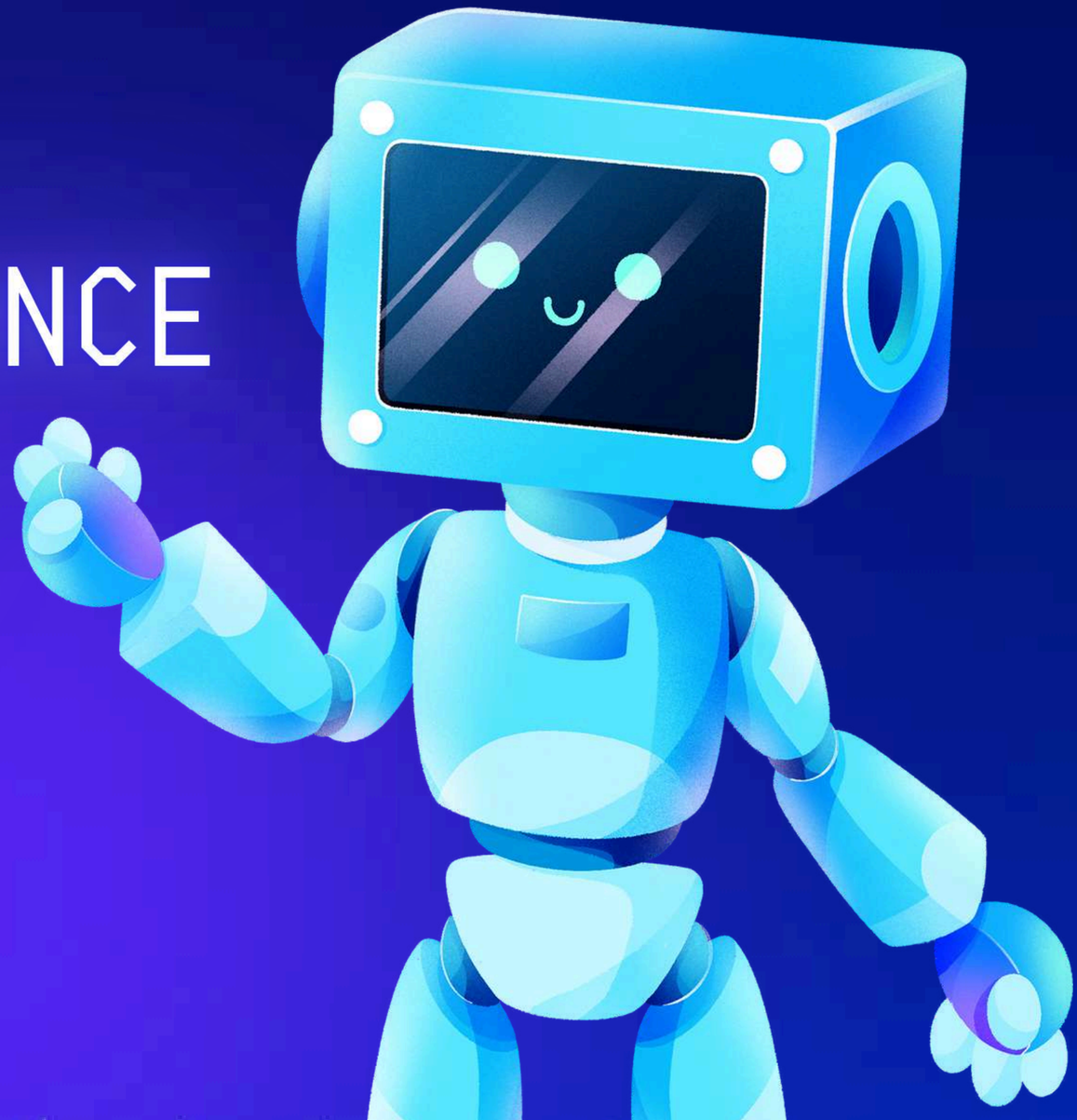
MACHINE LEARNING

STUDENT PERFORMANCE PREDICTION



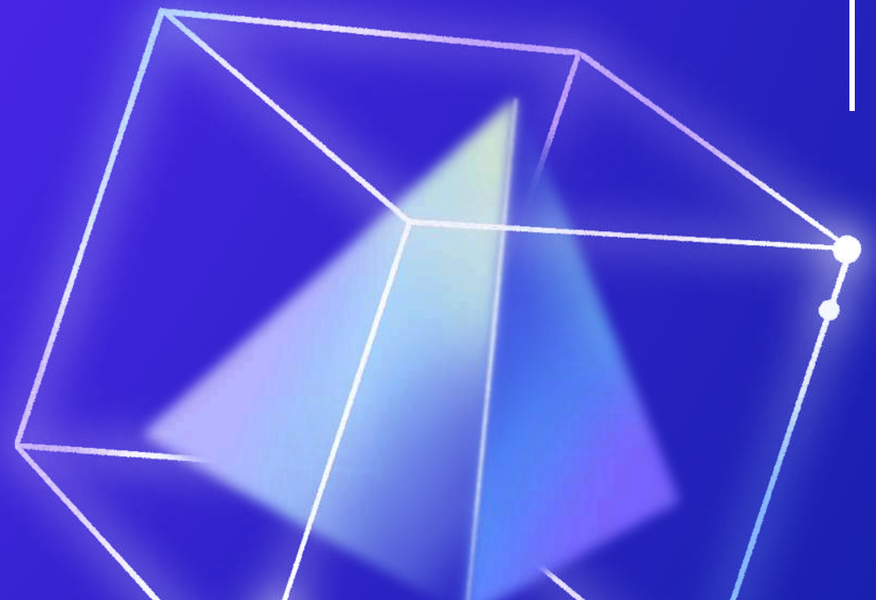
BY:

