Module Title:	6COSC017C - Machine Learning
Assessment:	Coursework 1
Student IDs:	00015775
Submission Deadline:	05.11.2025
Module Marker's Name:	
Feedback	

Contents Page:

Introduction

Based on the selected dataset of student performance, the purpose of this project is to predict the final year grades, also known as G3, in Math or Portuguese subjects in secondary education in Portugal. The original dataset was collected during 2008 and comes with its own papers describing the datasets' details and uploaded to UC Irvine Machine Learning Repository.

The link of dataset: https://archive.ics.uci.edu/dataset/320/student+performance

Digital Object Identifier (DOI): 10.24432/C5TG7T

For this dataset, as per authors of Cortez and Gonçalves Silva (2008), three types of analysis are suitable. The first one is binary classification, if the G3 mark is above or equal to 10, then a student will pass, otherwise fail the subjects. The second one is a 5-level classification with 1-being excellent/very good and 5 failing. Lastly, is regression analysis, with numerical outputs for G3 ranging from 0 to 20, inclusively. For this project, the regression analysis will be the main problem to solve, by trying to predict the approximate scores of the student's performance for the G3 grade based on selected feature inputs, which a user can specify from the website, without the need to download the application.

The website for this project is hosted on the Streamlit Community Cloud, and the link along with the github repository is publicly available.

GitHub Repository: https://github.com/00015775/MLDA-CW1-15775

Streamlit App link: _____

However, do note that if no people visit the website within certain days, approximately from 3 to 7 days without any visitors, then the link will switch to sleeping/hibernation mode. While this does not mean that the website/link will stop working, but it does mean that it can take up from 30 seconds to a couple of minutes until the website is fully awake and is functional to use. Hence, give it some minutes for it to wake up.

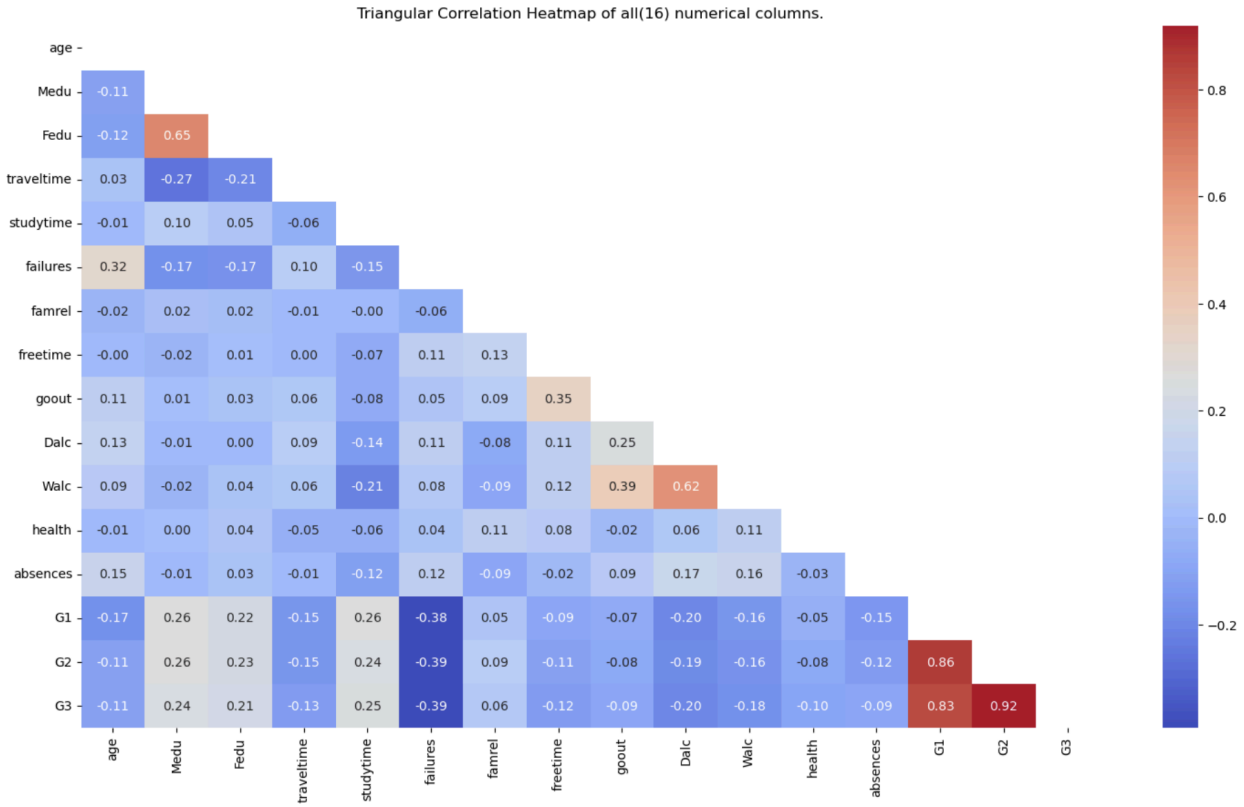
Description of Exploratory Data Analysis

