STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS

Akshay Kumar Boddu (6371439, Abodd010@fiu.edu)
Lahari Chowdary Narra (6380886, lnarr001@fiu.edu)
Sai Varshini Papineni (6368046, spapi007@fiu.edu)
Kolli Saibabu (6364112, skoll014@fiu.edu)

I. ABSTRACT

In education, student retention is an important concern. Although intervention programs increase retention rates, they need more knowledge about the student's academic performance. While educational standards have increased recently, over decades, the data shows a significant percentage of student attrition. The Machine Learning domain Extracts high-level knowledge from raw data through learning and offers intriguing, automated resources that can support the field of education. Performance prediction comes into play here and becomes significant. The application of machine learning to forecast academic literature often discusses performance student dropout rates. Predicting dropout rates in online or virtual learning is a popular area of study in such trials because data is readily available and dropout rates are high. Studies often employ dropout or performance predictions in circumstances unrelated to virtual learning.

The goal is to find the most effective predictive system. In other words, the goal is to figure out whether machine learning is a practical method of forecasting student performance or dropout rates. This thesis focuses, specifically on education, and predicting student performance is the goal. Student data is used to build a model that can decide, based on other characteristics, whether the people would succeed. Initially, the input is the training data set.

There are two distinct data sets, each with unique categories of data. The tabular style of these data sets shows that each row corresponds to a student, and every column, or variable, has specific data about a student, including age, gender, history of family, or health information. Also, the variable the algorithm is trying to forecast is a column standing for the student's achievement.

II. INTRODUCTION

Using the datasets present with the internet we are going to use the datasets that focus on the student mainly on how he focuses on his studies daily, how much time he spends on other activities, and how his geographical measures act on his performance.
Each student has his own attribute levels and other features that affect his studies some of them are intellectuals and can attend to their studies in a short period and can focus on them, while some of them focus on long hours but cannot focus.
This addresses the issue here by identifying how much time a person is spending time on their studies, and what factors are influencing that person.

Even though attrition still robs many would-be professionals of their full potential in engineering education. This is where predictive analytics, a branch of machine learning with significant ramifications for forecasting academic paths, comes into play.

Although the use of predictive models to assess student performance and find possible dropouts is not new, it has gained new significance in the digital age due to the prevalence of virtual learning environments,

high dropout rates, and an abundance of data. Using supervised learning algorithms, this research aims to find the most effective prediction systems, standing at the nexus of machine learning and education. Verifying machine learning's efficacy as a predictor of student performance or dropout risk is the main goal. The predictive analytics used in this thesis goes beyond academic research to serve as a useful tool with a wide range of applications, from predicting fraud to identifying contemporary trends sent in education and user behavior analysis. This thesis, which is specifically focused on education, aims to develop a predictive model that forecasts academic outcomes using student data.

II.1. Overview of Machine Learning

Machine learning is a subfield of artificial intelligence (AI), that allows systems to learn from their experiences and enhance their functionality without performing programs explicitly. The aim is to develop computer programs with autonomous learning and data access capabilities. Learning can begin with instructions, personal experience, or data observations. Finding data patterns is this technique's aim to help future scenarios make better decisions. Learning can begin with instructions, personal experience, or data observations. Finding data patterns is this technique's aim to help future scenarios make better decisions. Enabling computers to learn on their own, without direct human aid, and adapting their behavior as needed is the main aim. Numerous machine learning methods exist, and they are often divided into three categories:

Supervised

Unsupervised,

and Semi-supervised.

II.2. Predictive Analytics

Predictive analytics uses a model built from comparable historical data to forecast future events and behaviors in previously unseen data. Predictive analytics is applicable in various domains, including finance, education, healthcare, and law. Its method is standardized. Machine learning algorithms use data that has already been gathered to prove the connections between various data attributes. Based on found patterns, the resulting model can forecast features of upcoming data.

