UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MACHINE LEARNING WITH PYTHON

Module: ST3189 (Machine Learning)

UOL Student Number: 220627669

Page Count: 10 (Excluding Cover Page, Table of Contents and References)

Table of Contents

TASK 1 : UNSUPERVISED LEARNING

1.1 INTRODUCTION

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets, by discovering hidden patterns or data groupings without the need for human intervention. Unsupervised learning models are utilized for three main tasks: clustering, association, and dimensionality reduction. (IBM, 2021)

The dataset used for this task is the 'Mall Customer Segmentation' dataset from Kaggle. It includes customer attributes like age, annual income, and spending scores. The aim is to identify different customer segments by dividing the dataset into distinct clusters with similar characteristics and spending habits. This enables businesses to optimize product offerings and develop targeted marketing strategies.

1.2 EXISTING LITERATURE

A study conducted by (Singhal, 2020) concluded that women exhibit higher spending scores than men, while men earn a higher annual income than women in general. It was also found that young and senior women have higher spending score values than young and senior men. Additionally, K-Means clustering was applied by (Jetir, 2024) on age, spending score and annual income features to segment customers into 5 distinct groups. The customers in these segments were labelled accordingly, and marketing strategies were recommended for each group.

1.3 RESEARCH QUESTIONS

1. Do women typically spend more than men, even though men earn higher annual incomes on average?
2. How do spending scores differ across the age categories of males and females?
3. What customer segments can be identified through different clustering methods, and how can these insights inform business strategies?

1.4 EXPLORATORY DATA ANALYSIS (EDA)

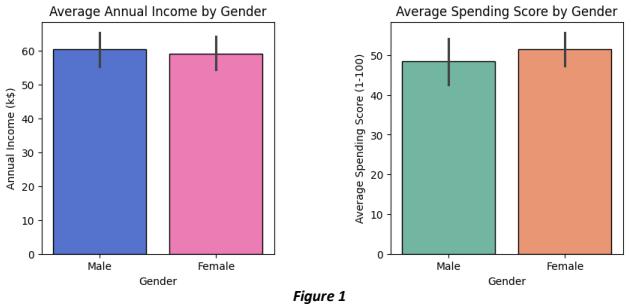


Figure 1

Figure 2

Figure 1 illustrates that males, on average, earn slightly higher annual incomes than females. It also reveals that females have greater average spending scores than males. This aligns with previous findings, confirming that women tend to spend more despite men having higher annual incomes.

The violin plots in figure 2 show the distributions of annual income and spending score by gender. Though males earn slightly more on average, the two distributions are fairly similar with most incomes between 25k and 100k. In contrast, females have higher average spending scores and a wider distribution, with more individuals in both high and low spending categories. Males exhibit a more balanced spending distribution.

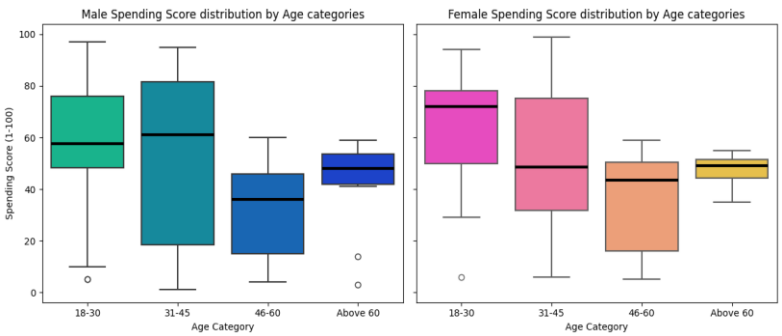


Figure 3

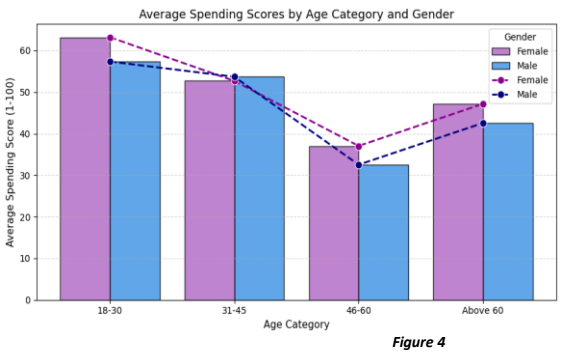


Figure 4

Figures 3 and 4 display the spending score distributions and average spending scores of males and females, according to four age categories: 18-30, 31-45, 46-60 and above 60.

We see that females in the 18-30, 46-60 and above 60 categories maintain higher median and average spending scores, which aligns with previous findings that young and senior women outspend the young and senior men. However, the middle-aged groups (31-45) are more balanced, with men having higher median scores and slightly higher average scores. Also, both genders experience a noticeable drop in spending in the 46-60 age category.

1.5 UNSUPERVISED MODELS

Principal Component Analysis (PCA) is a technique for reducing dimensionality by minimising the number of variables in a dataset, while retaining as much information as feasible. It transforms the original correlated variables into a set of new, independent variables called principal components. (Premanand, 2024)

However, PCA was not utilised in this analysis as the mall customer dataset contains only a few key features which are easily interpretable. With low dimensionality and no severe multicollinearity, PCA adds little value in this case.

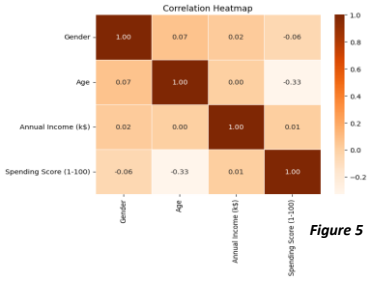
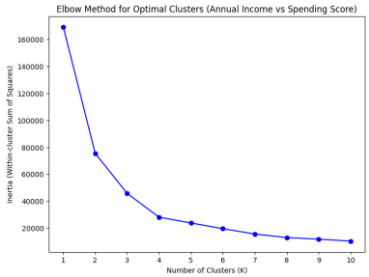


Figure 5

1.5.1 K-MEANS CLUSTERING

K-means is a centroid-based clustering algorithm that data assigns points to clusters based on distance. The objective is to minimize the sum of distances between points and their respective cluster centroids. (Sharma, 2025)

Figure 6

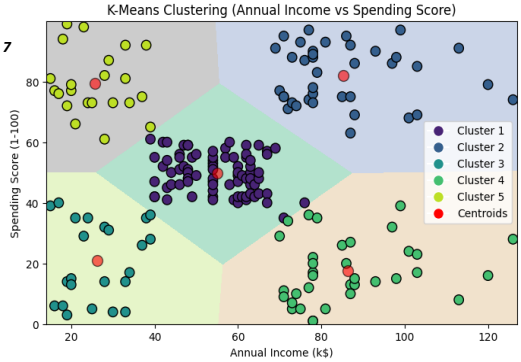


An elbow plot is plotted, displaying the within-cluster sum of squares (WCSS) values against their corresponding K values. The optimal K value is identified at the point where the graph forms an elbow. (Saji, 2025)
Beyond this point, rate of decrease of WCSS starts to slow down, indicating that further addition of clusters provides only minimal improvement in clustering quality.

Annual income and spending score were selected as the pair of features to form clusters, and the optimal number of clusters (K) was found to be 5 - as indicated in figure 6.

