STUDENT GRADE ANALYSIS & PREDICTION

Summary

In this project, we use a python programming language to apply data mining techniques and machine learning algorithms to predict student performance. We used various regression models to evaluate a feature on the dataset for which we assumed it affected student performance. The model was created using an existing dataset that included a variety of parameters and the final grades. Age, Mother Education, Father Education, study time, going out with friends, and the number of previous failures were the factors examined. The findings demonstrate that family education, previous failures, and others impact students' grades. Predicting student academic performance allows instructors to know how well or poorly their students will perform in their classes, allowing them to take proactive steps to increase student learning.

Keywords: Big data, data mining technique, student performance, Regression Models

1. Introduction

Big data technologies are used to extract valuable and meaningful information from big volumes of wide, variety, veracity, and fast-growing data [1]. One of the uses of big data in education is to predict student performance. It is considered the oldest and most popular data mining application in education. In this study, we develop a model that predicts student performance and tests the impact of factors on student performance. We first downloaded the dataset from UCI [2] and then split the dataset into a training dataset and a test dataset. The dataset uses the training dataset to build the models and the test dataset to validate the models. The main factors selected to test the impact on student performance were Age, mother education, father education, going out with friends, and the number of past failures. We use these factors to build various regression models of the G3 final grade response variable. Table 1 shows a description of the dataset variable model. G3 grades are excluded from the test dataset. Results were generated based on predictive models showing that only learning time and absenteeism could affect student performance. Several tasks, namely classification, and regression are used to build predictive modeling. Focus on regression. There are two main factors in predicting student performance: attributes and prediction methods.

2. Dataset Description

The dataset requirements for this study are met by the UCI repository [2]. The dataset contains student grades at the intermediate level of two Portuguese schools. Data attributes include student grades, social, demographics, and school characteristics. Final grade G3 is the target attribute. There are 33 attributes and 649 instances. Table 1 describes some of the attributes used for measuring students' performance

Table 1. List of attributes that measure students' performance

Features	Type	Description
Age	Integer	student's age (numeric: from 15 to 22)

Fedu	Integer	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - “5th to 9th grade, 3 - “Secondary education or 4 - “Higher education)
Medu	Integer	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - “5th to 9th grade, 3 - “Secondary education or 4 - “Higher education)
Go out	Integer	going out with friends (numeric: from 1 - very low to 5 - very high)
Failures	Integer	number of past class failures (numeric: n if 1<=n<3, else 4)

3. Methodology

This methodology corresponds to various phases of the model. It consists of data collection, data preprocessing, training, and test dataset generation, building a model, and detailed prediction of the stages described below.

3.1 Data Collection

The data was obtained by the UCI repository from Paulo Cortez of the University of Minho in Portugal [2], and it includes student grades, demographic, sociological, and school-related variables. It was collected through school reports and surveys.

3.2 Data Preprocessing

Cleaning, variable manipulation, and other preprocessing methods are used to prepare the dataset for data mining techniques at this stage. The data was required for preprocessing for this model to select the element that we believe impacts student performance. Because we are attempting to build a model and make predictions, we are concentrating our efforts on choosing the response variable. We observed that there are no missing values for this dataset.

```
In [11]: # Check for any missing values for these attributes
         students.isnull().any()

Out[11]: school        False
         sex           False
         age           False
         address       False
         famsize       False
         Pstatus       False
         Medu          False
```

3.3 Data Analysis

The dataset is now ready to use in a prediction model, but it requires more analysis and description of their relationships with student grades to comprehend the grades distribution. In this section, we

attempted to comprehend the relationships between grades and variables and evaluate the factors' correlations to demonstrate their impact on grade prediction. Table 2 describes the correlation of variables with grade 3. And however, before we choose our response variables, we are constructing a hypothesis to test the relationship between grades and the factors listed below:

1. How much % students going to school 1 and school 2.
2. No of Male and Female students
3. Age of Students
4. Do Students from rural areas and urban areas
5. The impact of Previous Failures on student performance
6. The impact of parental education on student achievement
7. The impact of age on a student's performance
8. The impact of parental education on student achievement

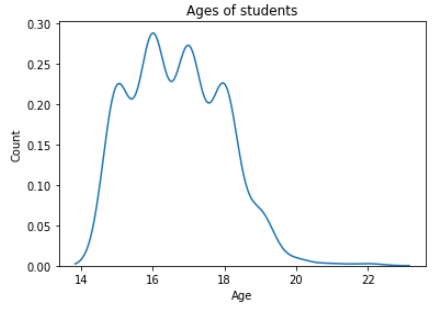
```
In [12]: # Exploratory Data Analysis
         # How much % of students going to school 1 and school 2
         students['school'].value_counts()

Out[12]: GP     349
         MS      46
         Name: school, dtype: int64
```

