Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Loan Prediction Eligibility

Supervised By:

Dr. Kirandeep Singh

Submitted By:

Prianshu Kumar – 2210990681

Priyanshu Bindal – 2210990686

Rajat Saini – 2210990706

Rajat Sharma - 2210990707

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:

➢ *Predicting Loan Eligibility with Machine Learning: A Python Project:*

In today's dynamic financial landscape, accurately assessing loan eligibility is crucial for both lenders and borrowers. Loan providers seek to minimize risk while ensuring a healthy loan portfolio, while borrowers desire efficient access to credit. This project delves into the world of Artificial Intelligence for Machine Learning (AIML) to develop a model that predicts loan eligibility using Python's powerful libraries: pandas, NumPy, and Matplotlib.

Our journey begins with acquiring a dataset containing loan applicant information. This data serves as the foundation for the model, encompassing details like income, credit history, demographics, and the most critical factor - loan approval status (approved or rejected). However, raw data seldom presents itself in a perfect state. The next step involves data preprocessing, a meticulous process akin to cleaning and organizing a workspace. We'll address missing values, identify and potentially remove outliers (extreme data points), and transform categorical features (like marital status or profession) into a format suitable for machine learning algorithms.

Beyond basic cleaning, feature engineering can be employed to create new features that might hold hidden predictive power. This could involve combining existing features (e.g., debt-to-income ratio) or deriving entirely new ones based on domain knowledge. Imagine incorporating an applicant's profession stability into the model - a factor not explicitly present in the original data but potentially valuable for predicting loan repayment likelihood.

Now comes the heart of the project: model selection and training. We'll explore various machine learning algorithms, each with its unique strengths and weaknesses. Common contenders include Logistic Regression, a workhorse for classification tasks; Decision Trees, known for their interpretability; Random Forests, which leverage the power of ensemble learning; and Support Vector Machines, adept at handling high-dimensional data. By training these models on the preprocessed data, we essentially teach them to identify patterns that distinguish loan-worthy applicants from those posing a higher risk.

Evaluating the model's performance plays a critical role. We'll employ metrics like accuracy, precision, recall, and F1-score to gauge the model's effectiveness in predicting loan eligibility. Accuracy tells us the overall proportion of correct predictions, while precision and recall delve deeper, revealing how well the model identifies true positives (approved loans predicted correctly) and avoids false positives (rejected loans predicted as approved). The F1-score provides a balanced view, combining precision and recall.

Finally, with a well-trained and evaluated model, we can consider deploying it into a production environment. This allows the model to make loan eligibility predictions for new loan applicants, potentially streamlining the loan approval process for lenders while offering borrowers a faster and more data-driven experience.

This project offers a glimpse into the power of AIML for loan eligibility prediction. By leveraging machine learning techniques, we can build models that assist lenders in making informed decisions and empower borrowers with a clearer understanding of their loan approval prospects. The potential benefits extend beyond individual applications, contributing to a more efficient and data-driven financial ecosystem.

1.2 Background

This project tackles the challenge of loan eligibility prediction using the power of Artificial Intelligence for Machine Learning (AIML). We'll leverage Python's versatile libraries – pandas for data manipulation, NumPy for numerical computations, Matplotlib for data visualization, and scikit-learn (commonly referred to as sklearn) for building and evaluating our machine learning models.

Our quest begins with acquiring a dataset brimming with loan applicant information. This data serves as the lifeblood of our models, encompassing details like income, credit history, demographics, and the critical factor – loan approval status (approved or rejected). However, raw data rarely arrives in a pristine state. The initial phase involves meticulous data preprocessing, akin to cleaning and organizing a workspace. We'll address missing values, identify and potentially remove outliers (extreme data points), and transform categorical features (like marital status or profession) into a format that machine learning algorithms can readily understand.

Exploring the Algorithm Arsenal:

This project delves into three popular machine learning algorithms, each with its unique approach to classification tasks:

1. **Decision Trees:** Imagine a series of yes-or-no questions leading to a final decision – that's the essence of decision trees. They are known for their interpretability, allowing us to visualize the decision-making process and understand the key factors influencing loan eligibility predictions.
2. **Logistic Regression:** This workhorse algorithm estimates the probability of an event occurring, in our case, loan approval. It excels at modeling linear relationships between features and the target variable (loan status).
3. **Naive Bayes:** This probabilistic classifier operates under the assumption of independence between features. It's efficient and performs well with categorical data, making it a strong contender for loan eligibility prediction where applicant demographics often play a role.
4. **SVM:** SVM stands for **Support Vector Machine**. It's a powerful supervised learning algorithm used for classification tasks in machine learning. Here's a breakdown of how it works in the context of classification:

Classification in Machine Learning:

Imagine you have a dataset containing emails categorized as spam or not spam. A classification algorithm learns from this data to build a model that can then categorize new unseen emails as spam or not spam.

Training and Evaluation:

By employing sklearn's functionalities, we'll train each model on the preprocessed data. This involves feeding the features (applicant information) and the target variable (loan status) to the algorithms, allowing them to learn the underlying patterns that differentiate approved and rejected loans.

But how do we know which model performs best? Here's where evaluation metrics come into play. We'll employ metrics like accuracy, precision, recall, and F1-score to assess each model's effectiveness. Accuracy tells us the overall proportion of correct predictions, while precision and recall delve deeper, revealing how well the model identifies true positives (approved loans predicted correctly) and avoids false positives (rejected loans predicted as approved). The F1-score provides a balanced view, combining precision and recall.

Visualization with Matplotlib:

Matplotlib, a powerful Python library, allows us to visualize the performance of our models. We can create informative graphs and charts to compare the accuracy, precision, recall, and F1-score of each algorithm. This visual representation helps us identify the champion model – the one that delivers the most accurate and reliable loan eligibility predictions.

1.3 Objective

The objective of this project is to develop a machine learning model capable of predicting loan eligibility for potential borrowers. This will be achieved by leveraging the power of Artificial Intelligence for Machine Learning (AIML) within a Python environment.

