P P SAVANI UNIVERSITY

School of Engineering

Minor Project

On

Student Performance - Data mining using Machine Learning Algorithm

Academic Year: 2022-23

Student's Full Name	Dhaduk Jenish Jaysukhbhai Gandhi Shubham Dilipbhai
Enrollment No	20SE02ML010 20SE02ML014
Branch	Machine Learning and Artificial Intelligence
Semester	6th

Supervised by

Mr. Kaushal Singh
P P Savani School of Engineering

PPSU
P P SAVANI UNIVERSITY
School of Engineering

CERTIFICATE

This is to certify that Mr. Dhaduk Jenish Jaysukhbhai,

Enrollment No. 20SE02ML010. from the Department of Machine Learning and Artificial Intelligence , has successfully completed the Minor Project on the Student Performance - Data mining using Machine Learning Algorithm during Academic Year 2022-23.

Date:

Name and Sign of Supervisor

Dean, SOE

PPSU

P P SAVANI UNIVERSITY

School of Engineering

CERTIFICATE

This is to certify that Mr. Gandhi Shubham Dilipbhai,

Enrollment No. 20SE02ML014. from the Department of Machine Learning and Artificial Intelligence , has successfully completed the Minor Project on the Student Performance - Data mining using Machine Learning Algorithm during Academic Year 2022-23.

Date:

Name and Sign of Supervisor

Dean, SOE

ACKNOWLEDGEMENT

We feel elated in manifesting our sense of gratitude to our project guide(s) Mr. Kaushal Singh. He has been a constant source of inspiration for us and we are very deeply thankful to him for his support and valuable advice

We extremely grateful to our Departmental staff members, Lab technicians and Non-teaching staff members for their extreme help throughout our project.

It is indeed with a great pleasure and immense sense of gratitude that we acknowledge the help of these individuals. We are highly indebted to our Dean **Dr. Niraj Shah**, Dean, School of Engineering, P P Savani University, for the facilities provided to accomplish this MINOR PROJECT.

Finally, we express our thanks to all of our friends who helped us in successful completion of this project.

Jenish Dhaduk 20SE02ML010

Shubham Gandhi 20SE02ML014

ABSTRACT

Student performance analysis can be used to identify patterns in the marks obtained by students and to draw useful conclusions from the same.We have used a machine learning model to obtain such correlations and patterns. This dataset used here consists of student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects:

Mathematics (mat) and Portuguese language (por)In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).We have classified the students into three categories, "good", "fair", and "poor", according to their final exam performance. We analyzed a few parameters that have an impact on students' final performance, including Romantic Status, Alcohol Consumption, Parents Education Level, Frequency Of Going Out, Desire Of Higher Education and Living Area. We have created machine learning models to predict students' final performance classification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURE

CHAPTER 1

INTRODUCTION TO PROJECT

MOTIVATION

- Universities today are operating in a very complex and highly competitive environment. The main challenge for modern universities is to deeply analyze their students' performance, to identify their uniqueness and to build a strategy for further development and future actions.
- University management should focus more on the profile of admitted students, getting aware of the different types and specific students' characteristics based on the received data.
- Hence there is a need for an efficient model for student performance analysis.

PROBLEM DOMAIN

To build an efficient model to predict the antecedent grade of the students based on their previous grade and cross verify the same using Chi square test. To analyse student performance based on parameters like Romantic Status, Alcohol Consumption, Parents Education Level, Frequency Of Going Out, Desire Of Higher Education and Living Area. Implementation of various Classification techniques such as Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Logistic Regression Classifier, and perform a comparison study.

AIM & OBJECTIVES

- To classify the students into three categories, "good", "fair", and "poor", according to their final exam performance and cross verify using Chi square test.
- To graphically show the correlation between parameters like Romantic Status, Alcohol Consumption, Parents Education Level, Frequency Of Going Out, Desire Of Higher Education, Living Area and their effects on Student Performance.
- To predict Grade 3 of the students based on their performance in Grade 1 and 2.
- To conduct a comparison study by implementing different classification models and find the best suited model.

CHAPTER 2
LITERATURE REVIEW

DATASET USED

This data displays student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

ATTRIBUTES

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

Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' 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<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
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)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend 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)

DATA PREPROCESSING

The datasets used are processed to check for null values,duplicates and invalid values.

We observe that there are no such irregularities in the dataset,implying that data is already clean and processed. Since both datasets have the same set of attributes and have similar kinds of data ,the both datasets are merged.

We take an additional step to remove all the null or duplicate indices to avoid errors and improve efficiency.

The dataset is now prepared for processing.

