MACHINE LEARNING WITH PYTHON

Module : ST3189 Machine Learning

UoL Student Number : 210495821

Page Count : 10 (Excluding the Cover page , Table of contents and References)

Table of Contents

Task 1 : Unsupervised Learning

1.1 Introduction

Unsupervised learning, also known as unsupervised machine learning, uses machine learning (ML) algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention (IBM, 2024). The dataset chosen for this task was the mall customer segmentation dataset.

1.2 Literature Review

A study in an article showed that time and money spent at the mall was significantly high among female as compared to male consumers. Consequently, the results attributed that personal attributes and shopping mall attractiveness factors played a crucial role in influencing customer shopping behavior amongst the mall shoppers (Ankit Katrodia, 2018).

1.3 Research Questions

1. Is there a difference in spending between males and females?
2. Does Age have an impact on the spending score at malls
3. What are the clusters (homogenous groups) that can be identified in this dataset?

1.4 Exploratory Data Analysis (EDA)

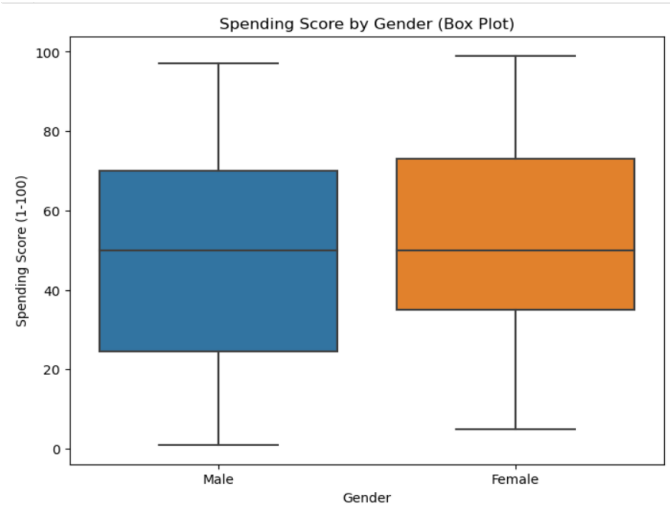


Figure 1

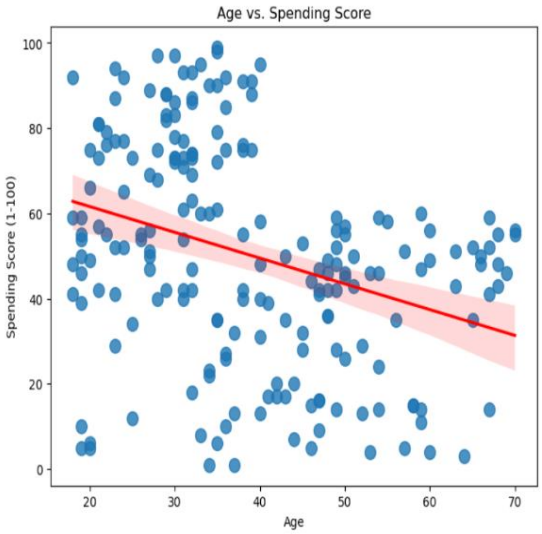


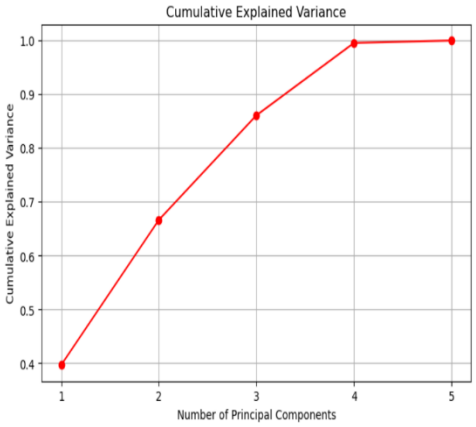
Figure 2

Based on Figure 1, we can see that the spending score between males and females seems to be visually similar with females have an average spending score of 51 whilst males have 49.

On the other hand we see that the age vs spending score seems to have a downward relationship with a correlation of -0.33 meaning the older people get, the lesser their spending score shopping malls is.

1.5 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends (Jaadi, 2024). To decide the number of principal components required a cumulative variance graph was plotted and was showed as follows.