II.3. Regression

In machine learning, regression is a statistical technique used to model and examine the relationships between one or more independent (predictor) variables and a dependent (target) variable. Predicting the value of the required variable from the values of the independent variables is the main goal.

Terminology in Regression:

- The variable being estimated or predicted is known as the dependent variable.
- Independent variables are those that are used to forecast the dependent variable's value.
- The regression that models the relationship between variables as a straight line is known as linear regression.

- A coefficient shows how much of an independent variable is used to predict a dependent variable.
- The difference between the values predicted by the model and the observed values is known as the residuals.
- Overfitting is a modeling error that happens when a function fits a small number of data points too closely.
- When a model is too basic to find the underlying pattern in the data, it is said to be underfitting.
- When independent variables in a regression analysis have a high degree of correlation, this is known as multicollinearity.

Steps in Building a Regression Model:

- **Data Collection:** Gather the relevant data that will be used to train the regression model.
- **Data Preprocessing**: Clean the data by handling missing values, and outliers, and normalizing or standardizing the data if necessary.
- **Exploratory Data Analysis:** Analyze the data to find patterns, trends, and relationships between variables.
- **Feature Selection:** Choose the relevant independent variables most predictive of the dependent variable.
- **Model Selection:** Choose the proper type of regression model based on the nature of the data and the problem (linear, polynomial, etc.).
- **Model Training:** Use the selected features to train the model using a machine learning algorithm.
- **Model Evaluation:** Assess the model's performance using evaluation metrics like MSE, RMSE, MAE, or R-squared.
- **Model Tuning:** Perfect the model by adjusting its parameters to improve performance.
- **Model Validation:** Test the model on a new, unseen dataset to check its generalization ability.
- **Deployment:** Implement the model in a real-world application for predictive analysis

II.4. Regression Algorithms

The commonly used Regression algorithms are:

Linear Regression:

- ***Primary Use*:** Linear Regression is fundamentally a regression model.
- ***Description:*** It is used to predict the value of a dependent variable based on the value(s) of one or more independent variables, assuming a linear relationship between them.
- ***Application:*** Ideal for situations where the relationship between variables will be linear.

Support Vector Machine (SVM):

- ***Primary Use:*** SVM is traditionally known for classification, but it can also be used for regression, known as Support Vector Regression (SVR).
- ***Description:*** In SVR, the algorithm tries to fit the best line within a threshold error margin, focusing on minimizing the error within a certain tolerance level.
- ***Application:*** Useful in scenarios where a margin of tolerance is essential in the prediction.

Advantages:

- efficient with intricate datasets.
- Practical in high-dimensional environments.
- For the decision functions, various kernel functions can be specified, and custom kernels can also be specified.

Challenges:

- SVM sensitivity to noise and outliers in the dataset may affect where the hyperplane is placed.
- The kernel function that is selected affects how well SVM performs. Selecting the right kernel and fine-tuning its parameters can be difficult.

Random Forest:

- ***Primary Use:*** Random Forest can be used for both classification and regression.
- ***Description:*** In regression tasks, it builds multiple decision trees and merges them together to get a more exact and stable prediction.
- ***Application:*** Highly effective in scenarios with complex data structures, offering robustness against overfitting.

Advantages:

- Decreased likelihood of overfitting.
- Supplies adaptability.
- Simple determination of feature importance.
- able to handle high dimensionality and large datasets.

Challenges:

- lengthy procedure
- more resources are needed to store data

III. Existing System

Underfitting problems arise when a poorly constructed prediction model does not sufficiently find patterns in the training dataset. On the other hand, overfitting could happen if the system is overly restructured to fit every detail in the training dataset. Adding a new layer to the neural network, changing some of the parameters, or adding more neurons to the hidden layer can all be used to address overfitting.

