UNIVERSITY OF LONDON | LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MACHINE LEARNING WITH PYTHON

MODULE: Machine Learning (ST3189)

UOL STUDENT NUMBER: 210475649

PAGE COUNT – 10 (Excluding Cover Page, Table of Contents and Bibliography)

TABLE OF CONTENTS

TASK 1: UNSUPERVISED LEARNING

INTRODUCTION

Unsupervised learning refers to a machine learning approach where the model learns patterns and structures from input data without explicit supervision or labeled outcomes (Brownlee, 2023).

The dataset used for this task is the 'Mall Customer Segmentation' dataset obtained from the Kaggle. The aim is to perform customer segmentation to identify groups of customers with similar characteristics and behaviors, facilitating targeted marketing strategies.

EXISTING LITERATURE

In the mall owner's study, it was identified that women exhibit higher spending scores while men generally have higher annual incomes (Singhal, 2020). In a research done, it was found that younger females are the primary spenders, suggesting a focus on targeted marketing strategies to leverage their higher spending scores and drive sales growth (Joberri, 2023).

RESEARCH QUESTIONS

1. *Do women tend to have higher spending scores than men, despite men having higher annual incomes on average?*
2. *Are younger females identified as the primary spenders in this market?*
3. *How do the characteristics of each cluster identified through K-Means and Hierarchical clustering differ, and what insights do these differences offer regarding customer segmentation?*

Exploratory Data Analysis (EDA)

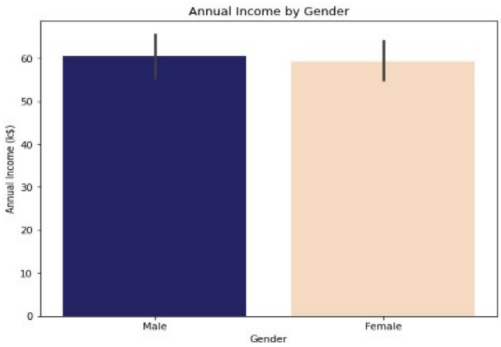

Figure 1

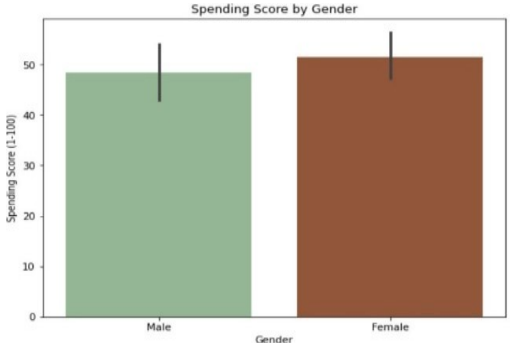

Figure 2

From Figure 1, it is evident that, on average, males tend to have a slightly higher annual incomes compared to females. Based on Figure 2, it can be seen that females tend to exhibit slightly higher spending scores compared to males. As observed in the previous study, our analysis confirms that women tend to have higher spending scores, while men typically have higher annual incomes.

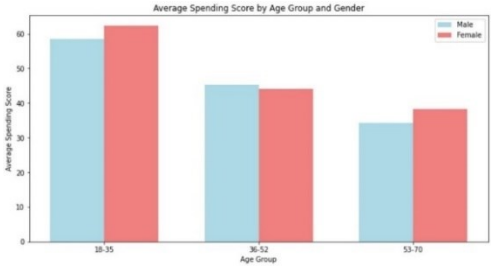

Figure 3

The bar plot further illustrates that overall females tend to have a slightly higher spending score. Specifically, among females aged 18-35, the average spending score is the highest, confirming previous research indicating that younger females are the primary spenders.

CLUSTERING

K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The objective is to minimize the sum of distances between data points and their assigned centroid, grouping similar points together (Sharma, 2023).

Hierarchical clustering refers to a method of grouping objects based on their similarity. It begins by treating each object as its own cluster and progressively merges clusters until only one remains, forming a dendrogram that visually represents the hierarchical relationships between clusters (Karabiber, 2024).

Both K-means and Hierarchical clustering were used to segment customers based on their income and spending behavior, aiming to enhance future marketing strategies.

