Machine Learning Models

STUDENT GRADE PREDICTION

Devi Priya Bijosh Mohan
Student No:c1056531

PCP Assignment 2

Abstract

In recent years, the educational level of the Portuguese population has significantly improved, particularly in disciplines such as mathematics and language. The purpose of this study is to apply machine learning techniques to estimate a secondary school student's final grade. The research begins with an exploratory data analysis(EDA) to determine the influences on a student's final grade, such as age, family, social, and personal life. Multiple regression model and classification models are built to predict the students' performance.

Table of Content

1.Introduction

Positive social behaviours have been shown to be important in supporting academic achievement in studies on educational outcomes over the last two decades. Educators, parents, students, and other members of the educational community believe that today's school must teach more than basic skills. Students' social-emotional competency, character, health, and civic involvement must all be improved in today's schools (Cristóvão et al., 2017). Many stakeholders, including students, professors, and academic institutes, are interested in predicting students' performance in a specific course or an entire programme. Student performance prediction has been shown to be effective in predicting at-risk pupils and dropout rates which is also utilised to create early warning systems and personalised suggestion systems to help students learn better (R. Alamri & B. Alharbi).

Business intelligence collects a large amount of data that contains relevant information such as trends and patterns that may be utilised to improve decision-making and increase success. Since the human experts have limitations to overlook crucial details, automated technologies are used for decision making purpose. (Paulo Cortez & Silva Alice)

2.Problem Definition and Approach

2.1 Task Definition

This project presents comprehensive analysis of machine learning techniques for predicting the student performance by creating multiple machine models.

Data Set

The data used in this study comes from two Portuguese secondary schools for two courses in Mathematics and Portuguese. Since the dataset was compiled from two different sources, mark reports and questionnaires, it includes the students' grades as well as socioeconomic aspects that may influence their performance. The data was integrated into two datasets related to Mathematics (with 395 records) and the Portuguese language (649 records) classes. The dataset link : Link to dataset

2.2 Approach

i) Exploratory Data Analysis(EDA) **ii) Model creation**(Random Forest Regression, Classification with Support Vector Machine (SVM), Random Forest, and Multi-Layer Perceptron Neural Networks models) and
iii) Model Evaluation

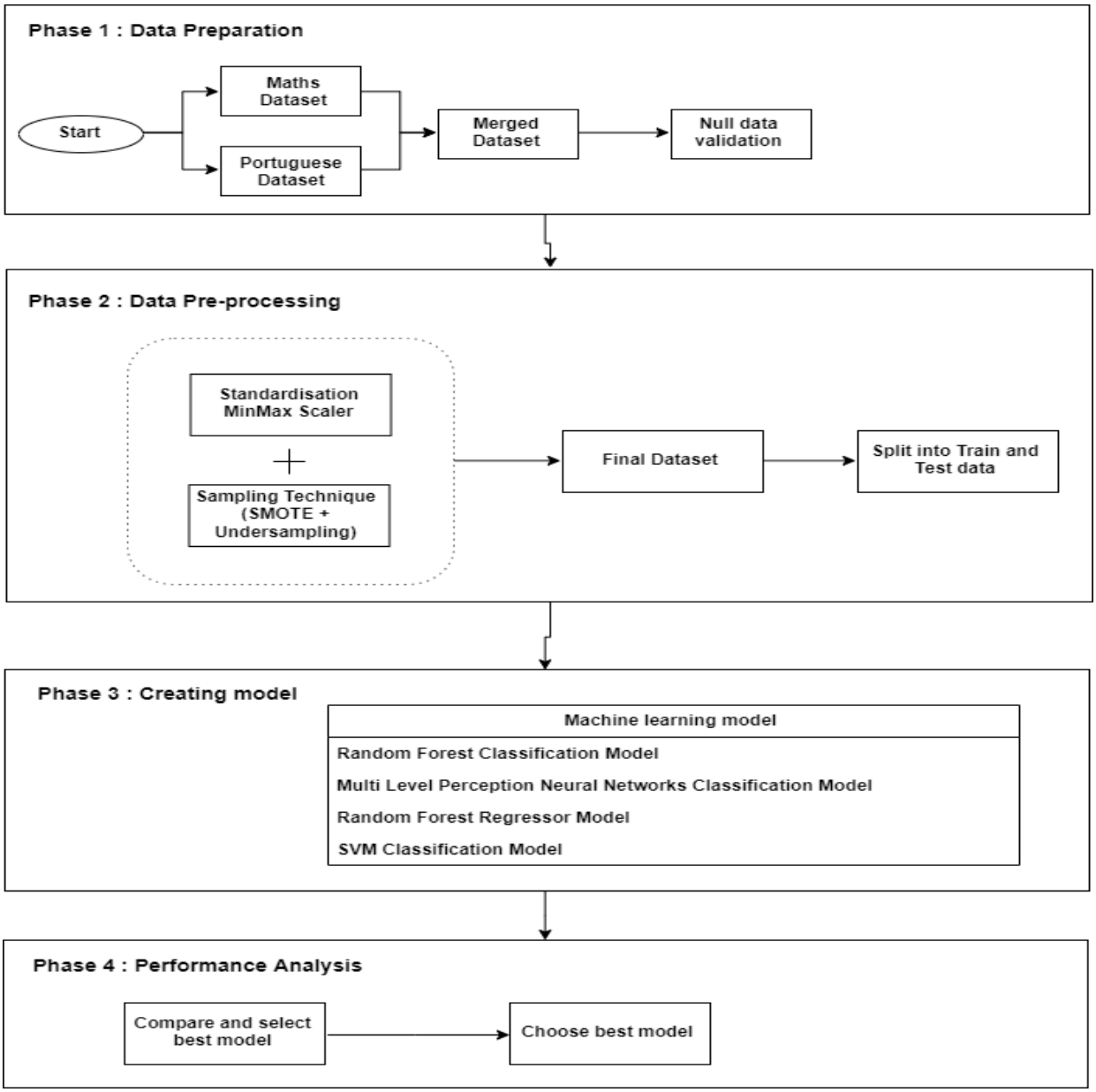
The stages include data preparation, data pre-processing, model construction, and performance evaluation. The datasets are combined to remove the duplicates. The data cleaning is not required due to no missing values.

In the data processing stage, the categorical values are converted to numerical using the label encoders, ordinal encoder and the train data set is scaled from 0 to 1 using the minMax scaler. The data is split into two parts: 77% is used to train the model, and remaining 33 % to test it. The regressor forecasts the students' final grade (G3). To visualise and see how the model fit the data, a residual plot is plotted. The Adjusted R2 and MSE values are used to assess the model's performance.

Based on the score, the classification models assign a category to the student (Fail, Good, or High). GridSearchcv and RandomSearchcv improve accuracy by optimising model parameters to find the best model. The problem of class imbalance is studied and solved utilising SMOTE and Random Undersampler techniques,

which are then fed back into the models to select the best model. The classification model's performance is evaluated using the confusion matrix. The phases are implemented as below.

Figure 1:Model creation framework



3.Results

3.1Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a way of evaluating datasets in order to summarise their key properties. Before modelling, EDA is used to see what the data can tell us.
The final dataset contains 662 observations with 33 variables, including the target, Final grade (G3), and additional 32 explanatory variables. While data cleaning phase, we found out that the data set is not having any null values. So no need of data cleaning.