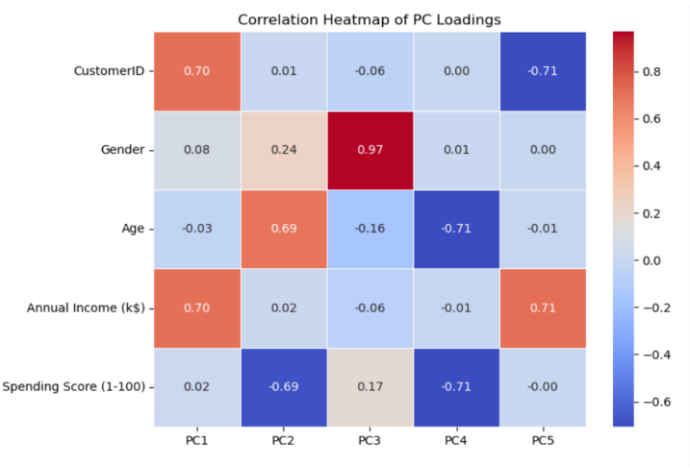


Figure 3 **Figure 4**

Based on figure 3 we can see the by the 4th principal component nearly 100% of the variation is explained thus it is ideal to have 4 principal components. Figure 4 shows the loadings (correlations) of the Principal components with the features in the dataset. PC1 has high loadings with Customer ID and Annual income meaning PC1 is primarily influenced by these two features. PC2 is influenced by Age and Spending score and PC3 captures the variation related to gender and PC4 once again is influenced by Age and spending score. We can omit PC5 since almost all the variation in the dataset is explained by the other Prinicipal components.

1.5 K-Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process (Javatpoint, 2024).

It was decided to segment based on two features so it is simpler and less complex to interpret, so the two pairs of features that were clustered were Age with Spending score and Annual income with spending score. To determine the optimal number of clusters needed the elbow method was used which involves plotting the variance explained by different numbers of clusters and identifying the “elbow”

point, where the rate of variance decreases sharply levels off, suggesting an appropriate cluster count for analysis or model training (Saji, 2024).

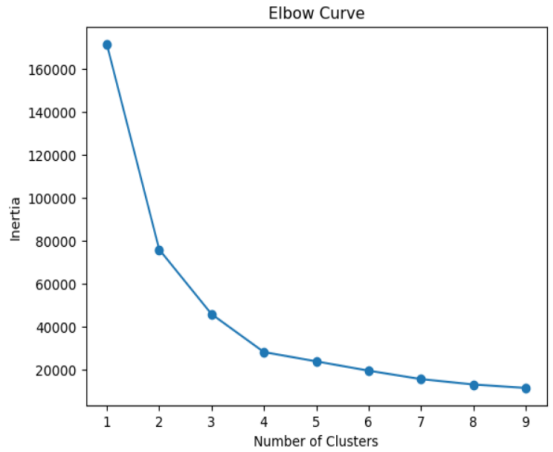


Figure 5

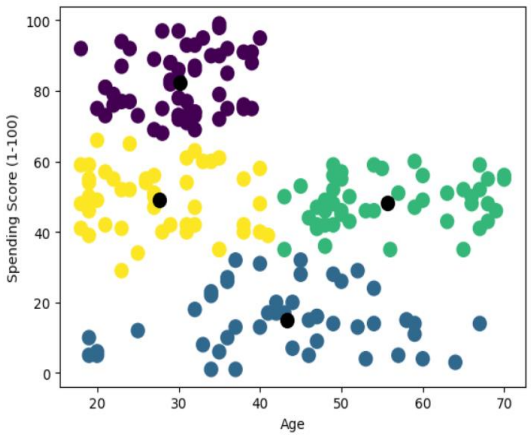


Figure 6

Based on figure 5, we see that curve begins to flatten at around 4 clusters therefore we decide to use 4 clusters to segment age and spending score and assign it to the cluster centroid which is ideally the mean value of that cluster (shown in black). Based on these results we can see there are a group of young people that spend less in malls and another group of young people that spend much more and a group of older people that spend as much as some young people.