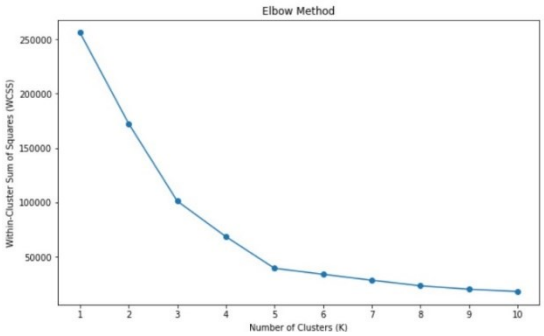

Figure 4: Elbow Graph

For K-Means clustering, the elbow method was used to determine the optimal number of clusters. The elbow point, which occurs around 5 clusters, indicates that this is the best number of clusters for our dataset. Beyond this point, the decrease in within-cluster sum of squares (WCSS) becomes minimal, suggesting that additional clusters may not substantially enhance cluster separation.

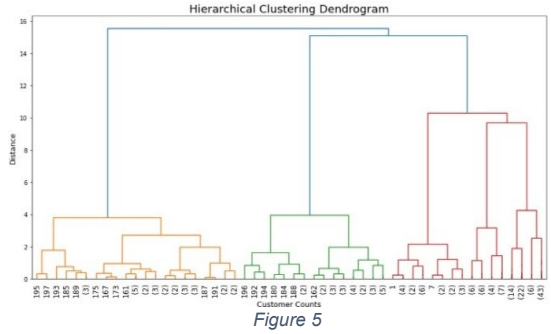

Figure 5

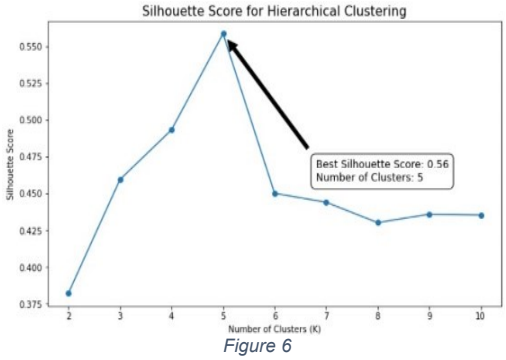

Figure 6

For Hierarchical clustering, the dendrogram was used to understand the hierarchical arrangement of clusters. Using silhouette scores, the optimal cluster number was identified as 5. This analysis suggests that 5 clusters effectively capture the underlying patterns in our data for this method as well.