ZOBIYA - 2022A7PS0033U

SHARON -2023A7PS0248U

TABLE OF CONTENTS

- Introduction 01
- Data Set 02
- Attributes in Dat Set 03
- EDA 04
- Cleaning the Data Set 05
- Model Selection 06
- Comparison 07



INTRODUCTION

In the field of education, understanding and improving student performance is effective for teaching and learning.

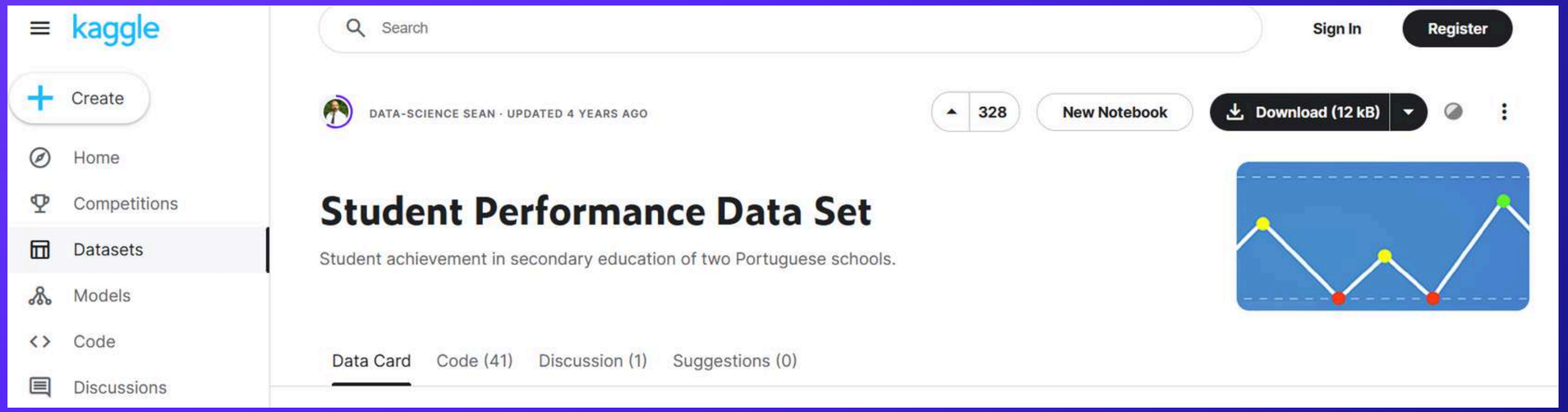
- The advent of advanced technologies, particularly in the realm of machine learning, educators now have a powerful tool at their disposal.

- Our project endeavors to harness the potential of machine learning algorithms to predict student performance, thereby providing educators with valuable insights to tailor their teaching strategies, identify at-risk students, and intervene proactively to ensure academic success for all.

DATA SET-STUDENT PERFORMANCE

The Data Set for Student Performance was taken from kaggle :
https://www.kaggle.com/datasets/larsen0966/student-performance-data-set

This data set talks about 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.



