Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Predicting Student Performance

Supervised By:

Md. Talib

Submitted By:

Varsha Kumari, 2210992514(G29)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

Abstract

In the realm of education, understanding and predicting student performance is crucial for effective personalized learning, early intervention, and academic support. This project proposes a machine learning framework to predict student performance based on various academic and non-academic factors. Leveraging a dataset encompassing student demographics, previous academic records, socio-economic indicators, and behavioral attributes, the proposed model employs a combination of supervised learning algorithms such as regression and ensemble techniques. Feature engineering techniques are applied to extract meaningful insights from the raw data, while model selection optimize predictive accuracy. Furthermore, interpretability methods are integrated to provide insights into the factors influencing student performance. The efficacy of the developed model is validated through performance metrics such as Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, R-squared , Adjusted R-squared. The outcomes of this project not only facilitate early identification of students at risk of academic underperformance but also provide actionable insights for educators and policymakers to enhance educational outcomes through targeted interventions and support mechanisms.

Table of Contents

1. Introduction

1.1 Background

The background of this project titled as "Predicting Student Performance" is to use machine learning algorithms to predict how well or poor, students are likely to perform in exams.

The dataset contains information about high school students' performance, influenced by several factors. It is consist of marks secured in various subjects by students in the United States.
This dataset is often used for educational data analysis and exploring factors influencing students' academic performance.

- The goal is to identify students who may need extra help or intervention before they reach the exam stage.
- It aims to improve the quality of education and decrease failure rates by providing insights into factors that influence student performance.
- The COVID-19 pandemic has highlighted the need for innovative solutions, as studies show student performance has decreased since then.
- Machine learning can be used to analyze previous student records and build models to predict future performance and guide decision-making.
- Factors like, parental-level of education, test preparation course, race/ethnicity, lunch type, etc can affect student performance.
- Predicting student performance is a challenging task in education systems, but machine learning provides an opportunity to tackle it.

So in summary, the background is to leverage machine learning to gain insights into student performance, identify at-risk students early, and ultimately improve educational outcomes, especially in light of recent challenges like the pandemic. The goal is to provide an innovative, data-driven solution to a longstanding problem in education.

1.2 Objectives

- Dentifying Factors Affecting Performance: Analyze various factors such as socio-economic status, demographic characteristics, parental involvement, study habits, attendance, extracurricular activities, and teaching methodologies to understand their impact on student performance.
- Understanding Academic Trends: Examine trends in student performance over time, across different subjects, grades, or cohorts to identify patterns, areas of improvement, and potential disparities.
- Predicting Student Outcomes: Develop predictive models to forecast student performance based on historical data and identified factors. These models can help in identifying at-risk students and implementing timely interventions to improve outcomes.
- Evaluating Interventions: Assess the effectiveness of interventions, programs, or policies implemented to improve student performance. Evaluate whether the interventions are achieving their intended outcomes and adjust strategies as needed

1.3 Significance

Predicting student performance through machine learning has several significant implications:

- Early Intervention: By analyzing various factors that influence student performance such as demographics, past academic records, and behavioral patterns, machine learning models can identify students who are at risk of underperforming. Early identification allows educators to intervene and provide additional support to these students, potentially preventing academic failure.
- Personalized Learning: Machine learning algorithms can analyze individual learning styles and preferences to tailor educational materials and teaching methods accordingly. This personalized approach can enhance student engagement and motivation, leading to improved performance.

- Resource Allocation: Schools and educational institutions can use predictive models to allocate resources more efficiently. By identifying students who are likely to require additional support, educators can allocate resources such as tutoring, counseling, or specialized programs to where they are most needed.
- Curriculum Development: Analysis of student performance data can provide insights into the effectiveness of different teaching methods and curriculum components. Machine learning can help identify areas where curriculum adjustments are needed to better align with student learning needs and improve overall academic outcomes.
- Policy Making: Predictive analytics in education can inform policy decisions at the institutional, district, or even national level. By understanding the factors that contribute to student success or failure, policymakers can implement targeted interventions and reforms to improve educational outcomes on a broader scale.

Overall, machine learning projects predicting student performance have the potential to revolutionize education by enabling proactive intervention, personalized learning experiences, and evidence-based decision-making.

2. Problem Definition and Requirements

2.1 Problem Statement

First, I wanted to list the questions I wanted to answer throughout the project.

I believe 60% success of a project depends on the understanding of the problem.

After going through the dataset thoroughly I came up with the following problems

which I wanted to solve. They are:

- Among the top scorer, are they equally good in math, reading, and writing?
- Among male and female students who are performing better?
- What is the distribution of reading, writing and math scores?
- What ethnicity is standing out?
- Does the parent's education level have any effect on their performance?
- What effect does 'test preparation' have on their performance?
- Does lunch type affect their performance?

2.2 Software Requirements

System software: Windows OS

Application Software: Google Colaboratoy , Google Chrome, Microsoft Excel Worksheet

2.3 Dataset link

https://drive.google.com/drive/folders/1cSDuqcKJPZhHxSS8ud4jQYgLpDjQpeQI?usp=sharing

3. Methodology

3.1 Support Vector Machines

Here's a simplified explanation of how prediction works in SVR:

- Training Phase: In the training phase, SVR learns to approximate the relationship between input variables and continuous target variables by finding a hyperplane (or hyperplanes) that best fits the training data within a specified margin of tolerance.
- Optimal Hyperplane: Similar to SVM, SVR aims to find the hyperplane(s) that maximize the margin while still fitting as many data points within a certain margin of error (epsilon).
- Loss Function: SVR uses a loss function that penalizes errors beyond a certain margin of tolerance (epsilon). Common loss functions include the epsilon-insensitive loss function or the Huber loss function.
- Kernel Trick (optional): As in SVM, SVR can utilize a kernel trick to map the input data into a higher-dimensional space, which can help capture more complex relationships between the input variables and the target variable.
- Prediction: To predict the target value of a new data point, SVR evaluates the learned function at that point. The predicted value is determined by the distance between the new data point and the hyperplane(s) learned during training, with consideration of the margin of tolerance (epsilon).
- Regression: Once the data is mapped to a higher-dimensional space (if necessary), SVR applies the learned function to predict the target variable for new data points.