One tactic to reduce overfitting is to fine-tune the prediction model until the best configuration is achieved, which ensures balance. For example, it is essential to perfect the number of hidden neurons. A model's ability to be the inherent complexity and diversity of the data may be limited by too few, while overfitting problems may

result from having too many hidden neurons. It's critical to find the point at which the structuring process ends, and further model improvements become insignificant.

Drawbacks:

- It may take some time to load the complete set of data.
- There is imprecision in the process.
- The analysis could try to move slowly.

IV. PROPOSED METHODOLOGY

IV.1. Overview

This research investigates students' performance prediction by using machine learning techniques. Real-world data from college reports and questionnaires will be gathered, including student grades, demographic data, social factors, and college-related features. Although earlier assessments substantially affect students' academic performance, an explanatory analysis reveals the existence of added relevant characteristics.

Using various algorithms, the dataset is processed to create prediction models. The predictions produced by these models will then be contrasted using common evaluation metrics like recall, accuracy, and precision. Notably, the comparison will center on feature selection. This thesis differs in that it examines various approaches in greater detail, including feature engineering and method selection. This innovative method aims to supply insightful information about the best practices for improving student performance predictions.

Advantages:

- prediction of student performance and possible identification of critical factors affecting the success or failure of an educational endeavor.
- enabling students to improve their academic performance.
- contribution to raising the rate of educational promotion.

IV. 2. Description of Modules

High-performance data analysis and data structure tools are available through the open-source Pandas Python library. Pandas are a widely used tool in data science and analytics, and it runs on top of NumPy. NumPy, a low-level data structure that eases multiple-dimensional arrays and a variety of mathematical array operations, is the foundation of Pandas. Robust time series functionality and faster alignment of tabular data are made possible by Pandas' higher-level interface. One of Pandas' most important data structures, the Data Frame, eases the storing and manipulation of tabular data in a two-dimensional structure.

Utilizing Scikit-Learn A Python implementation called Scikit-learn supplies a standard interface for a range of supervised and unsupervised learning algorithms. It is made available for both commercial and scholarly use under a simplified permissive BSD license. Since Scikit-learn is based on the SciPy (Scientific Python) library, installing SciPy is a

prerequisite for using Scikit-learn. This extensive stack includes NumPy for n-dimensional arrays, SciPy for basic scientific computing, Matplotlib for 2D/3D plotting, IPython for an enhanced interactive console, and Sympy for symbolic mathematics.

IV.2.1. Feasibility Study:

The first inquiry is aimed at evaluating the project's viability and estimating the likelihood that the system will be beneficial to the company. Evaluating the technical, operational, and financial viability of adding new modules and debugging current system components is the main goal. All systems would be possible in an ideal world with infinite resources and time. Certain factors are considered in the feasibility study:

IV.2.2Technical Feasibility:

Analyzing the technical complexities associated with the suggested system is necessary to assess its technical feasibility. This focuses on how well the current computer system can accommodate the planned additions. Python and its libraries supply the technical framework for developing Data Analytics applications. Notably, since Python and related libraries are freely available on the internet, no added software needs to be bought.

IV.2.3. Operational Feasibility:

Proposed projects are only practical if they can be converted into information systems that meet user operating requirements. One essential part of project implementation is operational viability. Given that the users are already familiar with the technologies involved, further training is not necessary, making the system operationally possible in this context. The system's intuitive user interface adds even more to its operational viability.

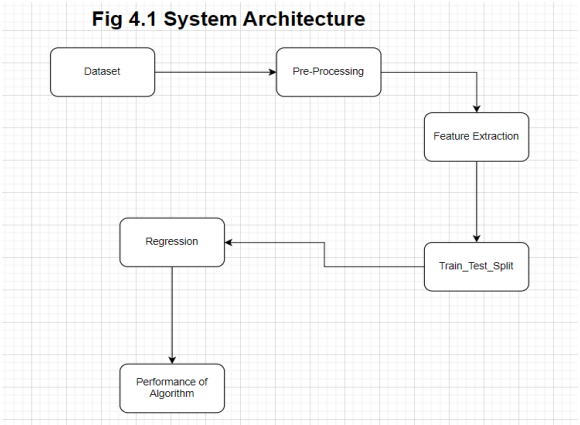
IV.2.4. Economic Feasibility:

Long-term returns, maintenance costs, and cost-benefit analysis are just a few of the variables considered by economic feasibility. Organizations can afford the minimal computing requirements of the suggested computer-based system. The system is considered economically workable without any added economic overheads.