ATTRIBUTES IN DATA SET

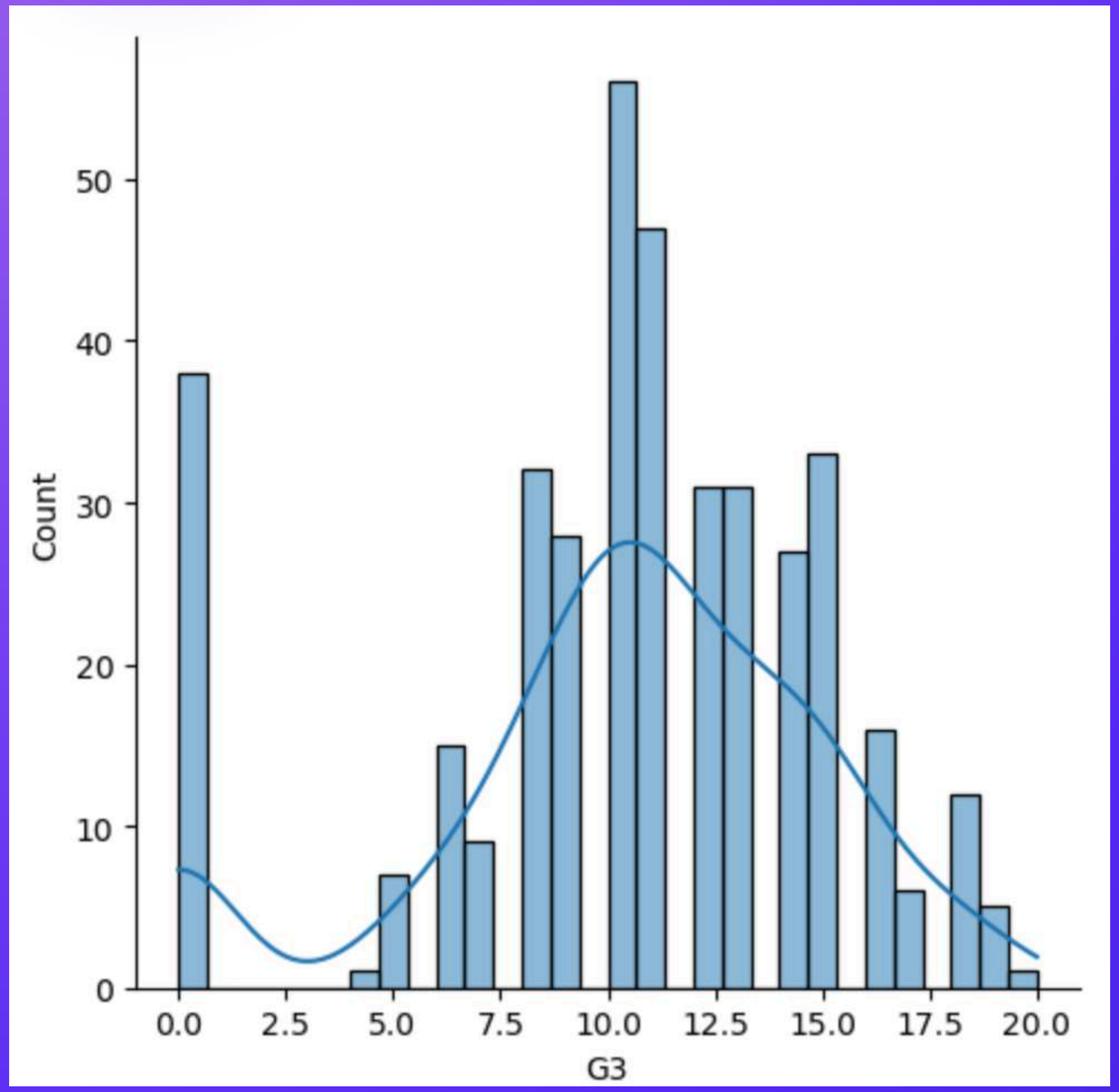
This data set consists of 33 columns which include the following:

- *school, sex, age, address* - These attributes deal with basic features of a student.
- *Famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian* - These attributes deal with the family of the student.
- *Traveltime, studytime, failures, schoolsup, famsup, paid*- These attributes deal with payement and study patterns of a student.
- *nursery, higher, internet* - These attributes deal with the school education
- *famrel,freetime,goout,activities* -These attributes deal with the recreational activities of a student.
- *health, absences* -These attributes deal with the health of student.
- *G1,G2,G3* - These attributes deal with the grades of the student.