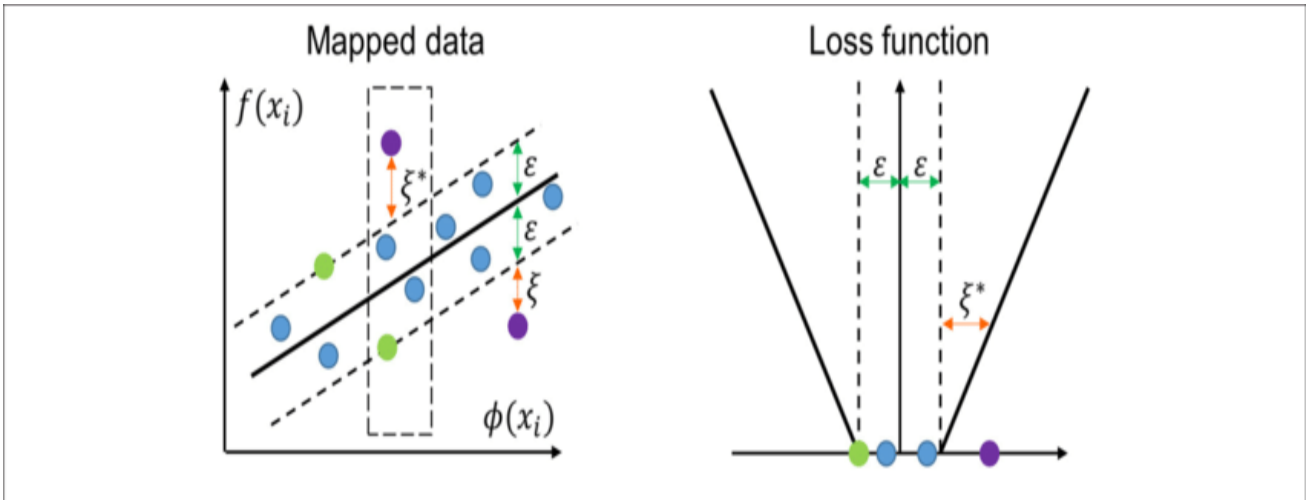

Fig.3.1.1-Support Vector Regression using Linear Kernel

3.2 Decision Trees

Decision Tree Regressor is a supervised learning algorithm used for regression tasks. It works by recursively partitioning the feature space into smaller regions and making predictions based on the average of the target variable within each region.

Here's a brief overview of how it works:

- Splitting Criteria: Decision trees make decisions by splitting the feature space into subsets based on certain criteria. The criteria typically used for splitting include minimizing variance or maximizing information gain.
- Recursive Partitioning: Starting from the root node, the decision tree recursively splits the data into smaller subsets based on the chosen splitting criteria. Each split creates branches that lead to child nodes.
- Leaf Nodes and Predictions: The process continues until a stopping criterion is met, such as reaching a maximum tree depth or having a minimum number of samples in each leaf

node. At this point, the terminal nodes are called leaf nodes. Each leaf node contains a prediction value, typically the mean or median of the target variable within that node's subset.

- Prediction: To make a prediction for a new data point, the decision tree traverses the tree from the root node down to a leaf node, following the path determined by the feature values of the data point. The prediction for the data point is then based on the prediction value stored in the leaf node.
- Model Interpretability: One of the key advantages of decision trees is their interpretability. The structure of the tree can be visualized and understood easily, making it intuitive to interpret how the model makes predictions.

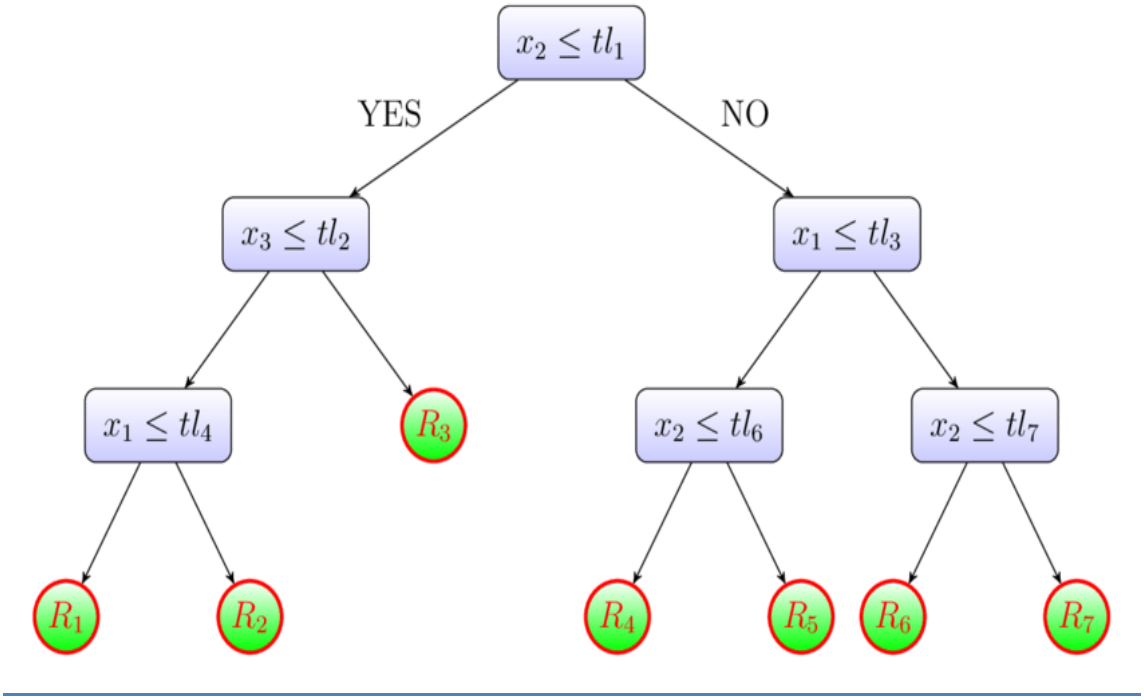

Fig.3.2.1-Decision Tree Regressor

3.3 Random Forests

Prediction in a Random Forest Regressor involves aggregating the predictions of multiple decision trees trained on different subsets of the data and features to produce a final prediction for the target variable. This ensemble approach typically results in robust and accurate predictions.

