Nitte Meenakshi Institute of Technology
Department of Computer Science and Engineering 18CS54 Data Mining
Project Report

STUDENT PERFORMANCE ANALYSIS

1NT19CS103, 1NT19CS113, 1NT19CS181, 1NT19CS200
Madhumitha R, Meghana Reddy, Shreya Shettar, Tejashree Krishna Murthy

Table of Contents

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.

1. INTRODUCTION

1.1 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.

1.2 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.

1.3 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.

2. DATA SOURCE AND DATA QUALITY

2.1 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)
`Pstatus`	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)

2.2 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.

3. METHODS & MODELS

3.1 DATA MINING QUESTIONS

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

1. Do the following parameters affect student performance?
    i. Alcohol Consumption
    ii. Romantic Status
    iii. Parents Education Level
    iv. Frequency Of Going Out
    v. Desire Of Higher Education
    vi. Living in Urban vs Rural

3.2 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)
```

3.3 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.

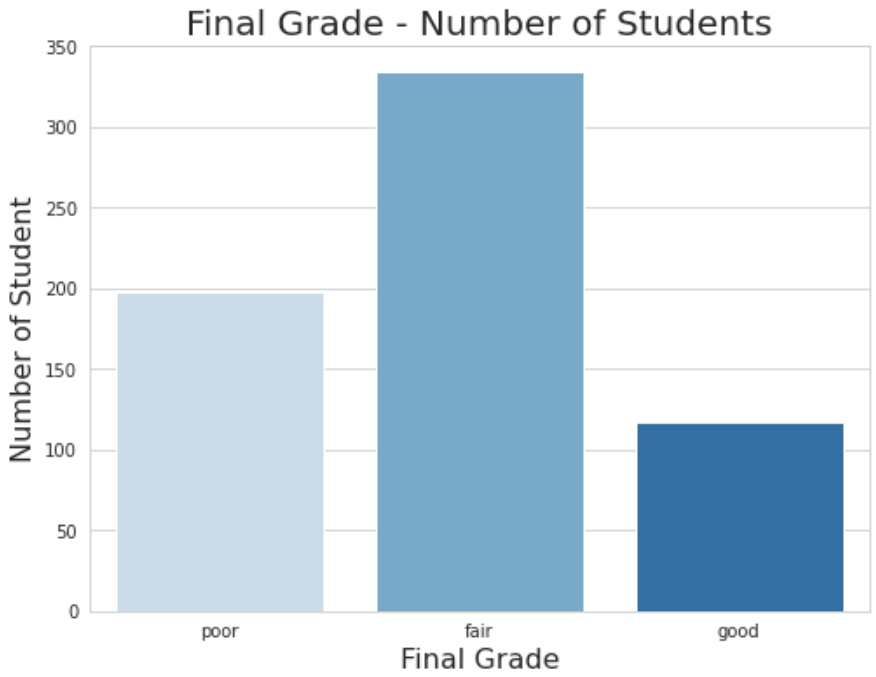
- **Support Vector Classifier**- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to 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.

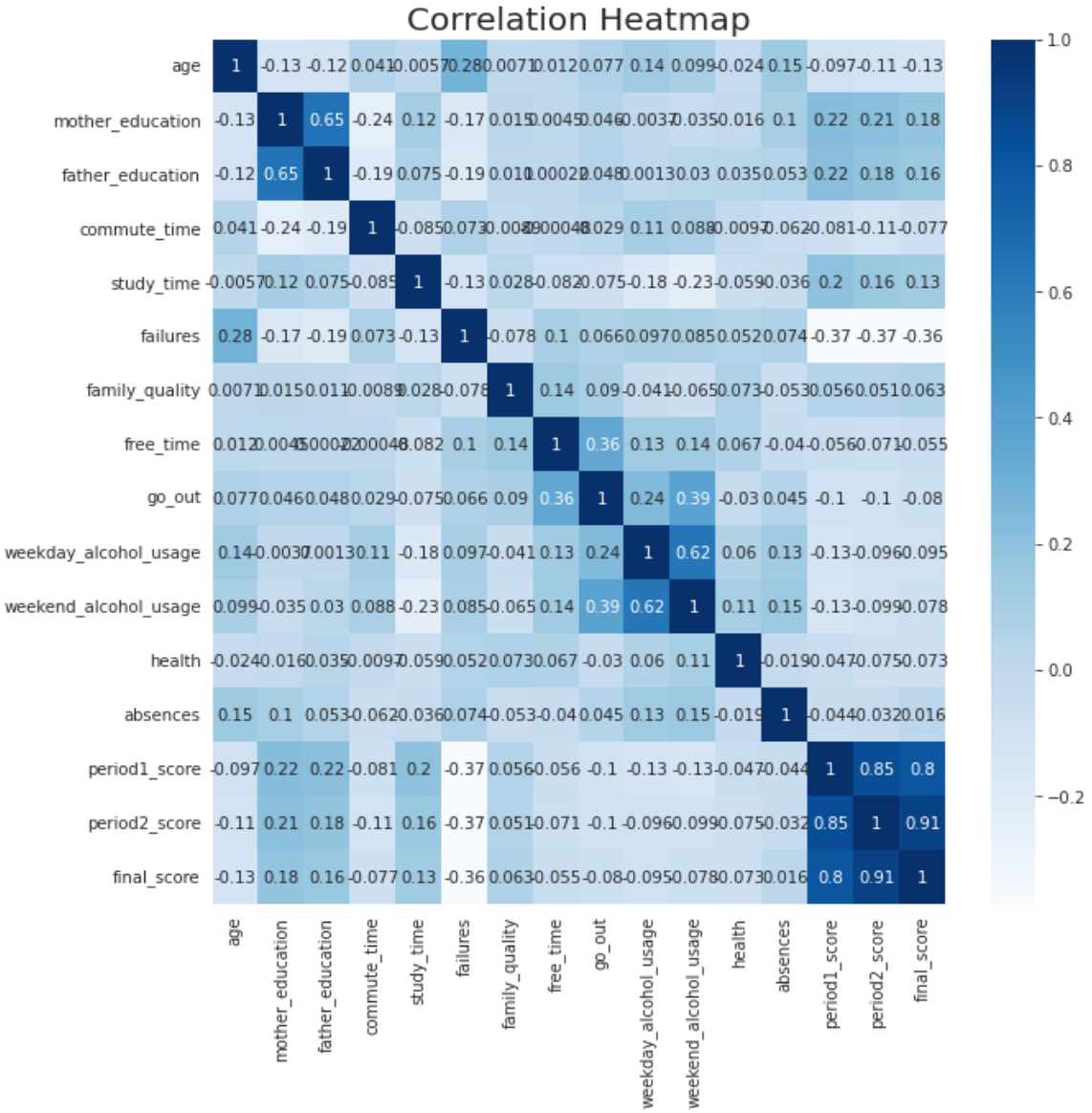
4. MODEL EVALUATION & DISCUSSION (with necessary visualizations)

I) Final grade distribution:

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



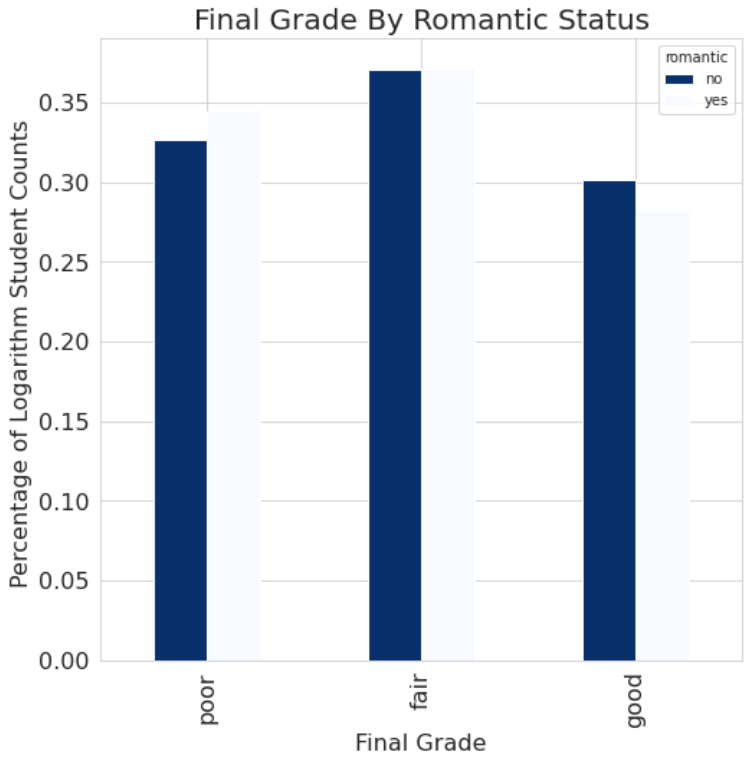
A correlation heatmap shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. The color of the cell is proportional to the number of measurements that match the dimensional value.