By acquiring a dataset containing loan applicant details and their corresponding loan approval status (approved or rejected), we can embark on the data preprocessing stage. This crucial step involves cleaning the data, addressing missing values, identifying and potentially removing outliers, and transforming categorical features into a format suitable for machine learning algorithms.

Following data preparation, we will train each chosen algorithm using sklearn. The models will be trained on the preprocessed data, essentially learning the patterns that distinguish loan-worthy applicants from those posing a higher risk.

Evaluation is paramount, and we will employ metrics like accuracy, precision, recall, and F1-score to assess the effectiveness of each model. These metrics provide insights into the model's ability to correctly predict loan eligibility and avoid false positives (rejected loans predicted as approved).

Finally, with a well-trained and evaluated model, we can explore deployment possibilities. This would allow the model to make loan eligibility predictions for new applicants, potentially streamlining the loan approval process for lenders and offering borrowers a faster, more data-driven experience.

In essence, this project aims to harness the power of machine learning to create a valuable tool for the financial sector. By building models that predict loan eligibility, we can contribute to informed decision-making by lenders and empower borrowers with a clearer understanding of their loan prospects. This has the potential to create a more efficient and data-driven financial ecosystem for all parties involved.

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://www.kaggle.com/datasets/devzohaib/eligibility-prediction-for-loan

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.

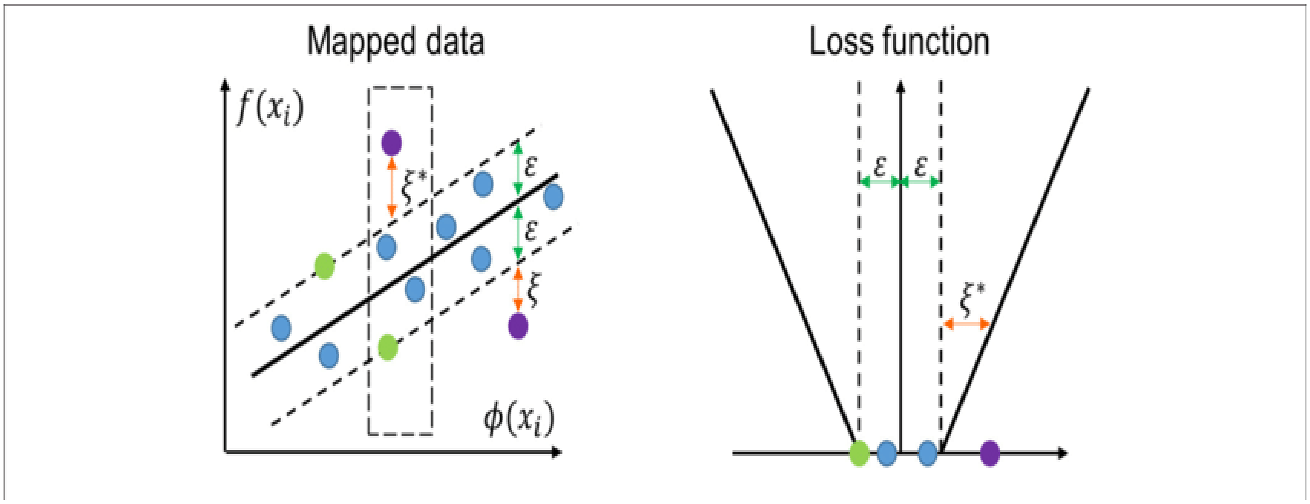


Fig.3.1.1-Support Vector Regression using Linear Kernel

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.

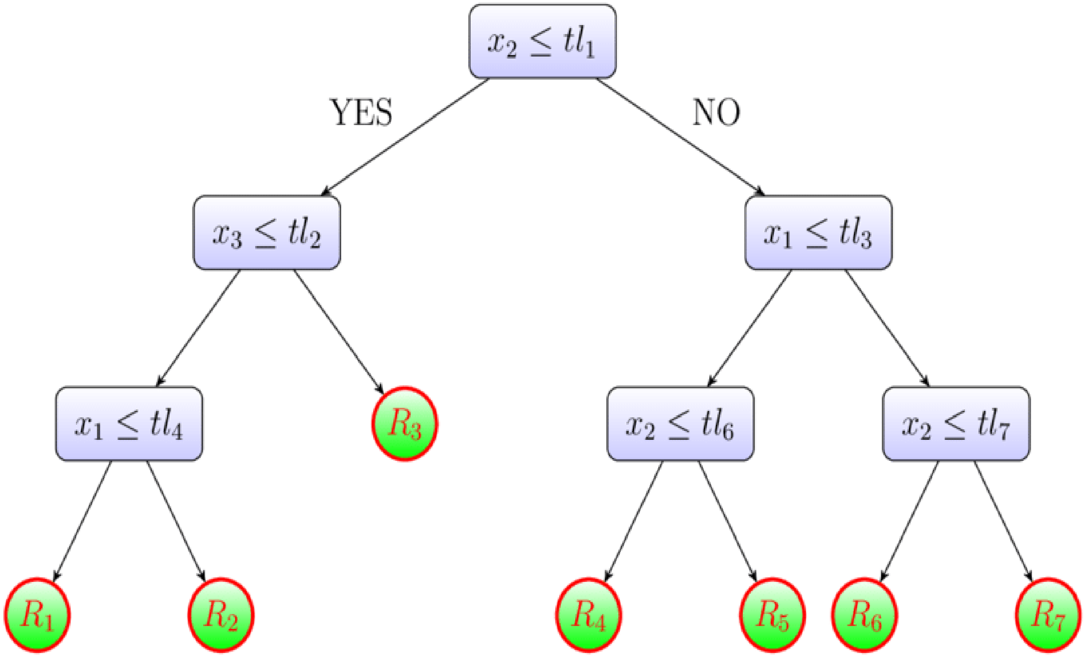


Fig.3.2.1-Decision Tree Regressor

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.