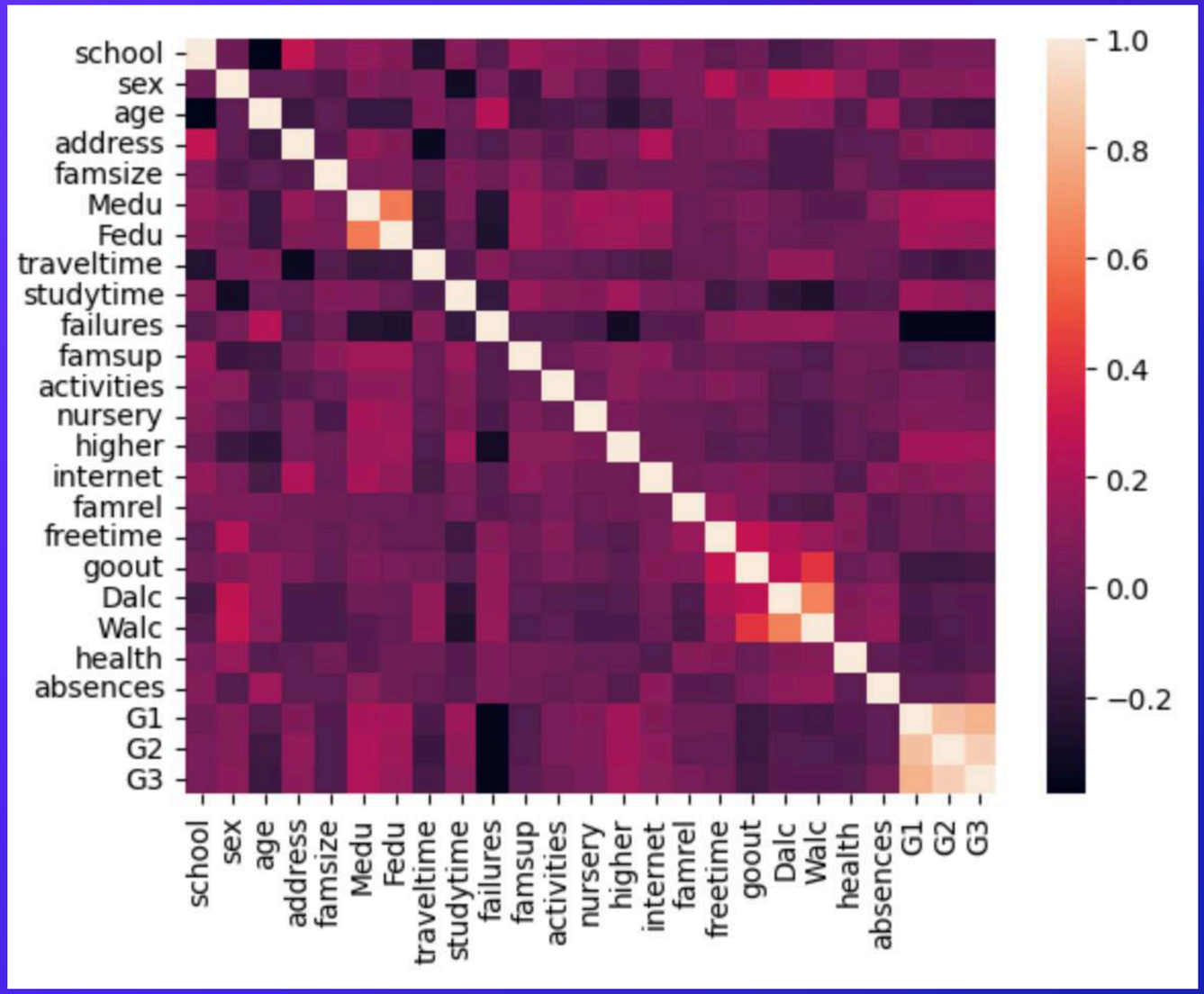
TARGET VARIABLE : G3 (Final Grade of the Student)

EDA

Helps you understand the given data set and figure out patterns within the data and rectify anomalies.



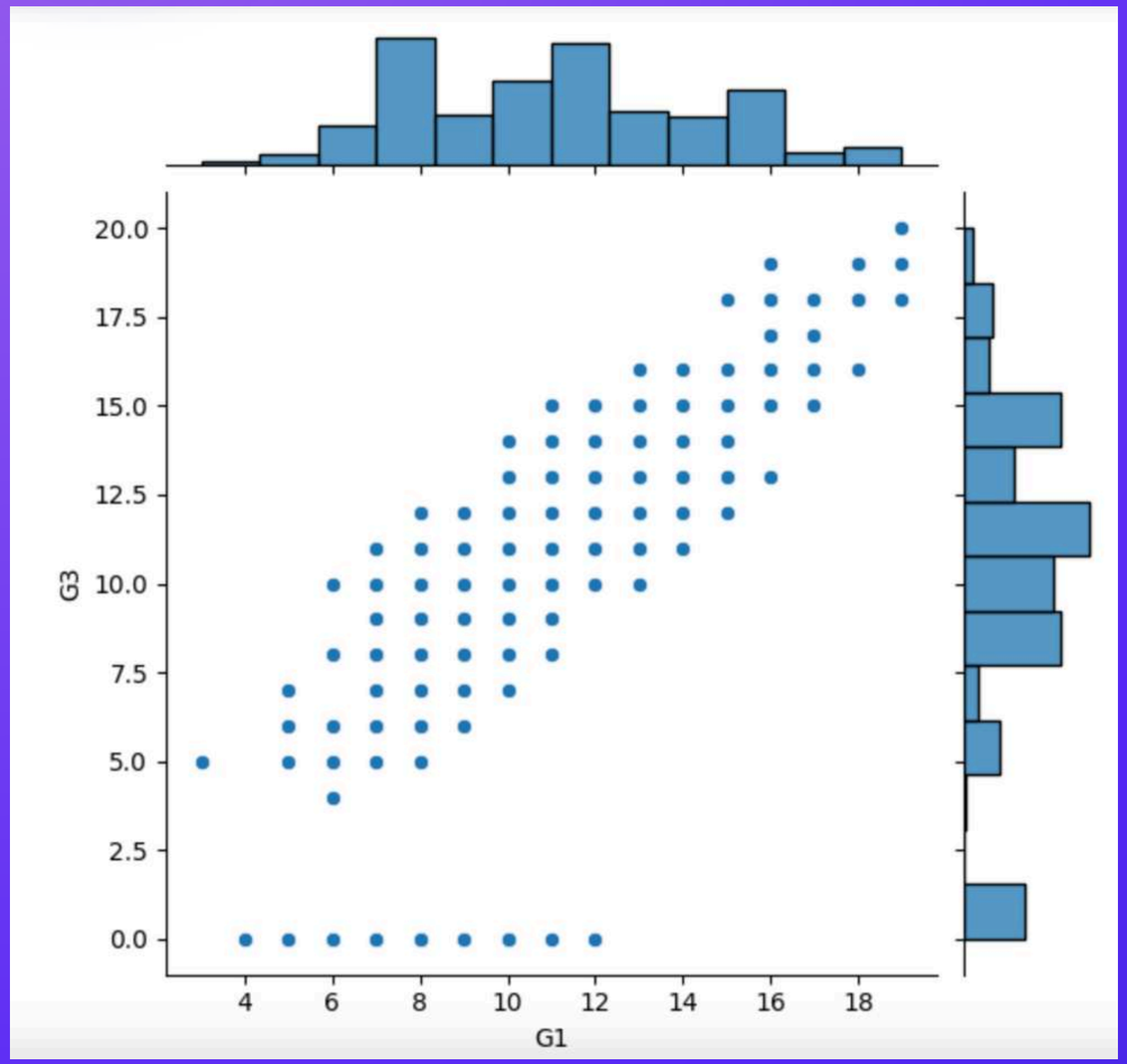
this graph shows the distribution of target variable G3