II) Parameters affecting Student Performance.

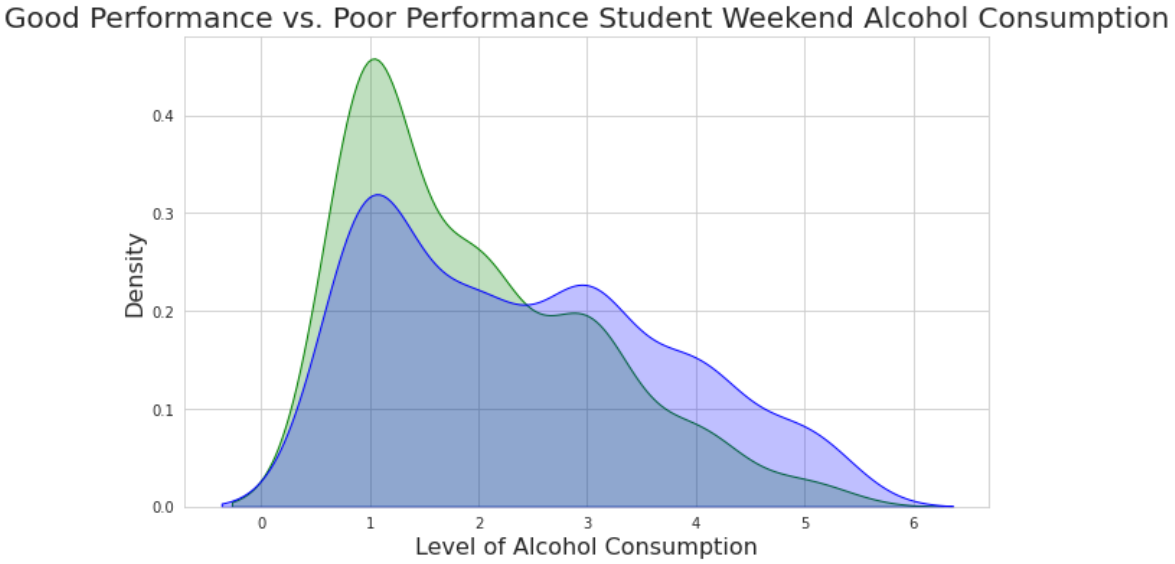
i) Romantic Status

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

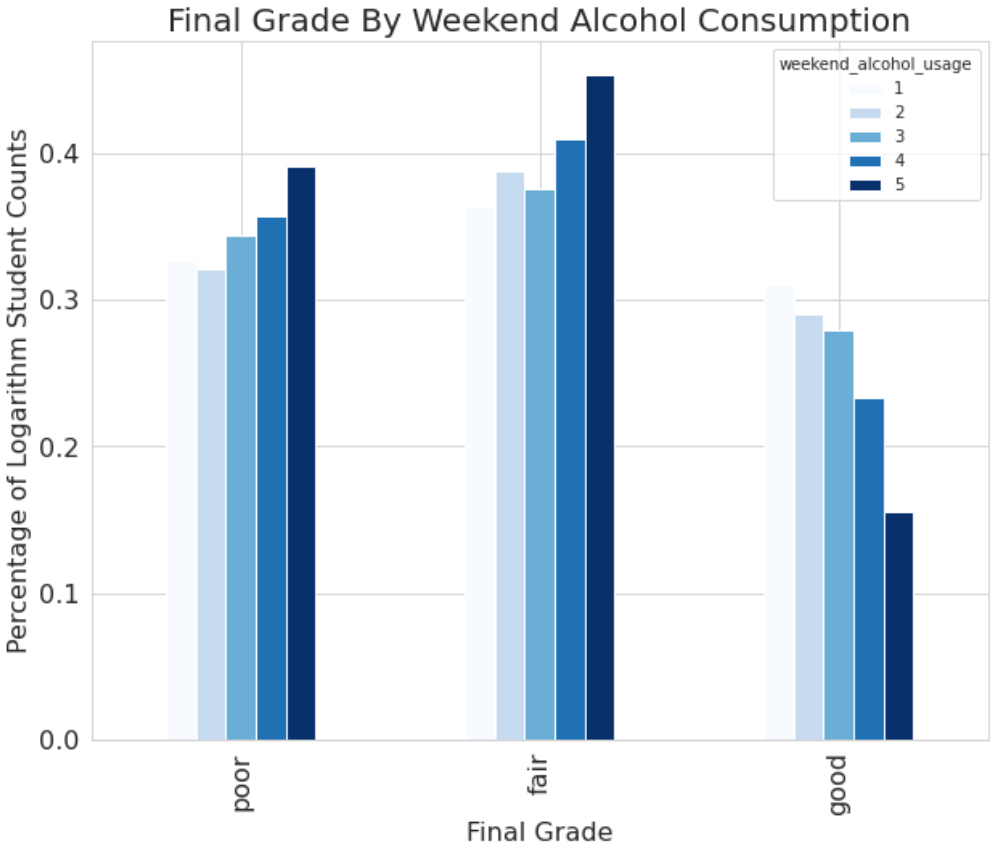


ii) Alcohol Consumption:

- 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.

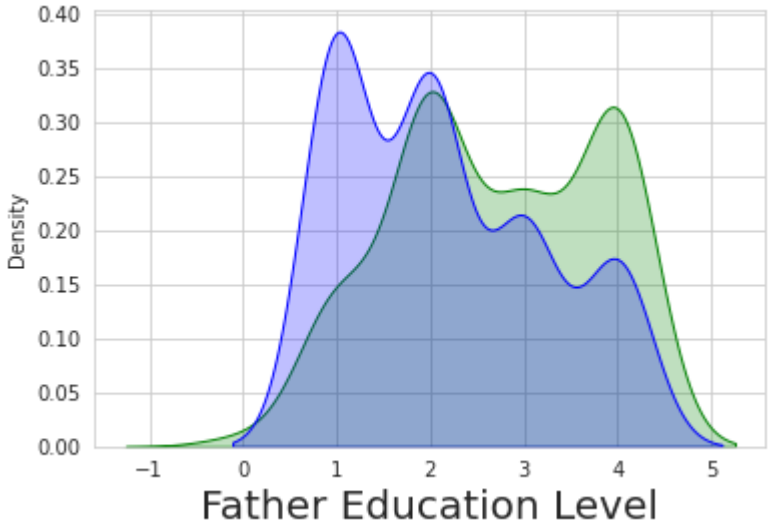


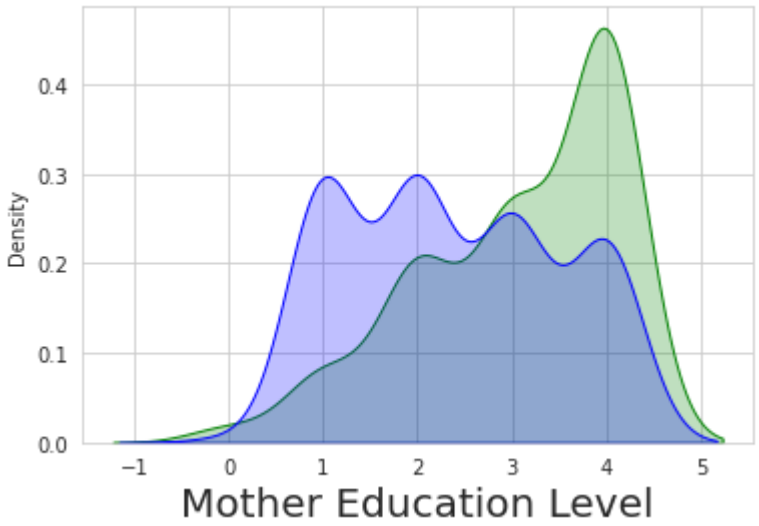
- The following graph shows that most of the students who consumed high levels of alcohol(Lvl 3,4,5) performed poorly/fairly.



iii) Parent's Education level:

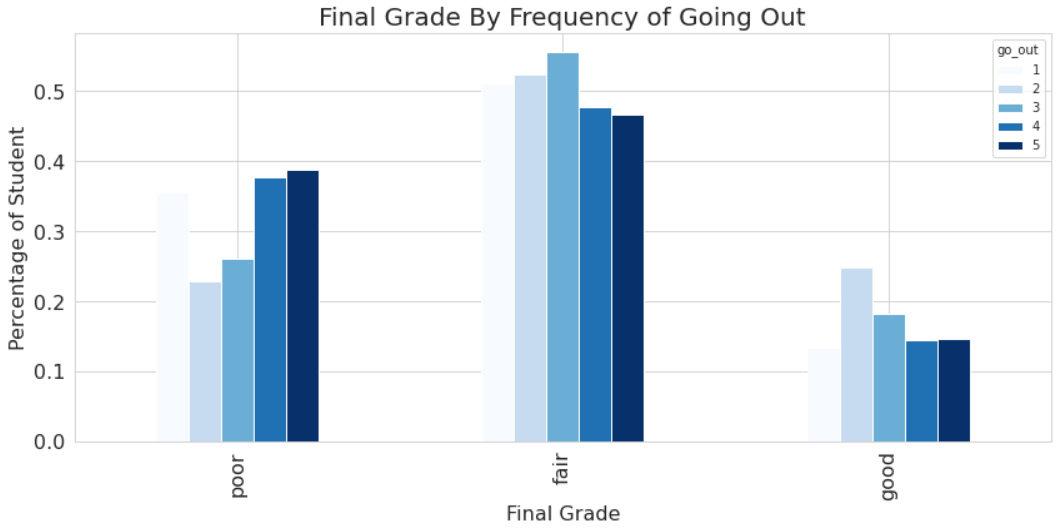
- Upon comparison it is found that the mother's education level has a significant impact on the child's performance as compared to the father's education level.





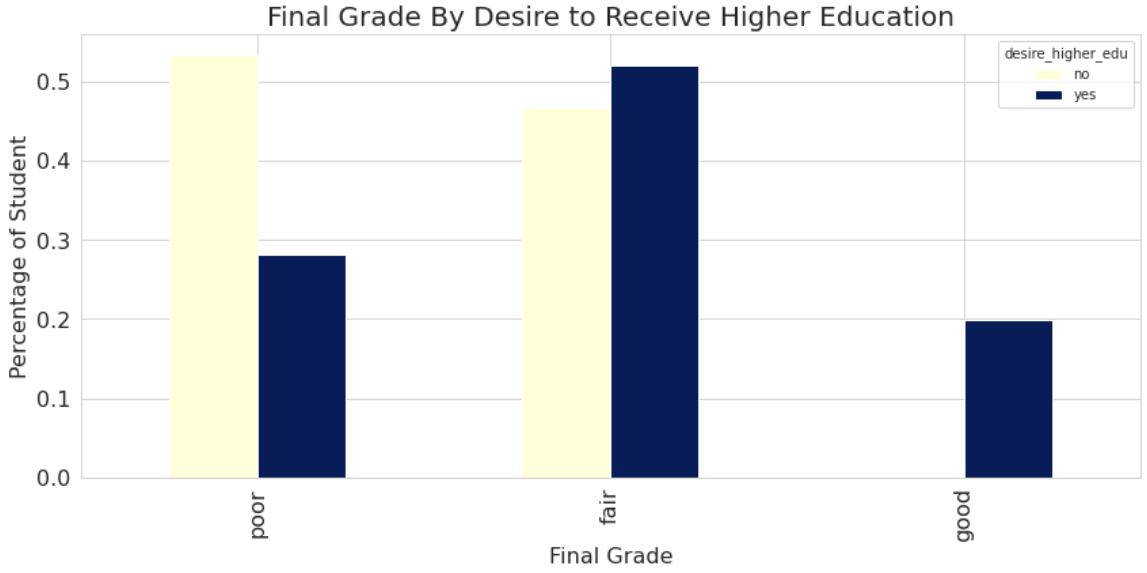
iv) Frequency of Going Out:

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