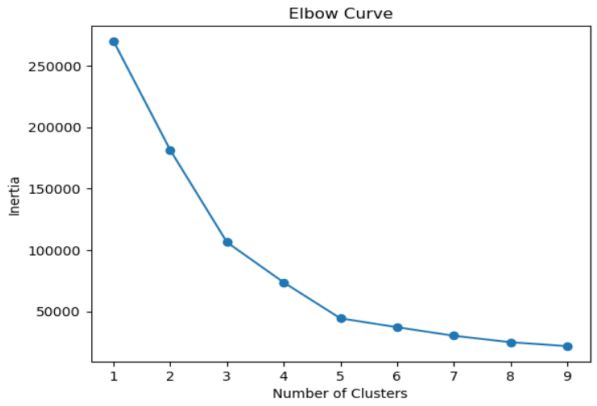


Figure 7

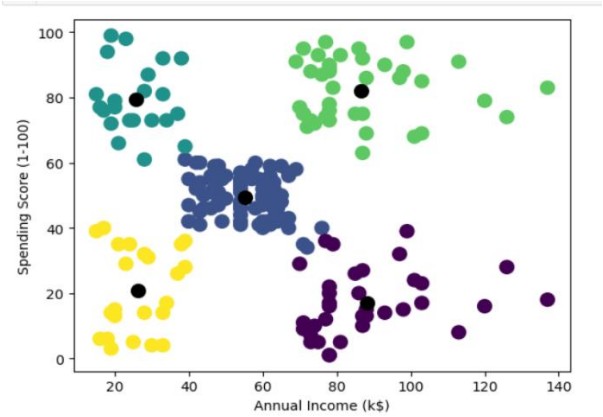


Figure 8

To segment Annual income and spending score it was found that the optimal number of clusters would be 5 and it was identified that there was a low income group with a high spending score and a high income group with a low spending score which goes against the general assumption that higher annual income translates to higher spending.

TASK 2 : Regression

2.1 Introduction

Regression analysis is a supervised learning method wherein the algorithm is trained with both input features and output labels. It helps in establishing a relationship among the variables by estimating how one variable affects the other (Kurama, 2024). The dataset used for this task was the data science salaries in jobs which was obtained from Kaggle and Regression models were created to predict the salaries of data science jobs.

2.2 Literature Review

Data scientists are in soaring demand and will continue to heighten even in the future. As the use of AI and ML expands, there will also be the importance of data privacy and reliability. (Sharma, 2024). According to reports, the United States is one of the highest-paying countries for data scientists. The movement of data scientists from other countries to the United States justifies this reality. In the US professionals in data science earns an annual salary of USD 120,000 which is significantly higher than any other country in the world (Careerera, 2022).

2.3 Research Questions

1. Has there been a spike in salaries of data science jobs over the past 3 years?
2. Does working remotely provide higher salaries?
3. Does the company location have an impact on salaries for data science jobs?

2.4 Exploratory Data Analysis (EDA)

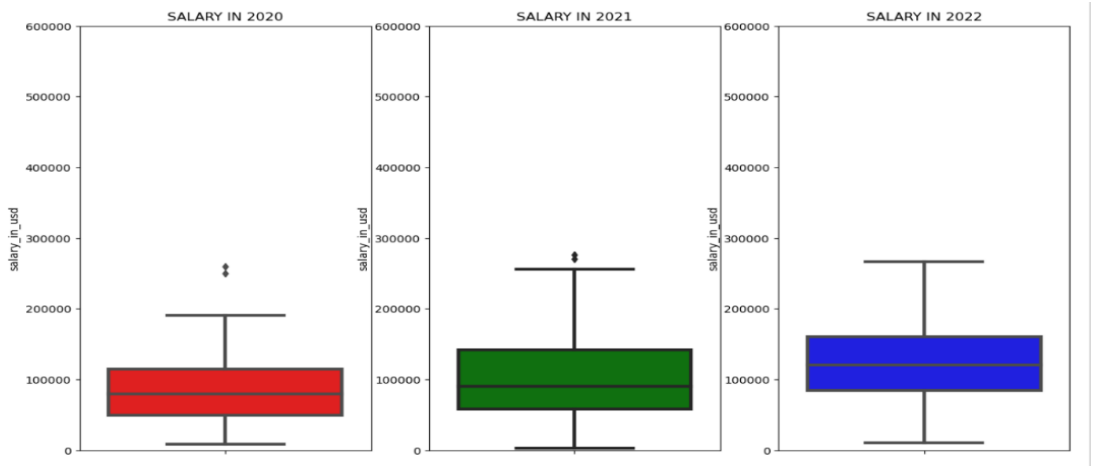


Figure 9

Based on the boxplots in figure 1 we can see the mean salary of data science jobs increasing with each year implying that salaries are increasing with each year. If we were to analyze this further, we can see there has been a 12% increase from 2020 to 2021 and 24% increase from 2021 to 2022 showing that salaries are indeed increasing with years.