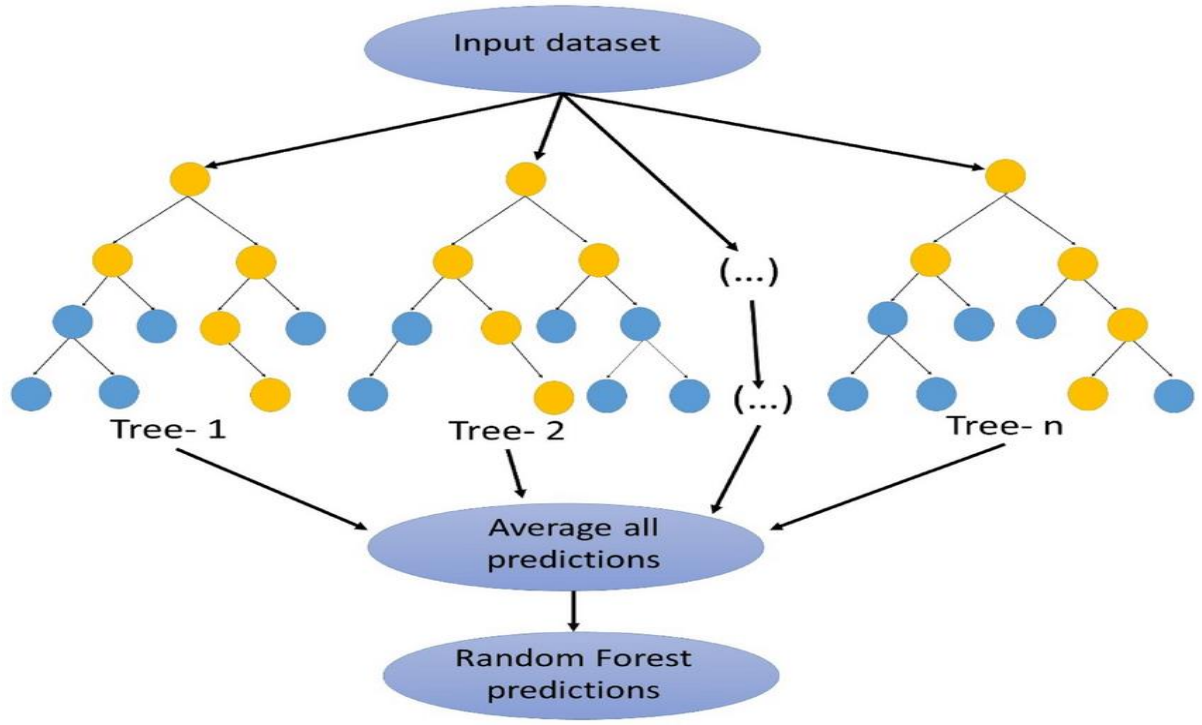

Fig.3.3.1-Random Forest Regressor

4.Results

4.1. The insight(s) found from the pie chart below

- More than 50% of the high scorers belong to Group D followed by the high scorers of Group E. Thus we can say that students from Group D and Group E are far more likely to perform great in exams.

- Group C has the least number of high scorers and Not a single student from Group A is a high scorer. Thus we can say that students from Group C are far less likely to perform great in exams whereas students from Group A has near to null possibility of performing great in exams.

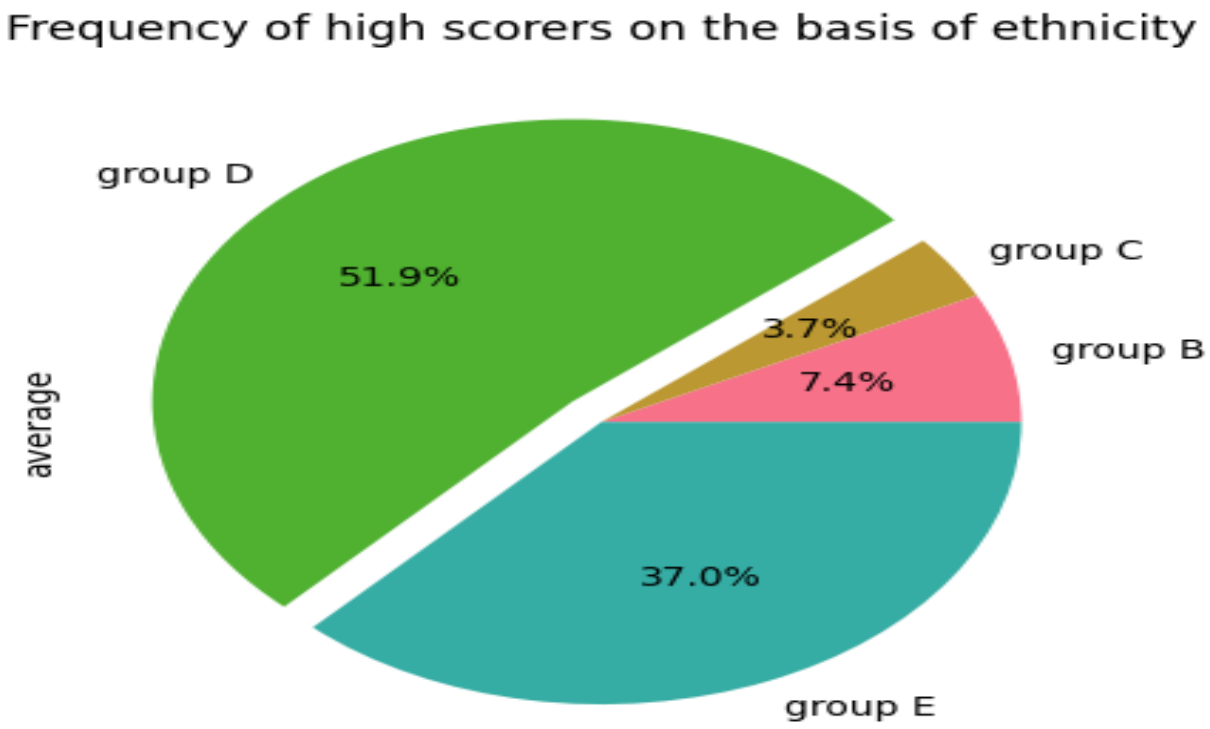

Fig.4.1.1-Pie Chart

4.2. The insight(s) found from the countplot below

Greatest number of low scorers belong to Group C.

Thus we can say that students from Group C are far more likely to perform poorly in exams.