The scatter plot below clearly visualizes the 5 distinct clusters, along with their corresponding decision boundaries and centroids:

Figure 7



Decision boundaries represent the lines or regions that separate different clusters in the feature space, indicating how the model classifies data points.

The bar chart in figure 8 displays the sizes of each cluster, where cluster 1 contains the most customers and clusters 3 and 5 have the least.

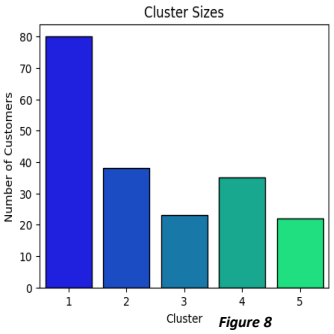


Figure 8

Cluster	
1	Low Income, High Spending
2	High Income, Low Spending
3	High Income, High Spending
4	Moderate Income & Spending
5	Low Income, Low Spending

The table contains descriptions of the K-Means clusters 1-5, which are aligned with the analysis conducted by (Jetir, 2024).
They classified clients with moderate income and spending scores as 'Standard clients', high income and high spending scores as 'Target customers', high income and low spending scores as 'Modest customers', low income and low spending as 'Modest customers', and low income and high spending as 'Careless consumers'.

In order to evaluate the quality of the K-Means clusters, the silhouette score was calculated and visualised in figure 9 :

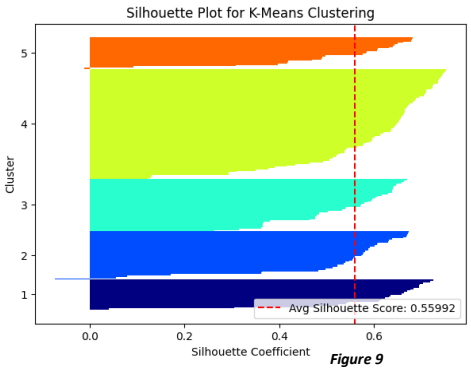


Figure 9

Silhouette scores evaluate how well data points are grouped within their assigned clusters, relative to points in other clusters. It is calculated for each data point and then averaged across all data points. (Ankita, 2025)
We see that the silhouette score for the K-Means clusters is 0.55992 ≈ 0.6, which indicates that the clusters are quite well-separated and distinct.

Next, K-Means clustering was performed using all three features: age, annual income and spending score.
A 3D scatter plot was generated to visualise the clustering results, as shown in figure 10 below.

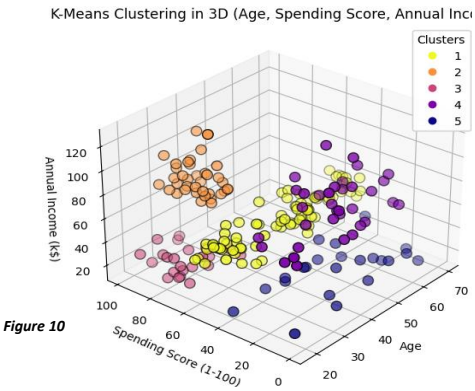


Figure 10

Figure 11 displays the mean values of the cluster features. Cluster 1 consists of middle-aged customers with moderate income and spending, Cluster 2 of young customers with high income and spending, Cluster 3 of young customers with low income but high spending, Cluster 4 of middle-aged customers with high income but low spending, and Cluster 5 of older customers with low income and spending. Slight differences are observed in our analysis compared to the study by (Jetir, 2024), relative to the customer segmentation by the three features.

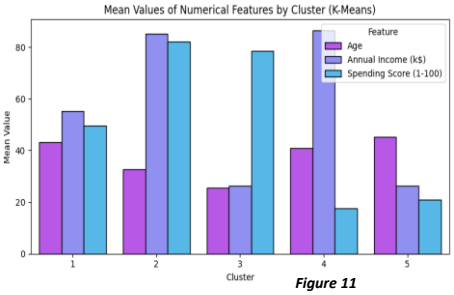


Figure 11

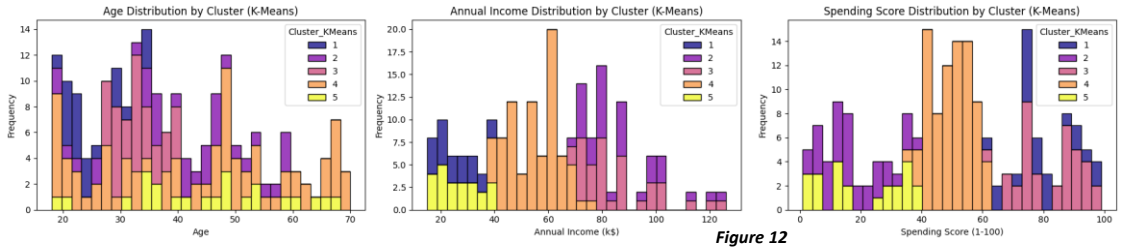


Figure 12

The distribution of the features in each cluster can be observed using figure 12.

Understanding these differences enables businesses to refine their marketing strategies more effectively. For Cluster 1, offer value-for-money products and loyalty programs; for Cluster 2, provide luxury and exclusive offerings; for Cluster 3, offer budget-friendly, trendy products with flexible payment options; for Cluster 4, promote spending with quality, exclusive products with personalized incentives; and for Cluster 5, focus on practical, affordable products with senior discounts and essentials.

1.5.2 HIERACHIAL (AGGLOMERATIVE) CLUSTERING

Hierarchical clustering is a technique used to group data points together based on how similar they are. It begins with each data point as its own separate cluster and then progressively merges or splits them based on their similarity. (Geeksforgeeks, 2025) In this analysis, Agglomerative clustering was used. It repeatedly merges clusters into larger ones until a single cluster remains.

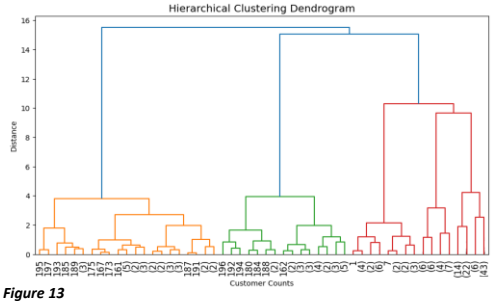


Figure 13

A dendrogram is used to visually represent the hierarchical relationships between clusters, as shown in figure 13.

Next, silhouette scores were used to determine the optimal number of clusters, which indicated the highest score at 5 clusters.
The scatter plot in figure 15 visualises the 5 distinct clusters and the silhouette plot in figure 16 checks the quality of clusters.