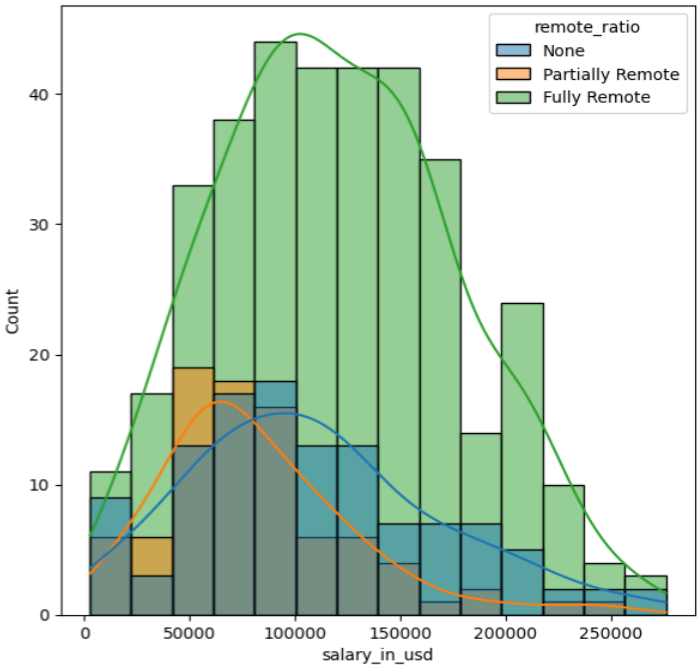


Figure 10

To answer the second question of whether remote working has an impact on salaries, A histogram was plotted and it showed that majority of the data science jobs were anyway fully remote and did have the highest mean salary of 120000 USD showing that working data science jobs remotely gives the provides the highest reward.

To find out which country provides the best salaries for data science jobs, we took the top 5 countries with the most amount of jobs available and analyzed their mean salaries . The results are as follows.

From figure 3 We can see The United States offers the highest salaries for data science occupations by quite a margin when compared to Countries like Canada and United Kingdom. However, since salaries are in USD, the standard of living may vary from country to country having an impact on the salaries.

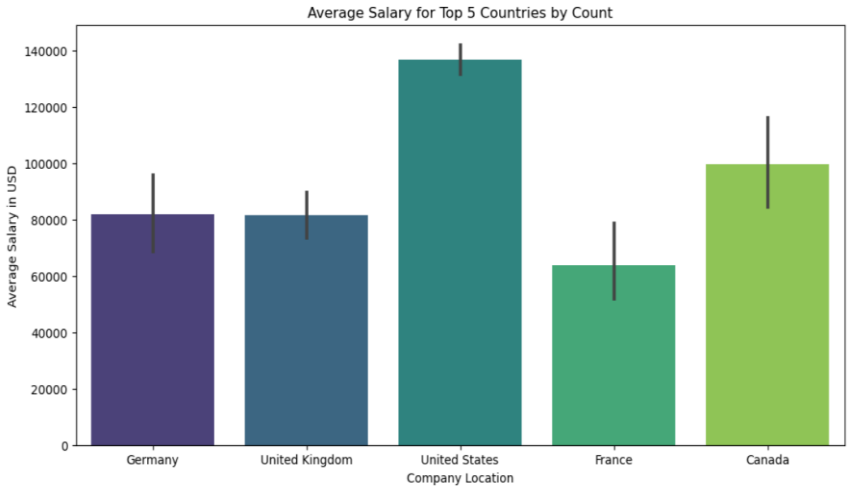


Figure 11

2.5 Regression Models

Since there were a large amount of data science jobs the job titles were further simplified 6 categories for easier analysis and the feature 'job_title' was replaced with 'Category'.