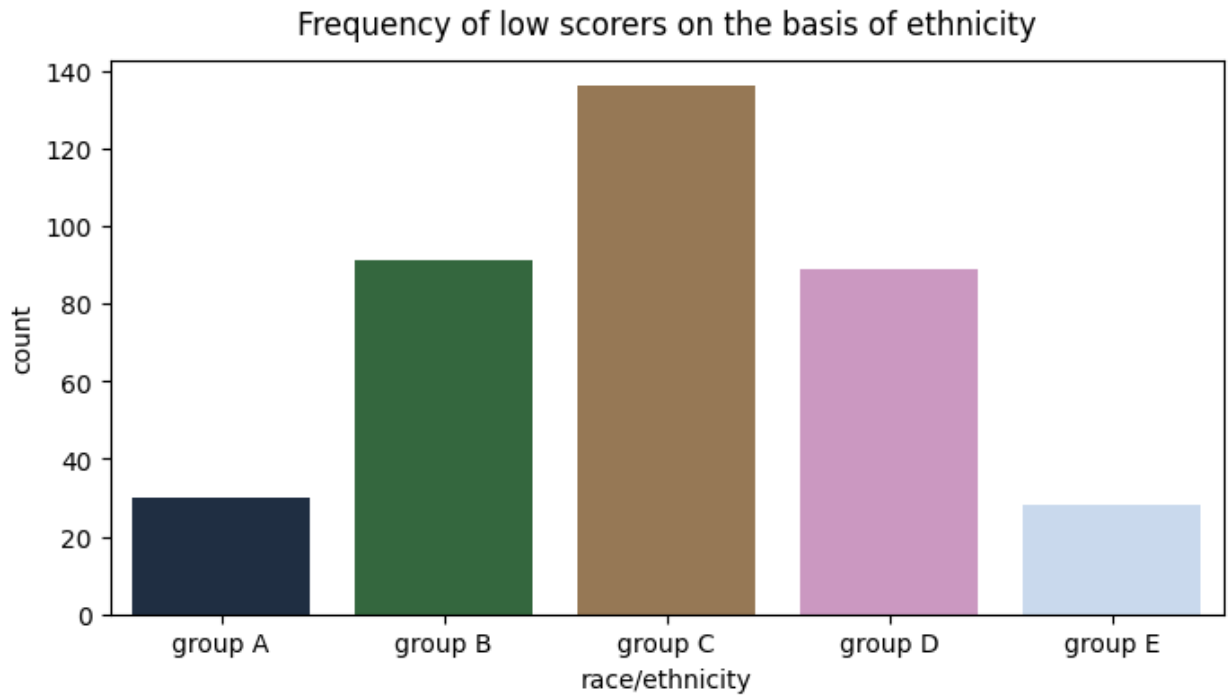

Fig.4.2.1-Countplot

4.3. The insight(s) found from the KDEplots below

- Male students have performed great in maths in comparison to the female students.
- Female students have performed great in reading and writing in comparison to the male students.

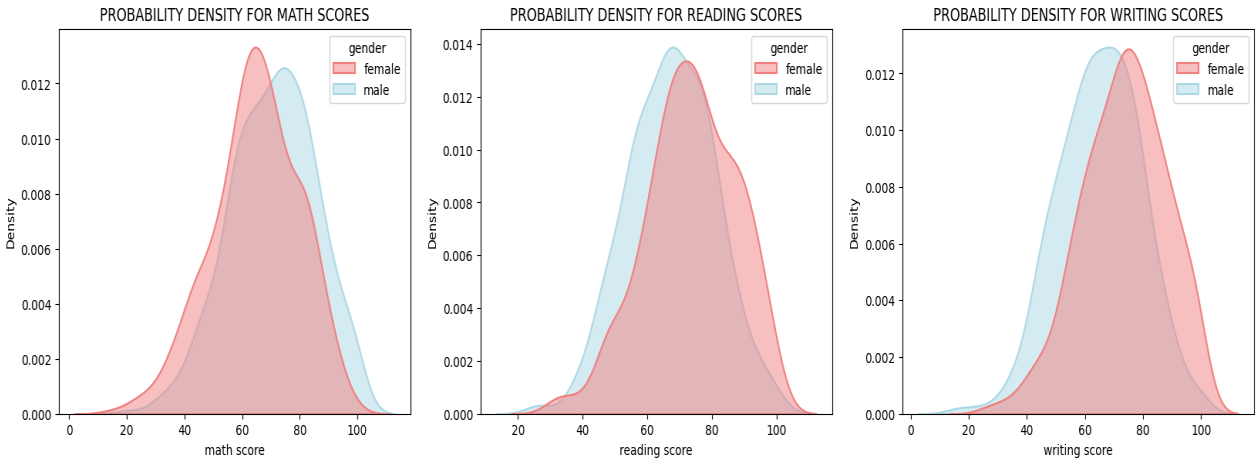

Fig.4.3.1-KDE Plots

4.4. The insight(s) found from the swarmplots below

- Test preparation course does not have much impact on the math score of students.
- But for the ones who have completed the test preparation course they are less likely to fall in the category of bottom scorers.
- Few of the students who did not opted for or complete test preparation course are the bottom scorers.