A statistical table is created to get the statistical values such as mean, median, minimum, maximum, kurtosis, skewness etc.

Table 3:Statistics table of the dataset variables

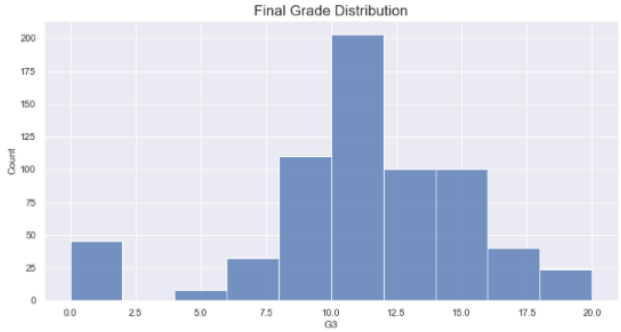
	Mean	Std	Minimum	First quartile	Median	Third quartile	Maximum	Skew	Kurtosis
Age	16.8127	1.2682	15.0	16.0	17.0	18.0	22.0	0.422084	-0.031293
Mother education	2.4924	1.1301	0.0	2.0	2.0	4.0	4.0	-0.006391	-1.254753
Father education	2.2931	1.0932	0.0	1.0	2.0	3.0	4.0	0.233855	-1.088070
Travel time	1.5650	0.7422	1.0	1.0	1.0	2.0	4.0	1.253308	1.164596
Study time	1.9275	0.8268	1.0	1.0	2.0	2.0	4.0	0.698354	0.039527
Failures	0.3323	0.7155	0.0	0.0	0.0	0.0	3.0	2.363761	5.113053
Family relation	3.9381	0.9412	1.0	4.0	4.0	5.0	5.0	-1.107387	1.416695
Free time	3.1843	1.0598	1.0	3.0	3.0	4.0	5.0	-0.189793	-0.431887
Go out	3.1722	1.1610	1.0	2.0	3.0	4.0	5.0	-0.002315	-0.833895
Daily alcohol	1.5045	0.9259	1.0	1.0	1.0	2.0	5.0	2.121290	4.231877
Weekly alcohol	2.2825	1.2891	1.0	1.0	2.0	3.0	5.0	0.624529	-0.800780
health	3.5317	1.4338	1.0	2.0	4.0	5.0	5.0	-0.492264	-1.102938
absences	4.9305	6.8529	0.0	0.0	3.0	8.0	75.0	3.843050	26.207011
G1	10.7281	3.0798	3.0	8.0	10.0	13.0	19.0	0.303780	-0.444839
G2	10.7085	3.5269	0.0	9.0	11.0	13.0	19.0	-0.415990	1.008060
G3	10.7251	4.1036	0.0	9.0	11.0	13.0	20.0	-0.805366	1.109024

From above table, we can understand that the average age of the students are 17, with the lowest of 15 and the highest of 22. Though some students had a high(3) failure rate in the past, the average failure rate is quite low, with 75% of students having no failures. When compared to weekend alcohol intake, daily alcohol consumption is quite minimal. Although the majority of students do not drink alcohol on a regular basis, there are a few students that consume high amount of alcohol. The average absence is five days, yet half of the pupils do not take any time off. About 75% of students took fewer than 10 days off, however some students took 75 days in the middle. The average G1 and G2 grades in the past are 3,3 and 4, respectively, some got 20 points.

Data Visualization

Matplotlib and Seaborn are the data visualisation libraries utilised in this project. The analysis is done based on different socio-economic factors. For the ease of analysis, the factors are grouped into different categories based on age, locations, time spend, social life and family attributes. Each criteria is analysed and given interpretations.

Figure 1: Final Grade Distribution



From this plot, that we can understand that about 40 students score very low grade <2, since the rest of the students score from 4.5 to 20, with an average of 10 to 12.

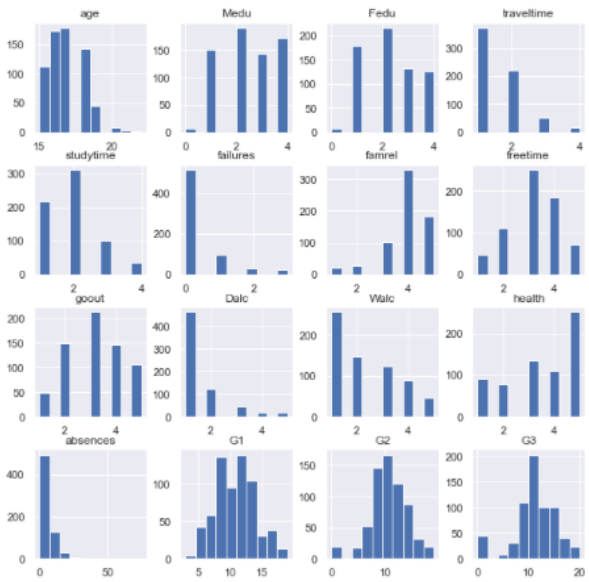


Figure 2: Histogram of variables

This shows the distribution of all the factors that affect final grade. The graph clearly plot G1,G2, G3 is almost normally distributed.

a)Students with different age groups

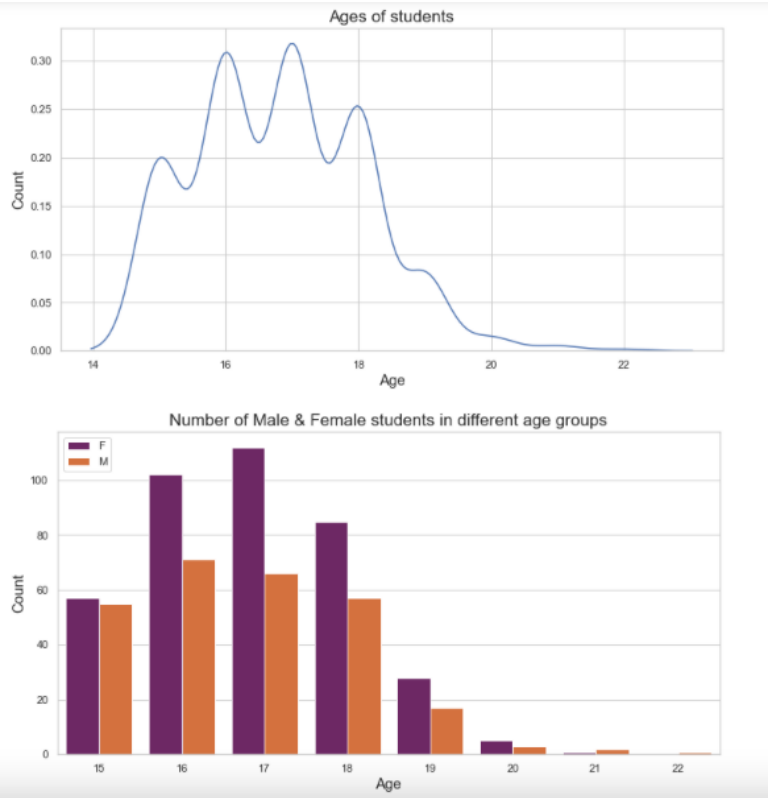


Figure 3 : Age distribution and plot

The student age seems to be ranging from 15-19, where gender distribution is pretty even in each age group. The age group above 19 may be year back students or dropouts.