The dataset has 649 rows and 33 columns. Out of 33 columns, 17 of them are of "*object*" data type, that is also known as "*string*" data type in Python, while the rest of 16 columns are of "*int64*" data type, integers. In terms of four levels of measurements, there are 16 nominal, 12 ordinal, 3 interval and 2 ratio data types.

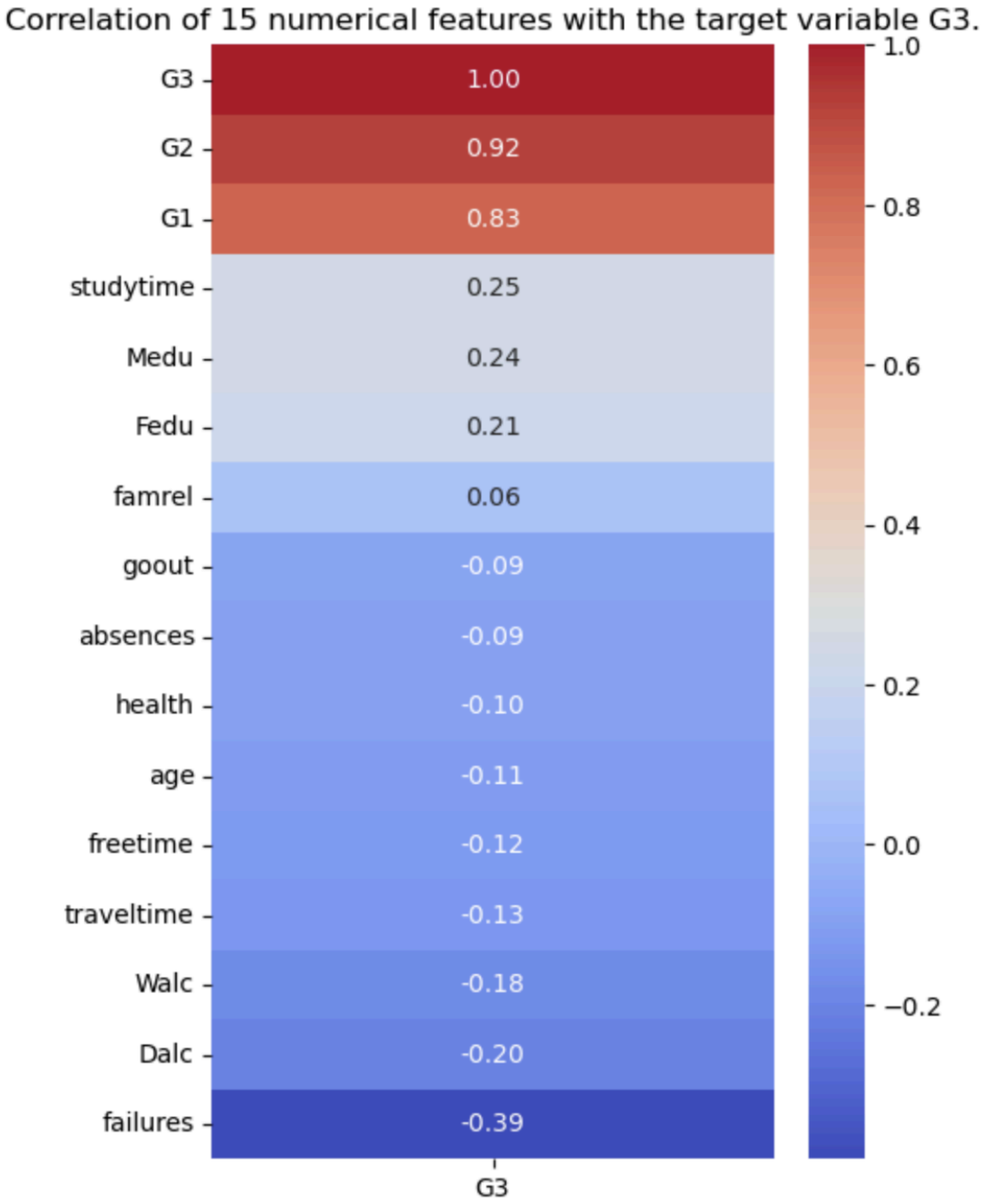
- **Nominal (16):** school, sex, address, Pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic
- **Ordinal (12):** famsize, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health
- **Interval (3):** G1, G2, G3
- **Ratio (2):** age, absences

Below is the triangular correlation heatmap, which shows how much each of the columns are correlated. Specifically this diagram shows the Pearson correlation matrix, the values for which range from -1 to 1, and it can only tell the linear relationship between two variables. So

even if the correlation coefficient shows as 0, it does not mean that there is no linear relationship/dependency between variables but rather it could mean that there is a non-linear relationship that a Pearson correlation coefficient is failing to detect.