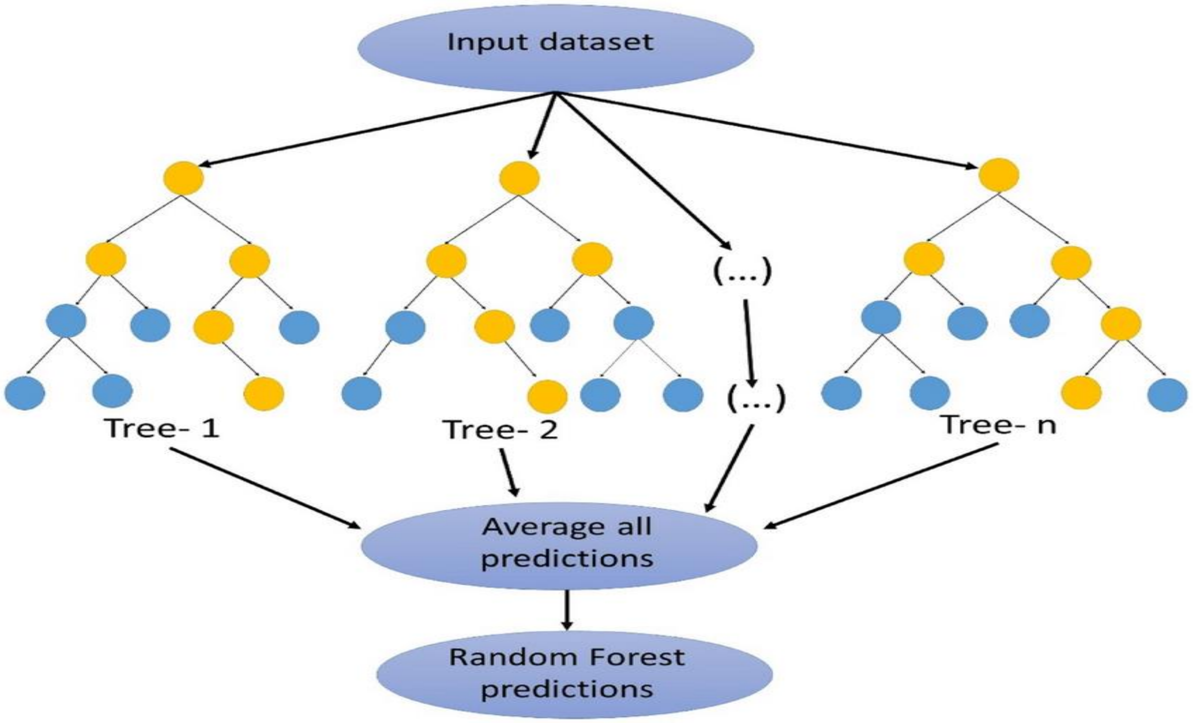


Fig.3.3.1-Random Forest Regressor

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.

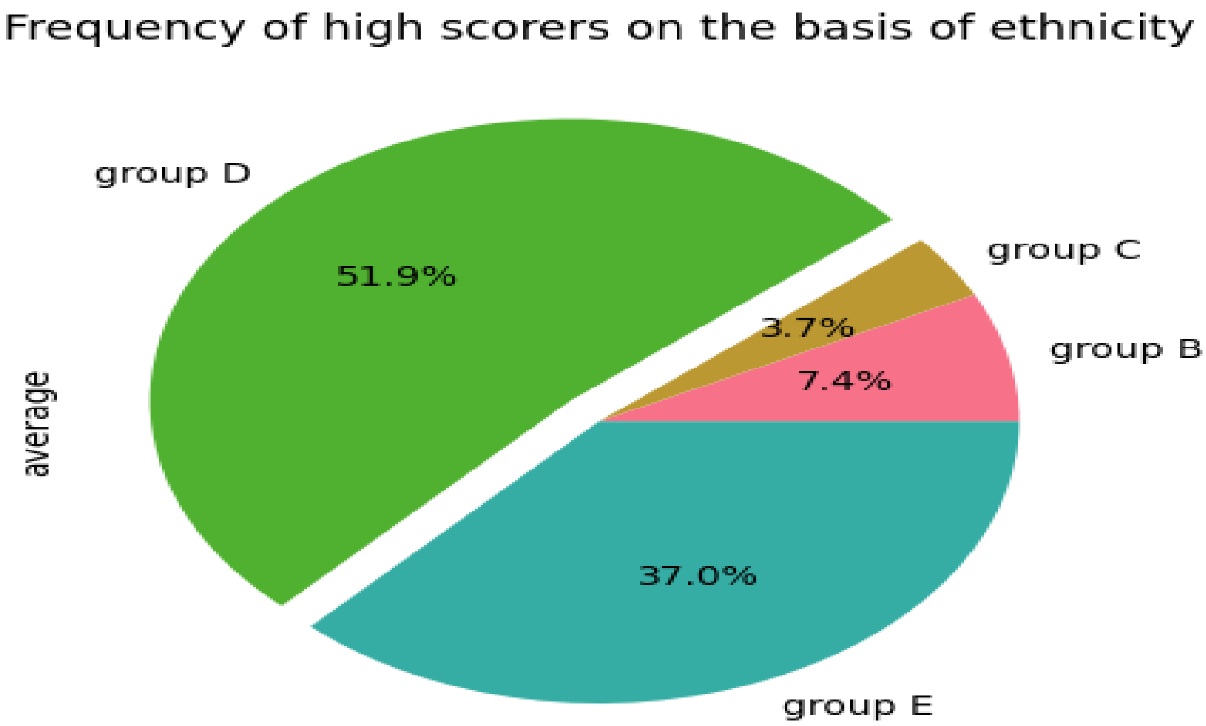


Fig.4.1.1-Pie Chart

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.