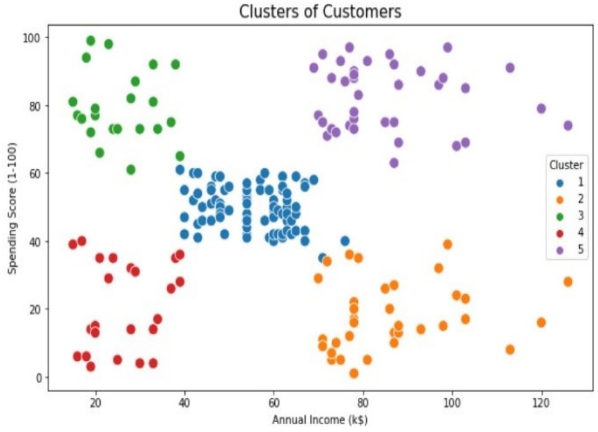

Figure 7: Scatter plot for K- Means clusters

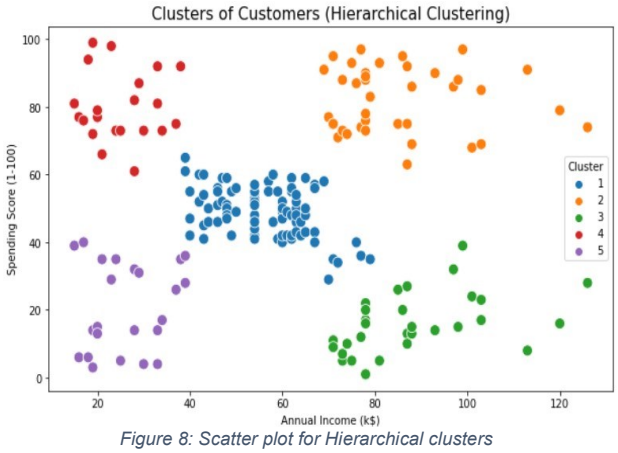

Figure 8: Scatter plot for Hierarchical clusters

In comparing K-Means and Hierarchical clustering results, the (blue) cluster 1 consistently depicts customers with moderate income and moderate spending across both methods. However, notable differences emerge in other clusters. In K-Means, the (red) cluster 4 represents lower income and lower spending, contrasting with Hierarchical clustering where it reflects lower income but higher spending. Similarly, K-Means (green) cluster 3 indicates lower income and higher spending, while in Hierarchical clustering, it signifies higher income but lower spending. The most noticeable difference is seen in the interpretation of the (purple) cluster 5, which in K-Means denotes high income and high spending, but in Hierarchical clustering, it represents lower income and lower spending.

The analysis of cluster sizes for both K-Means and Hierarchical clustering revealed consistent results, where Cluster 1 had the highest number of customers and Cluster 3 had the lowest out of the 5 clusters identified. After conducting further analysis using bar plots, we gained insights into how each cluster differs in terms of age, income, and spending behavior, as illustrated below.

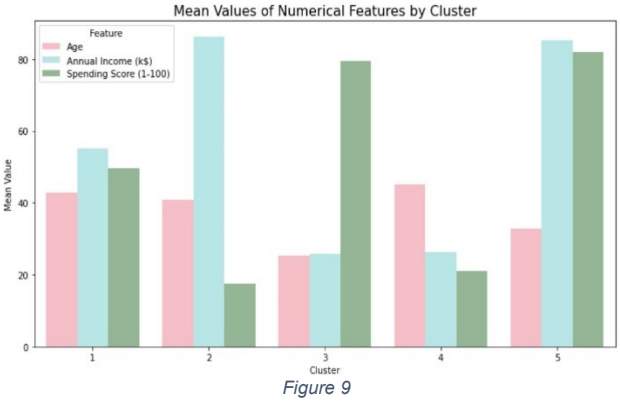

Figure 9

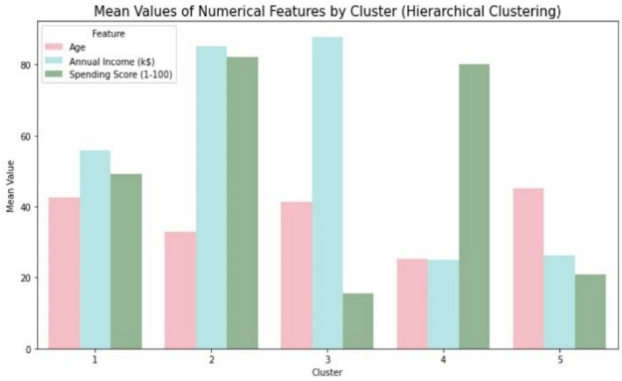

Figure 10

These results highlight subtle variations in customer segmentation between the K-Means and Hierarchical clustering methods. In K-Means clustering, Cluster 2 exhibits the highest annual income, while in Hierarchical clustering, this is observed in Cluster 3. Conversely, Cluster 3 in K-Means has the lowest annual income, whereas in Hierarchical clustering, it's Cluster 4. Regarding spending score, Cluster 5 has the highest score in K-Means, while in Hierarchical clustering, it's Cluster 2. Similarly, for age, Cluster 4 has the highest average age in K-Means, while in Hierarchical clustering, it's Cluster 5. These differences offer valuable insights for customer segmentation strategies.

Targeting Cluster 2 in K-Means could involve high-value offerings or luxury products. In Hierarchical clustering, the focus might be on retaining customers with high spending potential.

Efforts aimed at Cluster 3 in K-Means could concentrate on affordability or value-based promotions, while in Hierarchical clustering, the goal might be to uplift customers to increase spending.

Understanding these subtle differences enables businesses to tailor marketing strategies more precisely. Strategies for Cluster 4 in K-Means might cater to older demographics with specific preferences. In Hierarchical clustering, they could target younger customers with different needs and expectations.

By leveraging these insights, businesses can optimize engagement and revenue potential across diverse customer segments.

TASK 2: REGRESSION

INTRODUCTION

Supervised learning is a machine learning model where models are trained on labeled data to make predictions or decisions.