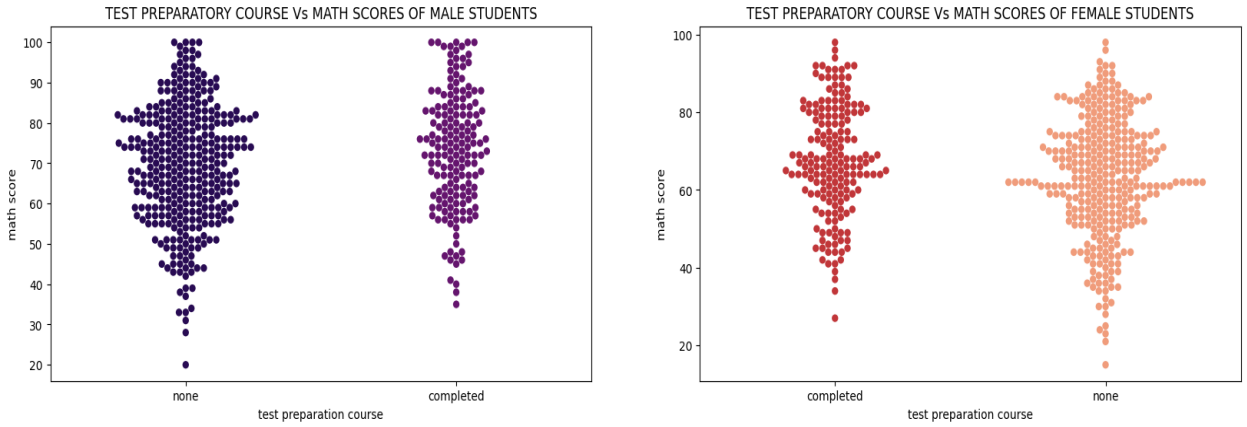

Fig.4.4.1-Swarmplots

4.5. The insight(s) found from the swarmplots below

- Male students(more in number) whose parents hold associate's degree or completed college , have fairly good math scores than others with different parental level of education.
- Male students(less in number) whose parents hold master's degree ,have scored good, none of them is a bottom scorer in maths.
- Male students(more in number) whose parents just completed their high school studies , are the bottom scorers in maths,though some of them have scored good marks as well.

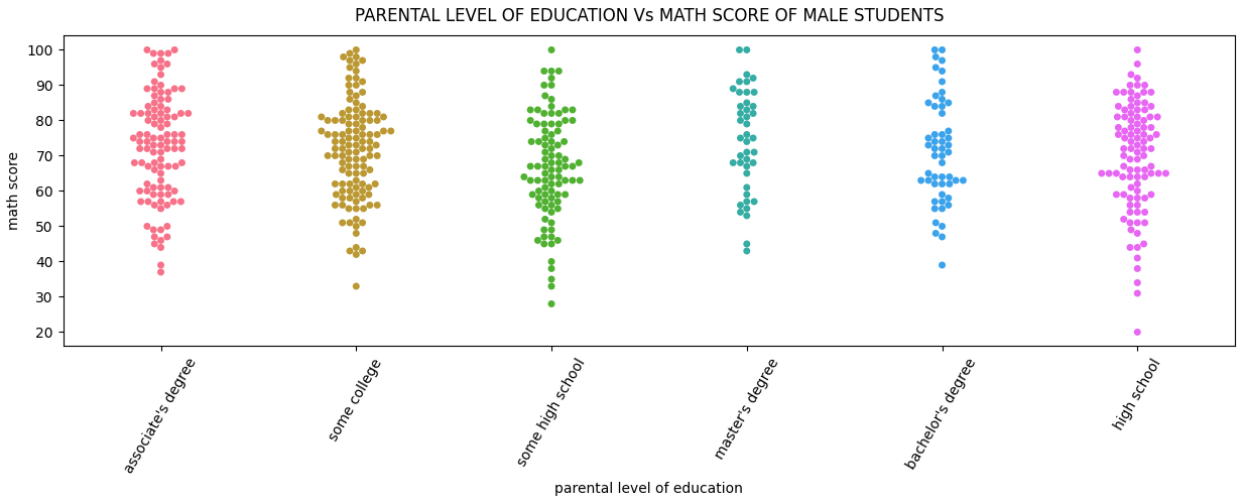

Fig.4.5.1-Swarmplot

4.6. The insight(s) found from the histplots below

- The students who have completed test preparatory course ,are the top scorers in reading and writing.
- The students who have not opted for any test preparatory course ,are the mid and bottom scorers in reading and writing.
- Female students are proportionally higher than the male students when we talk about the top scorers in reading and writing.

Fig.4.6.1-Histplots

4.7. The insight(s) found from the barplot below

- Female Students whose parents hold bachelor's, master's or associate's degree , are fairly the top scorers in writing and reading.

- Female Students whose parents have just completed their high school studies , are the mid/bottom scorers in writing and reading.

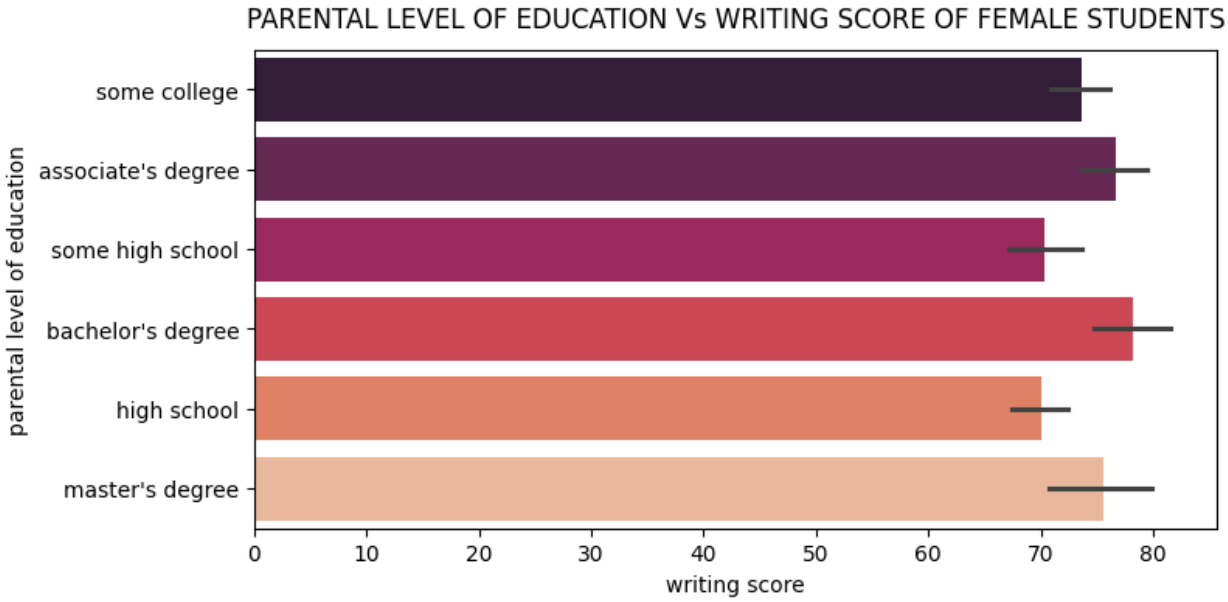

Fig.4.7.1-Barplot

4.8. The insight(s) found from the histplot below

- The count of female students scoring top average scores is higher than the male students.

- The count of male students scoring mid to bottom average scores is higher than female students.