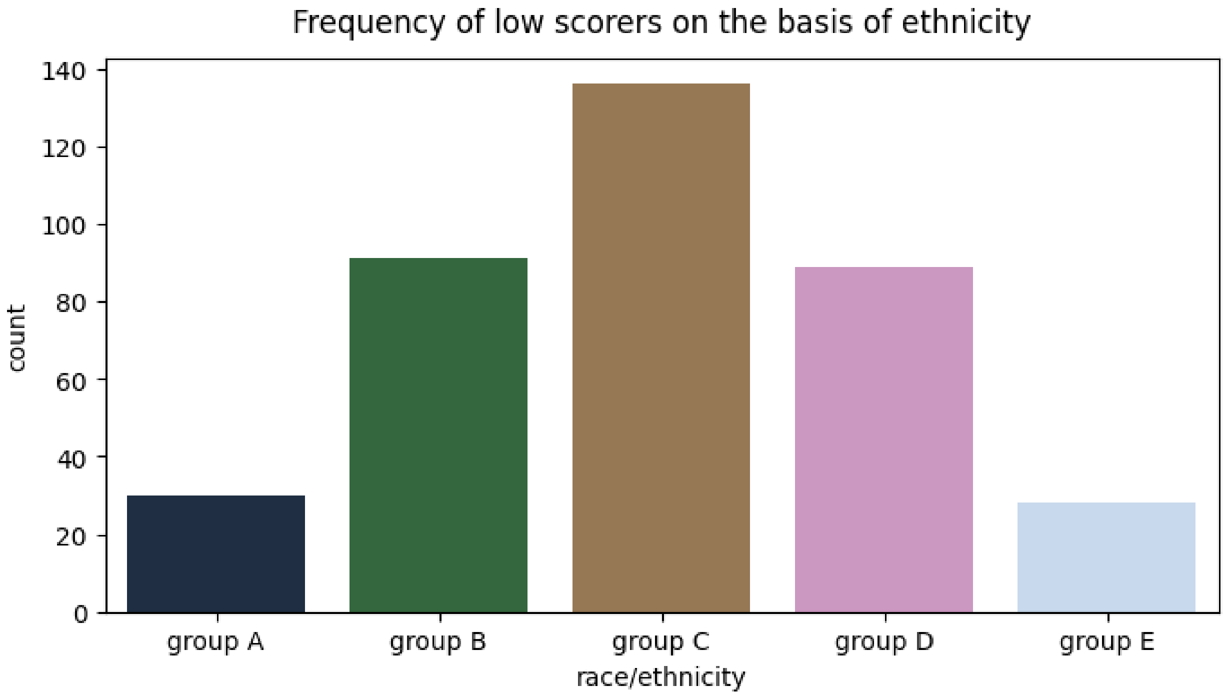


Fig.4.2.1-Countplot

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.

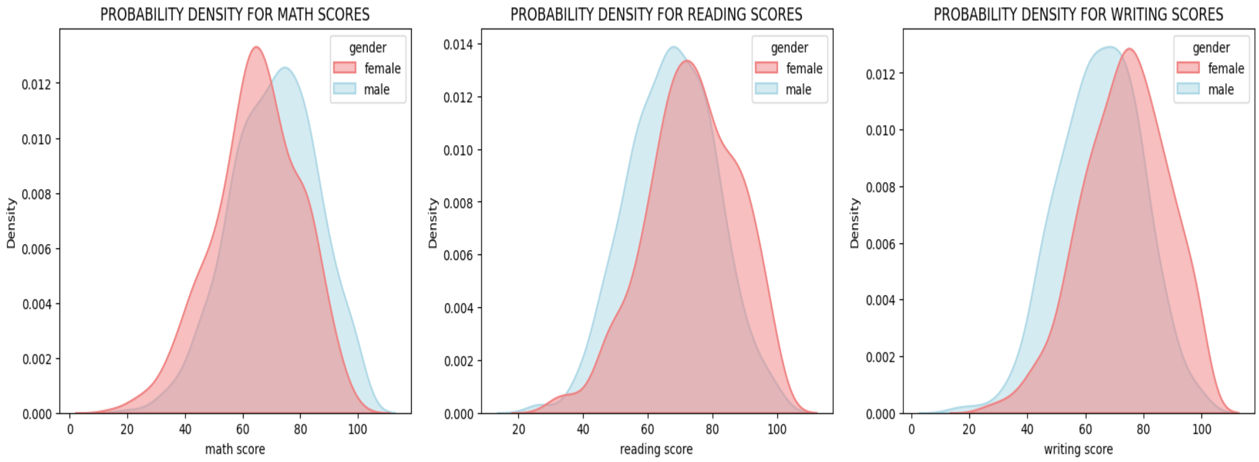


Fig.4.3.1-KDE Plots

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.

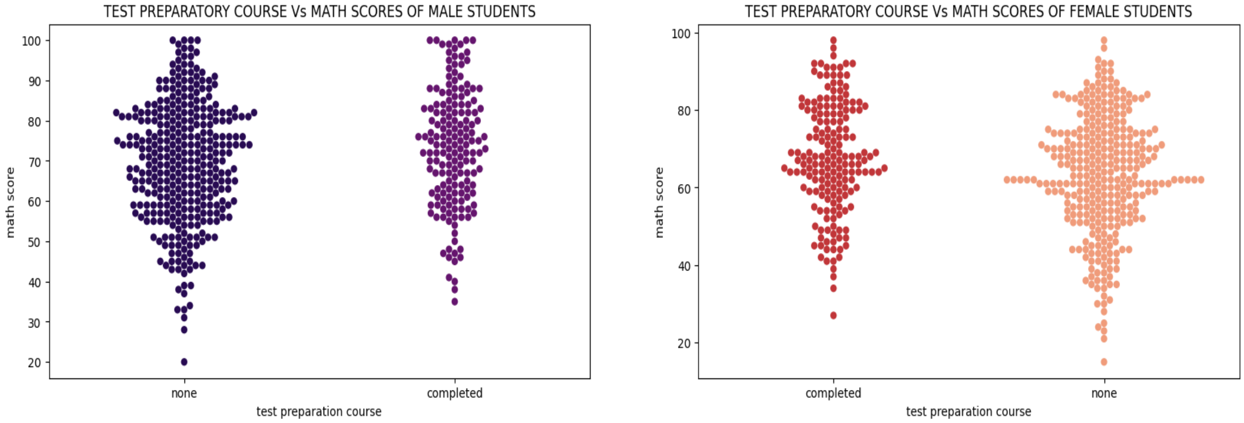


Fig.4.4.1-Swarmplots

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.