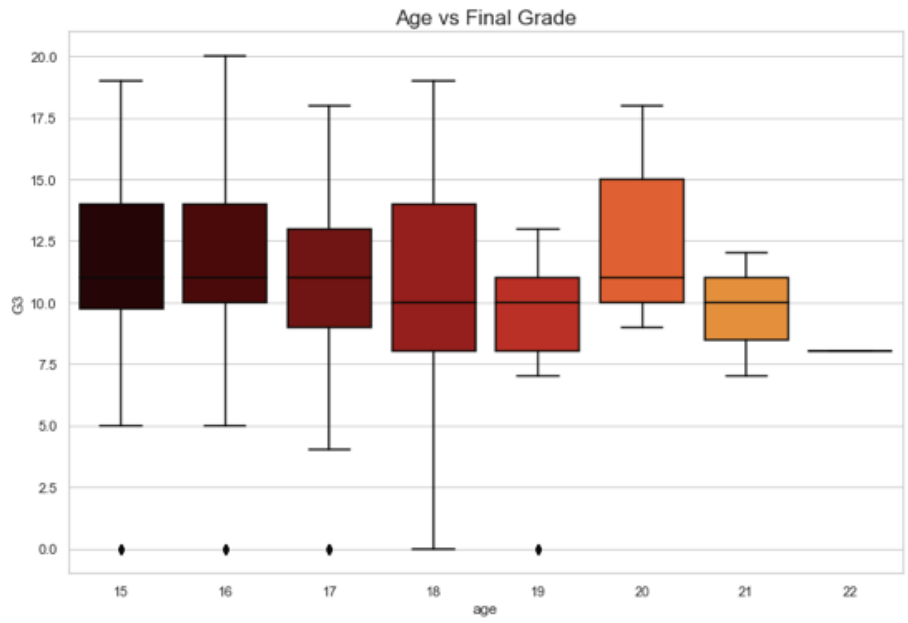


Figure 4 : Age vs Final grade plot

The above plot shows that the median grades of the three age groups(15,16,17) are similar. Also, 19 and 21 have the same median and also very small spread compared to others. Age group 20 seems to score highest grades among all.

b) Students from Urban & Rural Areas

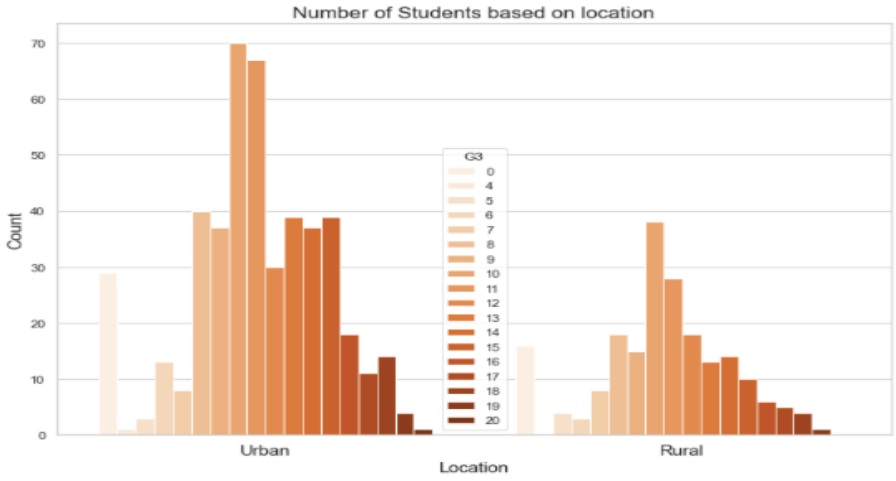


Figure 5 : Students distribution based on the location

Approximately 70% students come from urban region and 30% from rural region.

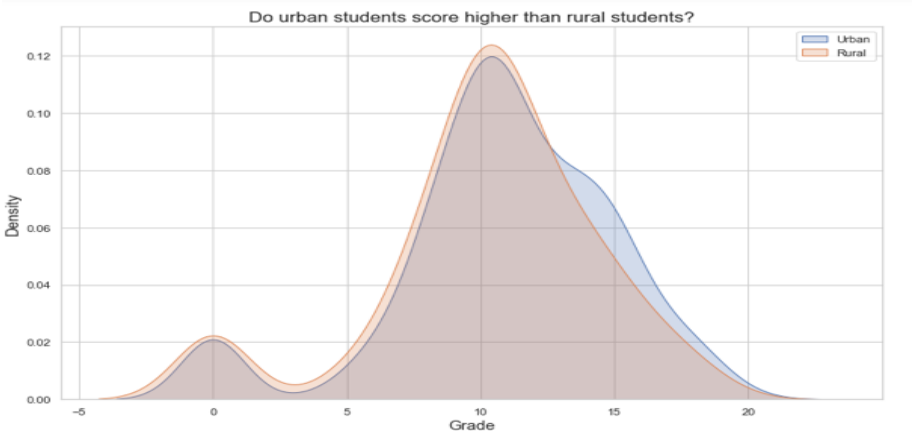


Figure 6 : Density plot of grade based on the location

From the above graph we can understand that there is not much difference between the grades based on location.