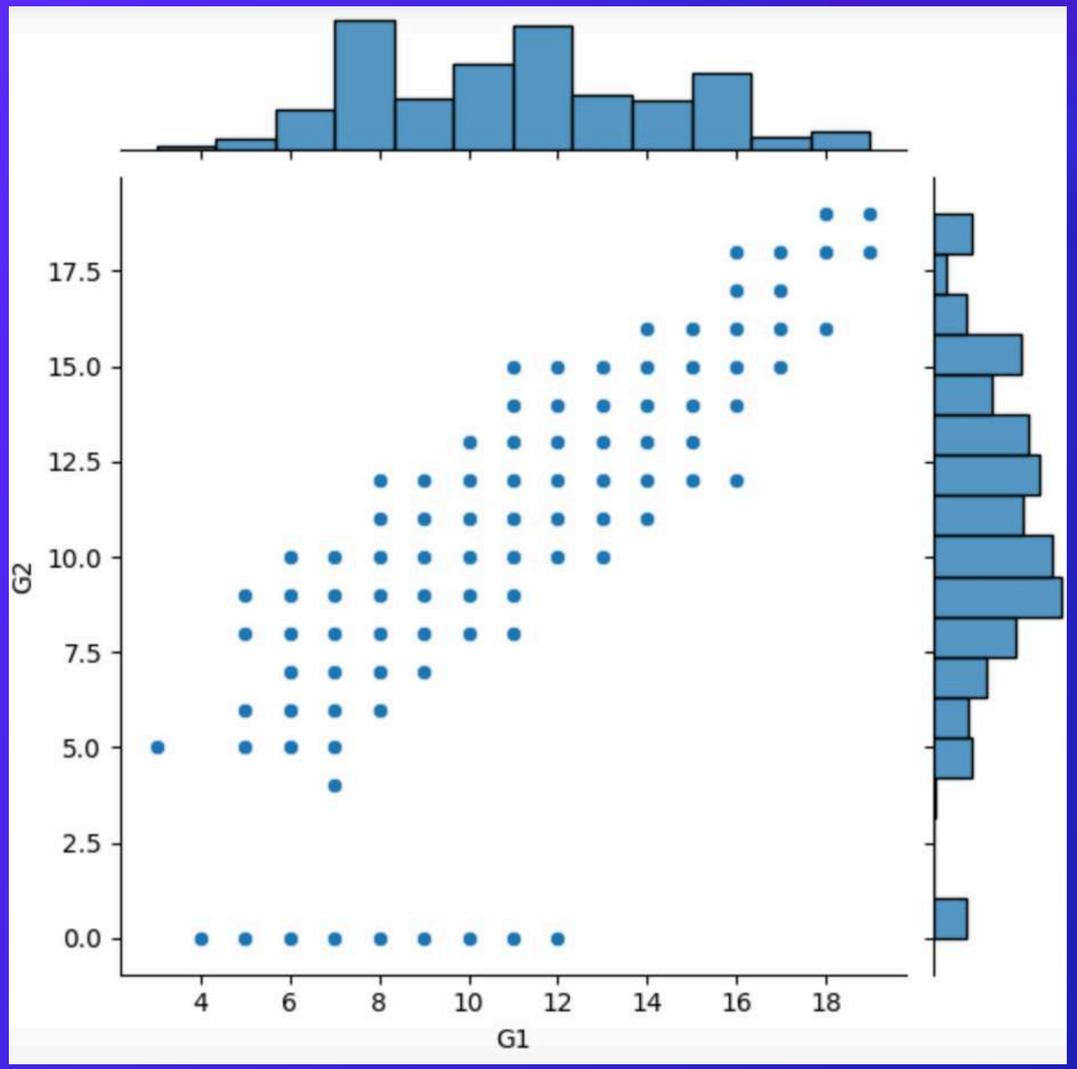
This heatmap shows the correlation between different columns in the dataset