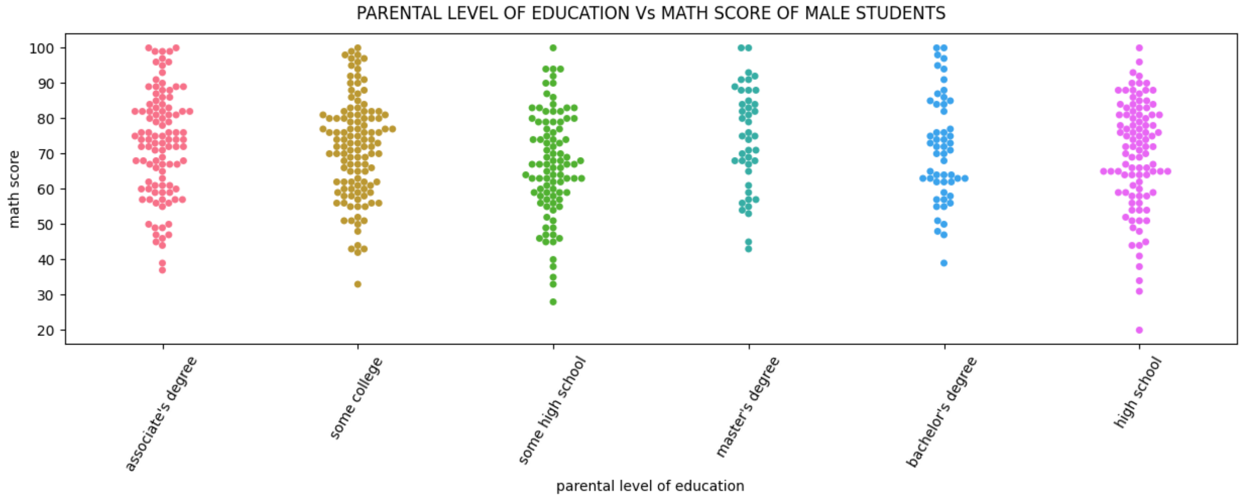


Fig.4.5.1-Swarmplot

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.

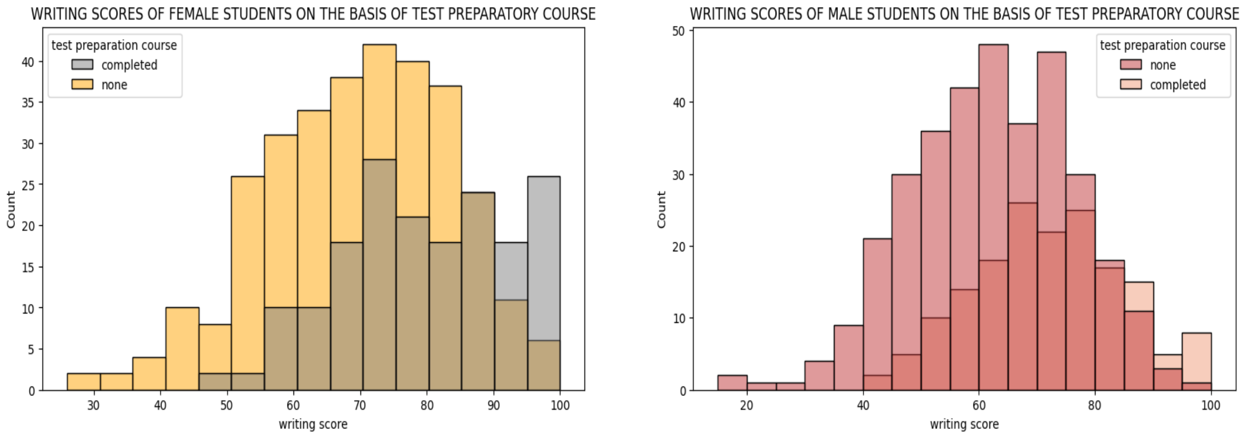


Fig.4.6.1-Histplots

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.

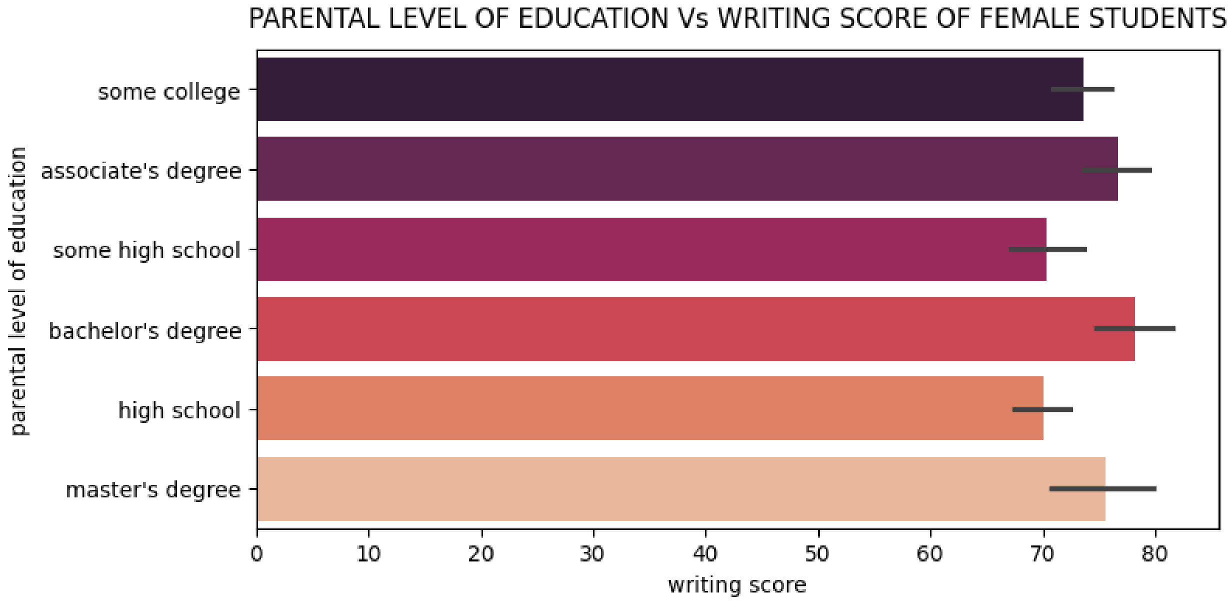


Fig.4.7.1-Barplot

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.