c)The Family Attributes

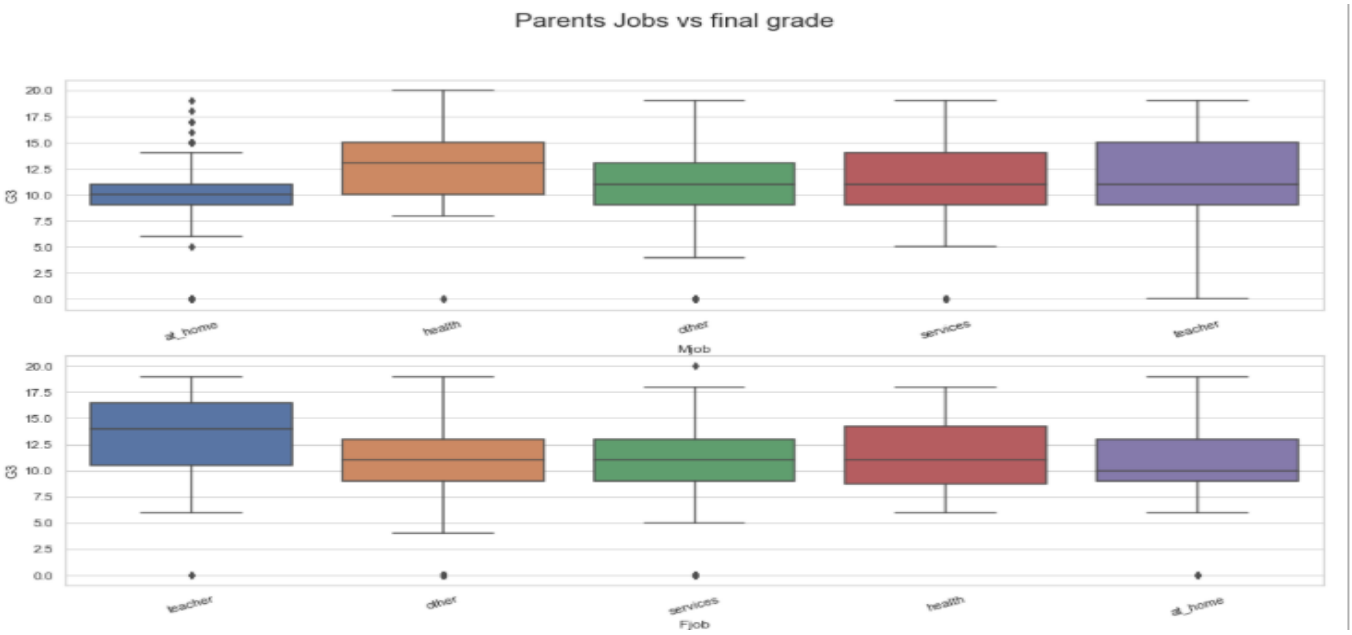


Figure 7 : Box plot of parents jobs vs final grade

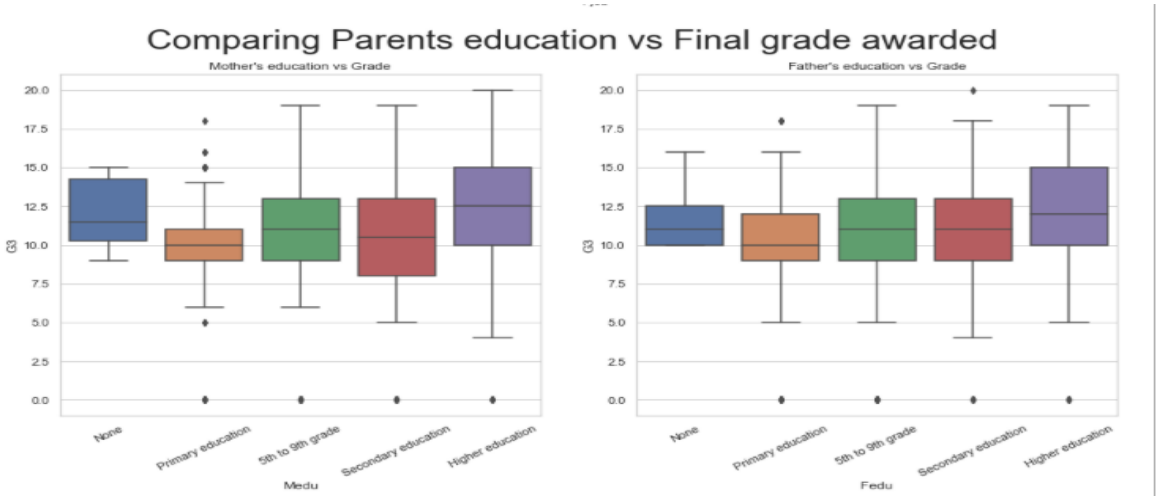


Figure 8 : Parents education vs Final grade

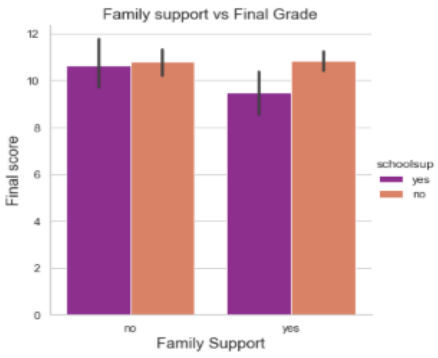


Figure 9 : Family support vs final grade

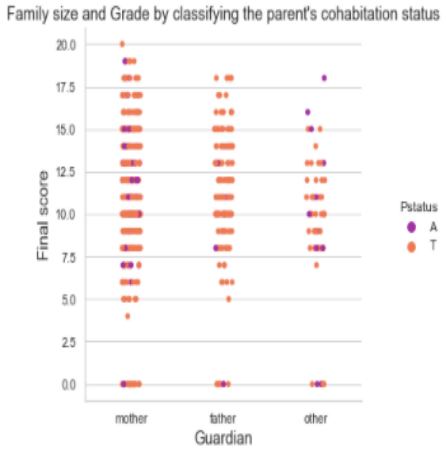


Figure 10 : Guardian vs final grade

Students' grades were affected by their parents' jobs and education. Students whose mothers or fathers work as teachers or in the health care field receive high grades. Students whose parents are unemployed received lower grades than others. Students with well-educated parents outperform their peers. The majority of pupils have their mother as their guardian, and they perform higher. There is no evidence that family and school support helps pupils achieve higher grades.

d)Social and Personal Life

Figure 12 : Romantic relationship vs final grade

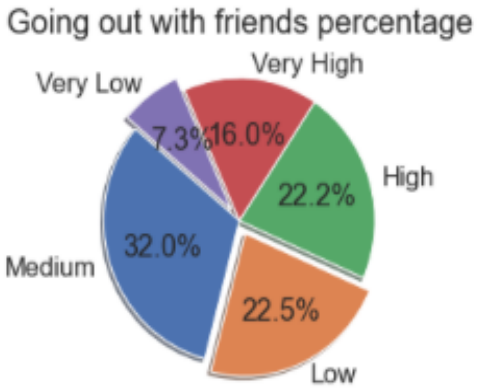


Figure 13 : Going out with friends distribution

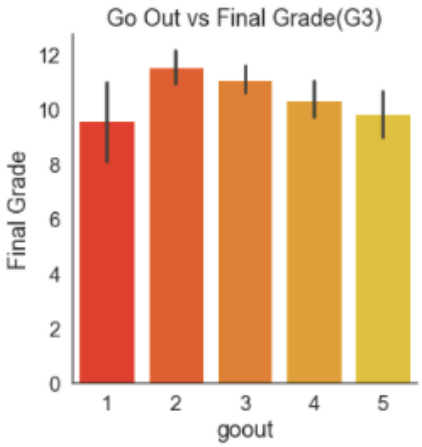


Figure 14 : Going out vs Final Grade

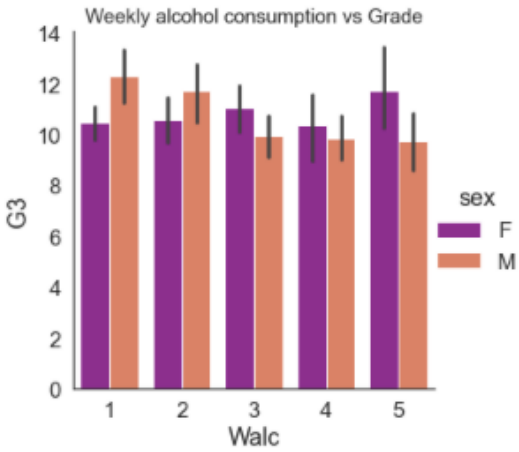


Figure 15 : Weekly alcohol consumption vs Final grade

Students who do not have a romantic relationship and with an average social life score higher. However, if they go out more often, their grade will be low. Students that consume less alcohol score higher than their peers. Surprisingly, female students who consume high amount of alcohol scores higher.

e) Past academic history:

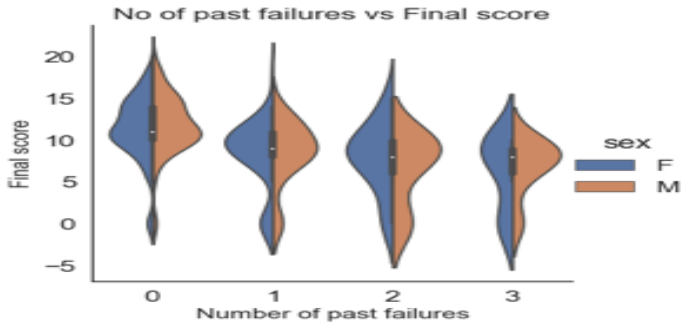


Figure 15 :Past failures vs final score

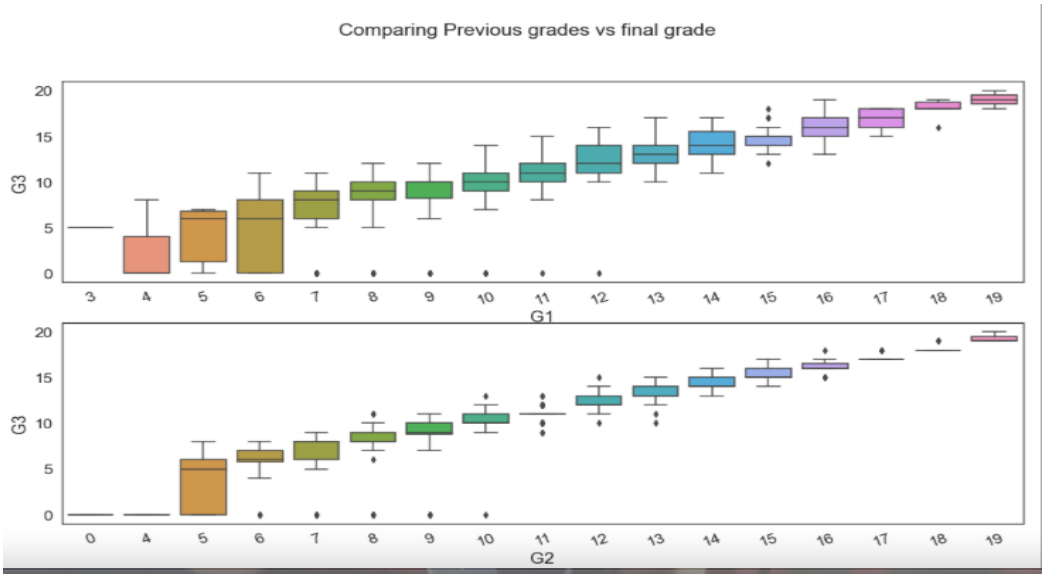
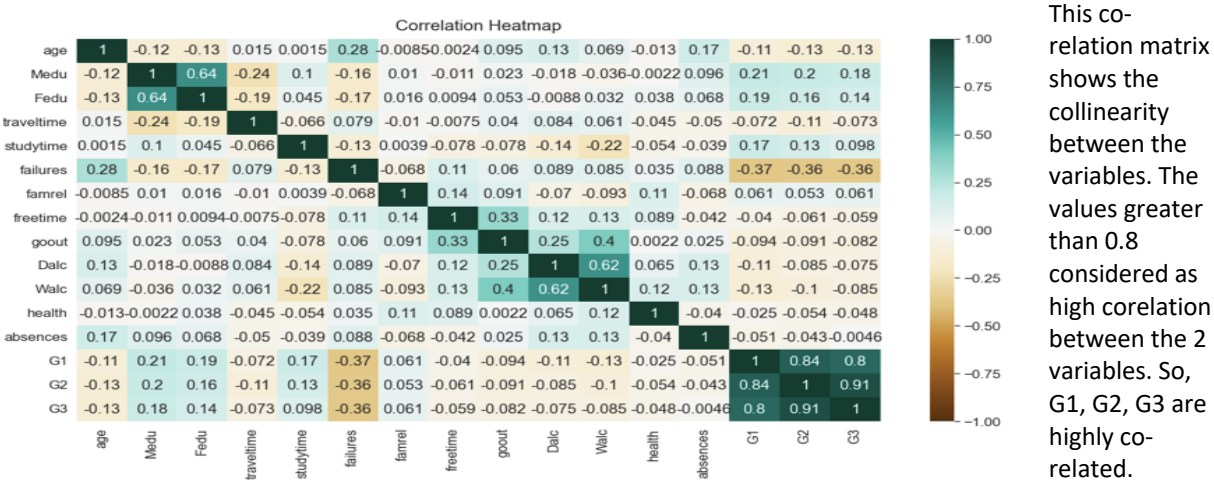


Figure 16 :Previous grades vs final grade

Students with no failures and strong past grades got good final grades.

Co-relation Matrix

Figure 17 : Correlation Heatmap



This co-relation matrix shows the collinearity between the variables. The values greater than 0.8 considered as high corelation between the 2 variables. So, G1, G2, G3 are highly co-related.

2.Model Creation

4.2.1. Random forest Regressor

The model has an R2 of 0.90 and an Adjusted R2 of 0.89. It does, however, have a high MSE of 1.8, which is quite high. Although the residual plots are evenly distributed, some residuals are irregularly distributed, resulting in a large mean square error. Randomserachcv is used to tune the parameters.

```
R2 value  :  0.82
Adjusted r2 :  0.82
Mean Squared Error: 3.01
Mean Absolute Error: 1.07
Root Mean Squared Error: 1.74
Accuracy : 0.82
```