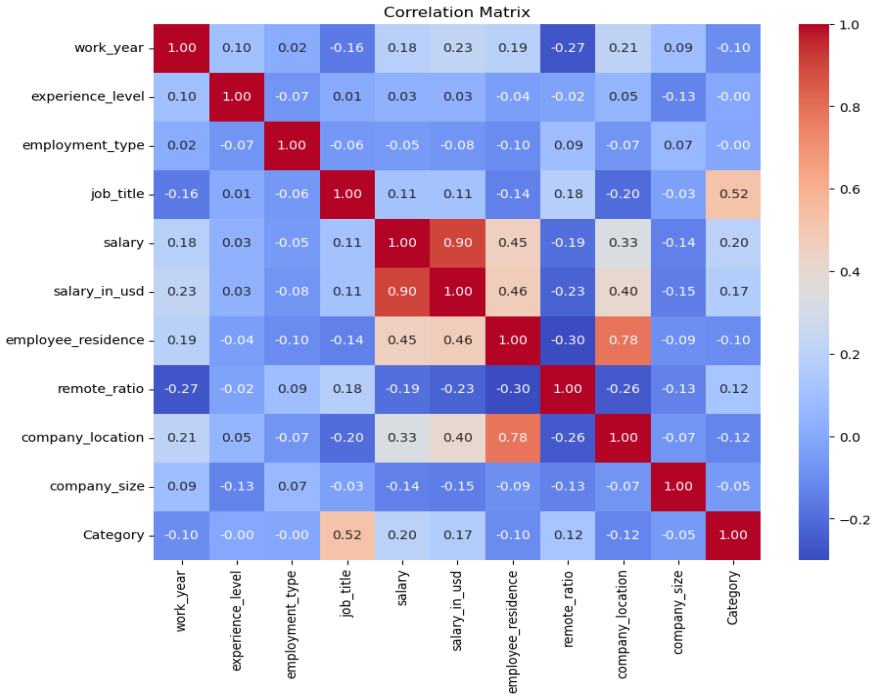


Figure 12

Based on correlation heatmap, it was decided to exclude the work year since data has been given for only 3 years and the salary columns since that is our predictor variable. The features were then scaled to bring the values to the same magnitude and data was split in an 80-20 ratio for train and test and a couple of models were trained on this data to see its effectiveness. The results are as follows.

Model	RMSE	R squared	**RMSE (after hyperparamtering tuning)**	**R squared (after hyperparamter tuning)**
Gradient boosting Regressor	39462.49	0.554	**30641.41**	**0.7013**
Random Forest Regressor	42890.18	0.473	**30536.4**	**0.7034**
XGBoost Regression	42954.61	0.472	**29553.25**	**0.722**

The RMSE (Root Mean Squared Error) measures the average deviation between the predicted values and the actual values and taking the square root of it. Lower values typically indicate better model performance.

The R^2 value on the other hand shows what percentage of the predictor variable (salary) is explained through the feature variables. A higher value ideally represents a better fit.

Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters (GeeksforGeeks, 2023). In this case we used a GridSearch technique on all 3 of these models and we can see the model performance increase immensely from the tuning.

TASK 3 : Classification

3.1 Introduction

Classification is a supervised machine learning process of categorizing a given set of input data into classes based on one or more variables (Ramakrishnan, 2023). The main objective of classification machine learning is to build a model that can accurately assign a label or category to a new observation based on its features (GeeksforGeeks, 2024). The dataset chosen for this task was students performances in exams and a variety of personal, social and economic factors that have an impact on them.

3.2 Literature Review

A study by the University of Sri Jayewardenepura, Sri Lanka showed that mothers' education levels made a significant contribution to the students' academic performance. However, English knowledge of the students becomes the second important factor which influences students' academic performance. Students with higher levels of attendance for lectures have positive effect towards their academic performance (Sriyalatha, 2024)

3.3 Research Questions

1. Do females tend to score higher marks than males?

2. Does the lunch provided to students have an impact on performance in exams

3. Does parents level of education affect student performances?

3.4 Exploratory Data Analysis (EDA)

Analyzing the data, we see that about 51.8% of the students were female and 48.2% of students were male showing that the gender distribution between males and females seem to be balanced. To answer the questions whether females perform better than males in exams, a bar plot was visualized showing how many males and females passed and failed.