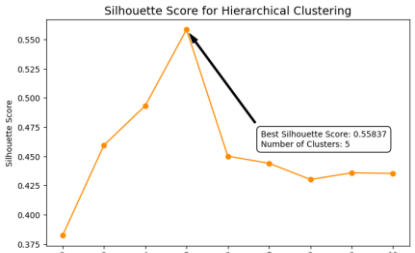


Figure 14

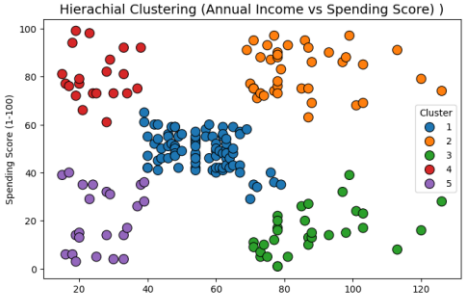


Figure 15

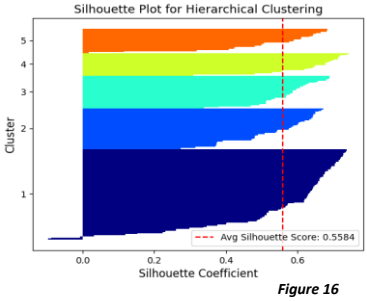


Figure 16

We see that the hierarchical clusters are consistent with K-Means in terms of the cluster sizes, mean values and distribution of features. However, the ordering of hierarchical clusters is different from K-Means, indicating variations in how data points are grouped and labelled across the two methods.
Furthermore, the average silhouette score of the hierarchical clusters is 0.55837 which is slightly lower than K-Means but still indicates a moderately good clustering structure.

1.5.3 GAUSSIAN MIXTURE MODEL

A Gaussian mixture model is a soft clustering technique where data points can belong to multiple clusters with a certain probability, unlike K-Means which assigns each point to a single cluster. It is used to determine the probability that a given data point belongs to a cluster. (Geeksforgeeks, 2025)

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are model selection criteria, that help identify the best number of components by penalizing for model complexity.

In order to determine the optimal number of clusters, AIC and BIC scores were plotted across different numbers of components and the minimum BIC score was observed at 5 components. Hence the optimal number of clusters was found to be 5, as illustrated in figure 17.

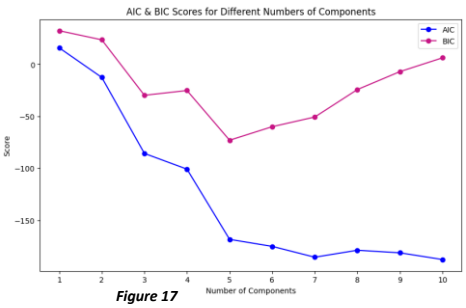


Figure 17

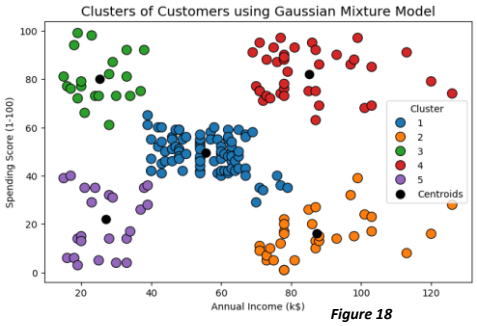


Figure 18

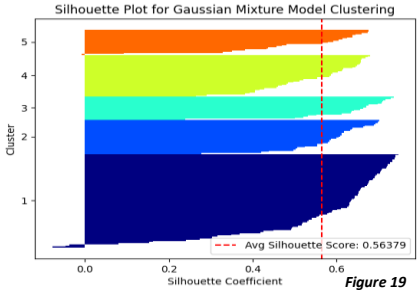


Figure 19

We see that the GMM clusters in figure 18 have different orders compared to both K-Means and Hierarchical clusters.
The cluster sizes remain the same, but the distribution of features and mean values of each cluster have now changed, indicating variations in customer segmentation.
The average silhouette score of the GMM clusters is 0.56379, which is a bit higher than the other 2 models. This indicates that GMM slightly outperforms both K-means and Hierarchical clustering in terms of defining well-separated clusters.

TASK 2: REGRESSION

2.1 INTRODUCTION

Supervised learning is a type of machine learning where algorithms are trained on labelled datasets to classify data or predict outcomes with accuracy. (Belcic, What is supervised learning?, n.d.)
Regression is a supervised learning technique where the goal is to predict a continuous numerical value based on one or more independent features. It finds relationships between variables so that predictions can be made. (Geeksforgeeks, 2025)

The dataset used for this regression task is the 'Vehicle Dataset' available on Kaggle. The goal is to build a predictive model that precisely forecasts the selling prices of used cars (in INR), based on a variety of feature attributes, helping buyers and sellers make informed pricing decisions.

2.2 EXISTING LITERATURE

According to the study conducted by (Ferizqa, 2023), when the age of a car and the number of previous owners increases, the selling price tends to decrease. It was also found that diesel-fuelled cars and automatic transmission vehicles are priced higher than petrol-fuelled cars and manual ones, while vehicles sold by trustmark dealers tend to be more expensive than those listed by individual sellers. Furthermore, the research done by (IJRASET, 2022) suggests that the Random Forest Regressor model is very well-suited for this task, with an extremely high accuracy of 98%.

2.3 RESEARCH QUESTIONS

1. Explore the variation of car prices based on their age and the number of previous owners, and examine how it relates to factors like kilometres driven and mileage.
2. How do vehicle specifications, such as fuel type and transmission type, influence selling prices?
3. Does the type of seller affect vehicle pricing?
4. What is the best regression model to predict the prices of used cars?

2.4 EXPLORATORY DATA ANALYSIS (EDA)

A new column called 'Car_Age' was created, which is the difference between the present year and year of manufacture.

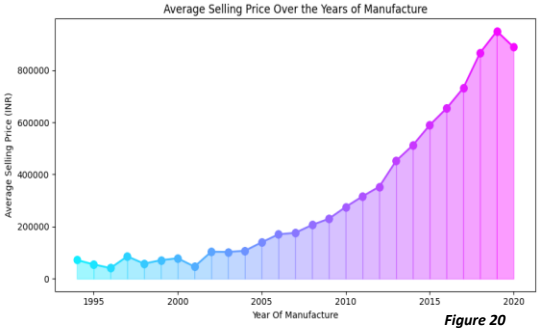


Figure 20

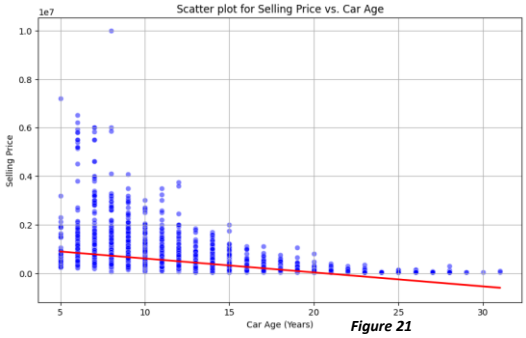
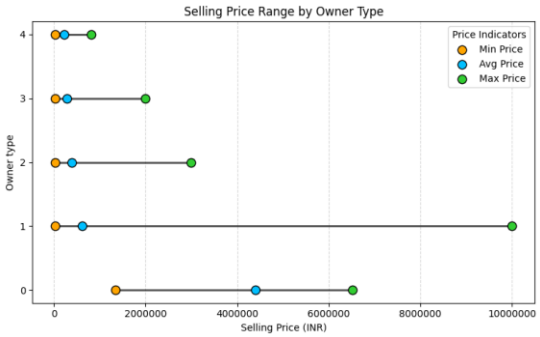


Figure 21

We see in figure 20 that newer cars are more expensive, as the average selling price increases as the year of manufacture becomes more recent.

The scatter plot further illustrates the clear negative relationship between car age and selling price.