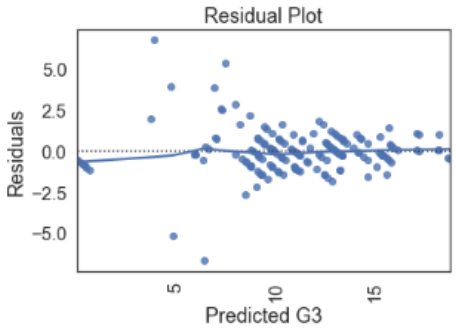


Figure 18 : Predicted vs residual plots

Feature Importance Graph

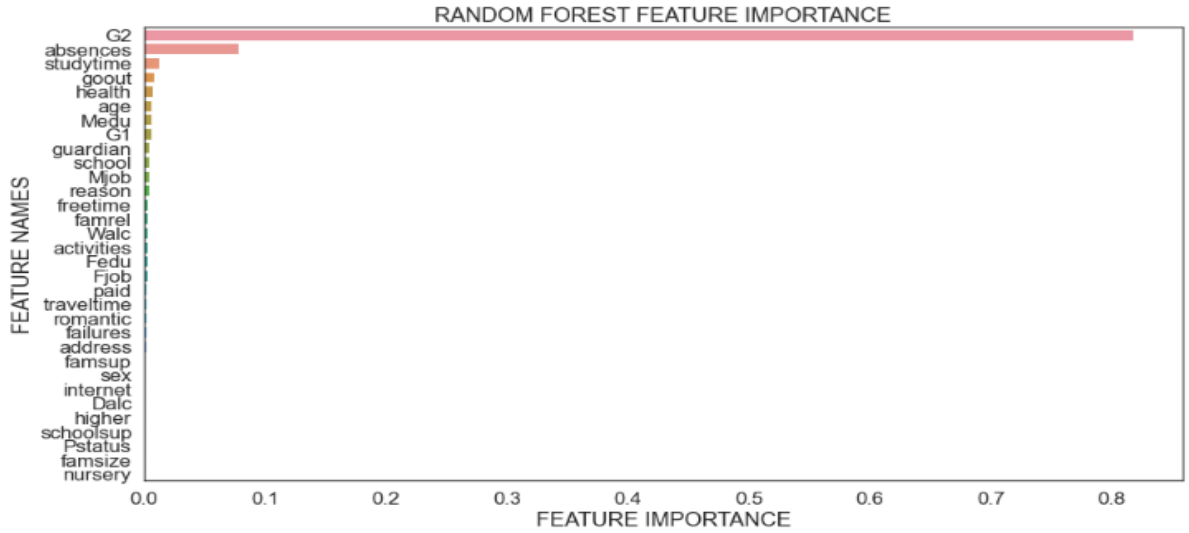


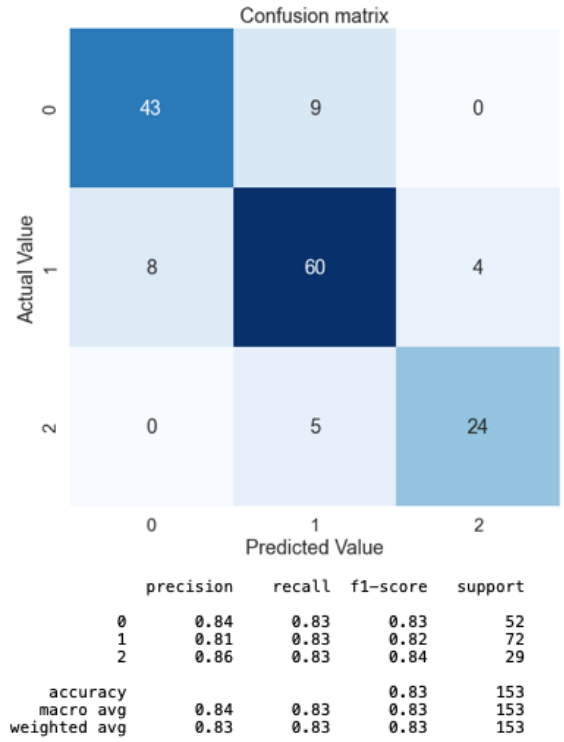
Figure 19 : Random Forest Feature Importance Graph

The above graph shows the importance of each features on the prediction, where we can see that the G2 plays significance role in the prediction and absence is the second factor which affects. The rests are having very little role in the prediction.

4.2.2 .Classification: 3 classification models are created.

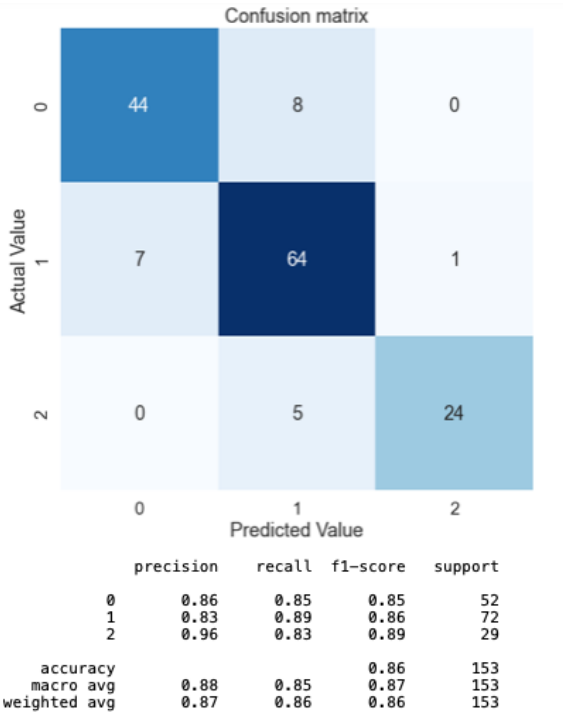
SVM MODEL

Random Forest classification Model



	precision	recall	f1-score	support
0	0.84	0.83	0.83	52
1	0.81	0.83	0.82	72
2	0.86	0.83	0.84	29
accuracy			0.83	153
macro avg	0.84	0.83	0.83	153
weighted avg	0.83	0.83	0.83	153

Figure 20 : Confusion matrix of SVM model



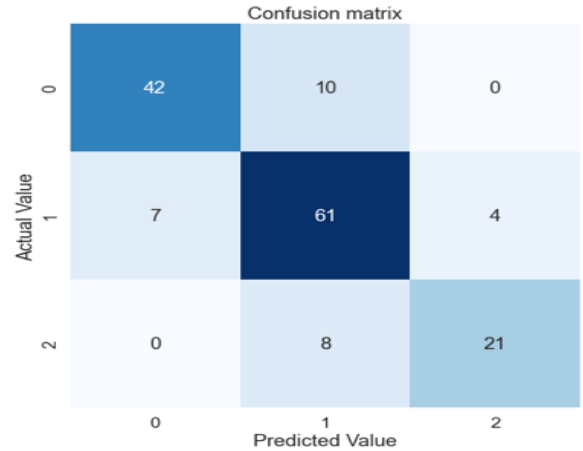
	precision	recall	f1-score	support
0	0.86	0.85	0.85	52
1	0.83	0.89	0.86	72
2	0.96	0.83	0.89	29
accuracy			0.86	153
macro avg	0.88	0.85	0.87	153
weighted avg	0.87	0.86	0.86	153

Figure 21 : Confusion matrix of Random forest model

MLP Classifier without sampling

Figure 22 : Confusion matrix

Accuracy (Imbalanced Data:): 81.05%



	precision	recall	f1-score	support
0	0.86	0.81	0.83	52
1	0.77	0.85	0.81	72
2	0.84	0.72	0.78	29
accuracy			0.81	153
macro avg	0.82	0.79	0.81	153
weighted avg	0.81	0.81	0.81	153

When comparing all three models random forest classifier have high accuracy and also the true positive and negative values.

Analyse the class imbalance issue

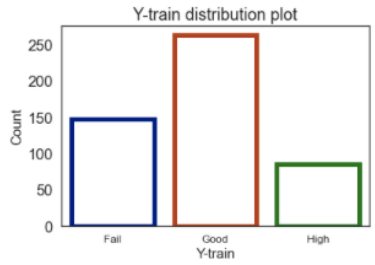


Figure 20: Class imbalance plot

This shows very high class imbalance issue. So we are using SMOTE and Random under sampler to solve it.