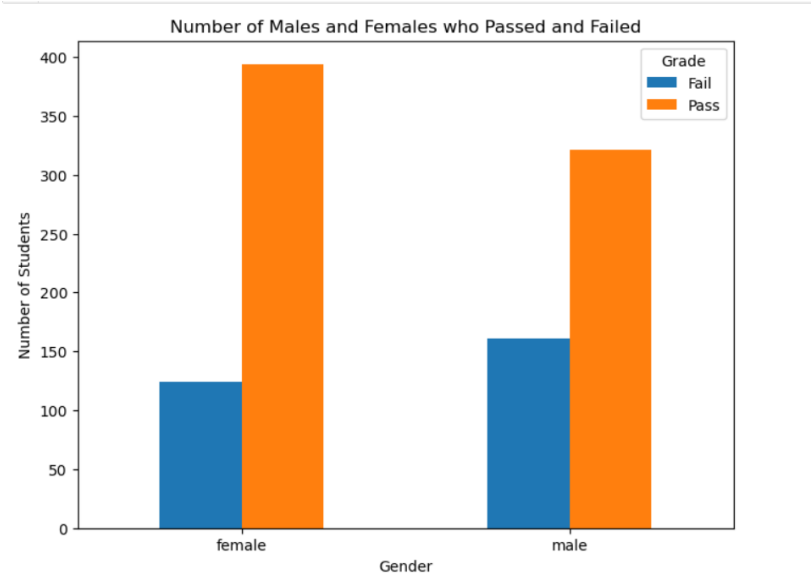


Figure 13

From Figure 5 we can see that more females tend to pass more than males. If we were to investigate this further about 76% of the females passed and 67% of the males passed showing that female pass rates are higher.

To see whether the lunch provided to students have an impact on exam performance a histogram was plotted showing the Average scores with a standard and a free lunch. Looking at the plot visually we see that standard lunch seems to have a higher average score than the free lunch. Calculating the percentages we see that about 80% of the students who had standard lunch had passed compared to the 56% pass rate of the free lunch students proving that the lunch provided has an impact on exam performances.

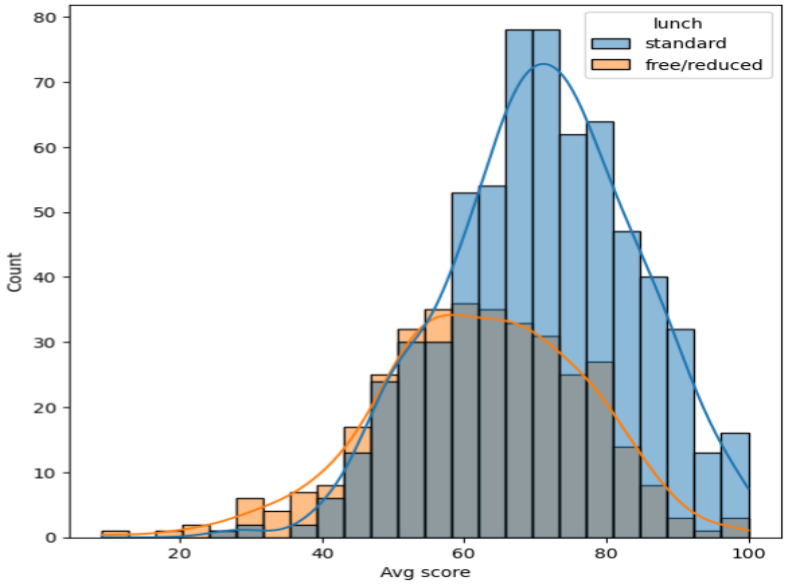


Figure 14

Next we want to check whether parental level of education had an impact on students grades since it was already proven that mothers' education levels have an impact on student grades it was decided to inspect this further.