Regression analysis identifies how changes in one or more independent variables affect the dependent variable. It provides estimates, predictions, and insights into how variables interact, allowing for better control and understanding of their relationships in the data (Mehta, 2023).

The regression analysis is based on the 'Boston House Prices' dataset obtained from Kaggle, featuring 14 variables, including the continuous target variable representing median housing prices.

EXISTING LITERATURE

In prior research using the Boston Housing dataset, it was identified that the number of rooms (RM) and the percentage of lower status of the population (LSTAT) as significant features (Mason, 2019).

RESEARCH QUESTIONS

1. *Does the level of crime in Boston correlate with the value of homes, suggesting a potential impact of safety on property values?*
2. *How does the percentage of lower status of the population (LSTAT) affect median housing prices (MEDV)?*
3. *How do different models perform in predicting the median home values (MEDV)?*

Exploratory Data Analysis (EDA)

In the data cleaning process, the two columns 'CHAS' and 'ZN' were excluded from the analysis. The 'ZN' column was excluded as its values were primarily clustered around zero within the middle 50% of the data, indicating limited usefulness for predicting 'MEDV'. Similarly, the 'CHAS' column, containing only zeros, lacks diversity and is therefore excluded.

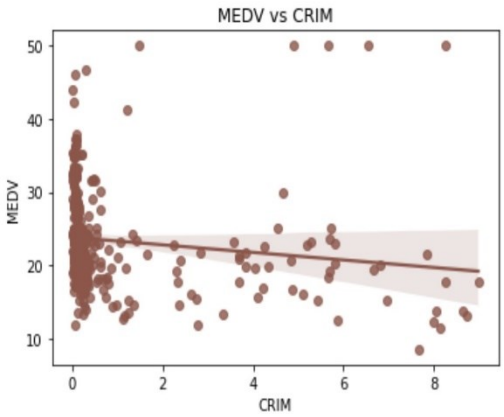

Figure 11: MEDV vs CRIM Scatter Plot

Using a scatter plot, we analyzed the relationship between the per capita crime rate by town (CRIM) and the median value of owner-occupied homes (MEDV) in Boston. The graph showed a negative correlation between these two variables, with a correlation coefficient of -0.15. This suggests that as the crime rate increases, the median value of homes tends to decrease. Therefore, there appears to be a potential impact of safety on property values in Boston, suggesting that areas with higher crime rates may experience lower home values.

The figure on the right shows how the median value of owner-occupied homes (MEDV) varies with the percentage of lower status population (LSTAT). It's evident from the scatter plot and the trend line that there's a strong negative correlation between these two variables. The correlation coefficient of -0.68 confirms this relationship, indicating that as the percentage of lower status population increases, the median value of homes tends to decrease significantly. This suggests that areas with a higher percentage of lower status population may have lower property values.

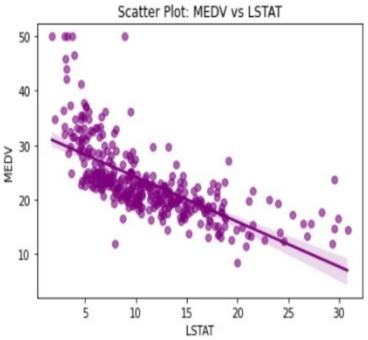

Figure 12: MEDV vs LSTAT Scatter Plot

Optimizing feature selection

Figure 13: Heatmap

As the next step, a heatmap was generated to examine correlation values among the feature variables. A correlation coefficient of 0.9 or higher was considered indicative of high correlation. The heatmap confirmed that only CRIM and RAD displayed a high correlation coefficient of 0.9, indicating a strong relationship. This suggests that the per capita crime rate by town (CRIM) and index of accessibility to radial highways (RAD) move together strongly in the same direction.

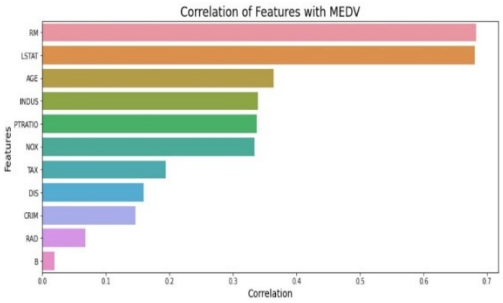

Figure 14: Correlation of features

The average number of rooms per dwelling (RM) shows the strongest correlation with MEDV, while the percentage of the lower status of the population (LSTAT) also demonstrates a significant correlation with MEDV, highlighting their importance as features. Consistent with the previous study, our analysis confirms the significance of RM and LSTAT in predicting house prices, as represented in the correlation plot.