But it is more interesting to see how much each of the input features are correlated with the target output G3. Note that as of now, only the numerical features are used to represent this linear relationship and categorical ones are not included.

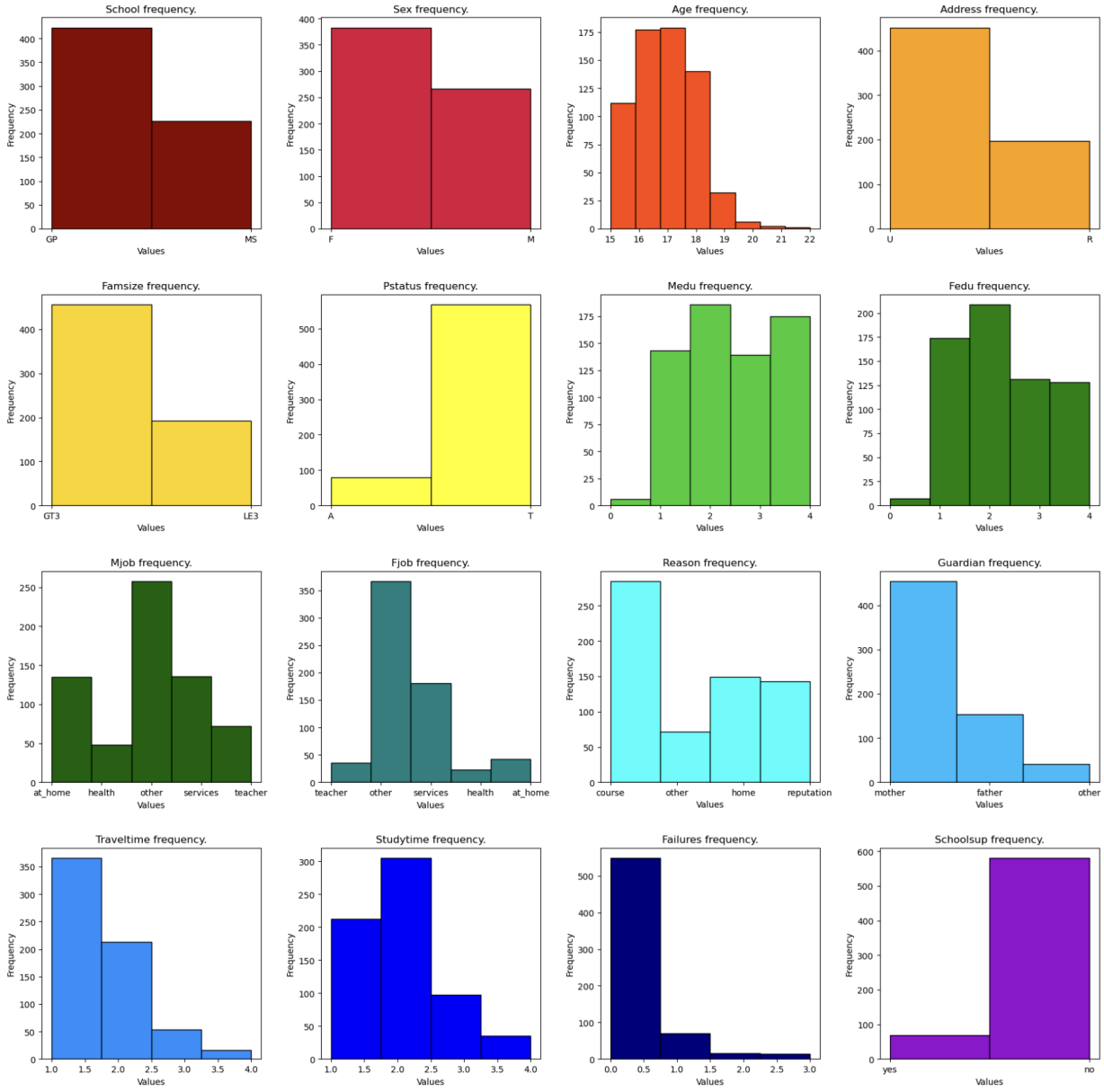