EDA

Also reveals missing data and allows you to visualize the data, making it easier to understand.

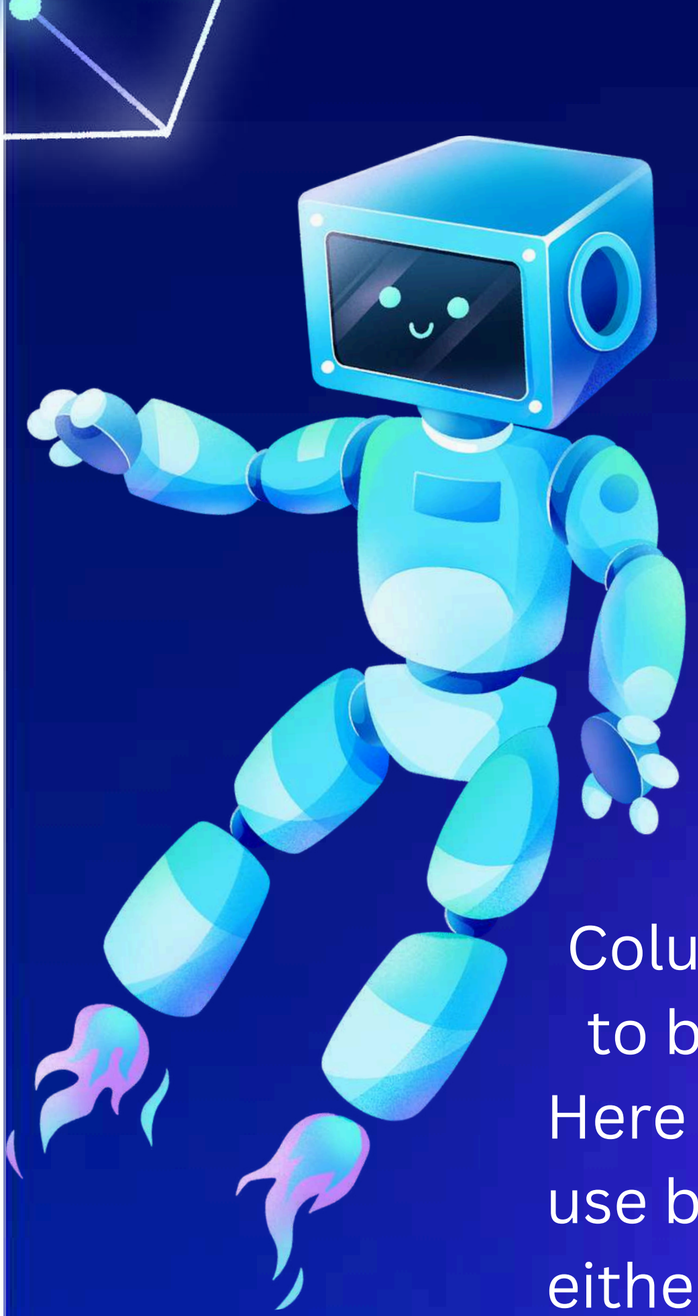


Correlation between columns G1 and G3



correlation between G1 and G2

BINARY ENCODING



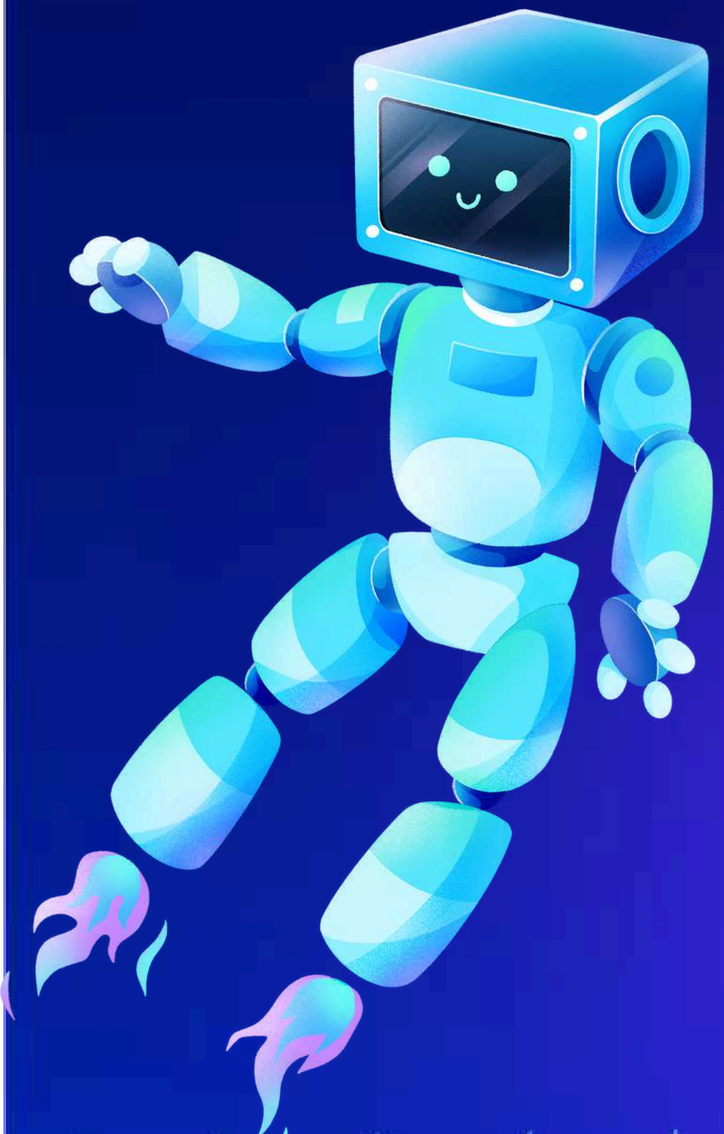
Process of converting non-numeric data to numeric form as machine learning algorithms ususally require numeric data.

```
studentperf['internet'] = studentperf['internet'].map({'yes': 1, 'no': 0})
studentperf['higher'] = studentperf['higher'].map({'yes': 1, 'no': 0})
studentperf['nursery'] = studentperf['nursery'].map({'yes': 1, 'no': 0})
studentperf['activities'] = studentperf['activities'].map({'yes': 1, 'no': 0})
studentperf['famsup'] = studentperf['famsup'].map({'yes': 1, 'no': 0})
studentperf['school'] = studentperf['school'].map({'GP': 1, 'MS': 0})
studentperf['sex'] = studentperf['sex'].map({'M': 1, 'F': 0})
studentperf['address'] = studentperf['address'].map({'U': 1, 'R': 0})