It has a correlation coefficient of -0.4273, suggesting a moderate negative correlation. This means that as the car gets older, its market value generally declines, but the relationship is not perfectly linear—other factors may also influence price variations.



No. of previous owners	Min Price	Average Price	Max Price
0	1350000	4403800	6523000
1	30000	620511	10000000
2	35000	393886	3000000
3	29999	289751	2000000
4	31000	231970	810000

Figure 22

Figure 22 compares the range and average values of selling prices by each owner type, and reveals that the average price decreases as the number of owners increases (as noted by (Ferizqa, 2023)).

There is a significant drop in the vehicle's value from 0 to 1 owner (due to depreciation), with first-time owners (owner type = 0) exhibiting the widest price range.

The negative effect on selling price by factors like car age and number of owners can be further justified using attributes such as the number of kilometres driven and mileage of the car, since higher usage leads to increased wear-and-tear and lower fuel efficiency.

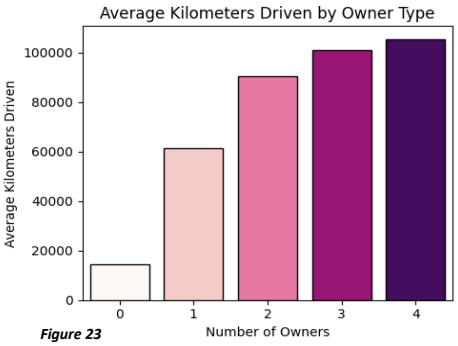


Figure 23

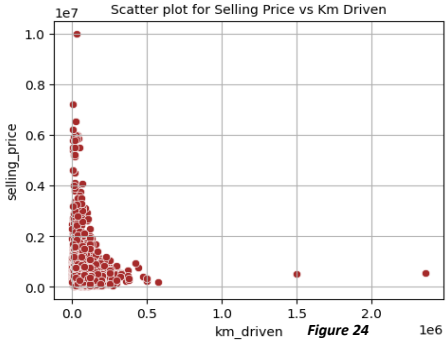


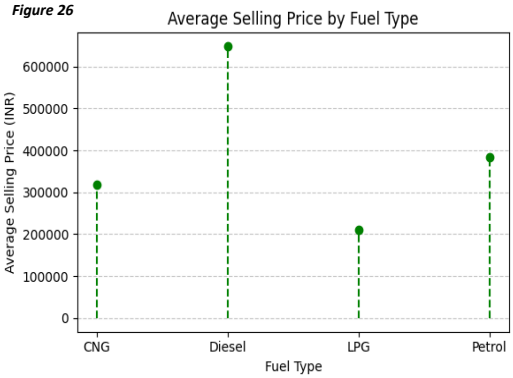
Figure 24

Figure 25

As shown in the boxplot, the average number of kilometres driven increases with the number of previous owners.
The effect of kilometres driven on selling price was evaluated using a scatter plot (figure 24), and a weak negative correlation is evident with a correlation coefficient of -0.1613. The scatter plot for selling price vs mileage (figure 25) also depicts a weak negative correlation of -0.1087. Hence, it can be concluded that cars with more kilometres driven and higher mileages tend to sell for less, but other factors likely matter more.

Next, we analyse how vehicle specifications such as fuel type and transmission type, influence selling prices:

Figure 26



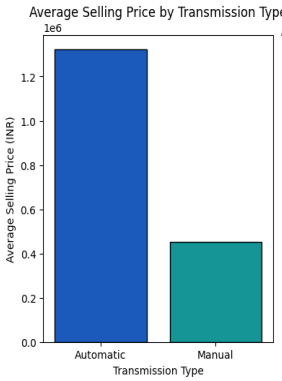


Figure 27

On average, diesel cars are distinctly the most expensive, whereas LPG vehicles are the cheapest. This may be due to factors such as better fuel efficiency, higher torque and longer lifespan that boost the market value of diesel vehicles.

Automatic transmission cars are also significantly pricier than manual vehicles on average, possibly due to their advanced

Finally, we explore how the seller type affects the prices of cars. The graph below displays the range and average values of selling prices by each seller type, where cars sold by dealers are the most expensive and the ones sold by individuals are the cheapest on average.

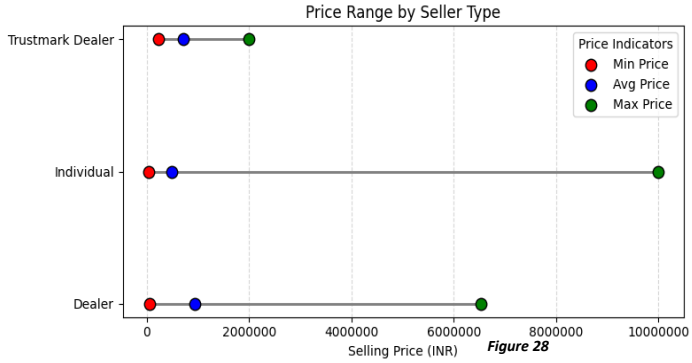


Figure 28

The high dealership prices could be possibly justified by the trust and credibility associated with it, as well as the additional services and potential warranties provided. We also see that individual sellers exhibit the widest price range.

Hence, it can be concluded that our analysis aligns with the research findings of (Ferizqa, 2023).

2.5 FEATURE SELECTION

Feature importance scores are used to determine the relative importance of each feature in a dataset, when building a predictive model. They help to rank the features based on how much they contribute to the final prediction. (Azaria, 2022)

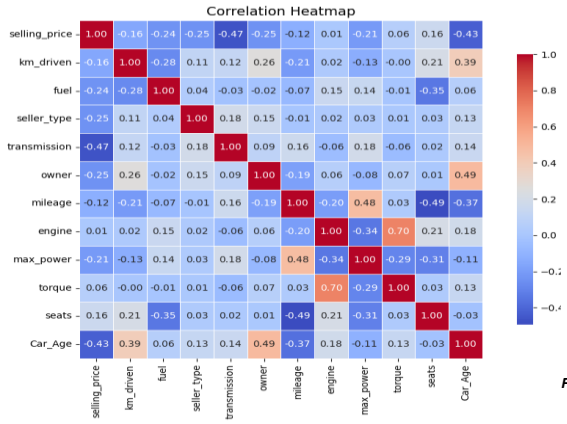


Figure 29

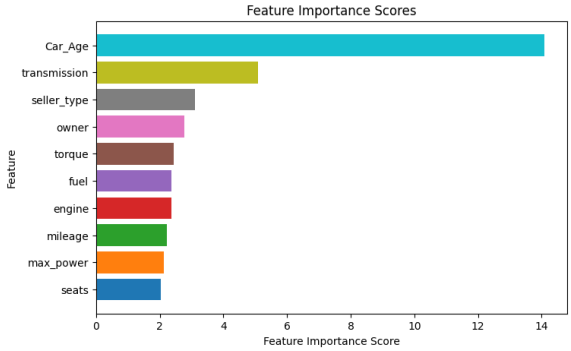


Figure 30

A correlation heatmap was plotted to display the correlation coefficients between the variables in the dataset, and identify potential multicollinearity.
High correlation among predictor variables in a regression model is known as multicollinearity, and can distort the interpretation of individual effects. However, in our case, the correlations among features are low to moderate, so we do not need to remove any variables when fitting our regression models.