To mitigate multicollinearity issues in our model, features such as RAD and B were dropped as they were considered least important.

REGRESSION MODELS

Various regression models were applied to the dataset to predict the median value of house prices. The dataset was split into training and testing data in a ratio of 80:20 and the models were then tested.

Model	R Squared	Mean Squared Error	Mean Absolute Error	Root Mean Squared Error
GradientBoostingRegressor	0.898002	5.03714	1.67041	2.24436
RandomForestRegressor	0.873182	6.26289	1.86787	2.50258
XGBRegressor	0.872336	6.30464	1.8057	2.5109
LinearRegression	0.695502	15.0375	2.91077	3.87782
DecisionTreeRegressor	0.678474	15.8785	2.65211	3.98478
KNeighborsRegressor	0.637458	17.904	2.37972	4.23132

When comparing the accuracy of the results, it was clear that Gradient Boosting Regressor, Random Forest Regressor, XGB Regressor, and Linear Regression were the most appropriate models for this dataset from the above table, so these were further explored.

The parameters described above can be explained as follows:

- R^2 score - This is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model (Taylor). A higher R^2 value signifies a better fit of the model to the data.

- Mean Squared Error (MSE) - This is the average of the squares of the errors, which are the differences between actual values and predicted values

- Mean Absolute Error (MAE) - This is a measure of errors between paired observations expressing the same phenomenon. It is the average absolute error between actual and predicted values. Absolute error, also known as L1 loss, is a row-level error calculation where the non-negative difference between the prediction and the actual is calculated (Allwright, 2022).

- Root Mean Squared Error (RMSE) - This is the square root of the mean of the square of all of the error. RMSE is considered an excellent general-purpose error metric for numerical predictions.

Gradient Boosting Regressor is the best fit model, demonstrating the highest R^2 score and the lowest MAE, MSE and RMSE. Following hyperparameter tuning the model's performance further improved, yielding an enhanced R^2 score of 0.902532, along with further reductions in MAE, MSE, and RMSE. This indicates that the hyperparameter tuning process refined the model's predictive capability, leading to enhanced accuracy in predicting median house prices.

Hyperparameter tuning for the Random Forest Regressor and XGB Regressor did not result in an increased R2 score, suggesting no improvement in the models' performance.

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.
3. Error terms have a constant variance (Homoscedasticity).
4. Error terms are normally distributed (Normality).

The following graphs were plotted to help understand whether any violations to the assumption of linearity exist.

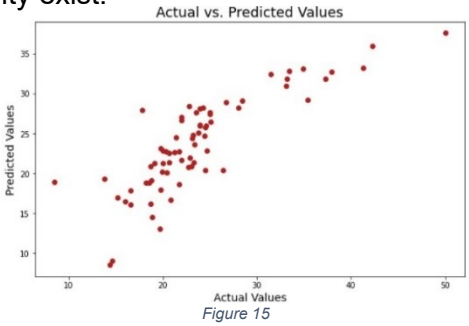

Figure 15

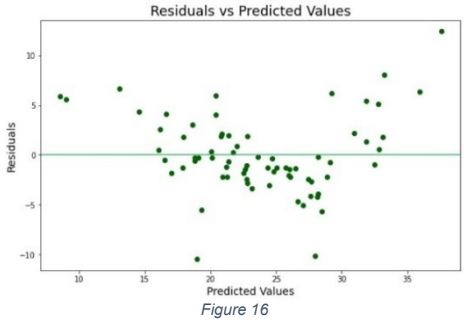

Figure 16

Although the Figure 15 does not show a perfect alignment along a straight line, it still indicates that the relationship between the actual and predicted values satisfies the linearity assumption for the model. Then the Residuals vs Predicted Values were plotted using a scatter plot (Figure 16). The green line, centered at zero, indicates that the error terms are approximately zero bias, satisfying the assumption of homoscedasticity (constant variance).

