Mining Student Data to Predict Result

Ahmed, Z.
BSc CSE
American International University, Bangladesh
Dhaka, Bangladesh
20-42085-1@student.aiub.edu

Hossain, N.
BSc CSE
American International University, Bangladesh
Dhaka, Bangladesh
20-41982-1@student.aiub.com

Shanto, S. S.
BSc CSE
American International University, Bangladesh
Dhaka, Bangladesh
20-43074-1@student.aiub.edu

Haider, F.
BSc CSE
American International University, Bangladesh
Dhaka, Bangladesh
20-42087-1@student.aiub.com

***Abstract*— This project aims to develop a predictive model to predict the academic performance of mining students using student data. The data will include student profiles, such as age, gender, study time, activities, educational background, and test scores. The model will then be used to predict the academic performance of mining students. Additionally, the model will also be used to identify factors that are most predictive of academic success and to identify potential areas of improvement in the mining program. We analyze a dataset of student records from a high school in the Portugal. To conduct this research, supervised and numeric data were collected from Kaggle having 395 instances, 28 attributes and 1 class attribute. The dataset has no missing values. The final dataset was generated by eliminating outliers and scaling the data. We used python programming for implementation. Data analysis helped to know the optimum algorithm for our chosen dataset by identifying the significant reasons of pass or fail among the students. We used Naïve bayes, KNN, SVM, Kernel SVM, ANN, Logistic Regression, Decision tree and Random Forest classifiers to build different models. The training and test set ratio was 75:25. Among them, Decision tree and Random Forest algorithm performed best with 100% accuracy on test dataset. Lastly, ROC curve and AUC are also compared among applied algorithms. A dataset with large number of instances would be compatible to analyze the performance metrices more accurately.**

Keywords—instances, attributes, Naïve Bayes, Decision Tree, Random Forest, ANN, KNN, Logistic Regression, SVM, Kernel SVM, preprocessing, ROC, AUC.

I. INTRODUCTION

The use of data mining in the field of education has become increasingly popular in recent years due to the availability of large volumes of student data. With this data, it is possible to make predictions about the outcomes of student performance. Data mining can provide insights into student behavior, academic performance, and other factors that can help educators make better decisions about how to support their students. The secondary student achievement statistics from two Portuguese schools is presented here. the early school leaving rate in Portugal for students aged 18 to 24 was 40% in 2006, compared to the average rate for the European Union of barely 15%. Since they give the foundational information for success in the remaining school subjects, failing in the core classes of mathematics and Portuguese (the native language) is very serious. In this study, we'll examine recent actual data from two secondary schools in Portugal. The goal is to forecast student performance and, if possible, to pinpoint the critical factors that influence academic success or failure. We will compare several models and also evaluate them on basic of their accuracy and other performance metrics.

1. Model 1-Naïve Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It's a statistical classifier that performs probabilistic prediction, such as it predicts class membership probabilities. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. This model is easy to implement but there has an assumption of class conditional independence and therefore happens loss of accuracy in the result. Practically, this model's dependencies exist among variables.

2. Model 2-KNN

K Nearest Neighbor (KNN) is a machine learning algorithm used for both regression and classification tasks. It is a supervised learning algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The idea is that similar cases have similar outcomes and dissimilar cases have dissimilar outcomes. KNN is based on feature similarity: choosing the k-nearest neighbors based on their distance to a given data point. The algorithm then predicts the class of the data point based on the classes of the k-nearest neighbors.

3. Model 3-SVM

Supervised learning models called Support Vector Machines (SVMs) are employed in both classification and regression problems. Finding the optimum hyperplane to split a dataset into two classes is the foundation of SVMs. It considers the training data and applies an algorithm to produce an ideal hyperplane that optimizes the margin

between the two classes in order to distinguish between two classes. By using kernels that may raise the dataset's dimensionality, SVMs can also be utilized for non-linear classification tasks.