A feature importance score bar graph in figure 30 was then plotted to visualize the most influential factors in predicting car prices. We see that car age, transmission, seller type and number of previous owners are the most important variables to predict the selling price. The number of seats in the car seems to provide the least contribution to the analysis.

2.6 REGRESSION MODELS

Various models were applied to the dataset to determine the most accurate model for predicting car prices. The data was standardized using Standard scaler to ensure uniform feature scaling, and then split into training and testing sets in an 80:20 ratio. The following metrics are used to evaluate model performance*:*

R-squared - represents the proportion of variability in the dependent variable that is explained by the regression model.

MSE (Mean Squared Error) - calculates the average of squared differences between actual and predicted values in the dataset.

RMSE (Root Mean Squared Error) - the square root of MSE, representing the standard deviation of residuals (prediction errors).

MAE (Mean Absolute Error) – computes the average of absolute differences between actual and predicted values in the model.

The below table shows the models used for this task and their corresponding evaluation metric values:

Model	R Squared	Mean Squared Error	Mean Absolute Error	Root Mean Squared Error
XGBRegressor	0.903892	2.27212e+10	78809	150735
GradientBoostingRegressor	0.898114	2.40871e+10	94920.2	155200
RandomForestRegressor	0.888308	2.64054e+10	83412.3	162497
DecisionTreeRegressor	0.827669	4.07411e+10	103039	201844
KNeighborsRegressor	0.729498	6.39501e+10	115396	252884
LinearRegression	0.484987	1.21755e+11	204176	348935

Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters, to enhance its performance. (Geeksforgeeks, 2025)
It was applied to the above models, resulting in improved accuracy for XGBoost (0.917846), Gradient Boosting (0.907297), and K-Nearest Neighbors (0.838825) models. However, it did not enhance the performance of the Random Forest and Decision Tree models, so the default configurations were used instead.

Due to their high R squared values and low error values, XGBoost and Gradient Boosting regressor models can be identified as the best models for predicting the selling price of cars. In contrast, Linear Regression performs the worst, exhibiting poor predictive accuracy and high error metrics.
It is noteworthy that the research conducted by (IJRASET, 2022) utilised the Random Forest Regressor model, which supposedly displayed a very high prediction accuracy of 98%. However, in this analysis, the model was able to explain only 87% of the variance.

For each of the models, a scatter plot for actual vs predicted values, scatter plot for residual vs predicted values, and a distribution graph for residuals were plotted. The graphs of our best performing model (XGB regressor) are given below:

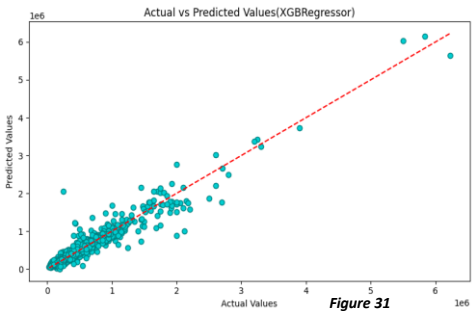


Figure 31

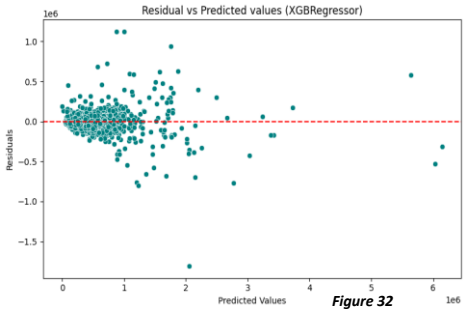


Figure 32

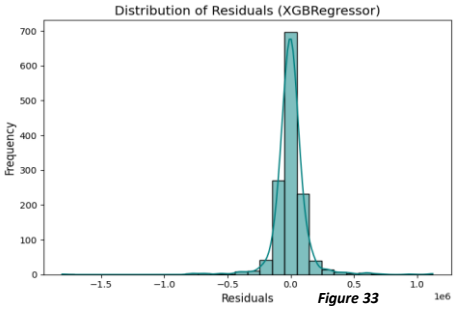


Figure 33

2.6.1 MULTIPLE LINEAR REGRESSION: Prior to fitting the model to the dataset, the following assumptions were considered;

1. A linear relationship between the dependent variable and the independent variables (Linearity).
2. Error terms are independent (No Autocorrelation)
3. Error terms have a constant variance (Homoscedasticity).
4. Error terms are normally distributed (Normality).

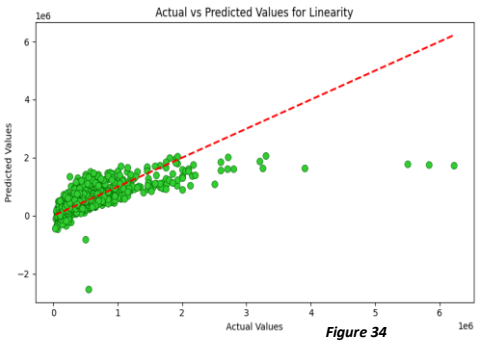


Figure 34

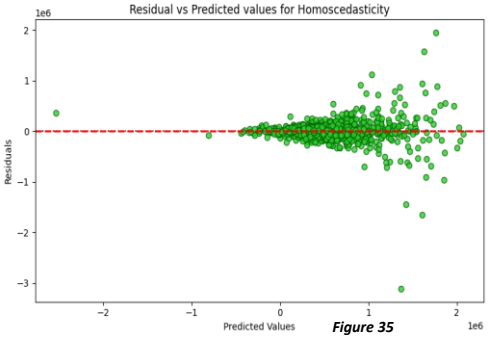


Figure 35

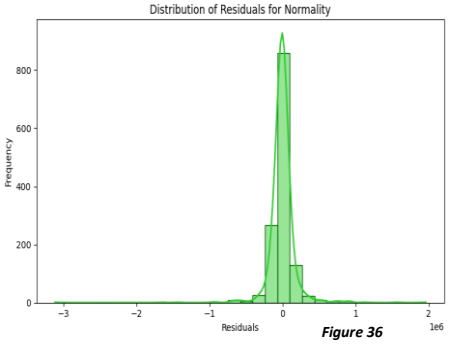


Figure 36

Although figure 34 does not show a perfect straight-line fit, it still suggests that the relationship between actual and predicted values aligns with the model's linearity assumption. In figure 35, the red line centered at zero indicates that the error terms are roughly unbiased, validating the assumption of homoscedasticity. The histogram shows that the errors follow an approximately normal distribution, confirming that this assumption is satisfied as well. Hence, the linear regression model can be used to predict car prices.

TASK 3: CLASSIFICATION

3.1 INTRODUCTION

Classification is a supervised learning method that sorts data points into predefined groups called classes, based on their features or characteristics. (Belcic, IBM, 2024)
The dataset selected for this task is the 'Heart Failure Prediction' dataset obtained from the Kaggle. The final goal of this task is to classify individuals as either having heart disease or being healthy, based on various health indicators.