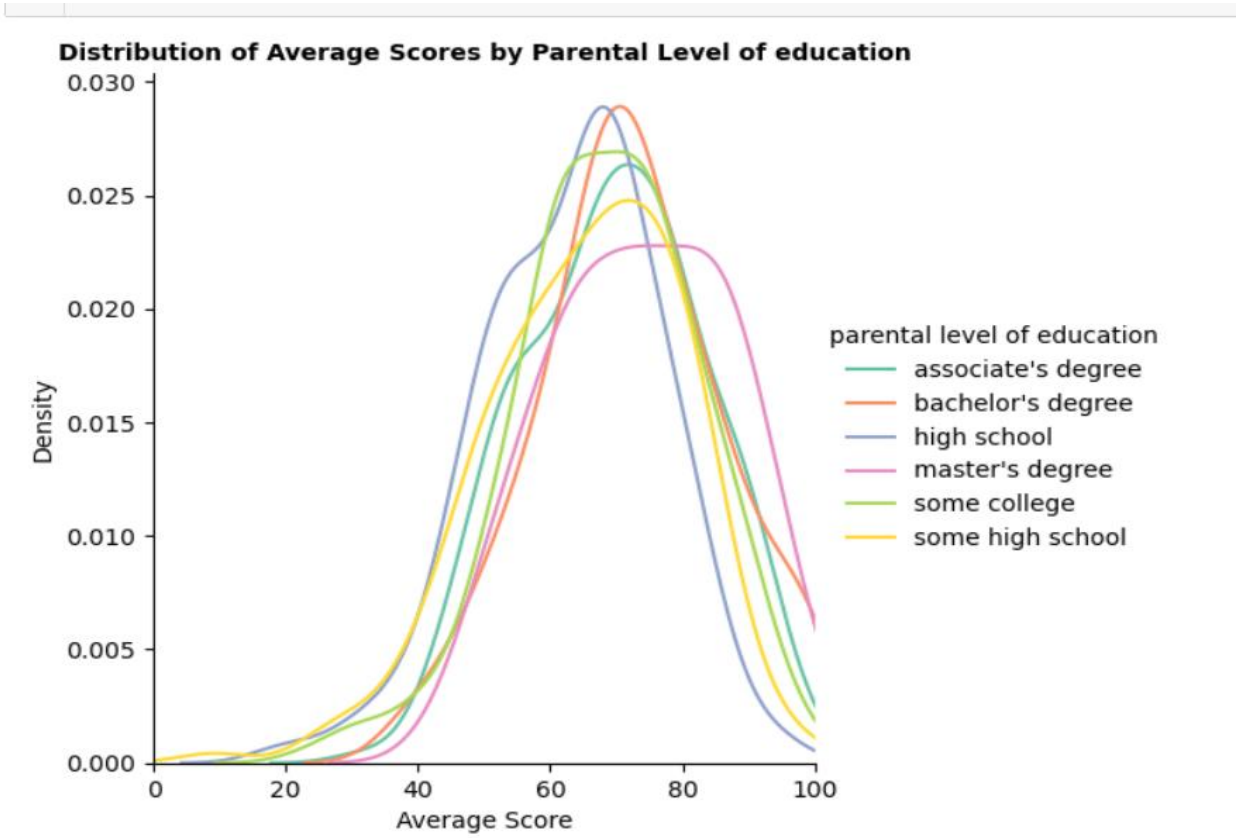


Figure 15

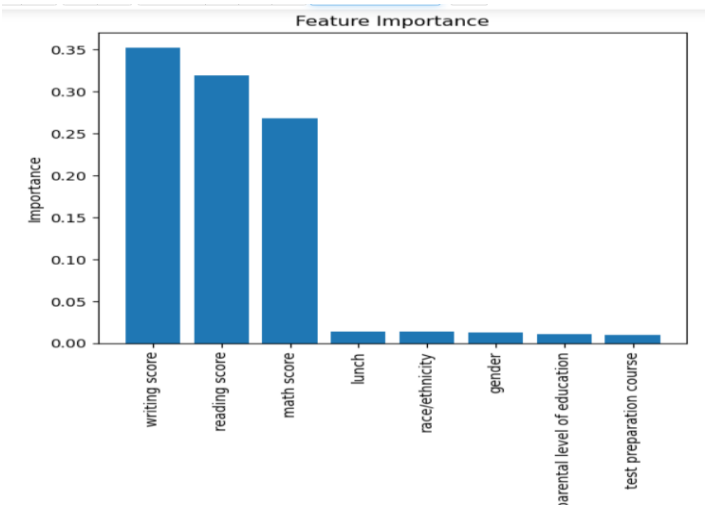
Based on this distribution plot we can see that parents who seem the have a master's degree seem to have higher average score than someone with a just a bachelor's degree or high school degree.

Analyzing the pass rates for each parental level of education we see that about 83% of students from parents who have a master's degree pass the exams compared to 61% to 64% pass rates from parents whose highest level of education is high school. Showing that higher parents education tend to translate into better student grades but may not always be the case.

3.3 Classification models

The target variable was decided to be the grade column which was decided by taking the average of the math, reading and English scores and putting pass if marks are greater than 60 else fail.

3.4 Feature Importance



Based on this it can be seen that writing, reading and math scores have the biggest impact on grade thus they were kept as features and variables like total score and average score were omitted to prevent multi collinearity within the model.

A couple of models were taken into test after splitting the data to a 80:20 ratio and the accuracy results were as follows.

We can see that Xgboost classifier and the decision tree classifer gives the highest accuracy of a near 100 score. This maybe due to the complexity of the task being too low .

Model	Accuracy Score
Xgboost Classifier	99.5%
Decision Tree Classifier	99.5%
Logistic Regression	98.5%
Random Forest Classifier	98.5%

The confusion matrices can also be visualized to get a better understanding of the performances of the models.