V. SYSTEM DESIGN

Methodology Overview

Each step is crucial in building a reliable and robust machine learning model. The arrows show the typical flow of data and the order of operations in a project.



The steps involved are:

- Data Collection
- Pre-Processing
- Feature Extraction
- Train_Test_split
- Regression

- Measuring Performance of Algorithms

V.2. DATASET

By using performance metric A collection of observations known as the test set will be used to check the model performance. From the test data, there are no observations found in the training data. It will be complicated to make a decision on whether the algorithm has learned to generalize from the training set or has just memorized it if the test set does have examples from the training set. We used Kaggle and other websites to collect the data. And used a website to generate the test data

Here are some of the websites we collected the data from
datasets

Website to generate data

This is how the dataset looks like in the raw format, we combined the two datasets and Pre-Processing on it.

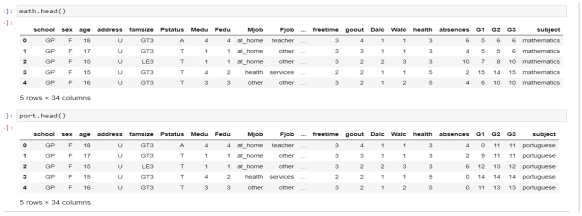


Fig 4.2 Datasets

V.3. PRE-PROCESSING

We performed Data Pre-Processing since the data was already cleaned but we had to make some changes according to our Model and for the feature extraction we had to add some of the added columns, we converted the categorical data into the numerical data using the dummies,

Here are the Pre-Processing steps involved in our project

-> Mapping Categorical values into Numerical values

-> Travel_time, Study_time into 1: '<2h', 2: '2-5h', 3: '5-10h', 4: '>10h'

-> And using the dummies we converted the remaining values to the numerical values

Before the pre-processing

Fig 4.3 data before the pre-processing

After the Pre-Processing

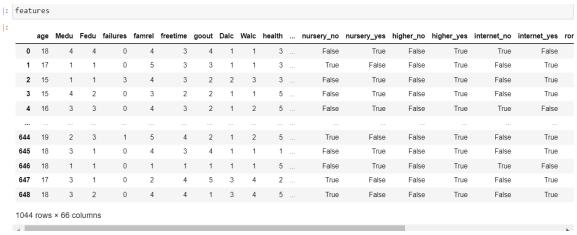


Fig 4.4 Dataset after pre-processing

V.4. FEATURE EXTRACTION

Now we have the data ready with us and we must extract the Features that are needed for the Model, after we extracted the Features, we received are below