4. Model 4-Kernel SVM

Kernel SVM is a type of Support Vector Machine (SVM) which uses kernels to map the original data points to a higher-dimensional space in order to make it possible to perform non-linear classification. It is a supervised learning algorithm, which is used for classification and regression problems. Kernel SVM can be used for both linear and non-linear classification. In linear classification, the data points are separated with a straight line and in non-linear classification, the data points are separated with a non-linear curve or a hyperplane. The kernel function is used to map the data points to a higher-dimensional space so that the data points can be separated in a non-linear way. The kernel functions used are the Radial Basis Function (RBF), the Polynomial Kernel, and the Sigmoid Kernel.

5. Model 5-Logistic Regression

Logistic regression is a type of statistical analysis used in machine learning. It is a supervised learning algorithm used to predict a binary outcome based on a set of independent variables. It is used to estimate the probability of an event occurring, and is often used for classification problems. Logistic regression produces a probability score between 0 and 1, and uses a decision threshold to assign a class label (e.g., 0 or 1). It is used in a wide variety of applications, including medical diagnosis, marketing, finance, and risk management.

6. Model 6-ANN

An artificial neural network (ANN) is a computational model inspired by biological neural networks, which are networks of neurons found in the brains of animals. Artificial neural networks are composed of interconnected nodes, which are mathematical functions that take in a set of inputs and produce a set of outputs. The connections between nodes are weighted, and the output of a node is determined by the weighted sum of its inputs. ANNs are used to solve complex problems in fields such as machine learning, image recognition, natural language processing, and robotics.

7. Model 7-Decision Tree

Decision trees are machine learning algorithms that use a tree-like model of decisions and their possible consequences. They are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Decision trees use a tree structure to represent a number of possible outcomes of a particular decision. Each branch of the tree represents a possible decision, and each leaf node represents the result of that decision. The algorithm works by considering every possible outcome of a decision and choosing the one with the greatest overall benefit.

8. Model 8-Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to create a more robust and accurate model. It works by randomly selecting a subset of features from the data set, building a decision tree based on those features, and then repeating the process with different sets of features. The resulting collection of decision trees is then used to make predictions, with the average of the results forming the final prediction. Random forests are popular in both classification and regression tasks due to their accuracy and robustness.

II. METHODOLOGY

A. Flowchart of the Proposed Solution

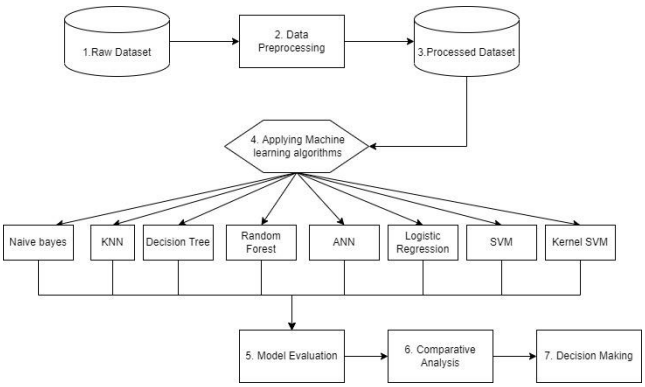

Fig: Flowchart of the solution

After preprocessing the data eight algorithms will be applied. Then, these eight models will be evaluated. Based on the comparative analysis, our final decision will be made.

B. Methods Used

In this research, we have used dataset collected from Kaggle. Initial dataset is a supervised dataset, has a numeric data type. Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or supervise algorithms into classifying data or predicting outcomes accurately. Supervised learning maps an input to an output based on example input-output pairs. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. It infers a function from labeled training data consisting of a set of training examples.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
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).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Table 1: The preprocessed student related variables