CHAPTER 3

SYSTEM DESIGN AND DIAGRAMS

DATA MINING QUESTIONS

1. Prediction of Grade 3 using Grade 1 and Grade 2

2. Do the following parameters affect student performance?

    1. Alcohol Consumption
    2. Romantic Status
    3. Parents Education Level
    4. Frequency Of Going Out
    5. Desire Of Higher Education vi. Living in Urban vs Rural

DATA MINING ALGORITHMS

Decision Tree:

```
msl=[]
for i in range(1,58):

tree = DecisionTreeClassifier(min_samples_leaf=i) t= tree.fit(X_train, y_train)
ts=t.score(X_test, y_test)
msl.append(ts)

msl = pd.Series(msl) msl.where(msl==msl.max()).dropna()
```

Random Forest Classifier:

```
ne=[]
for i in range(1,58):

forest = RandomForestClassifier()

f = forest.fit(X_train, y_train)
fs = f.score(X_test, y_test) ne.append(fs)

ne = pd.Series(ne) ne.where(ne==ne.max()).dropna()


ne=[]
for i in range(1,58):
forest = RandomForestClassifier(n_estimators=36, min_samples_leaf=i) f =
forest.fit(X_train, y_train)
fs = f.score(X_test, y_test)
ne.append(fs)
```

```
ne = pd.Series(ne)
ne.where(ne==ne.max()).dropna()
```

Support Vector Classification:

```
svc = SVC()
s= svc.fit(X_train, y_train)
```

Logistic regression:

```
ks=[]
for i in range(1,58):
        sk = SelectKBest(chi2, k=i)
        x_new = sk.fit_transform(X_train,y_train)
        x_new_test=sk.fit_transform(X_test,y_test) l = lr.fit(x_new, y_train)
        ll = l.score(x_new_test, y_test)
        ks.append(ll)
ks = pd.Series(ks)
ks = ks.reindex(list(range(1,58)))
```

```
#plot
plt.figure(figsize=(10,5))
ks.plot.line()
plt.title('Feature Selection', fontsize=20) plt.xlabel('Number of Feature Used',
fontsize=16) plt.ylabel('Prediction Accuracy', fontsize=16)
```

DATA MINING MODELS

So as per our analysis of data, our choices of model are-:

- **Decision Tree Classifier**- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.**sklearn.tree.DecisionTreeClassifier** is the class used.
- **Random Forest Classifier-** Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction**.sklearn.ensemble.RandomForestClassifier** is the class used.
- **Support Vector Classifier**- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the **future. svm class is used from sklearn.**
- **Logistic Regression Classifier**- Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. **sklearn.linear_model.LogisticRegression** is the class used.

CHAPTER 4

IMPLEMENATION DETAILS

1. Final grade distribution:

The graph given below shows the distribution of students in each grade.