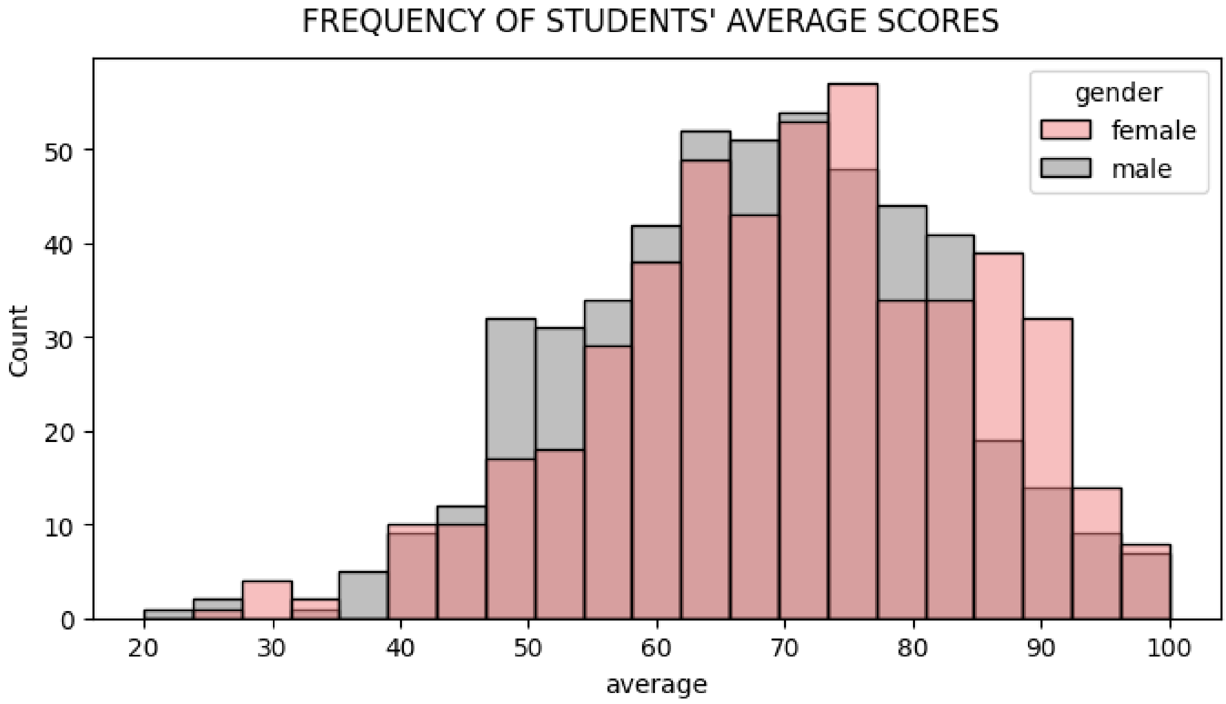


Fig.4.8.1-Histplot

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.

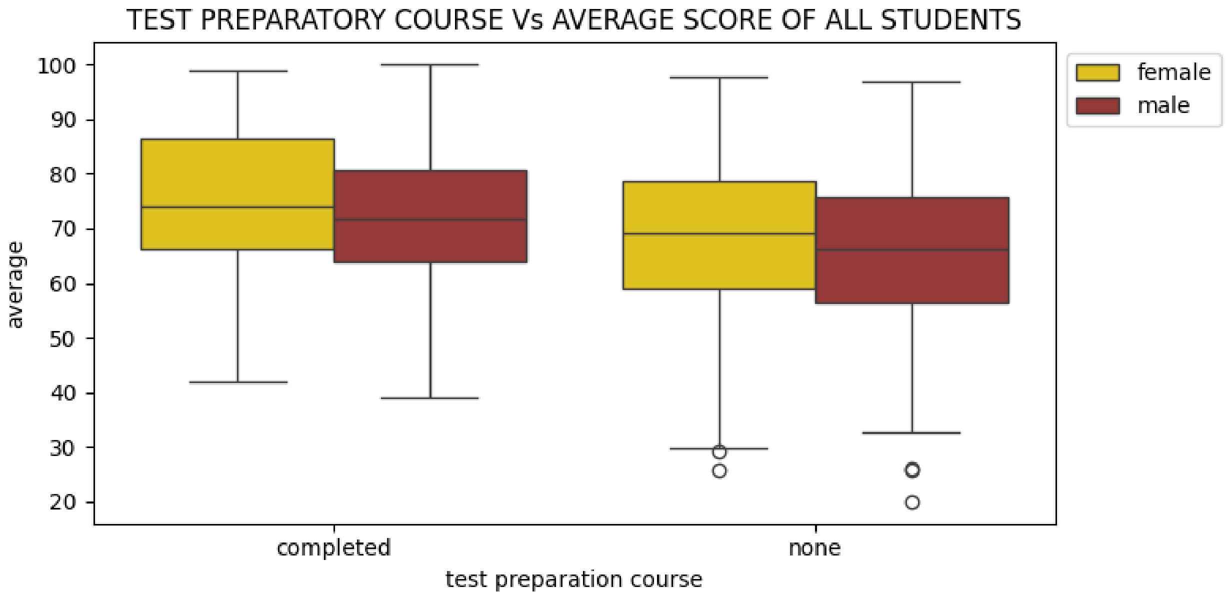


Fig.4.9.1-Boxplot

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.