Failures, health, absences, G1, G2

V.5. TRAIN_TEST_SPLIT

It will be required to split the data into test and training sets. Models are trained using the training set, and evaluation is done using the test set. We followed the 80-20% Train-Test split method on the dataset.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
```

Fig 4.5 Splitting the datasets

V.6. Regression

In this phase, a regression algorithm is applied to the training data to develop a model that can predict a continuous target variable based on one or more feature variables. Here we will predict the G3, the final grade of the student, based on the other features.

V.7. PERFORMANCE OF ALGORITHM

Metrics like Mean-Squared error, and R-squared error will be used to decide the Performance of the algorithms used in the Project (SVM, Random Forest, Linear Regression).

Mean-squared error: Less mean-squared error is generally considered as a better algorithm

R-Squared error: More the R-squared error better the algorithm.

VI. OUTPUT

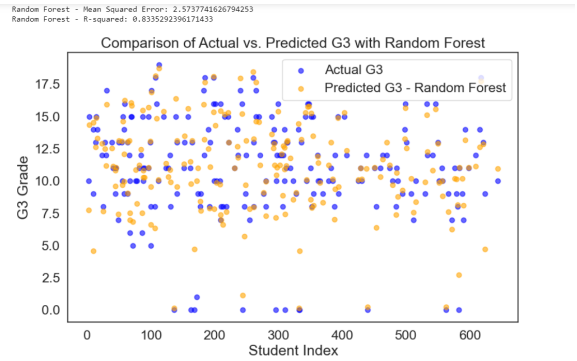


Fig 4.6 Prediction of the G3 Final Grades using Random Forest

VI.1 Comparison

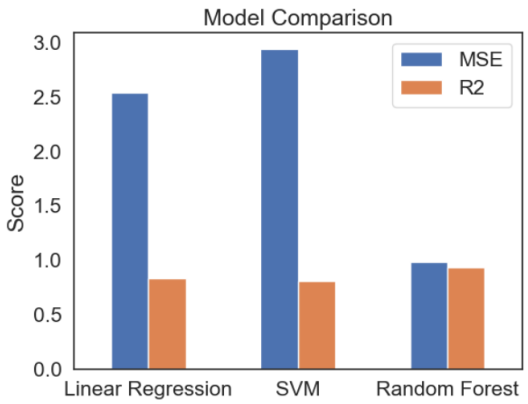


Fig 4.7 Comparision of the ML models with the MSE and R-Squared error.

VI.2 Exploratory Data Analysis

We performed Data Analysis using the data we identified the following trends from it Some of them are :

> Distribution of pass/ fail based on final grades.

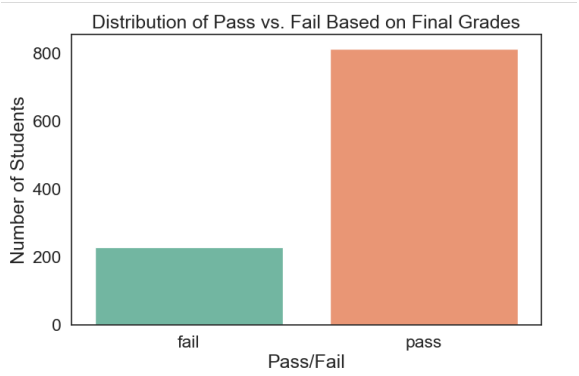


Fig4.8: Distribution of pass/ fail

>Final grades obtained based on the study time.

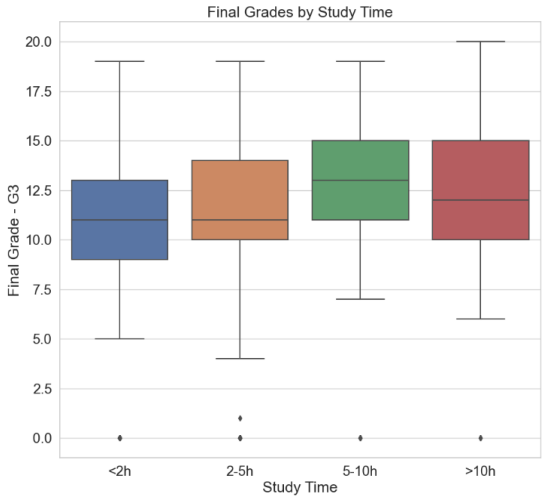


Fig 4.9: Final grades obtained based on study time.

> Comparison of grades b/w the males and females

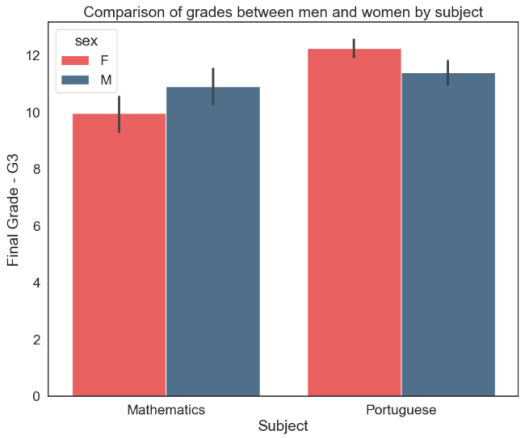


Fig 5.0 Comparision of grades

VI.3. CHALLENGES ENCOUNTERED

To obtain data that is possible for predicting grades, it is crucial to embark on a comprehensive and strategic data acquisition process. This involves finding and gathering relevant information that can serve as valuable inputs for a predictive model aimed at forecasting academic performance.

Performing trend analysis on the available data involves a comprehensive examination of historical information to find patterns, tendencies, or shifts over time. This analytical process aims to uncover insights into the direction and size of changes within the dataset.

VII. FUTURE ENHANCEMENTS

To enhance the learning experience for both students and tutors, it's essential to focus on several key areas that contribute to effective education and can make future enhancements by supplying the details of the

student and other factors that influence student education. The predictive model we developed can supply input of various attributes, including performance in two exams, to forecast a student's final grade. This innovative tool empowers students by allowing them to input their specific attributes, enabling a personalized prediction of their expected grades. By providing this predictive capability, students gain valuable insights into their academic trajectory, allowing them to proactively plan and tailor their study strategies accordingly and manage their studies. Collaborating with educational researchers to tailor tutoring interventions according to predicted users who are at risk of failing is a strategic approach aimed at enhancing academic support and improving overall educational outcomes.