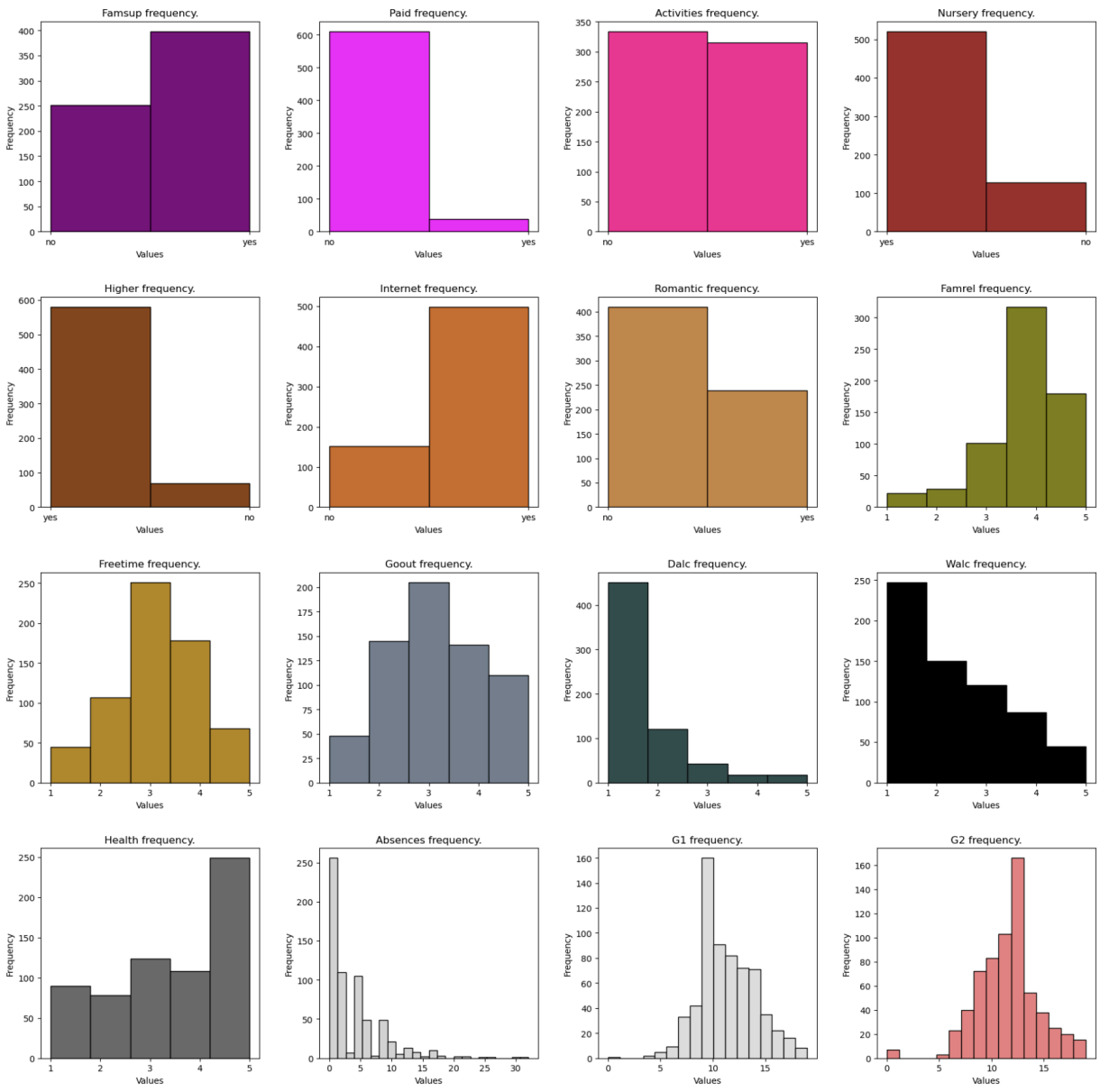
Here below are the summary statistics of the numerical features(16), this is created in order to get the high level aggregated values for each of the numerical columns, rather than seeing all the data at once.