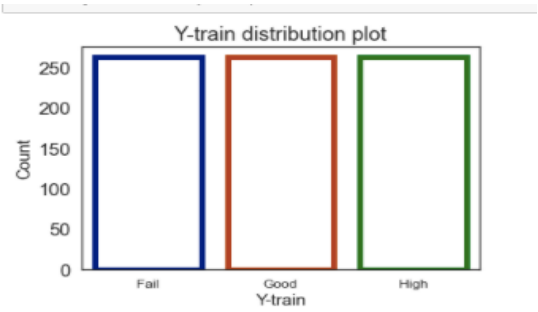


Figure 21: SMOTE resampled data

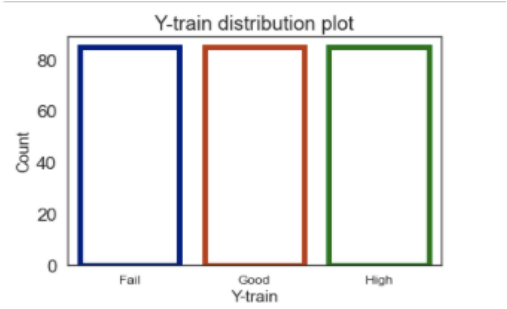
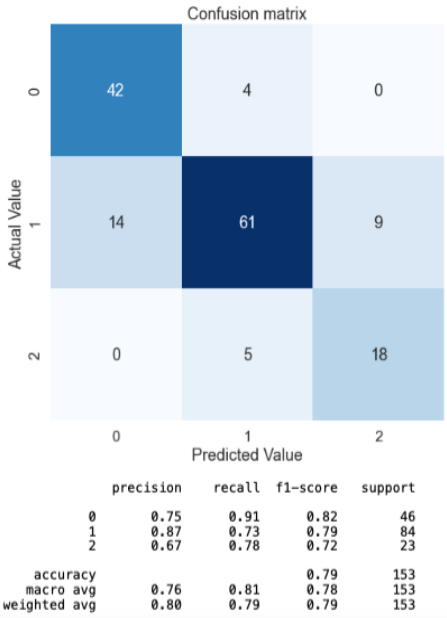


Figure 22: Random UnderSampler data

Results of the models after applying SMOTE

2.1.Classification: 3 classification models are created.

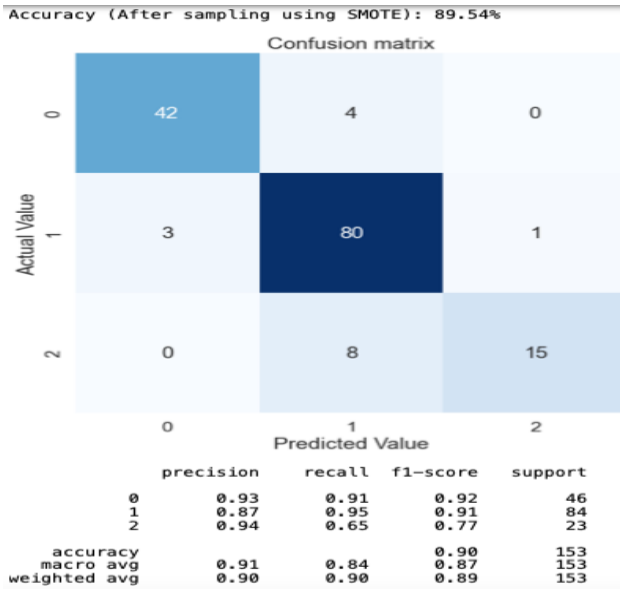
SVM MODEL



	precision	recall	f1-score	support
0	0.75	0.91	0.82	46
1	0.87	0.73	0.79	84
2	0.67	0.78	0.72	23
accuracy			0.79	153
macro avg	0.76	0.81	0.78	153
weighted avg	0.80	0.79	0.79	153

Figure 23 : Confusion matrix of SVM model

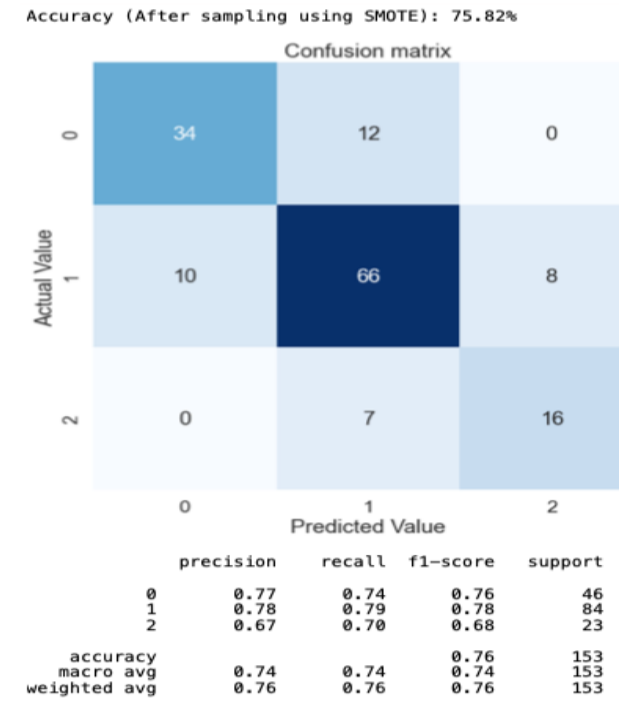
Random Forest classification Model



	precision	recall	f1-score	support
0	0.93	0.91	0.92	46
1	0.87	0.95	0.91	84
2	0.94	0.65	0.77	23
accuracy			0.90	153
macro avg	0.91	0.84	0.87	153
weighted avg	0.90	0.90	0.89	153

Figure 24 : Confusion matrix of Random forest model

MLP Classifier without sampling



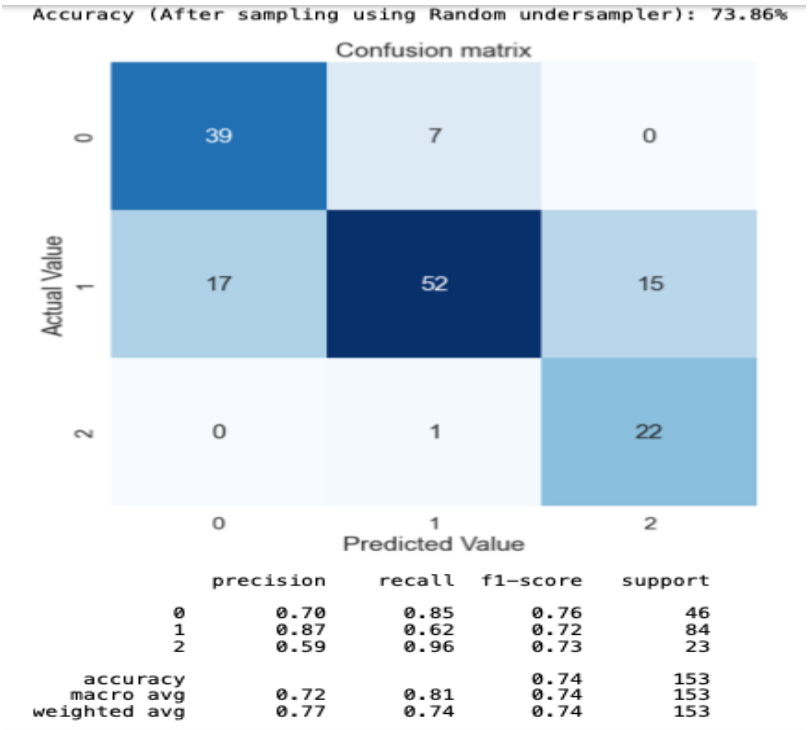
	precision	recall	f1-score	support
0	0.77	0.74	0.76	46
1	0.78	0.79	0.78	84
2	0.67	0.70	0.68	23
accuracy			0.76	153
macro avg	0.74	0.74	0.74	153
weighted avg	0.76	0.76	0.76	153

Figure 25 : Confusion matrix

When comparing all three models random forest classifier have high accuracy and also the true positive and negative values whereas MLP is having very low accuracy and true positive and negative values.