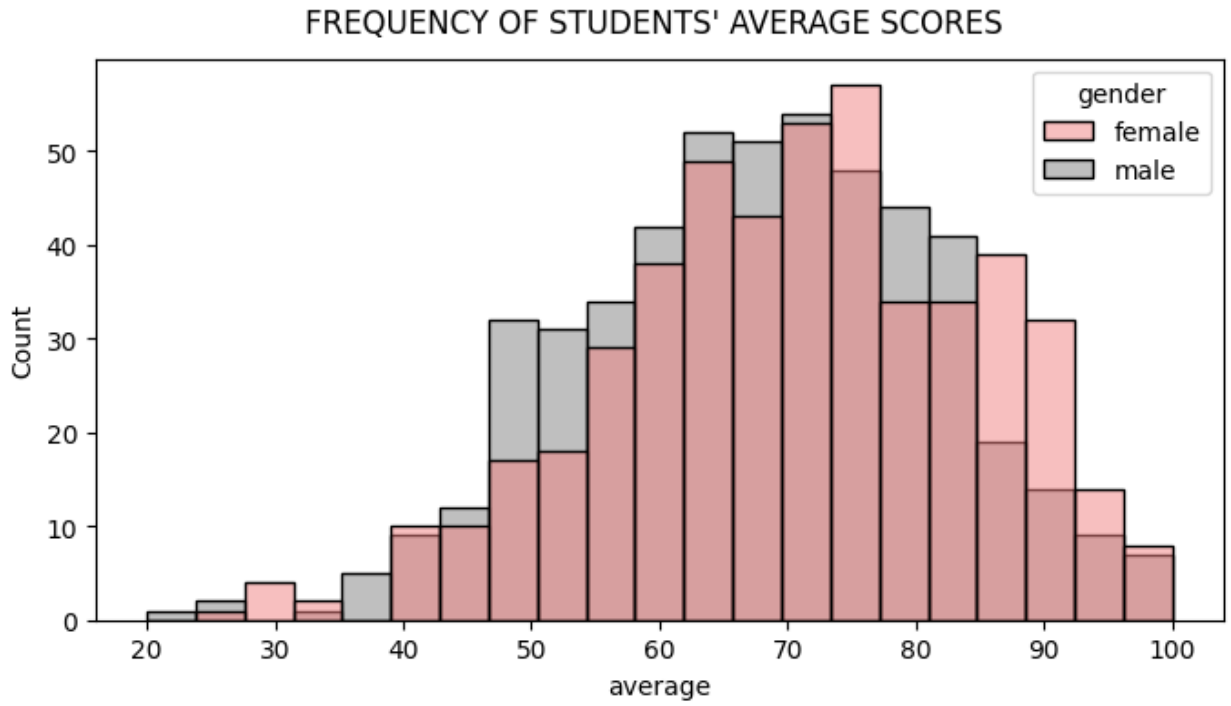

Fig.4.8.1-Histplot

4.9. The insight(s) found from the boxplots below

- Students who completed test preparation course , have secured higher average scores than the ones who did not opted for the course itself.

- Female students outshine male students in scoring higher average scores in both the conditions.

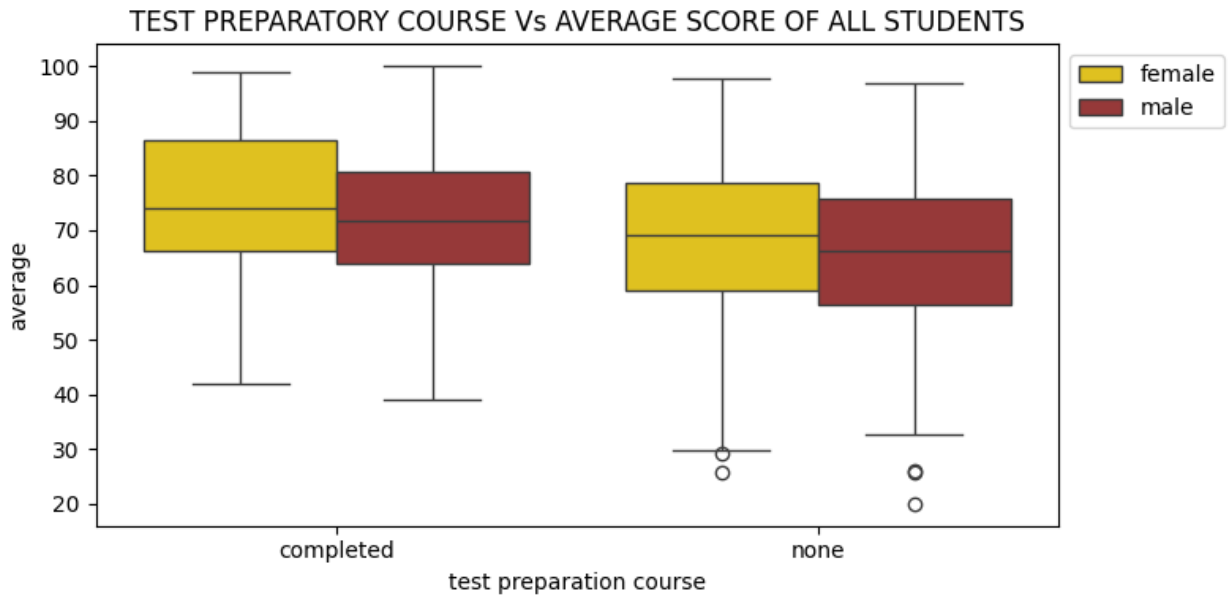

Fig.4.9.1-Boxplot

4.10. The insight(s) found from the barplot below

- Students whose parents hold associate's or bachelor's degree have secured higher average scores than others.

- Students whose parents have just come from high school , have relatively lower average scores than others.