studentperf['famsize'] = studentperf['famsize'].map({'GT3': 1, 'LE3': 0})
```

Columns like sex that have values male and female can be mapped to binary values 1 and 0 to make it understandable to the model. Here since there are only 2 unique values for each column we can use binary encoding but if there are more unique values we can use either label encoding or one-hot encoding.

REMOVING UNNECESSARY COLUMNS

Important to remove irrelevant information from dataset as it may affect the performance of the model, and it will result in faster training time.



```
studentperf.pop("reason")
studentperf.pop("guardian")
studentperf.pop("schoolsup")
studentperf.pop("paid")
studentperf.pop("romantic")
studentperf.pop("Pstatus")
studentperf.pop("Mjob")
studentperf.pop("Fjob")
```

MODEL SELECTION

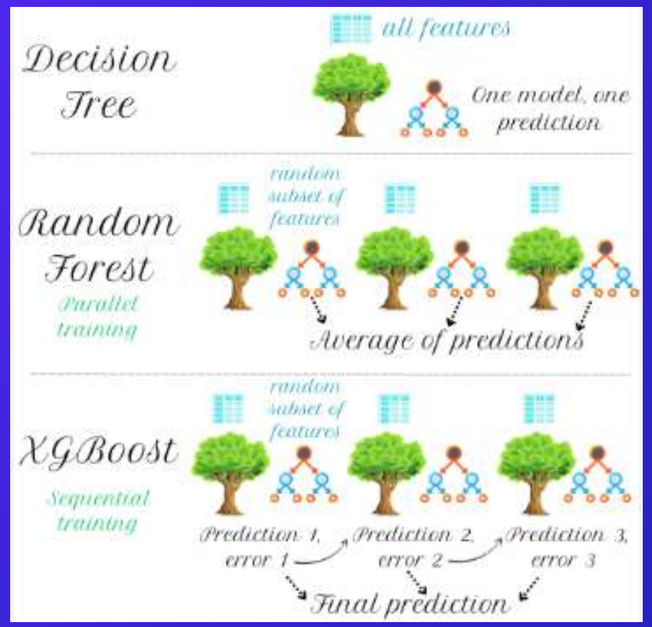
01

There are many Machine learning algorithms available but we should select a model that gives a high accuracy in determining the target variable.