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880	3.930663	3.180277	3.184900
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235	0.955717	1.051093	1.175766
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000

Dalc	Walc	health	absences	G1	G2	G3
649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
1.502311	2.280431	3.536210	3.659476	11.399076	11.570108	11.906009
0.924834	1.284380	1.446259	4.640759	2.745265	2.913639	3.230656
1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	2.000000	0.000000	10.000000	10.000000	10.000000
1.000000	2.000000	4.000000	2.000000	11.000000	11.000000	12.000000
2.000000	3.000000	5.000000	6.000000	13.000000	13.000000	14.000000
5.000000	5.000000	5.000000	32.000000	19.000000	19.000000	19.000000

Below are the frequency distribution diagrams for all input features, representing how often each of the unique values or groups of values appear in the dataset, so that we can see what values dominate and which does not. It also tells us whether the values are normally distributed or not. The different coloring is provided for individual diagrams so as not to confuse one with another, providing the visual difference.


School frequency.
Frequency
Values
GP
MS
Sex frequency.
F
M
Age frequency.
Address frequency.
U
R
Famsize frequency.
GT3
LE3
Pstatus frequency.
A
T
Medu frequency.
Fedu frequency.
Mjob frequency.
at_home
health
other
services
teacher
Fjob frequency.
Reason frequency.
course
other
home
reputation
Guardian frequency.
mother
father
other
Traveltime frequency.
Studytime frequency.
Failures frequency.
Schoolsup frequency.
yes
no



Dataset Preparation

As per the authors, the dataset does not have any missing, duplicate values nor outliers. As a self-reassurance, the checkups on any missing, duplicate, impossible value combinations were done and as expected, the dataset is clean. When it comes to feature-engineering, all the

categorical both ordinal and nominal input features were encoded using ordinal and one-hot-encoded methods respectively. After encoding, the data type of all the attributes converted into "*float64*". Since all numerical columns have meaningful and interpretable units and fall within certain range age (15-22), absences (0-32), failures (0-4), G1, G2 (0-20) and the rest fall within (0-4) or (1-5) range being a likert scale alike, there was no feature scaling applied for those numerically measurable columns of age, failures, absences, G1, G2. Also, all numerical inputs were already on a comparable and similar scale. Additionally, since the intention is to use simple tree-based models (*which do not require feature scaling as much, without depending on distance calculation*), no feature scaling for those five columns were applied. Without such scaling, model coefficients and feature selection results remain directly interpretable. However, feature scaling for those specific five columns can be considered in future works if non-linear models are to be used.

When it comes to feature selection, the choice was between statistical filter methods or wrapper methods, but given the limitations and benefits of both, the ultimate choice leaned towards embedded methods that have the best of both worlds. Embedded methods are more accurate than filter methods, but not as computationally expensive and not as overfitting as the wrapper methods. Since both wrapper and embedded methods are considered model-dependent, that is they can be applied only during the model training, the feature selection with embedded methods was done at the later stages of model training and not during the data preparation process. The dataset was splitted into 80% for testing and 20% for testing.

Justification of the choice of Machine Learning algorithms

For training, supervised machine learning was used due to its simple nature, higher accuracy and having more control compared to unsupervised learning. For hyperparameter tuning, the GridSearchCV was used to define the total number of decision trees to build for each of the tree-based models individually, otherwise known as "*n_estimators*". As defined earlier this is the regression problem, hence regression metrics, which are the MSE, MAE, R^2 and RMSE were used as evaluation metrics for each of the models. As discussed above, the embedded methods of feature selection were applied to each of the models separately and out of total 42 columns (*42 after encoding, 33 without encoding*), only 20 the most deterministic features were selected. As mentioned earlier there was no to minimal feature scaling done for the 5 numerical input features (*while leaving the rest as they are since they are all ranged between 1-5*), on the basis of that the tree-based algorithms, specifically Random Forest, XGBoost and LightGBM were trained on the same dataset and their MSE, MAE, R^2 and RMSE were evaluated. Below are the results of evaluation metrics of each model:

	Model	Selected_Features	MAE	MSE	RMSE	R2
1	RandomForest	20	0.756692	1.582632	1.258027	0.837707
2	XGBoost	20	0.748204	1.626565	1.275369	0.833202
3	LightGBM	20	0.780075	1.708428	1.307069	0.824807

Conclusion

Confirming the findings of Cortez and Gonçalves Silva (2008), by using past school grades (first and second periods), demographic, social and other school related data, all three models with the accuracies ranging from 74% to 83% could predict the student performance. The limitation of this project is that it boldly assumes that there is a linear relationship between the variables, while completely disregarding the possibility of existence of non-linear dependency. The future works can be tailored towards exploring and training with non-linear algorithms, which could potentially yield a much better performance and accuracy compared to linear models. Also, the further proper feature scaling can be applied to all numerical inputs before training on non-linear models.

References

Cortez, P. and Gonçalves Silva, A.M., (2008). *Using data mining to predict secondary school student performance.* [online] Available at: https://api.semanticscholar.org/CorpusID:16621299 [Accessed 22 November 2025].