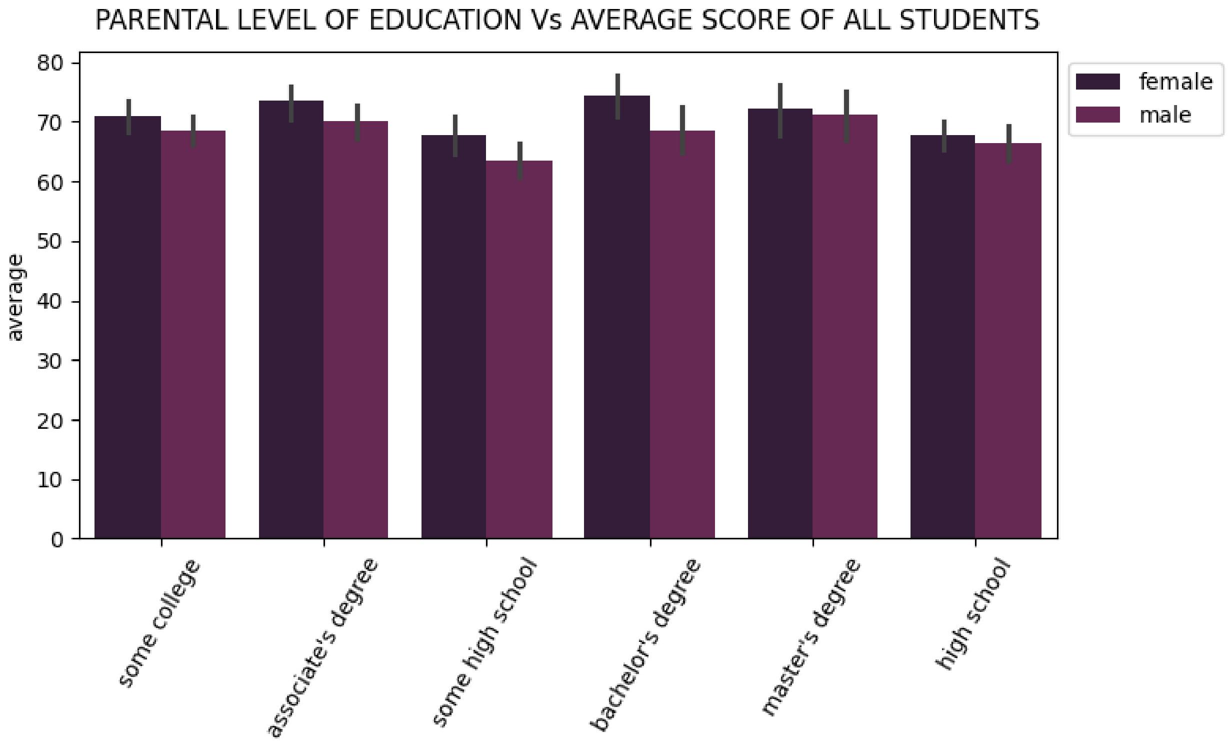


Fig.4.10.1-Barplot

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.

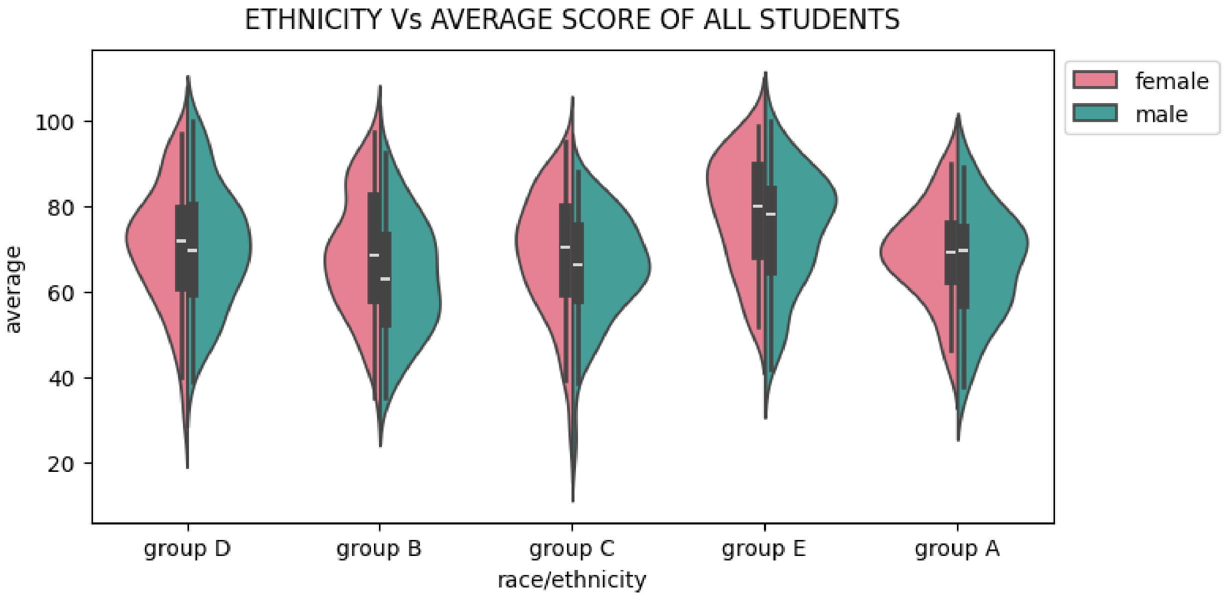


Fig.4.11.1-Violinplot

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.