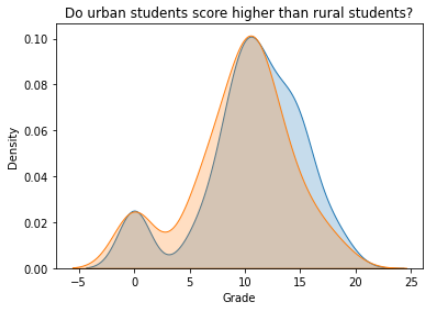
```
In [13]: # Number of Male and female students
         female_students = len(students[students['sex'] == 'F'])
         print(" No of female students",female_students)
         male_students = len(students[students['sex'] == 'M'])
         print(" No of male students",male_students)

          No of female students 208
          No of male students 187
```

```
# Age of students
plot = sns.kdeplot(students['age'])    # Kernel Density Estimations
plot.axes.set_title('Ages of students')
plot.set_xlabel('Age')
plot.set_ylabel('Count')
plt.show()
#Observation:Plot shows the median grades of the three age groups are similar
#Age groups: 15,16,17
```

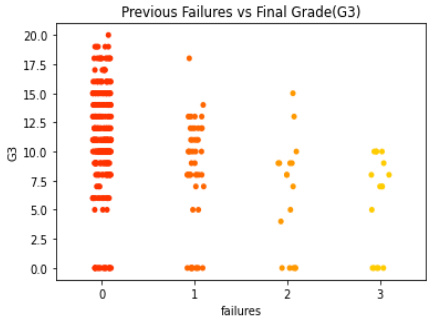



```
# Do urban students perform better than rural students?
# Grade distribution by address
sns.kdeplot(students.loc[students['address'] == 'U', 'G3'], label='Urban', shade = True)
sns.kdeplot(students.loc[students['address'] == 'R', 'G3'], label='Rural', shade = True)
plt.title('Do urban students score higher than rural students?')
plt.xlabel('Grade');
plt.ylabel('Density')
plt.show()
#Observation:Graph clearly shows
#There is not much difference between the grades based on location.
```



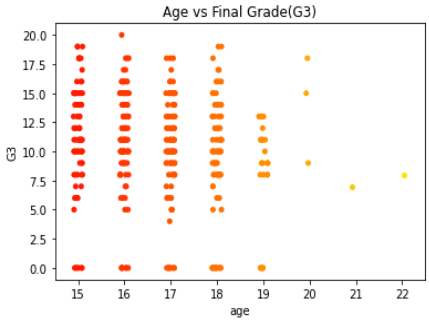
```
# Failures
plot = sns.stripplot(x=students['failures'],y=students['G3'],palette='autumn')
plot.axes.set_title('Previous Failures vs Final Grade(G3)')
# Observation: Student with less previous failures usually score higher
```

```
Text(0.5, 1.0, 'Previous Failures vs Final Grade(G3)')
```



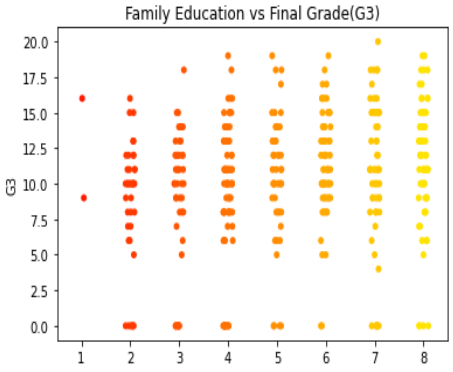

```
# Does age affetcs final grade
plot = sns.stripplot(x=students['age'],y=students['G3'],palette='autumn')
plot.axes.set_title('Age vs Final Grade(G3)')
# Observation:
# Age group 20 seems to score highest grades among all.
```

```
Text(0.5, 1.0, 'Age vs Final Grade(G3)')
```



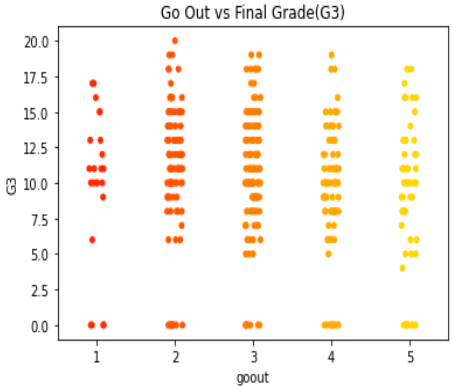

```
# Family Education Attribute i,e Mother Education and Father Education
family_education = students['Fedu'] + students['Medu']
plot = sns.stripplot(x=family_education,y=students['G3'],palette='autumn')
plot.axes.set_title('Family Education vs Final Grade(G3)')
# Observation: Educated Families results in highest grade
```

```
Text(0.5, 1.0, 'Family Education vs Final Grade(G3)')
```




```
# Going out
plot = sns.stripplot(x=students['goout'],y=students['G3'],palette='autumn')
plot.axes.set_title('Go Out vs Final Grade(G3)')
# Observation: Students goes out lott scores less
```

```
Text(0.5, 1.0, 'Go Out vs Final Grade(G3)')
```