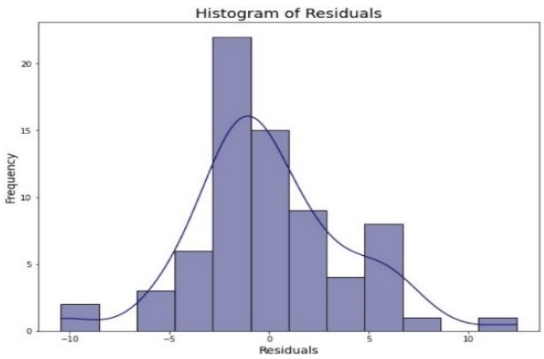

Figure 17

The histogram indicates that the errors are somewhat normally distributed, confirming that this assumption is also met.

TASK 3: CLASSIFICATION

INTRODUCTION

Classification is also a supervised learning method in machine learning where the goal is to categorize input data into predefined classes or categories based on their features.

The dataset selected for this task is the 'Heart Failure Prediction' dataset obtained from the Kaggle. The dataset aims to classify individuals as either having heart disease or being normal based on various health indicators. give the sentence in a better way

EXISTING LITERATURE

A previous study conducted on this dataset found that the majority of patients were male and that males were more likely to be diagnosed with heart disease than females. Additionally, about half of the patients had asymptomatic chest pain, and most of these patients were also diagnosed with heart disease (Ozcan, 2023). A research done on the prevalence and impact of cardiovascular diseases globally revealed that Logistic Regression Classifier model achieved the highest accuracy of 85.05% among four classification machine learning models created (Kothadia, 2022).

RESEARCH QUESTIONS

1. *How does the prevalence of heart disease vary across different age groups and between genders?*
2. *What is the relationship between chest pain type and the likelihood of having heart disease?*
3. *What is the best classification model to classify individuals with heart disease and those who are normal?*

Exploratory Data Analysis (EDA)

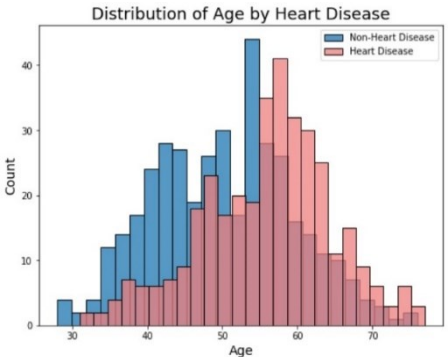

Figure 18

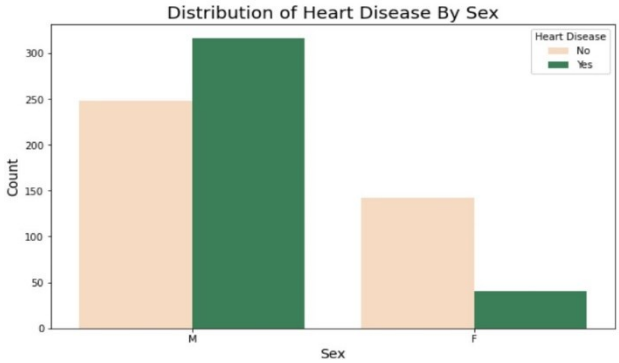

Figure 19

From the histogram, it can be seen that the prevalence of heart disease increases with age, peaking around 60 years old. Younger age groups show a higher proportion of individuals being normal h, suggesting that it is less common among younger individuals. There is a significant increase in heart disease cases starting from age 50, with a peak around 60, followed by a gradual decrease. The Figure 19 shows a higher prevalence of heart disease in males compared to females. Hence, it's quite evident that both age and gender significantly influence the likelihood of heart disease, with individuals aged 55 and above and males being at higher risk.

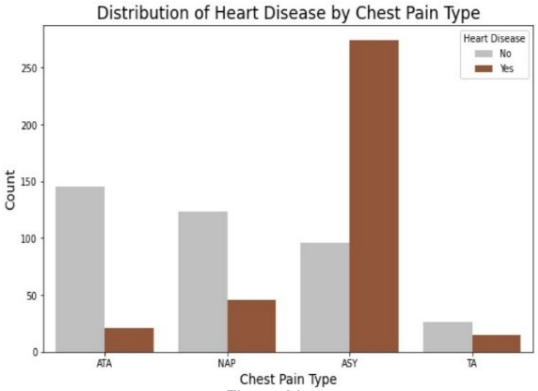

Figure 20

It is clearly seen that individuals with ASY (Asymptomatic) chest pain have the highest number of heart disease cases, while those with TA (Typical Angina) have the lowest. This suggests that many heart disease cases occur among those who do not experience chest pain symptoms, highlighting the importance of including asymptomatic individuals in heart disease screening and prevention efforts.