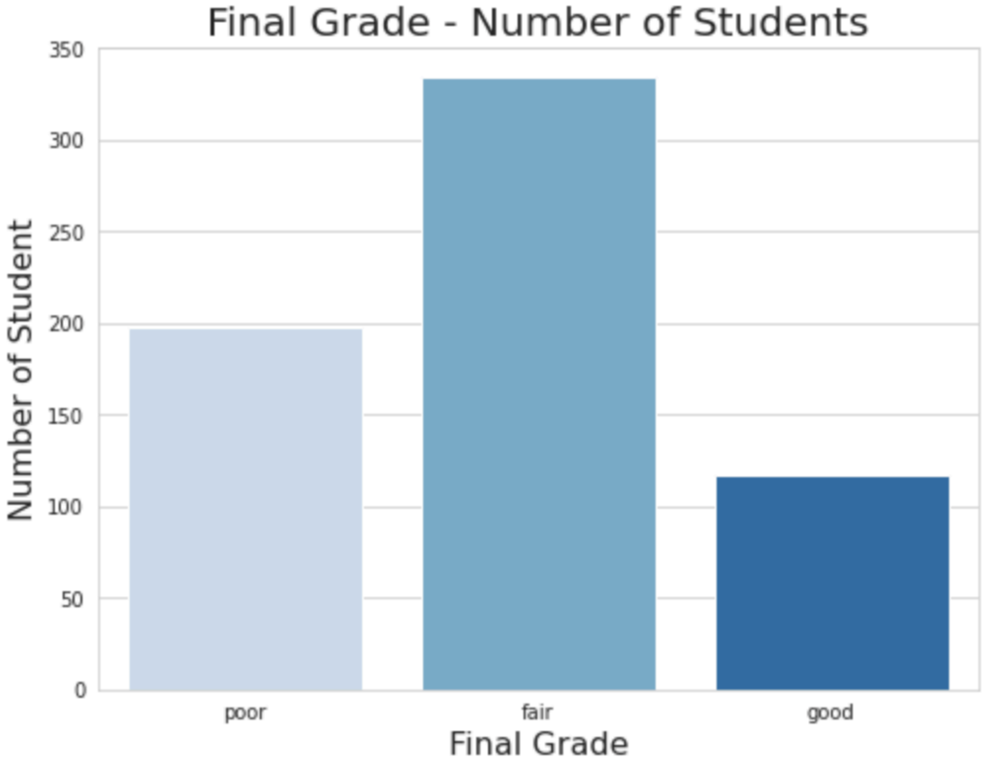

fig. 1 Final Grade

A correlation heatmap shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. The color of the cell is proportional to the number of measurements that match the dimensional value.

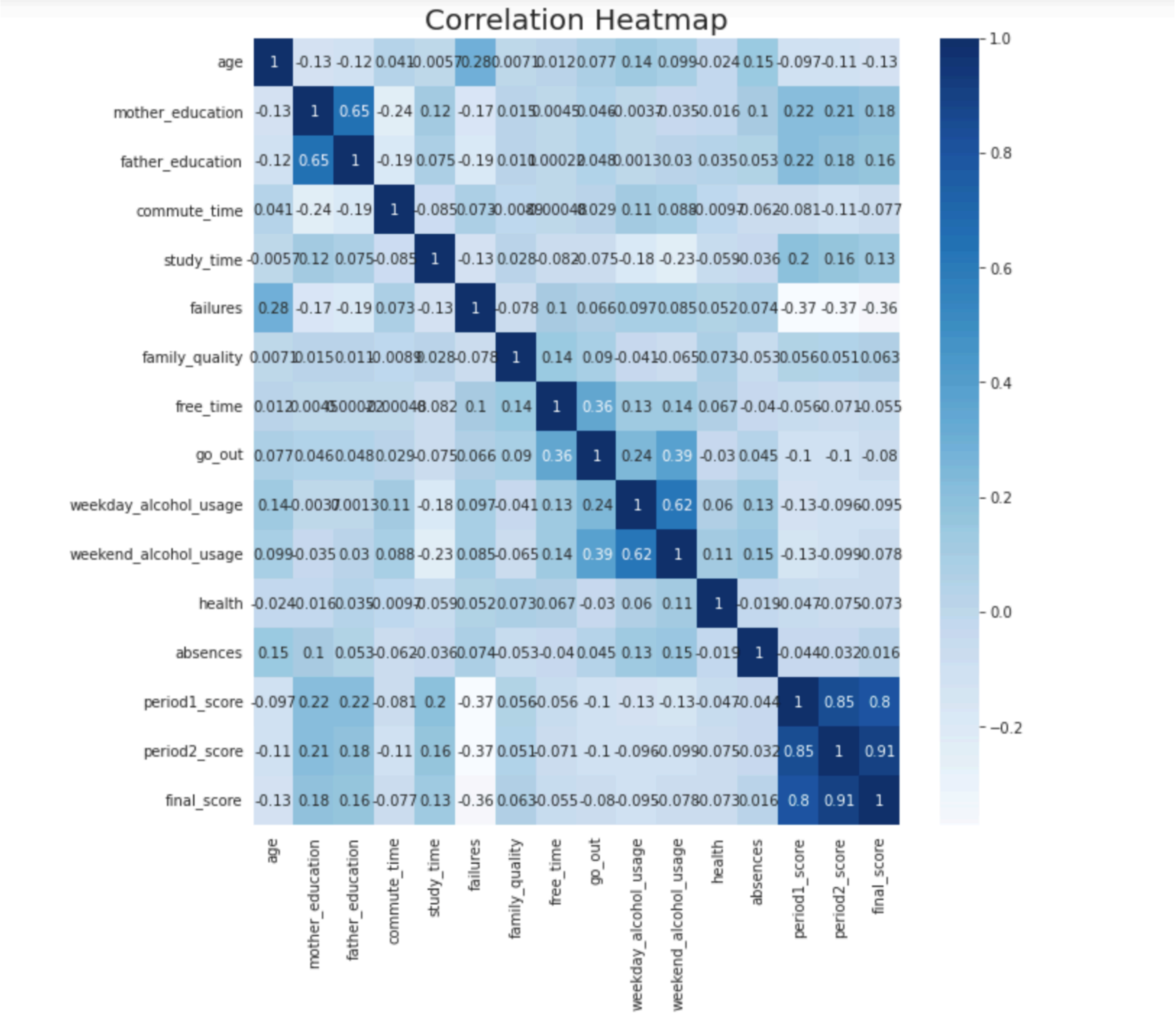

Fig. 2 Correlation Heatmap

2. Parameters affecting Student Performance.

1. Romantic Status

- We infer from the graph below that Romantic Status has a negative impact on the student's performance.