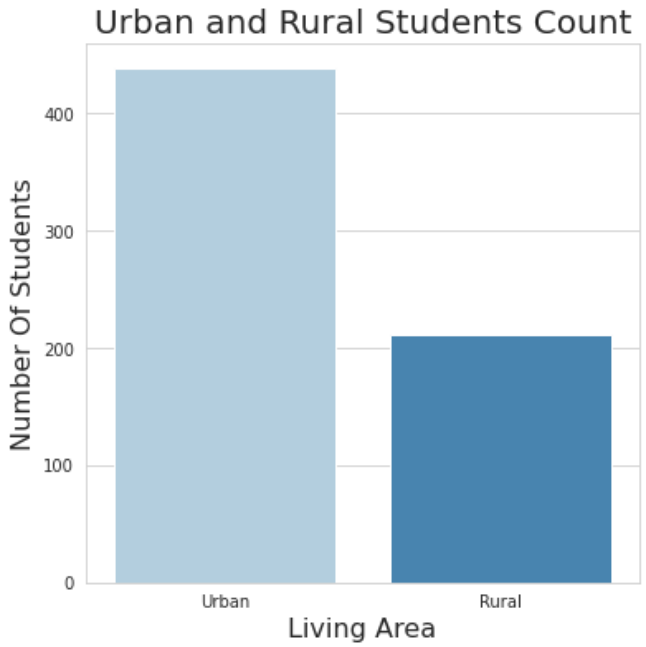
v) Desire to pursue higher education

- 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”.

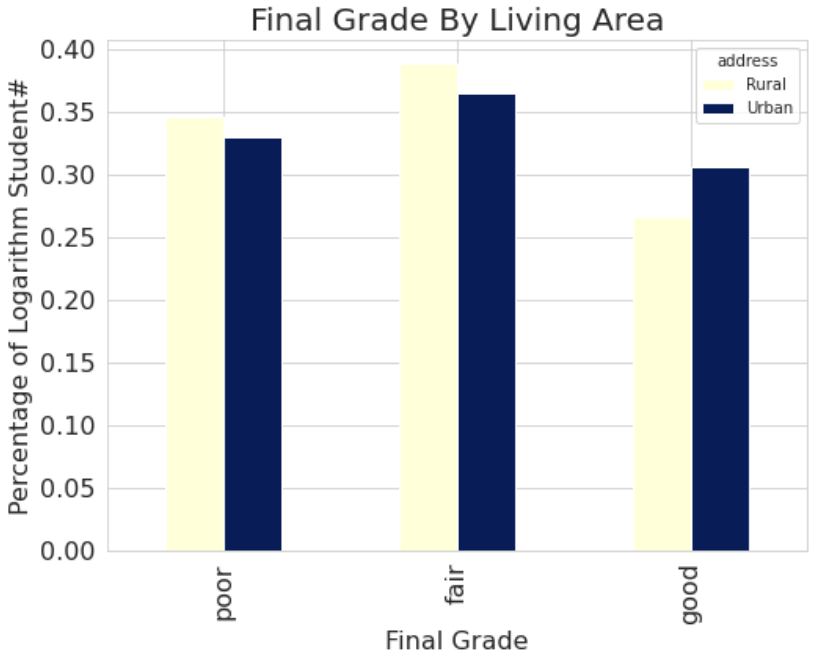


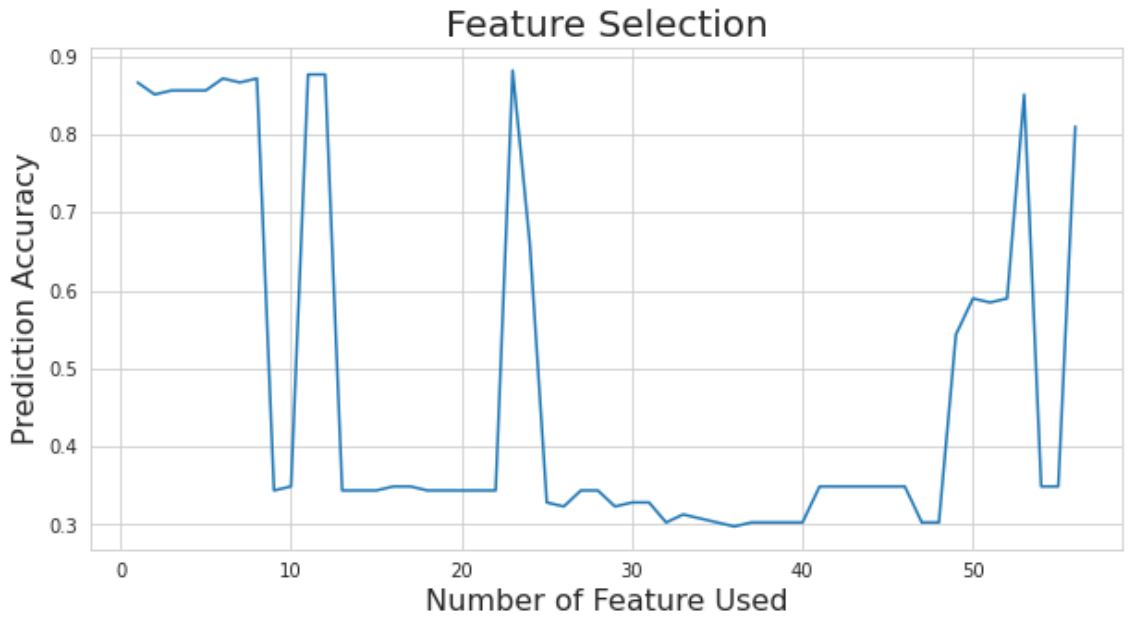
vi) Living in Urban Vs Rural Area

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



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





MODEL	MODEL SCORE	CROSS VALIDATION
Decision Tree	0.8810572	0.87179487
Random Forest	0.9845814	0.87179487
SVC	0.8656387	0.84102564
Logistic Regression	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.

5. CONCLUSION & FUTURE DIRECTION

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

6. REFLECTION PORTFOLIO

By working on this project we learned the following:

- **Teamwork** is essential because it helps us maintain an enjoyable work environment. The more we all work close to each other, the more we get to know each other's style of