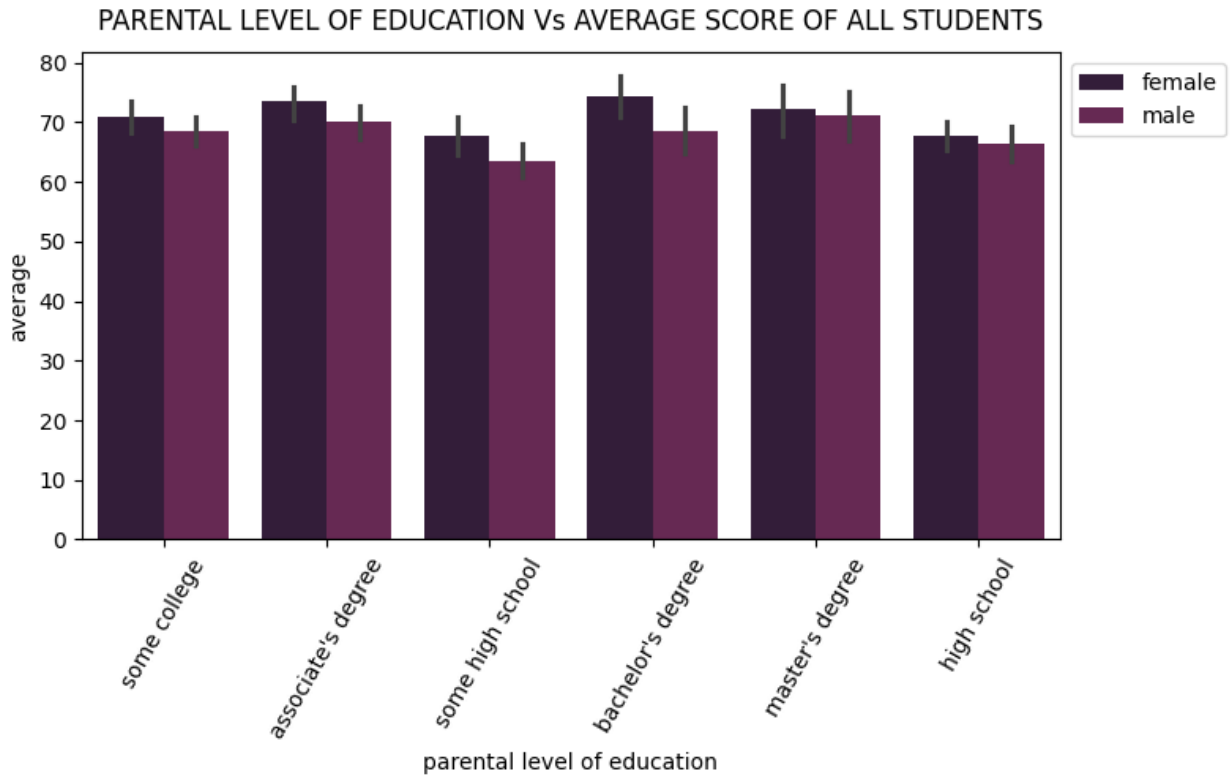

Fig.4.10.1-Barplot

4.11. The insight(s) found from the violinplots below

- Students belonging to Group D and Group E have secured higher average scores than others.

- Students belonging to Group C have secured relatively lower average scores than others.

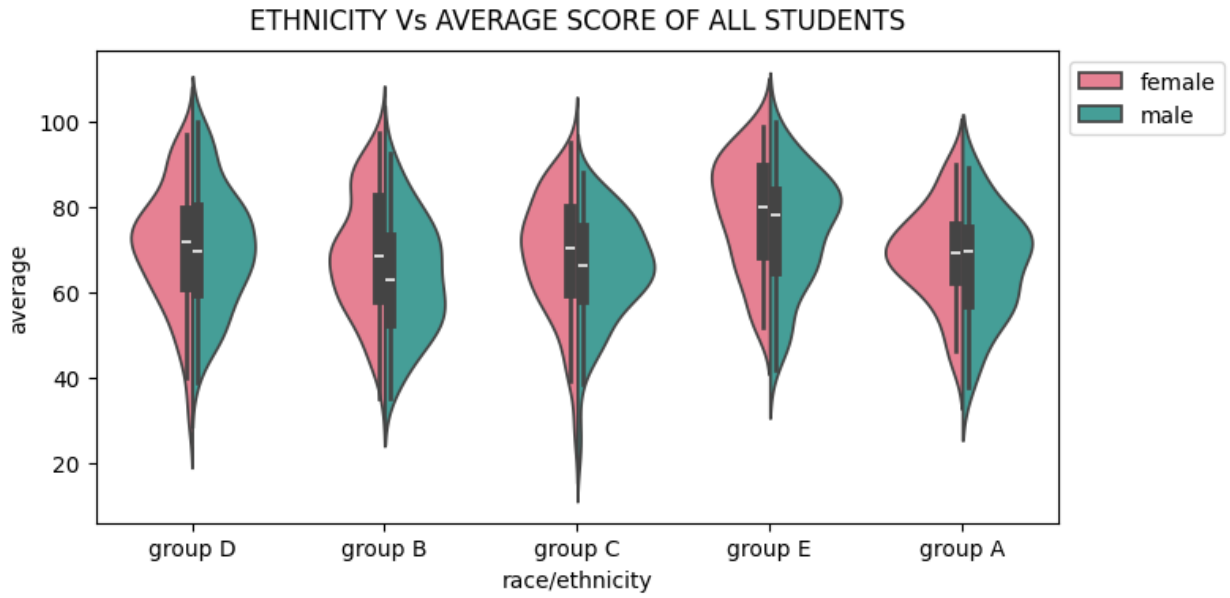

Fig.4.11.1-Violinplot

4.12. The insight(s) found from the swarmplots below

- The density of Students who had standard lunch and secured higher average scores is more than the ones who had free/reduced lunch.

- The density of students taking free lunch and scoring better average scores is little lower than the ones who had standard lunch.we can say that the type of lunch has no significant impact on average scores of students.