Although G1 and G2 are period grades of a student and are highly correlated to the final grade G3, we drop them. It is more challenging to predict G3 without G2 and G1, but such prediction is much more helpful because we want to find other factors that affect the grade. A linear regression model, SVM model, and Random Forest model were built with those variables to add value to the final grade.

Table 2. Correlation of variable with Grade 3

G3	1.000000
failures	0.360415
Medu	0.217147
age	0.161579
Fedu	0.152457
goout	0.132791

3.4 Generating Training and Test Dataset

When building a machine learning model, it is beneficial to train the model on a subset of the dataset. The other phase is to test or validate the model; we split our data so that 80% of it will be used to train the model and 20% will be used to validate it. The training dataset is used to construct models in which the model's variables can be tweaked, and the resulting models are applied to the test dataset to provide an unbiased evaluation of a model built on the training dataset.

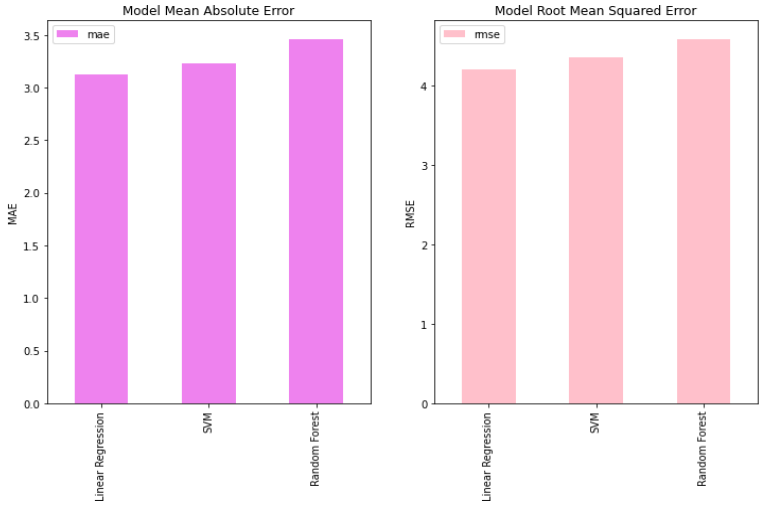
3.5 Model Generation

We evaluated several machine learning models such as the Linear Regression model, SVM model, and Random Forest model by training on the training set and testing on the testing set. And calculated Mean Absolute error and Root mean squared error to choose the best model for predicting. Table 3 represents the Mean Absolute Error and Root Mean Squared Values for Linear Regression, SVM, and Random Forest models. Figure 1 shows the Mean Absolute Error & Model Root Mean Squared Error for three models. As we compare both we can clearly see linear regression is performing the best in both cases. A linear regression model is a mathematical equation used to approximate reality before making predictions based on that approximation. Depending on this dataset, the model's factors that can affect the student's performance are the Failures, going out with friends, and family education.

Table 3. MAE and RMSE Values for various models

	MAE	RMSE
Linear Regression	3.128279	4.208638
Random Forest Model	3.464151	4.592805
SVM	3.226823	4.363263

Figure 1. Mean Absolute error and Root Mean squared error



4. Conclusion

Predicting student performance often helps educators and learners improve their learning and teaching processes. This study tests how factors affect student performance, depending on the existing dataset, which contains many factors influencing final grades. The elements selected were Age, Mother Education, Father Education, Going out with friends, and several past failures. Test their effect using a linear regression model in the final grade G3 of the response variables. As a result, Failures, Family education, and going out with friends affect student performance are determined. Datasets are old and require more detailed information, and can collect more data through surveys at local schools, universities, or universities; more effective and better results are obtained

5. References

[1] Al-Kabi, M. N., & Jirjees, J. M. (2019). Survey of Big Data applications: health, education, business & finance, and security & privacy. JIS&T (Journal of Information Studies & Technology), 2018(2), 12.

[2]. UCI, (2014) Student Performance Data Set
https://archive.ics.uci.edu/ml/datasets/student+performance