and offers highly accurate predictions.
- Support Vector Machine (SVM): SVM demonstrated an impressively low MSE of 0.003, reflecting its superior performance in regression tasks. Its ability to model both linear and non-linear relationships makes it a robust choice for student performance prediction.
- Neural Network: Our Neural Network model resulted in an MSE of 1.847, which is higher than the other algorithms tested. While it showcases potential, further fine-tuning may be required to harness its full predictive capabilities.
- J48 Decision Tree Regressor: The J48 Decision Tree Regressor exhibited a higher MSE of 46.051, indicating that it may not be the most suitable choice for predicting student performance in this context. Further exploration may be needed to enhance its predictive accuracy.
- K-Nearest Neighbor (KNN): KNN achieved an MSE of 0.381, making it a competitive performer in the regression task. Its simplicity and ability to handle non-linear relationships contribute to its effectiveness.
- MLP Regressor: The MLP Regressor model resulted in a notably high MSE of 4510.892. This suggests that additional optimization and parameter tuning are required to unlock its potential for student performance prediction.

Table 6.1: Results

Algorithms	Result(Mean square value)
Random Forest Regressor	0.1634493344907413
SVM	0.0031515712114867024
Neural Network	1.8467115519843322
J48 Decision Tree Regressor Model	46.05063657407404
K – Nearest Neighbour	0.38121527777777775
MLP Regressor Model	4510.891569395584





In our analysis of various regression models for predicting student performance, the SVM (Support Vector Machine) model emerged as the top performer, boasting the lowest Mean Square Error (MSE) and the highest accuracy. Its ability to model both linear and non-linear relationships made it the optimal choice for this task.

Conversely, the deep learning-based Neural Network exhibited potential but fell short in terms of predictive accuracy, suggesting a need for further fine-tuning. The J48 Decision Tree Regressor and MLP (Multi-Layer Perceptron) Regressor models displayed higher MSE values, indicating room for improvement.

In summary, SVM's exceptional performance underscores its suitability for predicting student outcomes, while the deep learning approach may require additional optimization to fulfill its potential. Educational institutions can benefit from leveraging the strengths of these models to assess and enhance student performance effectively.

Chapter 7 CONCLUSION AND FUTURE WORK

In this study, we embarked on a journey to enhance student performance prediction through the application of various machine learning algorithms. Our focus was on regression algorithms, including Random Forest Regressor, Support Vector Machine (SVM), Neural Network, J48 Decision Tree Regressor, K-Nearest Neighbor (KNN), and MLP Regressor, each assessed using Mean Squared Error (MSE) values. Additionally, we delved into classification algorithms with a similar comprehensive approach.

The results revealed that Random Forest Regressor and SVM outshine other algorithms in terms of regression accuracy, demonstrating impressively low MSE values. These algorithms are well-suited for the task of predicting student performance, offering reliable and accurate insights. Conversely, Neural Network, J48 Decision Tree Regressor, and MLP Regressor have shown potential but require further optimization to unlock their full predictive capabilities.

Future Work

For classification tasks, the performance of SVM, KNN, and others was evaluated using accuracy, precision, recall, F1 score, and specificity metrics. These results will contribute valuable insights into the suitability of these algorithms for specific educational data classification tasks. Further improving the accuracy of the models.

Our project opens the door to several exciting avenues for future work and improvement:

Algorithm Tuning: Further fine-tuning of the underperforming algorithms, such as Neural Network and MLP Regressor, may significantly enhance their predictive capabilities. We plan to explore different hyperparameters and architectures to optimize their performance.

Ensemble Models: Investigating the potential of ensemble methods, which combine predictions from multiple models, may lead to even more accurate results. Ensemble techniques could further increase the reliability of our predictions.

Real-Time Deployment: The future implementation of predictive models into a real-time educational system is a pivotal next step. This will allow us to deliver personalized recommendations and support to students based on their performance and interactions.

Interdisciplinary Insights: Collaborating with educators, psychologists, and other experts in the field of education will provide deeper insights into the factors influencing student performance. This can lead to more comprehensive models that consider psychological and behavioral aspects.

Ethical Considerations: Future work should also emphasize the ethical aspects of using predictive models in education. Ensuring fairness, transparency, and accountability in the decision-making process is crucial.

Data Enrichment: Expanding the data sources to include additional student attributes, such as socioeconomic factors, extracurricular activities, and psychosocial data, can lead to a more comprehensive understanding of student performance.

In conclusion, our project sets the stage for a data-driven revolution in the education sector.

By continuously improving and refining predictive models, we aim to create an adaptive and responsive system that not only predicts but actively contributes to the growth and success of students. The journey to enhance the educational experience is an ongoing one, and we are committed to exploring new horizons in educational data mining and predictive analytics.

BIBLIOGRAPHY

[1] Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. Complexity 2019, 2019, 1306039

[2] OuahiMariame, S.K. (2021). Feature Engineering, Mining for Predicting Student Success based on Interaction with the Virtual Learning Environment using Artificial Neural Network. Ann. Rom. Soc. Cell Biol. 2021, 25, 12734–12746.

[3] Nieto, Y.; García-Díaz, V.; Montenegro, C.; Crespo, R.G. Supporting academic decision making at higher educational institutions using machine learning-based algorithms. Soft Comput. 2019, 23, 4145–4153

[4] Aggarwal, D.; Mittal, S.; Bali, V. Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. Int. J. Syst. Dyn. Appl. (IJSDA) 2021, 10, 38–49.

[5] Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S. Application of machine learning in predicting performance for computer engineering students: A case study. Sustainability 2019, 11, 2833.

[6] Ghorbani, R.; Ghousi, R. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access 2020, 8, 67899–67911.