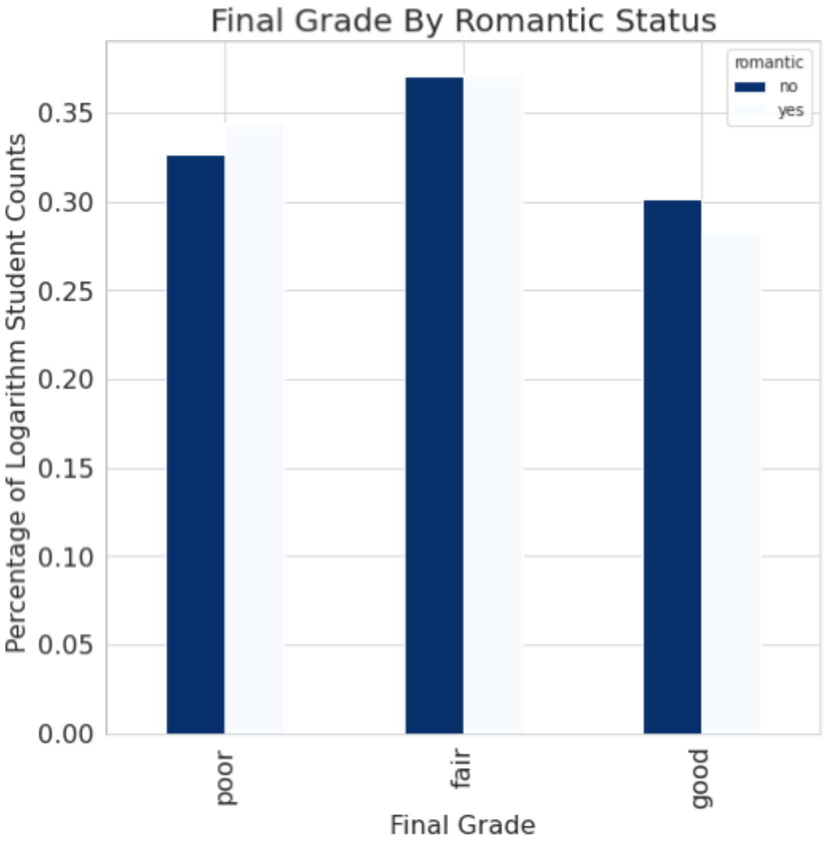

Fig. 3 Final Grade by Romantic Status

2. Alcohol Consumption:

- From the graph we can infer that the maximum number of students consume low levels of alcohol and perform better which is evidently seen below.

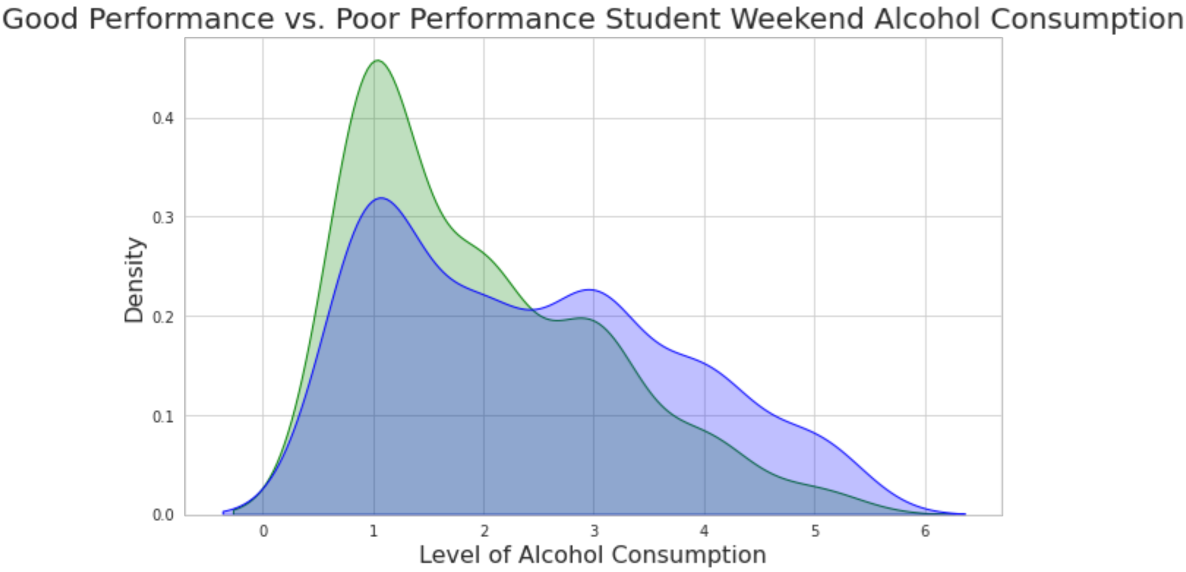

Fig. 4 Good Performance vs. Poor Performance Student Weekend Alcohol Consumption

- The following graph shows that most of the students who consumed high levels of alcohol(Lvl 3,4,5) performed poorly/fairly.