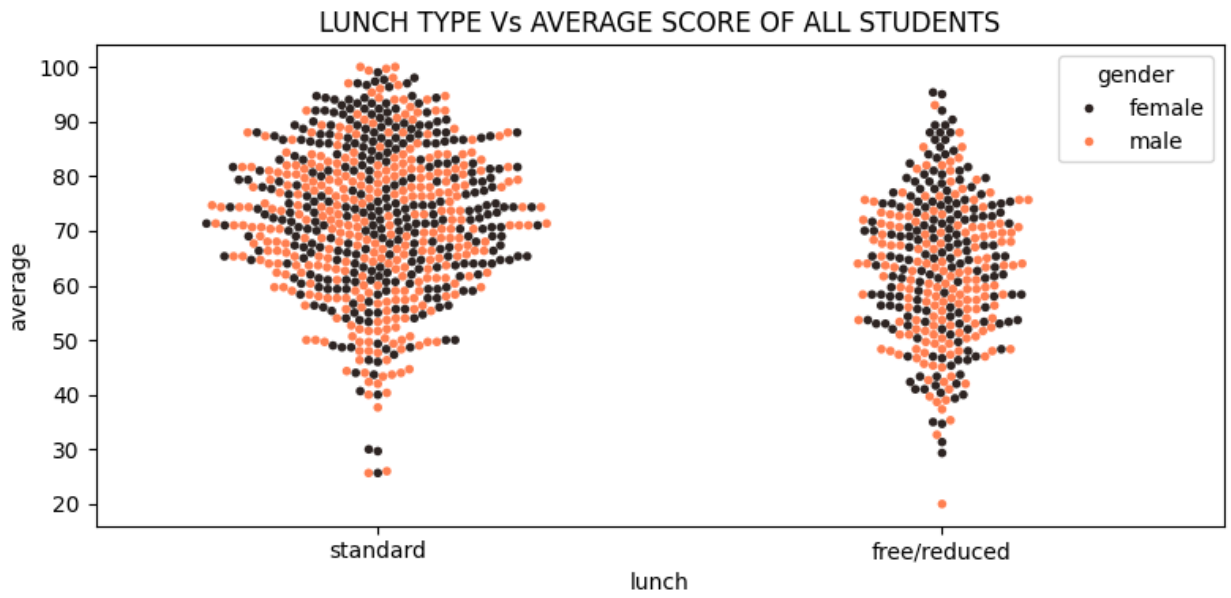

Fig.4.12.1-Swarmplot

4.13. The insight(s) found from the heatmap below

- Students whose parents hold bachelor's or associate's degree are more likely to score greater average scores than others.

- Students whose parents have just come from some high school,are a little more likely to score lesser average scores than others.

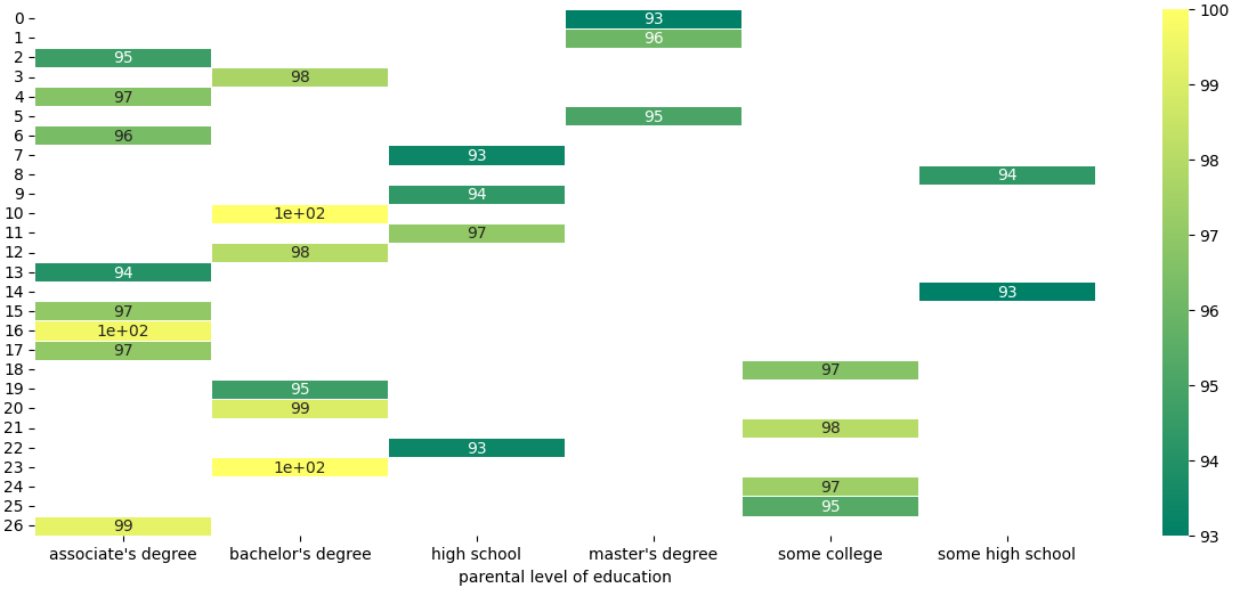

Fig.4.13.1-Heatmap

4.14. Evaluation metrics score chart & Model Selection

After above observations, I choose the final model , Support Vector Regression with lowest MAE,MSE,RMSE and highest R2 score values that performs exceptionally well in predicting students' performance.

	Model Name	MAE	MSE	RMSE	R2	Adjusted R2
0	Support Vector Regression	0.035282	0.001916	0.043771	0.999990	0.999990
1	Decision Tree Regressor	1.103333	2.180000	1.476482	0.989049	0.988469
2	Random Forest Regressor	0.561333	0.670444	0.818807	0.996632	0.996454

Fig.4.14.1-Evaluation Metrics Score Chart

5.References

- https://colab.research.google.com/drive/1ksmroQtN_KoCeJzpzPgGAbLv0UG_6G_a?usp=sharing#scrollTo=nBYaR4P0HZln
- https://colab.research.google.com/drive/14TP6tNzUT5M0YfgzwTMF_6WBuQkLUgXp?usp=sharing#scrollTo=OpSwoluyz5Wa
- https://colab.research.google.com/drive/1Hgb530U11qbP4iSf1I2Le0BRpgGUpb16#scrollTo=TtzGW880z5Ul
- https://colab.research.google.com/drive/1VDRu_jTyDkjnhbQXxmF2-4NjihbdSNfb
- https://colab.research.google.com/drive/1D3445tyCOcZmXBIVPX2rnQE-x15mk0r_#scrollTo=StK0mYCHudKC
- https://colab.research.google.com/drive/1We0qcY_28wr1q44boqVsK1-cdRpSL9VY#scrollTo=qAOV1VbS1c0o
- https://colab.research.google.com/drive/1OPd_XzW77o-CSfYLqi92Q2PsmTAwsIo4#scrollTo=EzamsUNXFJAn
- https://medium.com/aiskunks/categorical-data-encoding-techniques-d6296697a40f#:~:text=It%20refers%20to%20the%20process,with%20text%20or%20categorical%20variables.
- https://scikit-learn.org/stable/model_selection.html#model-selection