Fig: Individual histogram of the attributes

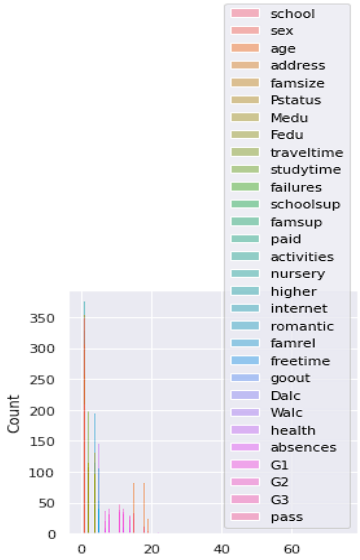

Fig: Histogram of the attributes altogether

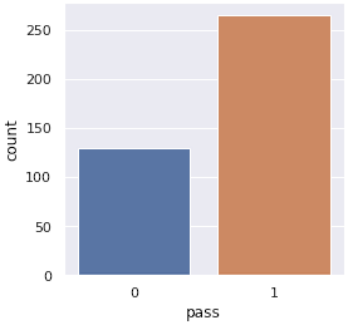

Fig: Bar plot of class attribute

We are seeing the bar plot of the class attribute and histogram of all attributes. From the bar plot we can see the number of failures is much less than number of passes.

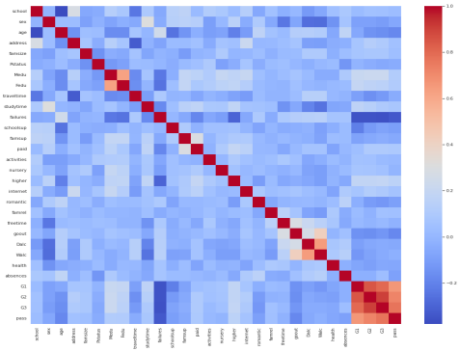

Fig: Correlation Heatmap

In the correlation heatmap the goal attributes G3 (final grade), G2 (second period grade), and G1 exhibit substantial connections (first period grade). Additionally, there is a strong correlation between goout, which refers to going out with friends, and both Dalc (day alcohol intake) and Walc (weekend alcohol consumption). There are also positive correlations between Fedu, or the father, and Medu, or the mother, in terms of schooling.

III. LITERATURE REVIEW

State-of-the-Arts

Data mining is the process of extracting and analyzing data from large data sets to discover patterns and trends. This process has been used in a variety of fields, including business, healthcare, and education. As the availability of data has increased, data mining has become increasingly important in education, particularly in predicting student performance and success. The purpose of this literature review is to summarize the current research on using data mining to predict student performance and success.

Data mining has been used to predict student performance in a variety of contexts. For example, the authors combine two different CNN models with different numbers of convolution layers and pooling layers to produce a single hybrid CNN model in the EDM field. A hybrid 2D CNN model by combining two different 2D CNN models to predict academic performance. Outperformed baseline models, such as K-nearest neighbor, Naïve Bayes, Decision trees, and Logistic regression, in terms of accuracy with an accuracy of 88%. Converted 1D numerical data into a 2D image [1].

Data mining has been used to predict student performance by comparing previous results and other factors. For example, authors compared six classification algorithms which include three basic algorithm decision trees, naive Bayes and PART and three ensemble algorithm bagging, boosting, and random forest. Analyzed various resulting factors (confusion matrix, error rate, rms., etc.) of these classifiers and compare the accuracy level [5].

A comprehensive analysis was conducted of machine learning techniques to predict the final student grades in the

first semester courses by improving the performance of predictive accuracy [2].

A predictive analytics model was proposed using supervised machine learning methods that predict the student's final grade (FG) based on their historical academic performance of studies. J48 was the best predictive analytics model with the highest prediction accuracy rate of 99.6% that could contribute to the early detection of students' dropout so that educators can remain the outstanding achievement in higher education [3].

The authors built a model to improve student performance based on predicted grades and enable instructors to identify such individuals who might need assistance in the courses. Collaborative Filtering (CF), Matrix Factorization (MF), and Restricted Boltzmann Machines (RBM) techniques were used. Among them RBM technique was found to be better than the other techniques used in predicting the students' performance in the particular course [4].