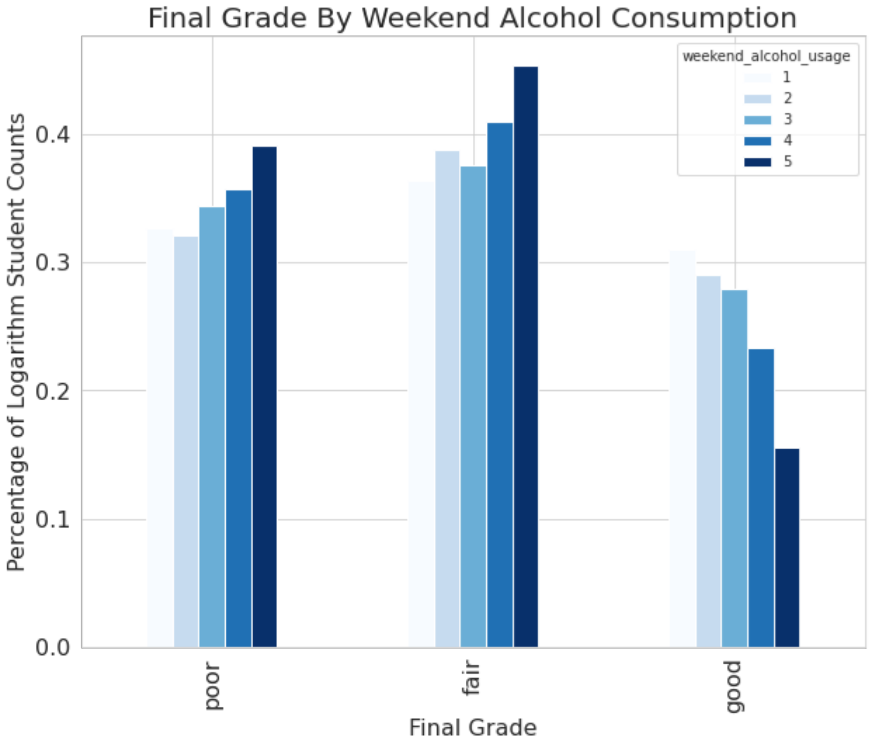

Fig. 5 Final Grade by Weekend Alcohol Consumption

3. Parent's Education level:

- Upon comparison it is found that the mother's education level has a significant impact on the child's performance as compared to the father's education level.

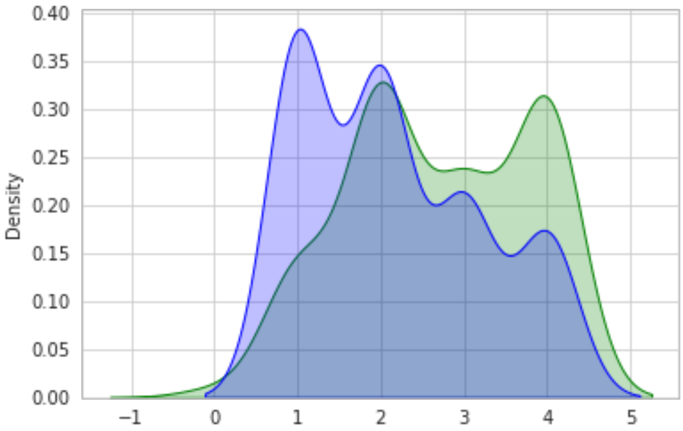

Fig. 6 Father Education Level

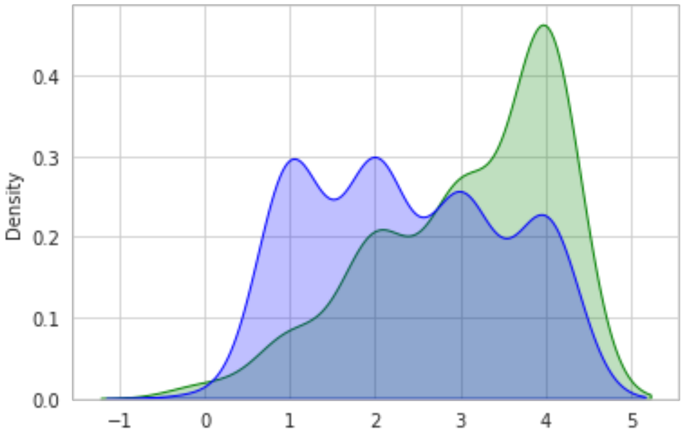

Fig. 7 Mother Education Level

4. Frequency of Going Out:

- It is observed from the bar graph below that those students who scarcely go out get “good” grades.