3.2 EXISTING LITERATURE

A previous study conducted by (Ozcan, 2022) found that men are more likely to be diagnosed with heart disease than women. Half of the patients experience asymptomatic chest pain (which occurs without symptoms) and most of them are diagnosed with heart disease. Also, patients who have exercise-induced chest pain are predominantly diagnosed with heart disease, whereas those without it tend to be healthy. Additionally, the slope of the peak exercise ST segment shows that heart disease patients typically have a flat slope, while healthy individuals show an upward slope.

Furthermore, a separate analysis done by (Arxiv, 2024) states that cholesterol has weak correlations with other variables, suggesting that it may have a little predictive power in the model. Logistic Regression and K-Nearest Neighbors algorithms were utilised, with K-Nearest Neighbors achieving a higher accuracy of 87% compared to 85% for Logistic Regression.

3.3 RESEARCH QUESTIONS

1. How does the occurrence of heart disease vary between genders and across different age groups?
2. How are chest pain type and exercise-induced angina linked to heart disease diagnosis?
3. How do cholesterol levels and maximum heart rate influence the likelihood of heart disease?
4. Which classification model is most effective for distinguishing between individuals with and without heart disease?

3.4 EXPLORATORY DATA ANALYSIS (EDA)

The dataset consists of 563 male patients (75.6%) and 182 female patients (24.2%). Regarding their health status, 52.3% of the total patients are classified as normal, while 47.7% are diagnosed with heart disease, indicating a relatively balanced distribution for our analysis.

The bar chart below (figure 37) displays the distribution of normal and heart disease patients among the two genders, clearly showing a higher prevalence of heart disease among males compared to females, and aligning with previous findings.

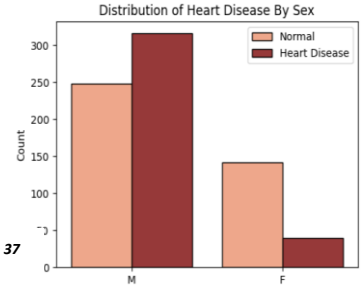


Figure 37

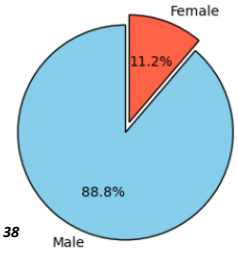


Figure 38

This trend is further highlighted in figure 38, which presents the gender breakdown of the total positive heart disease cases.
Only 11.2% of the positive heart disease cases are females, which could be a reflection of various factors such as biological differences, lifestyle choices, or the demographics of the dataset itself.

The diagrams below illustrate heart disease variation across different age groups. Figure 39 shows an increasing proportion of heart disease patients with age, peaking at around 60 years old. In contrast, younger age groups show a higher proportion of healthy individuals, indicating that the condition is less prevalent among them. There is a significant rise in heart disease cases from the 40-49 to 50-59 group, with a rapid drop after age 70 .

Therefore, it is evident that both age and gender play a major role in determining the risk of heart disease, with males and individuals aged 55 and older being more susceptible.

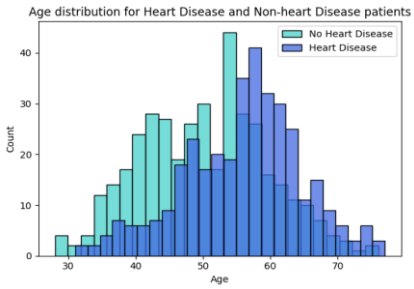


Figure 39

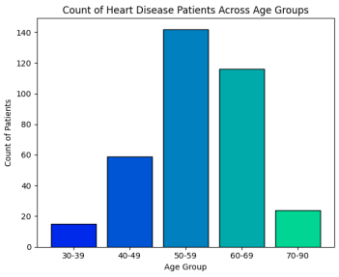


Figure 40

Next, we evaluate the relationship between chest pain type and the likelihood of being diagnosed with heart disease:

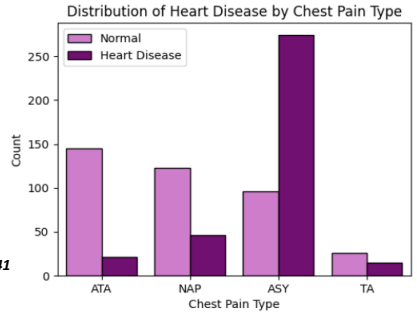


Figure 41

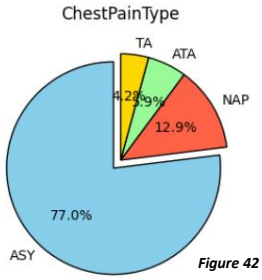


Figure 42

Figure 41 clearly shows that a significant proportion of patients experiencing asymptomatic chest pain (ASY) are diagnosed with heart disease.

In fact, 77% of all positive heart disease cases involve individuals who have experienced ASY, a condition that presents without obvious symptoms or warning signs.

This highlights the importance of proactively screening and addressing heart disease, even in individuals who may not exhibit obvious signs of the condition.

We further examined the relationship between exercise-induced angina and heart disease diagnosis, which also aligned with the findings of (Ozcan, 2022).

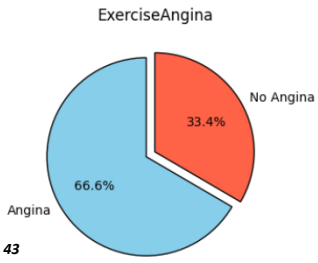


Figure 43

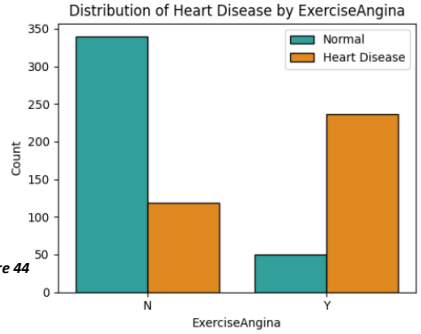


Figure 44

The data reveals that most patients do not experience exercise-induced angina and are generally classified as normal. However, among those who do experience it, the majority are diagnosed with heart disease, indicating a strong association between the two.
66.6% of all heart disease patients have exercise-induced angina, as shown in figure 43.

The ST segment in an ECG reflects the heart's electrical activity. In heart disease patients, the peak exercise ST segment is typically flat, indicating reduced blood flow or blocked arteries, which hinders the heart's response to stress.
In contrast, healthy individuals exhibit an upward-sloping ST segment, showing proper heart function and adaptation to physical activity.

Figure 45 illustrates this distinction, showing mostly flat peaks for heart disease patients and upward slopes for healthy individuals.

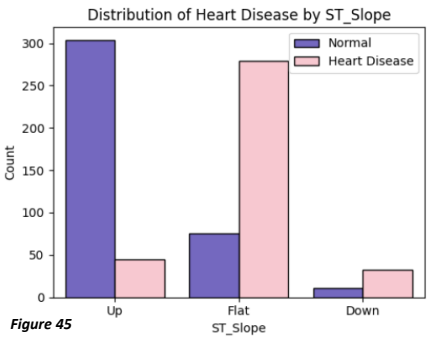


Figure 45

Finally, we explore how patients' cholesterol levels and maximum heart rates influence the likelihood of heart disease.