VIII. CONCLUSION

Expected results will play a crucial role in the assessment process for the G3 which will be assessed as a final grade, the attainment of specific marks will be decided by supplying a clear framework to gauge the performance. These expected outcomes serve as benchmarks against which the actual achievements and outcomes of the G3 are measured. The Random Forest algorithm unequivocally stood out as the most effective and robust solution. The evaluation process involved a meticulous examination of multiple models, considering factors such as mean squared error and R-squared error and overall performance on our specific dataset.

The Random Forest model highlighted superior performance compared to its counterparts, underscoring its effectiveness as a powerful predictive tool in the context of educational outcomes. What sets Random Forest apart is nonlinear relationships and capturing complex interactions will be handled by its ability to among features, and mitigate overfitting, all of which are crucial considerations in the nuanced landscape of educational data. The analysis of our data strongly suggests that the Random Forest algorithm outperforms the Linear Regression technique for our specific dataset. This conclusion is drawn based on the evaluation of many factors and metrics, emphasizing the superior suitability of Random Forest in handling the intricacies of our data. In employing a diverse range of machine learning techniques to predict the likelihood of student failures, our approach encompasses the use of various algorithms and methodologies to analyze and interpret complex data sets. The aim is to develop a comprehensive predictive model that can effectively find factors contributing to student failures and supply valuable insights for targeted interventions. In addition to that Using the Exploratory Data analysis we got some noticeable trends from the data we had, which is used for identifying the outliers, patterns, and relationships between the data anomalies. This often exposes the errors present in the data and identifies the missing values and outliers, Overall we got some really good insights from the data. We got similar results for the randomly generated data using the Mockaroo website,
Finally, we can conclude that for the linear data Linear algorithms like Linear

Regression or SVM, these kind of algorithms generate good results, but sometimes depending on the data and the Features it depends.

IX. REFERENCES

[1]http://archive.ics.uci.edu/ml/datasets/Student+Performance
[2]http://www3.dsi.uminho.pt/pcortez/student.pdf
[3]https://www.kaggle.com/code/gzanatta/data-analysis-of-schools-in-portugal-eda
[4] I. H. Witten, E. Frank, M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques". San Mateo, CA, USA: Morgan Kaufmann 2016

[5]https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

[6] https://www.mockaroo.com/