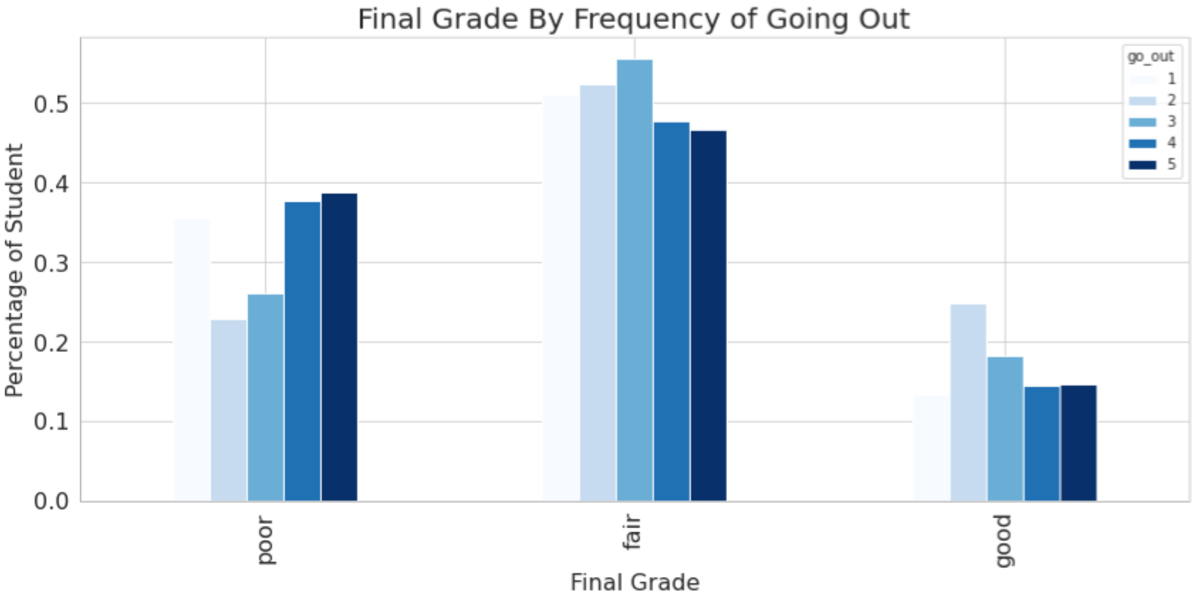

Fig. 8 Final Grade by Frequency of going out

5. Desire to pursue higher education

- From the below graph we can see that students who **DO NOT** have a desire to pursue higher education are more probable to score “poorly”.