In conclusion, data mining has been shown to be an effective tool for predicting student performance and success, particularly when analyzing student demographic, academic background, and survey data. As the availability of data increases, data mining is likely to become increasingly important in predicting student performance and success.

IV. FINDINGS AND ANALYSIS

A. Training Set – 75% & Testing Set – 25%

The dataset was divided into training and test sets, with 75% of the data utilized for training and 25% used for testing. The optimal model for our dataset was then determined by comparing the test results of various techniques.

```
[13] from sklearn.metrics import confusion_matrix, accuracy_score
     y_pred = classifier.predict(X_test)
     cm = confusion_matrix(y_test, y_pred)
     print(cm)
     accuracy_score(y_test, y_pred)

     [[30  7]
      [ 4 58]]
     0.8888888888888888
```

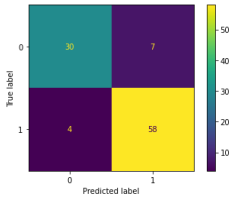

Fig: Prediction Accuracy and Confusion matrix for Naïve bayes classifier

Naïve Bayes classification algorithm was able to correctly predict 88 out of 99 instances which gave it 88% accuracy. Naïve bayes rightly classified 30 instances from class '0' and 58 instances from class '1'

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)

[[23 14]
 [ 2 60]]
0.8383838383838383
```

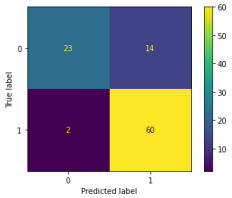

Fig: Prediction Accuracy and Confusion matrix for KNN classifier

K-nearest Neighbor classification algorithm was able to correctly predict 83 out of 99 instances which gave it 83% accuracy. KNN rightly classified 23 instances from class '0' and 60 instances from class '1'.

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)

[[33  4]
 [ 1 61]]
0.9494949494949495
```

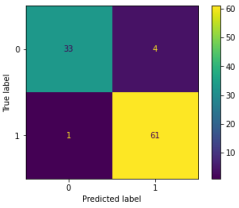

Fig: Prediction Accuracy and Confusion matrix for SVM classifier

SVM classification algorithm was able to correctly predict 94 out of 99 instances which gave it 94% accuracy. SVM rightly classified 33 instances from class '0' and 61 instances from class '1'.

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)

[[29  8]
 [ 3 59]]
0.8888888888888888
```

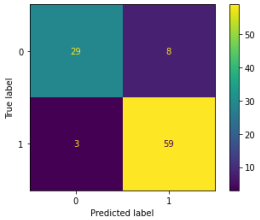

Fig: Prediction Accuracy and Confusion matrix for Kernel SVM classifier

Kernel SVM classification algorithm was able to correctly predict 88 out of 99 instances which gave it 89% accuracy. SVM rightly classified 29 instances from class '0' and 59 instances from class '1'.

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)

[[32  5]
 [ 2 60]]
0.9292929292929293
```

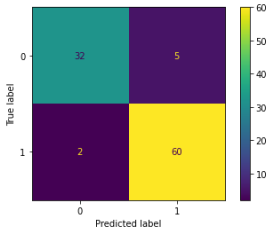

Fig: Prediction Accuracy and Confusion matrix for Logistic Regression classifier

Logistic Regression classification algorithm was able to correctly predict 92 out of 99 instances which gave it 92% accuracy. SVM rightly classified 32 instances from class '0' and 60 instances from class '1'.

```
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)
```

```
[[24  6]
 [ 2 47]]
0.898734177215189
```

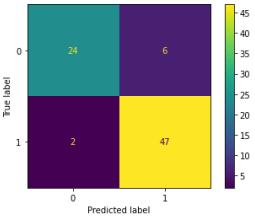

Fig: Prediction Accuracy and Confusion matrix for ANN classifier

ANN classification algorithm was able to correctly predict 89 out of 99 instances which gave it 90% accuracy. SVM rightly classified 24 instances from class '0' and 47 instances from class '1'.