As seen in the previous study, we also observed in this study that a higher prevalence of heart disease among males compared to females, with asymptomatic chest pain being the most common symptom associated with heart disease.

Optimizing feature selection

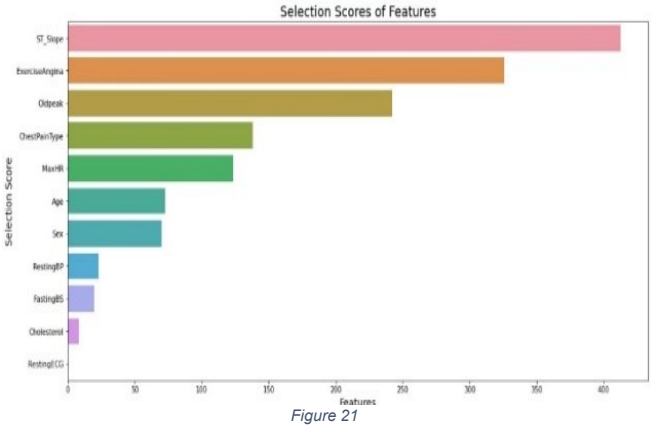

Figure 21

None of the features showed a strong linear relationship with a correlation coefficient very close to 1, from the heatmap. Therefore, none of the variables were dropped. Using the SelectKBest feature selection method, it was identified 'ST_Slope' as the most significant feature, as it attained the highest score indicating its importance as a feature. On the other hand, the selection score of 'RestingECG' was zero, indicating it was the least significant feature and therefore, was dropped from the analysis.

CLASSIFICATION MODELS

Various classification models were employed to classify individuals as either having heart disease or being normal. The dataset was split into training and testing data with an 80:20 ratio, followed by testing the models.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.873333	0.837838	0.898551	0.867133
Gradient Boosting	0.853333	0.821918	0.869565	0.84507
Random Forest	0.846667	0.802632	0.884058	0.841379
XGBClassifier	0.826667	0.779221	0.869565	0.821918
K-Nearest Neighbors	0.806667	0.777778	0.811594	0.794326
Decision Tree	0.733333	0.716418	0.695652	0.705882

As observed by the results above, Logistic Regression, Gradient Boosting, Random Forest and XGB Classifier have the highest accuracy scores. This suggests that these models provide accurate predictions, making them the best fit for further exploration.

- 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 (Paperspace).
- Random Forest - An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (MDPI, 2024).
- XGB Classifier - It's a powerful machine learning algorithm known for its efficiency, speed, and accuracy, which combines the predictions of multiple weak learners (GeeksforGeeks, 2023).
- Logistic Regression - A predictive modeling technique designed for binary classification tasks, distinguishing between two classes based on input features.

The evaluation metrics used for the model comparison can be explained as follows:

- Accuracy - measures the overall correctness of predictions made by a classifier.
- Precision - measures the proportion of true positive predictions among all positive predictions made by the classifier, indicating how many of the predicted positive instances are actually relevant (FasterCapital).
- Recall - measures the proportion of true positive predictions identified correctly out of all actual positive instances.
- F1-Score - A measure of a model's accuracy that combines precision and recall into a single metric, providing a better understanding of model performance (Kundu, 2022).
- A ROC curve (Receiver Operating Characteristic curve) - a graph showing the performance of a binary classifier across different threshold values, plotting the true positive rate against the false positive rate.
- Confusion Matrix - a table used to describe the performance of a model, displaying the counts of true positive, true negative, false positive, and false negative predictions.

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

Hyperparameter tuning was conducted for the logistic regression model, but the accuracy remained unchanged, indicating no improvement.