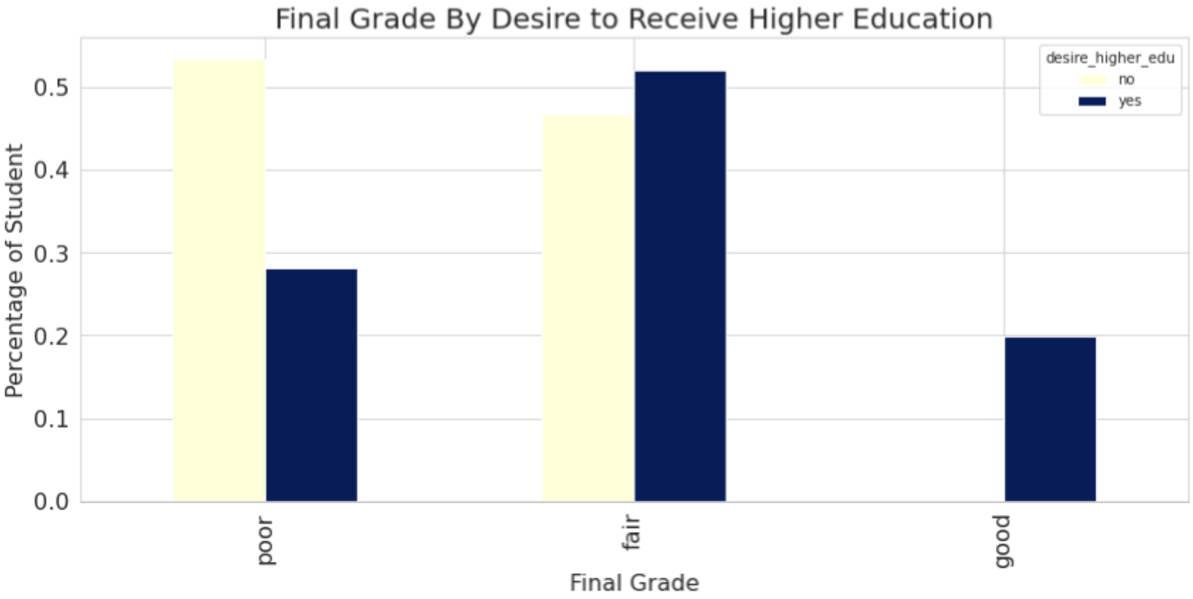

Fig. 9 Final Grade by Desire to Receive Higher Education

6. Living in Urban Vs Rural Area

- The graph below shows that maximum number of students belong to Urban areas.

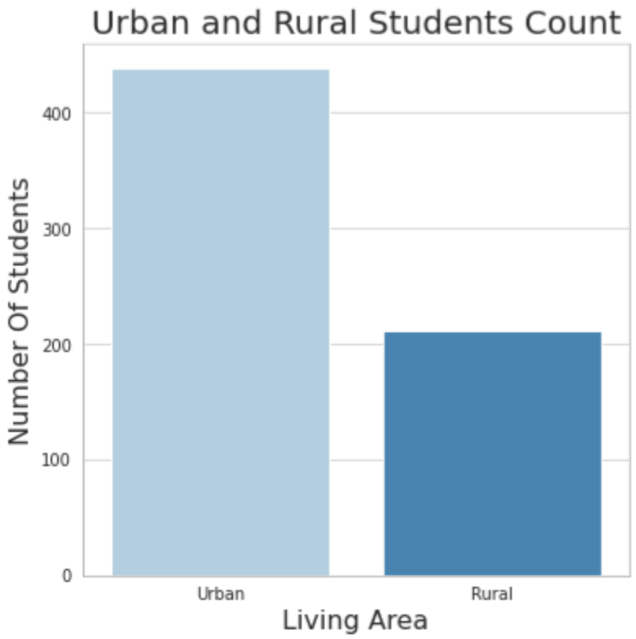

Fig. 10 Urban and Rural Student Count

- The following graph illustrates that students who perform well live in urban areas.