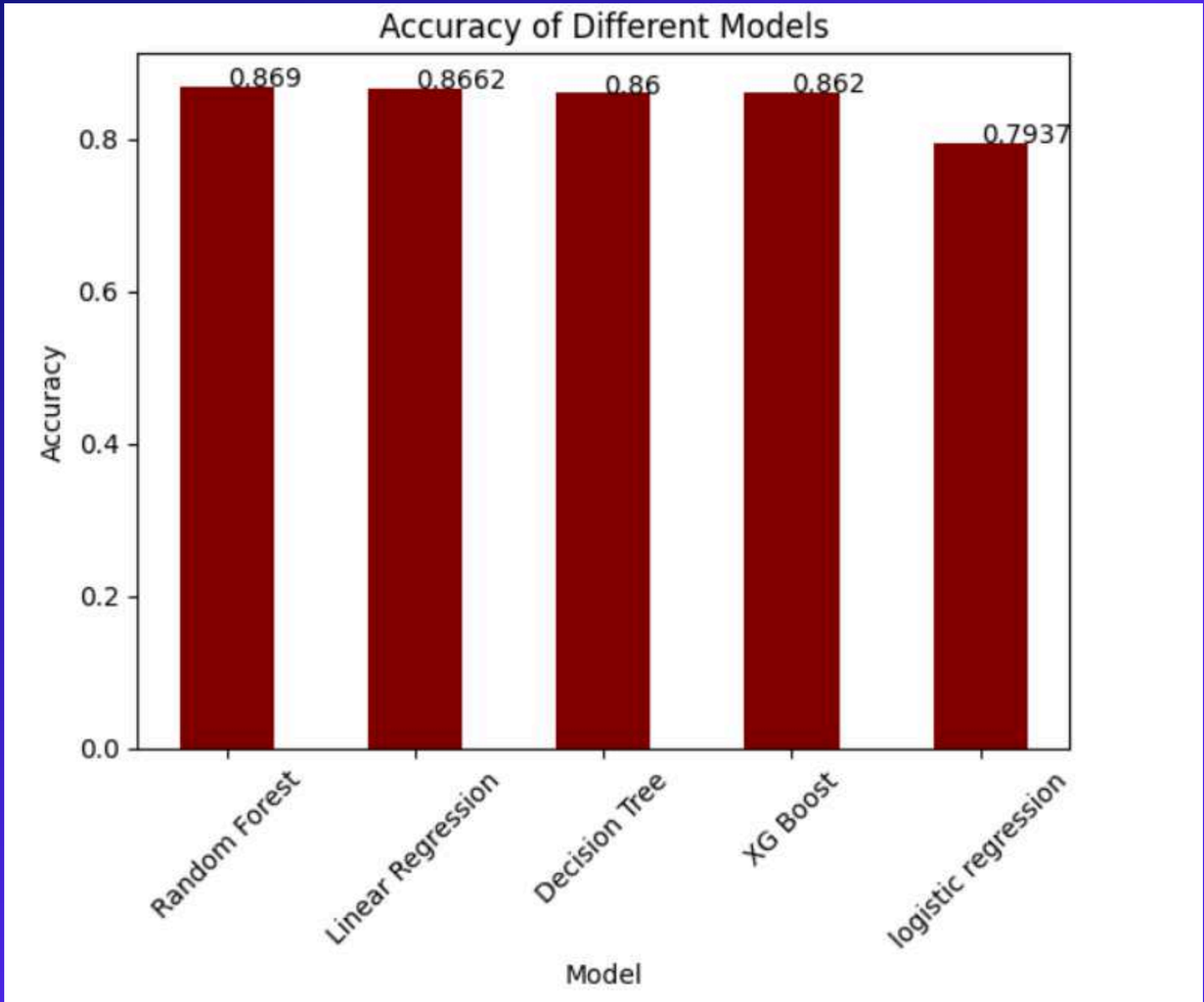
02

We tested 5 different Machine learning models and compared their R2 scores to determine the most effective model amongst them.These models include:

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- XG Boost



COMPARISON



As seen from the figure, Random Forest has the most accuracy in comparison to other models and logistic regression has the least accuracy. Hence , the best model to work on the data set of Student Performance Prediction is Random Forest

GITHUB LINK:
https://github.com/zobiyaFathima/zobiya-codes

STREAMLIT

User input features

School

GP

Age

16

15 22

G1

13

3 19

G2

17

0 19

Mother Education

2

0 4

Father Education

4

0 4

Internet

1

0 1

Familyrelation

5

1 5

freetime

2

1 5

Health

5

1 5

STUDENT PERFORMANCE PREDICTION APP

This app predicts the performance of a student based on various features pertaining to a student.This app finds the final grade G3 of a student.

Data set obtained from kaggle :

https : //www.kaggle.com/datasets/larsen0966/studentperformancedataset

User input Features

	ltime	studytime	failures	activities	internet	famrel	freetime	health	absences	G1	G2
0	1	4	1	1	1	5	2	5	10	13	17

Predicted Student Score-G3

value
17.6

THANK YOU!