working and pace at which each person works. The more we work together, the more we learn to live with each others' likes, dislikes, strengths, and weaknesses. Another important aspect of working together as a team is it increases our work efficiency.
- **Github** usage for this project which helped us with seamless collaboration without compromising the integrity of our project. Since multiple people worked together and collaborated on this project, with the help of Github we found it easy to keep track of revisions—who changed what, when, and where those files are stored. GitHub took care of this problem by keeping track of all the changes that have been pushed to the repository.
- **Python** provides great libraries to deal with data science applications. Its ease of use and simple syntax is one of the main reasons why Python is widely used in the scientific and research communities and thus makes it easy to adapt for people who do not have an engineering background. We learnt how to use various Python libraries effectively.
    - Pandas helped us by providing extremely streamlined forms of data representation. It helped us to analyze and understand data better. Simpler data representation facilitated better results. It also helped us save a lot of time by importing large amounts of data very fast.
    - Numpy helped us in performing a huge variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it also supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.
    - Matplotlib is a python library used for data visualization. It helped us understand the phrase "A picture is worth a thousand words" much better. This is the best library for 2-dimensional plotting with Python. Plots, Histograms, Error charts, Power spectra, Bar chart, Scatter Plots and many more could be created with the help of Matplotlib
    - Seaborn is a library in Python predominantly used for making statistical graphics. Seaborn is a library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- We understood how to efficiently develop Machine Learning models in Python using **Google Colaboratory**. It supports collaborative development and all the team members can share and concurrently edit the notebooks, even remotely. By using Colab, one can avoid downloading or installing all the libraries required. This saves space and time and since everyone is present at the same time, working on the same code, debugging becomes easier.

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

Colab link:
https://colab.research.google.com/drive/1X2aklfkJY3PUsUPu3yH_Owci8Uz10bQl?usp=sharing#scrollTo=sEui7k_HqA3j

c. Setup to execute the code (if required)

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