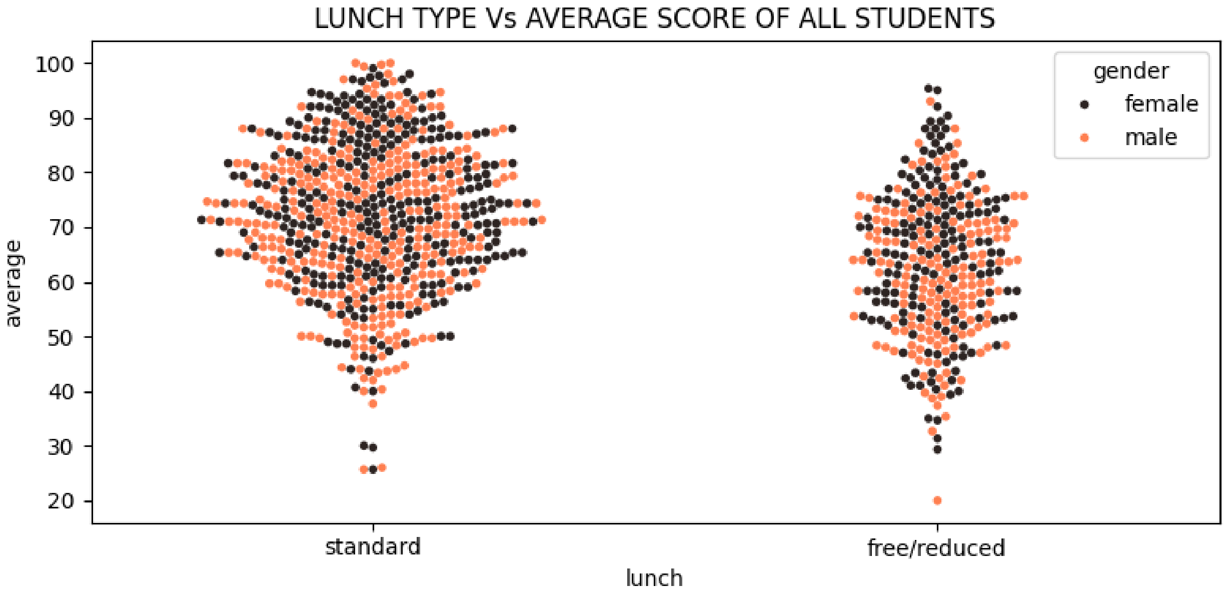


Fig.4.12.1-Swarmplot

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.

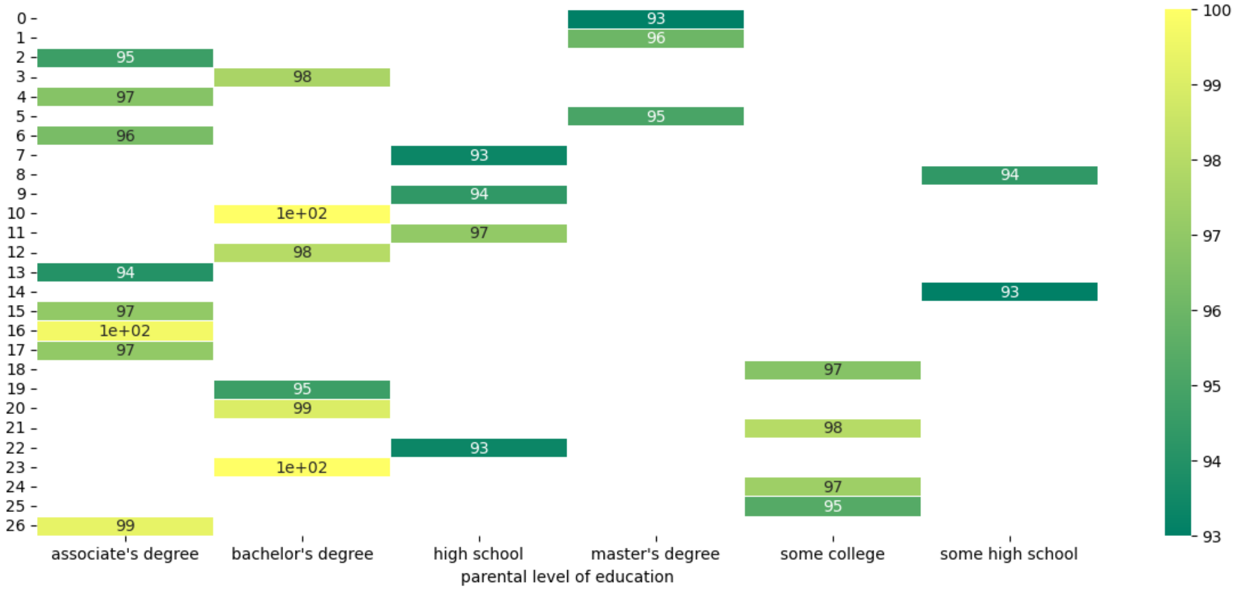
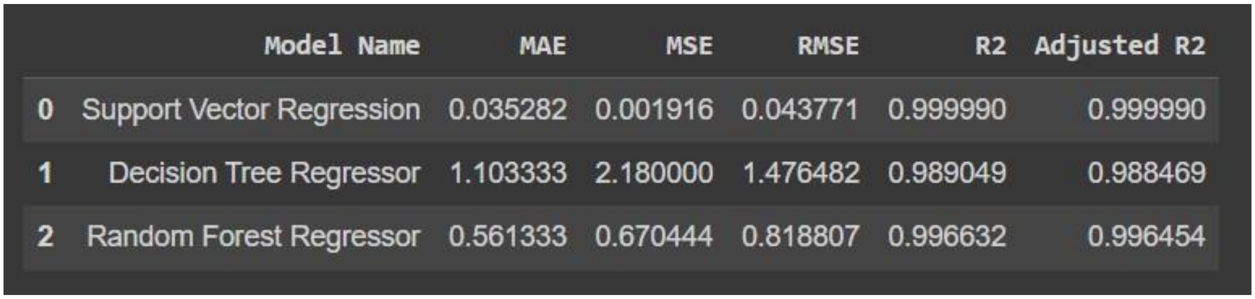


Fig.4.13.1-Heatmap

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