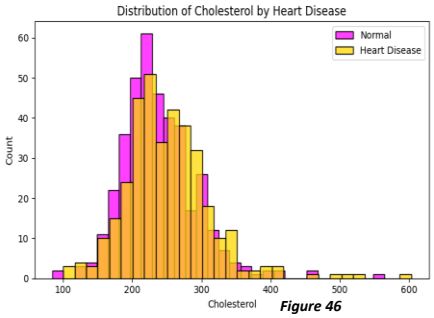


Figure 46

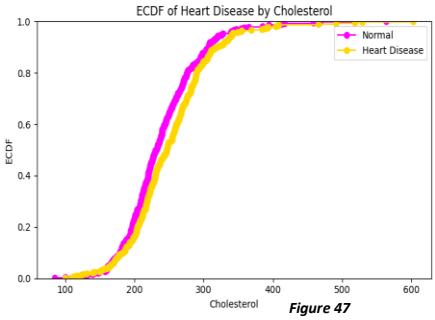


Figure 47

The cholesterol distributions of both groups peak between 200-250 mg/dL, but the heart disease group has a wider spread and more cases above 300 mg/dL. However, significant overlap suggests high cholesterol is common in heart disease patients but not a definitive predictor of the condition. The ECDF plot shows the heart disease curve slightly more to the right, indicating only slightly higher cholesterol levels than normal at the same cumulative probability.

Hence, cholesterol is not a reliable distinguishing factor.

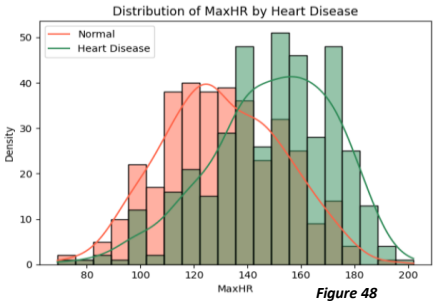


Figure 48

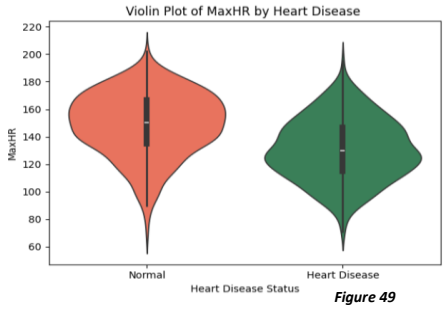


Figure 49

Healthy individuals generally have higher MaxHR values (140–180), while those with heart disease tend to have lower MaxHR values (100–140), indicating reduced cardiovascular efficiency. The distribution for healthy individuals is broader with a higher median (150–160), while heart disease patients show a narrower range with a lower median (120–130).This pattern suggests that MaxHR could be a useful indicator for diagnosing heart disease.

3.5 FEATURE SELECTION

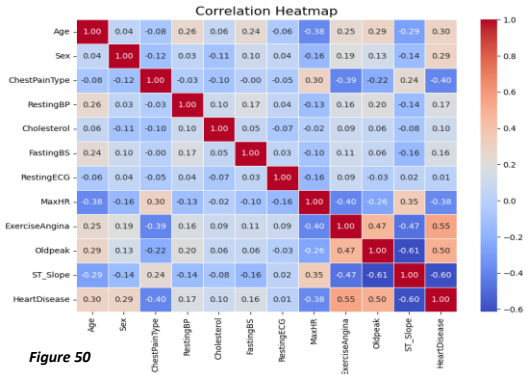


Figure 50

The heatmap analysis in figure 50 revealed no strong linear relationships between features, so no variables were removed based on correlation.

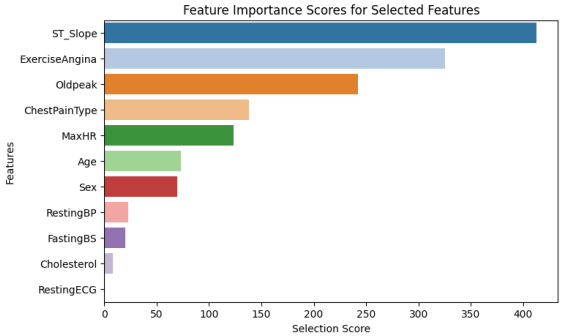


Figure 51

The feature importance score graph was then used to select the most significant features, where ST Slope and Exercise Angina were identified as the most important features due to their highest selection scores. In contrast, Resting EG had a selection score of zero, indicating that it was the least significant feature and was consequently dropped from the analysis. We see that Cholesterol also has a low importance score, which aligns with (Arxiv, 2024)'s discovery that it may have low predictive power in the model.

3.6 CLASSIFICATION MODELS

Several classification models were employed to classify individuals as either having heart disease or being. The dataset was divided into training and testing sets in an 80:20 ratio, after which the following models were evaluated:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.873333	0.837838	0.898551	0.867133
Gradient Boosting	0.86	0.833333	0.869565	0.851064
CatBoost	0.853333	0.821918	0.869565	0.84507
Random Forest	0.84	0.808219	0.855072	0.830986
XGBClassifier	0.826667	0.779221	0.869565	0.821918
K-Nearest Neighbors	0.806667	0.777778	0.811594	0.794326
Decision Tree	0.773333	0.761194	0.73913	0.75

The performance of the model is assessed using the following metrics :

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \qquad Recall = \frac{true\ positive}{true\ positive + false\ negative}$$

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ positive + false\ negative + true\ negative}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

- **Accuracy**- measures the proportion of correct predictions made by the model, considering both true positives and true negatives.
- **Precision** - measures the proportion of true positive predictions out of all positive predictions made by the classifier.
- **Recall** - measures the proportion of true positive predictions identified correctly out of all actual positive instances.
- **F1 score** - measures the model's performance by combining precision and recall into a single metric.

Since our dataset is relatively balanced (52:48), accuracy is a reliable metric for evaluating model performance as it is not biased toward the majority class.

Hyperparameter tuning was applied to all the above models, resulting in improved accuracy for Random Forest (0.86), XGB Classifier (0.86), K-Nearest Neighbors (0.8467) and Decision Tree (0.8133) models. However, it did not enhance the performance of the Logistic Regression, Cat Boost and Gradient Boosting models, so the default configurations were used instead.

Now all the models achieved over 80% accuracy, with Logistic Regression, Gradient Boosting, Random Forest and XGB Classifier standing out as the top performers with strong predictive capabilities.