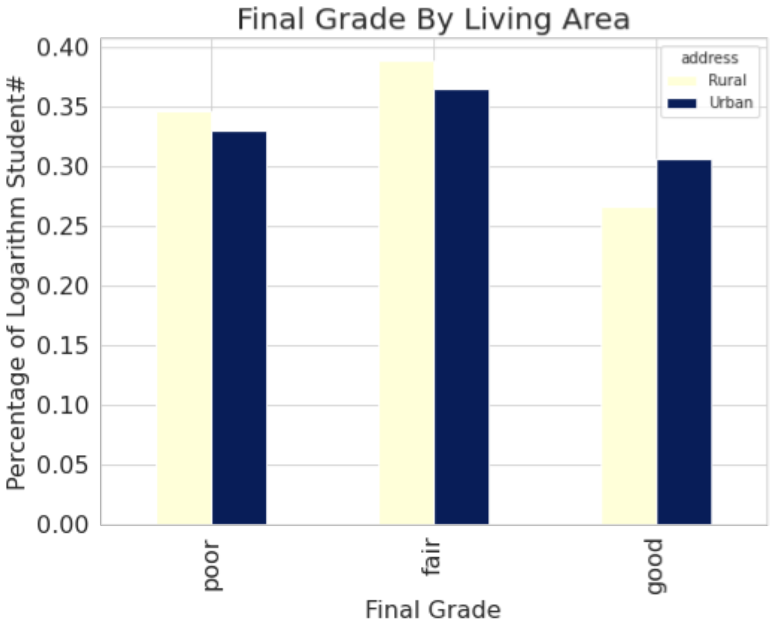

Fig. 11 Final Grade by Living Area

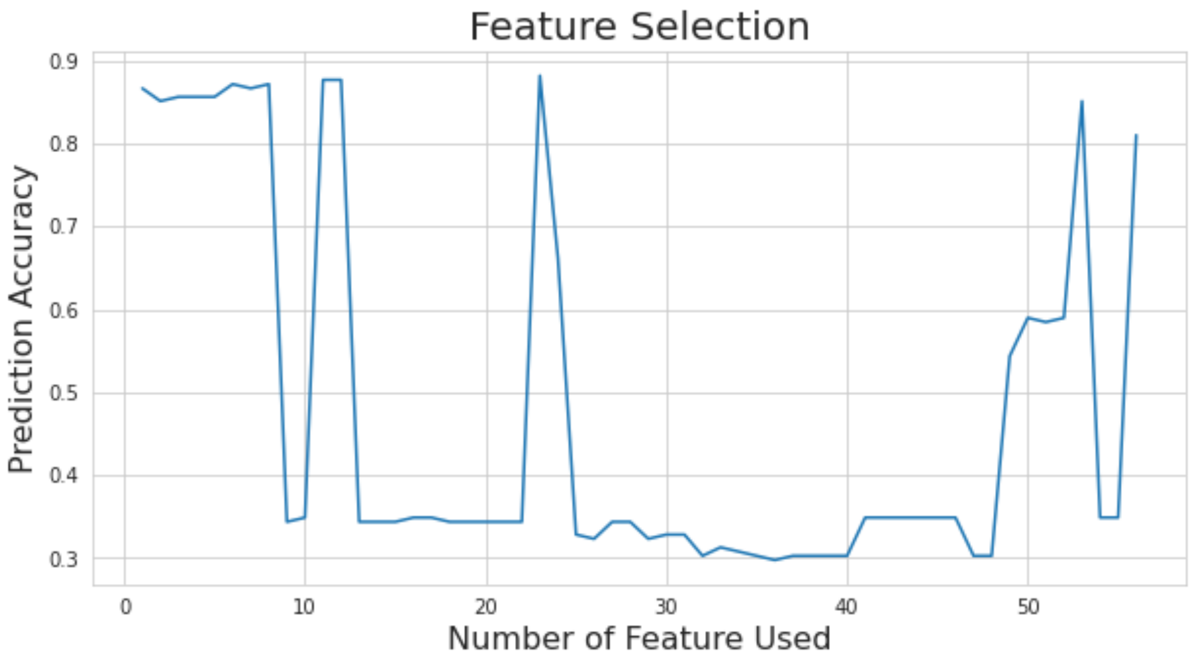

Fig. 12 Feature Selection

MODEL	MODEL SCORE	CROSS VALIDATION
Decision Tree	0.8810572	0.87179487
Random Forest	0.9845814	0.87179487
SVC	0.8656387	0.84102564
Logistic Regrassion	0.8854625	0.86666666

After running the model, the model score and the cross validation score are noted down. Out of these Random Forest Algorithm has the maximum values. Hence, we can conclude that the Random Forest Algorithm works best for this dataset.

CHAPTER 5
CONCLUSION AND FUTURE WORK

The machine learning model analyses the performance of the students in the dataset.The dataset was collected from two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por) in [Cortez and Silva, 2008].

We use the ML model to predict the third grade G3 using previously achieved grades G1 and G2.The data is prepared for the processing.The scientific behaviour of the students based on factors like romantic status,alcohol consumption, parent's education level, frequency of going out, desire to pursue higher education and living area.This paper proposes the application of data mining techniques to predict the final grades of students based on their historical data.Preprocessing operations on the dataset, categorizing the final grade field into five and two groups, increased the percentage of accurate estimates in the classification. The wrapper attribute selection method in all algorithms has led to a noticeable increase in accuracy rate. Overall, better accuracy rates were achieved with the Random Forest method for both mathematics and Portuguese dataset.The proposed method proves its worth from the achieved results and can be used in practice. Through these results, helping educational institutions in terms of staff and students is easy, predicting future data reduces education difficulties and helps to develop future plans for education policy.In the future, update features that are extracted may be needed and their weight is chosen carefully; by updating hidden layers in neural network, the system can be made more reliable

REFERENCES

1. sklearn classes:

https://scikit-learn.org/stable/index.html

2. Research papers:

https://www.hindawi.com/journals/complexity/2021/9958203/

https://pslcdatashop.web.cmu.edu/ResearchGoals

https://f1000research.com/articles/10-1144

3. Books:

Introduction to Data Mining - by Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Published by Pearson India Education Services Pvt Ltd.

APPENDICES

a. Link to the dataset chosen

http://archive.ics.uci.edu/ml/datasets/Student+Performance#

b. Python Codes Implemented

Jupyter Notebook link:

http://localhost:8888/notebooks/Model.ipynb

c. Setup to execute the code (if required)

No setup required since the project was primarily implemented on colab or Jupyter Notebook for easy accessibility.