Results of the models after applying Random Under sampler

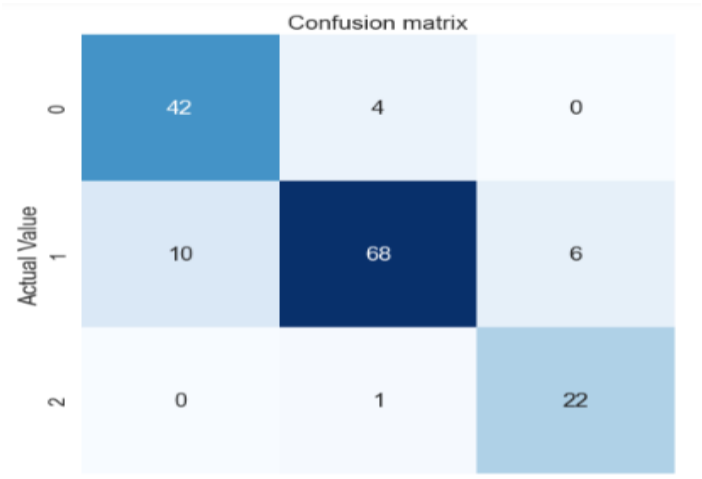
SVM MODEL



	precision	recall	f1-score	support
0	0.70	0.85	0.76	46
1	0.87	0.62	0.72	84
2	0.59	0.96	0.73	23
accuracy			0.74	153
macro avg	0.72	0.81	0.74	153
weighted avg	0.77	0.74	0.74	153

Figure 26 : Confusion matrix of SVM model

Random Forest classification Model

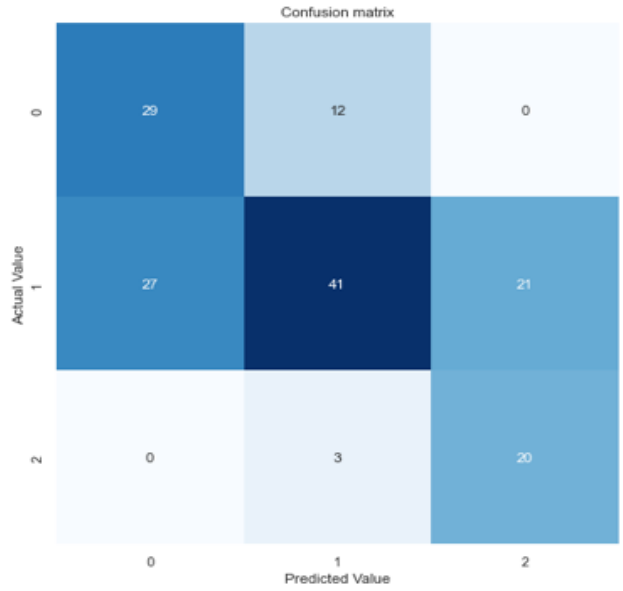


	precision	recall	f1-score	support
0	0.81	0.91	0.86	46
1	0.93	0.81	0.87	84
2	0.79	0.96	0.86	23
accuracy			0.86	153
macro avg	0.84	0.89	0.86	153
weighted avg	0.87	0.86	0.86	153

Figure 27 : Confusion matrix of Random forest model

MLP Classifier without sampling

```
Accuracy (After sampling using Random undersampler): 58.82%
```



	precision	recall	f1-score	support
0	0.52	0.71	0.60	41
1	0.73	0.46	0.57	89
2	0.49	0.87	0.62	23
accuracy			0.59	153
macro avg	0.58	0.68	0.60	153
weighted avg	0.64	0.59	0.58	153

Figure 28 : Confusion matrix

When comparing all three models random forest classifier have high accuracy and also the true positive and negative values.

5. Execution Steps

- Open main_module.py file
- Install the imported libraries.
- Run each line by line

6. Related Work

To evaluate student performance, an automated evaluation method has been proposed which uses a tree algorithm to accurately predict student performance (Dhilipan et al., 2021). The classification of the suggested system is done using Education Data Mining. The database is analysed using the clustering data mining technique. (Chilimbi et al., 2014)

Using student information from college enrolment, a novel learning model has been presented where the dataset is fed into machine learning algorithms that can be used to apply and predict academic success. They chose 13 algorithms from five different ML categories: Nave Bayes, SVM, MLP, IBK, Rules, and tree. (Li et al., 2013)

The binary/five-level classification and regression tasks were used to model the two main classes. Decision Trees, Random Forest, Neural Networks, and Support Vector Machines models were also examined with three input selections (Paulo Cortez & Silva Alice). The results suggest that if the first and/or second school period grades are available, excellent predictive accuracy can be reached.

7. Reflection

I applied scalers, hyper parameter methodologies, and sampling techniques appropriately in this study, which has helped me learn more about the relevance of pre-processed data. Outliers are readily visible in the EDA, and the multicollinearity heatmap displays some high values that indicate the possibility of multicollinearity. However, outliers and multicollinearity are not examined or avoided in order to improve the model's accuracy. The model can be made more accurate by dealing with outliers and co-related variables. In addition, I only concentrated on a few specific models. So in a future study, I will focus more on the aforementioned issues.

8. Conclusion

SVM stands for supervised machine learning, which transforms data using the kernel before determining an appropriate boundary between the various outputs. The random forest model builds a huge number of decision trees during training and provides the mean/mode of prediction of each tree. A MLP is a fully linked feedforward artificial neural network. SVM is better to work in binary classification problem, not better with multiple classification problems. The MLP is time consuming to perform and having a very low accuracy rate. But the random forest model is good in both multiple classification and regression with high accuracy and good Adjusted R2 rate(regression). Because of the high Mean square error rate, the best model to select is random forest classification model. Based on accuracy and performance, the random forest classifier with SMOTE over sampled data appears to be an excellent model compared to other models. Because the use of machine learning to forecast student performance at institutions will ultimately improve the decision support system, students' academic performance will increase in the future. Visualization was used to determine the factors that influence performance. Absence is listed as a major issue, however there appear to be some outliers, which could indicate that some students have dropped out. Travel time, age, romantic relationships, parental education, and employment status are other factors impacts the studies.

References

Buniyamin, N., bin Mat, U., & Arshad, P. M. (2015). Educational data mining for prediction and classification of engineering students achievement. Paper presented at the *2015 IEEE 7th International Conference on Engineering Education (ICEED),* 49-53.

Chilimbi, T., Suzue, Y., Apacible, J., & Kalyanaraman, K. (2014). Project adam: Building an efficient and scalable deep learning training system. Paper presented at the *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14),* 571-582.

Cristóvão, A. M., Candeias, A. A., & Verdasca, J. (2017). Social and Emotional Learning and Academic Achievement in Portuguese Schools: A Bibliometric Study. *Frontiers in Psychology, 8* https://www.frontiersin.org/article/10.3389/fpsyg.2017.01913

Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. Paper presented at the *IOP Conference Series: Materials Science and Engineering, , 1055*(1) 012122.

Paulo Cortez, & Silva Alice.*USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE.* http://www3.dsi.uminho.pt/pcortez/student.pdf

R. Alamri, & B. Alharbi. (2021). *Explainable Student Performance Prediction Models: A Systematic Review*10.1109/ACCESS.2021.3061368

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. https://doi.org/10.5281/zenodo.883859

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, *23*(7–8), 673–692.