- **Logistic Regression** - a classification model that uses several independent parameters to predict a binary-dependent outcome - i.e. whether a patient is diagnosed with heart disease or not. (Shah, 2023)
- **Gradient Boosting** - An iterative ensemble learning technique that minimizes a loss function by iteratively fitting new models to the negative gradient of the loss function, gradually improving prediction accuracy. (Kurama, 2024)
- **Random Forest** - a machine learning algorithm that creates an ensemble of multiple decision trees to reach a singular, more accurate prediction or result. (Donges, 2024)
- **XG Boost** - eXtreme Gradient Boosting is an ensemble learning method that combines the predictions of multiple weak learners (typically decision trees) to produce a strong predictive model. (APMonitor, 2023)
- **CatBoost classifier -** uses a combination of ordered boosting, random permutations and gradient-based optimization to achieve high performance on large and complex datasets with categorical features. (Oppermann, 2023)

The classification report and confusion matrix for our most accurate model (Logistic Regression) are given below:

(Confusion matrix is used to visualise the actual and predicted classes)

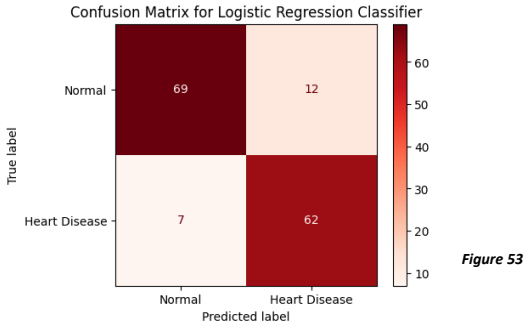


Figure 53

Classification Report for Logistic Regression:				
	precision	recall	f1-score	support
0	0.91	0.85	0.88	81
1	0.84	0.90	0.87	69
accuracy			0.87	150
macro avg	0.87	0.88	0.87	150
weighted avg	0.88	0.87	0.87	150

The metrics in the report suggest a fairly balanced performance for both groups, with a greater precision in identifying normal cases and a slightly higher recall for heart disease. The overall accuracy of 0.87 suggests good classification performance. The confusion matrix in figure 53 shows correctly identified 69 normal cases (true negatives) and 62 heart disease cases (true positives). However, it misclassified 12 normal cases as heart disease (false positives) and 7 heart disease cases as normal (false negatives), indicating room for improvement in detecting normal cases.

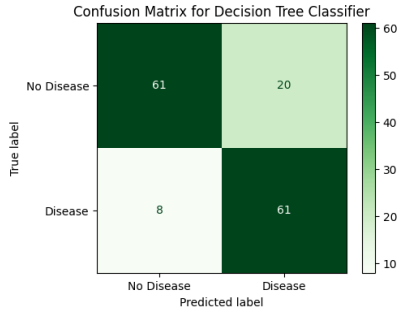


Figure 54

In comparison, the confusion matrix for our least accurate model, the Decision Tree Classifier, is shown in Figure 54. This model correctly classifies 61 normal cases (true negatives) and 61 heart disease cases (true positives). However, it misclassifies 20 normal cases as heart disease and 8 heart disease cases as normal, highlighting its inability to match the performance of the top model, which delivers more accurate classifications and fewer misclassifications.

The ROC curve shows the performance of a binary classifier across different threshold values, plotting the true positive rate against the false positive rate. A curve towards the top left corner signifies superior performance, and the area under the curve reflects overall model effectiveness, with higher values representing better discrimination power.

In this case, XGBoost has the highest AUC value of 0.92.
However, when considering all metrics, Logistic Regression proves to be the best overall model for distinguishing between heart disease and normal patients.

Further, as opposed to the previous study by (Arxiv, 2024), our Logistic Regression model exceeds the accuracy of K-Nearest Neighbors, which only achieved a tuned accuracy of 84.67%.

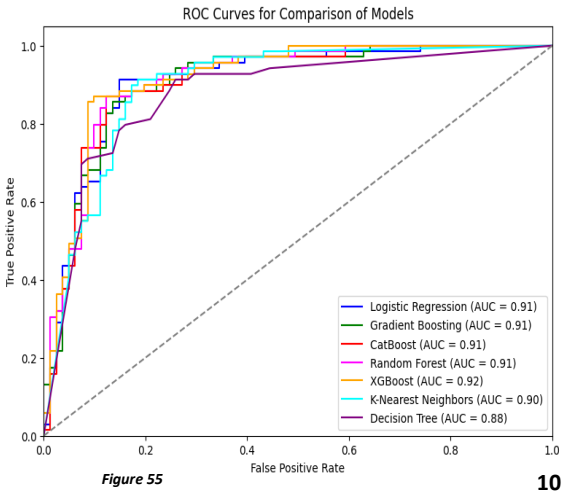


Figure 55

REFERENCES

Ankita. (2025, March 10). *K-Means: Getting the Optimal Number of Clusters*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

APMonitor. (2023, November 24). *Machine Learning for Engineers*. Retrieved from APMonitor: https://apmonitor.com/pds/index.php/Main/XGBoostClassifier

Arxiv. (2024, September 5). *Classification and Prediction of Heart Diseases using Machine Learning Algorithms*. Retrieved from arxiv: https://arxiv.org/html/2409.03697v1

Azaria, N. (2022, June 06). *Feature Importance: 7 Methods and a Quick Tutorial*. Retrieved from Aporia: https://www.aporia.com/learn/feature-importance/feature-importance-7-methods-and-a-quick-tutorial/

Belcic, I. (2024, October 15). Retrieved from IBM: https://www.ibm.com/think/topics/classification-machine-learning

Belcic, I. (n.d.). *What is supervised learning?* Retrieved from IBM: https://www.ibm.com/think/topics/supervised-learning

Donges, N. (2024, November 26). *Random Forest: A Complete Guide for Machine Learning*. Retrieved from Built In: https://builtin.com/data-science/random-forest-algorithm

Ferizqa, D. (2023, July 12). *Statistical Analysis: Used Car Price*. Retrieved from Medium: https://medium.com/@dsyafz/statistical-analysis-used-car-price-132b073439d5

Geeksforgeeks. (2025, February 27). *Gaussian Mixture Model*. Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/gaussian-mixture-model/

Geeksforgeeks. (2025, February 04). *Hierarchical Clustering in Machine Learning*. Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/hierarchical-clustering/

Geeksforgeeks. (2025, March 11). *Hyperparameter tuning*. Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/hyperparameter-tuning/

Geeksforgeeks. (2025, January 13). *Regression in machine learning*. Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/regression-in-machine-learning/

IBM. (2021, 09 23). Retrieved from IBM: https://www.ibm.com/think/topics/unsupervised-learning

IJRASET. (2022). *Prediction of Price for Cars Using Machine Learning*. Retrieved from Academia: https://www.academia.edu/82831177/Prediction_of_Price_for_Cars_Using_Machine_Learning

Jetir. (2024). *Customer Segmentation Using Machine Learning for Shopping Mall Customers.* Retrieved from https://www.jetir.org/papers/JETIR2409386.pdf

Kurama, V. (2024, September 13). *Gradient Boosting In Classification: Not a Black Box Anymore!* Retrieved from Digital Ocean: https://www.digitalocean.com/community/tutorials/gradient-boosting-for-classification

Oppermann, A. (2023, April 06). *What Is CatBoost?* Retrieved from Built In: https://builtin.com/machine-learning/catboost

Ozcan, M. (2022, December 21). *A classification and regression tree algorithm for heart disease modeling and prediction*. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S2772442522000703#b31

Premanand, S. (2024, November 6). *PCA in ML*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2022/07/principal-component-analysis-beginner-friendly/

Saji, B. (2025, March 10). *Elbow Method for Optimal Cluster Number 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/

Shah, D. (2023, March 31). *Logistic Regression: Definition, Use Cases, Implementation*. Retrieved from v7labs: https://www.v7labs.com/blog/logistic-regression

Sharma, P. (2025, January 7). *What is K-Means Clustering?* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#What_Is_K-Means_Clustering?

Singhal, M. (2020, Apri 10). *Analytics Vidhya*. Retrieved from Medium: https://medium.com/analytics-vidhya/mall-customers-cluster-analysis-b2ece6effdaa