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)
```

```
[[37  0]
 [ 0 62]]
1.0
```

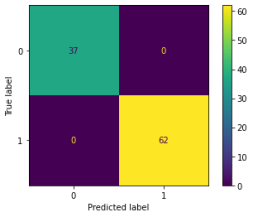

Fig: Prediction Accuracy and Confusion matrix for Decision Tree classifier

Decision Tree algorithm was able to correctly predict 99 out of 99 instances which gave it 100% accuracy.

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)
```

```
[[37  0]
 [ 0 62]]
1.0
```

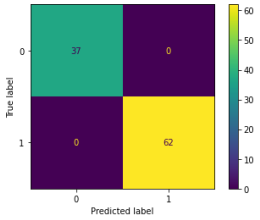

Fig: Prediction Accuracy and Confusion matrix for Random Forest classifier

Random forest algorithm was able to correctly predict 99 out of 99 instances which gave it 100% accuracy.

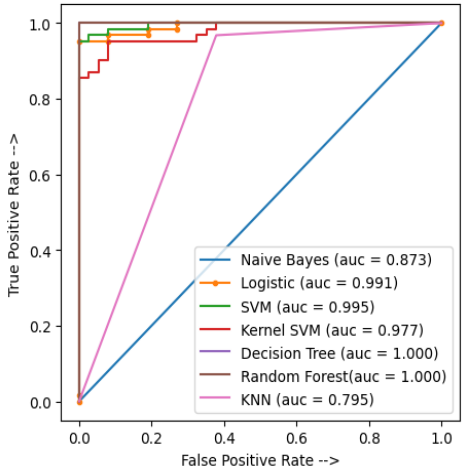

Fig: ROC Curve and AUC of applied classifiers

ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) are two related measures used to evaluate the performance of a binary classification model. ROC is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The AUC represents the degree or measure of separability, which is the ability of the model to distinguish between the two classes. It is calculated by measuring the area under the ROC curve. A model with an AUC of 1 has a perfect discrimination and a model with an AUC of 0.5 has no discrimination ability.

From the graphical representation we can interpret that Decision Tree and Random Forest are the best performing model with the AUC score of 1.

V. CONCLUSION

In this work, we have examined the prediction of secondary student grades in two key subjects (Mathematics and Portuguese) using prior school grades (first and second periods), demographic, sociological, and other educational variables. The obtained findings show that, if the first and/or second school period grades are known, a high prediction accuracy is achievable. However, a review of the information provided by the best predictive models has revealed that, in some circumstances, there are additional relevant features, such as school-related (e.g., absence rates, reasons for choosing a particular school, extracurricular school support), demographic (e.g., student age, parent employment and education), and social (e.g., going out with

friends, alcohol consumption) variables. We can see from the data that the random forest algorithm and decision tree both provided the best accuracy of 100% in test set. Thus, these two have been chosen as our classifiers. If an appropriate dataset becomes available in the future, we can categorize students as good, terrible, or average performers rather than just predicting their results.

REFERENCES

[1] M. J. M.-A. a. J. E. B. Sujan Poudyal *, "Prediction of Student Academic Performance Using a Hybrid 2D CNN Model," *MDPI*, pp. 50-71, 2022.

[2] A. S. R. I. O. K. E. H.-V. F. A. M. G. SITI DIANAH ABDUL BUJANG, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE ACCESS*, pp. 1-14, 2021.

[3] A. S. O. K. Siti Dianah Abdul Bujang, "A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning," *IOP Conference Series: Materials Science and Engineering*, pp. 13-21, 2021.

[4] J. Q. A. N. M. a. F. K. Zafar Iqbal*, "Machine Learning Based Student Grade Prediction: A Case Study," *Springer* , pp. 50-62, 2017.

[5] B. I. S. T. R. H. B. R. A. H. M. S. I. Khaledun Nahar, "Mining educational data to predict students performance," *Springer*, pp. 77-90, 2021.