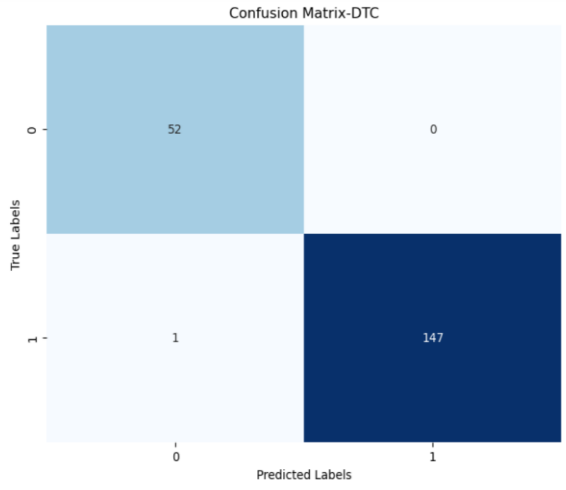


Figure 16

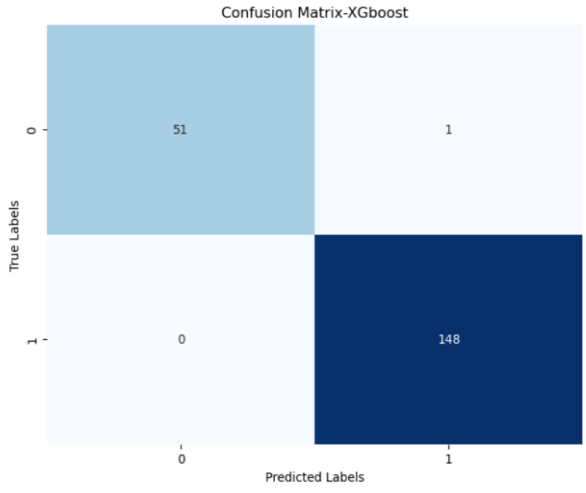


Figure 17

From this , we can see that out of the 200 data points , DTC was correctly able to predict 52 fail grades and 147 pass grades and the XGboost Classifier was able to correctly predict 51 fails and 148 passes.

References

Ankit Katrodia, M. N. (2018). *(PDF) Consumer Buying Behavior at Shopping Malls: Does Gender Matter?* Retrieved from https://www.researchgate.net/publication/338779361_Consumer_Buying_Behavior_at_Shopping_Malls_Does_Gender_Matter

Careerera. (2022, January 11). *Data Scientist Salaries Around the World*. Retrieved from Careerera.com: https://www.careerera.com/blog/data-scientist-salaries-around-the-world

GeeksforGeeks. (2023, December 07). *Hyperparameter tuning*. Retrieved from https://www.geeksforgeeks.org/hyperparameter-tuning/

GeeksforGeeks. (2024, January 24). *Getting started with Classification*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/getting-started-with-classification/

IBM. (2024). *What Is Unsupervised Learning?* Retrieved from IBM: https://www.ibm.com/topics/unsupervised-learning

Jaadi, Z. (2024). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Retrieved from Built In: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Javatpoint. (2024). *K-Means Clustering Algorithm - Javatpoint*. Retrieved from www.javatpoint.com: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Kurama, V. (2024). *Regression in Machine Learning: What It Is and Examples of Different Models*. Retrieved from Built In: https://builtin.com/data-science/regression-machine-learning

Ramakrishnan, M. (2023, 24 July). *What is Classification in Machine Learning and Why is it Important?* Retrieved from Emeritus Online Courses: https://emeritus.org/blog/artificial-intelligence-and-machine-learning-classification-in-machine-learning/

Saji, B. (2024, January 7th). *Elbow Method for Finding the Optimal Number of Clusters in K-Means*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/#:~:text=Elbow%20Method%20Definition&text=It%20involves%20plotting%20the%20variance,for%20analysis%20or%20model%20training.

Sharma, P. (2024, March 1). *Are Data Science Jobs In Demand in 2024? Prospects, And More*. Retrieved from IIM SKILLS: https://iimskills.com/are-data-science-jobs-in-demand/#:~:text=Data%20scientists%20are%20in%20soaring,continue%20making%20data%2Ddriven%20decisions.

Sriyalatha, M. A. (2024). *factors contributing to students' academic performance: a case ...* Retrieved from https://mgt.sjp.ac.lk/bec/wp-content/uploads/2017/08/Sri-Lankan-Journal-of-Business-Economics-SLJBE-Vol.-06-2016-Article-05.pdf