The classification report shows the precision for class 0 is 0.91, and for class 1 is 0.84, the model correctly identified 91% of normal cases and 84% of heart disease cases. In terms of recall, the model achieves 0.85 for class 0 and 0.90 for class 1 signify that the model captured 85% of normal cases and 90% of heart disease cases. The F1 score for normal cases is 0.88, and for heart disease, it's 0.87. These metrics collectively suggest a relatively balanced performance, with higher precision in detecting normal cases and slightly higher recall for heart disease. Overall accuracy is 0.87, suggesting the model performs well in correctly classifying both classes.

From the confusion matrix for Logistic Regression, it can be seen that the model correctly identified normal cases (true negatives) 69 times, but there were instances where it misclassified normal cases as heart disease (false positives) 12 times. Additionally, the model accurately predicted heart disease cases (true positives) 62 times, but it also incorrectly labeled some heart disease cases as normal (false negatives) 7 times. While the model shows good accuracy in detecting heart disease, there's room for improvement in correctly identifying normal cases.

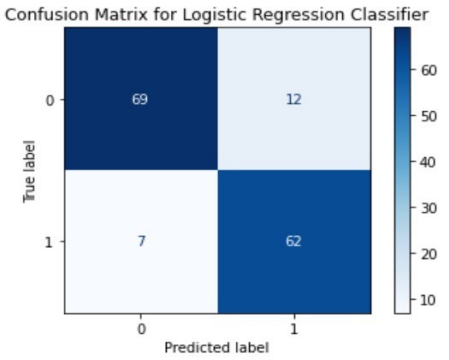

Figure 22

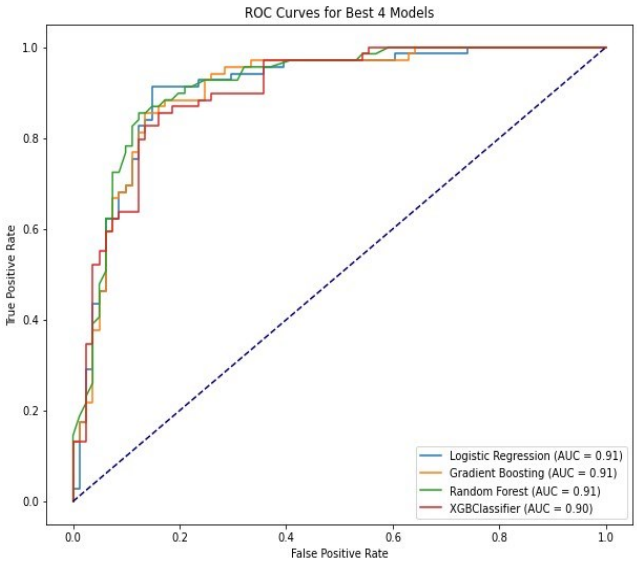

Figure 23

Among the four models analyzed, logistic regression, gradient boosting, and random forest classifiers demonstrated an AUC of 0.91, while the XGB classifier showed a slightly lower AUC of 0.90.

However, despite similarities with other models, logistic regression emerged as the best performer, with the highest precision, recall, and F1 score among the four models. This indicates that logistic regression effectively distinguishes between classes while maintaining a balance between precision and recall, making it the most suitable model for the classification task.

The best model for classifying individuals as either having heart disease or being normal was the Logistic Regression model, achieving an accuracy of 87.3%. This accuracy exceeded that of the previous study by (Kothadia, 2022).

BIBLIOGRAPHY

Allwright, S. (2022). How to interpret MAE (simply explained). Retrieved march 27, 2024, from https://stephenallwright.com/interpret-mae/

Brownlee, J. (2023). Supervised and Unsupervised Machine Learning Algorithms. Retrieved march 28, 2024, from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

FasterCapital. (n.d.). The Importance Of Residuals In Assessing Model Performance. Retrieved march 20, 2024, from https://fastercapital.com/topics/the-importance-of-residuals-in-assessing-model-performance.html

GeeksforGeeks. (2023). ML | XGBoost (eXtreme Gradient Boosting). Retrieved march 28, 2024, from https://www.geeksforgeeks.org/ml-xgboost-extreme-gradient-boosting/

Joberri. (2023). Mall Customers Segmentation And Clustering. Retrieved march 29, 2024, from https://medium.com/@iwojojo54/mall-customers-segmentation-and-clustering-57b0e5fbdb81

Karabiber, F. (2024). Hierarchical Clustering. Retrieved march 30, 2024, from https://www.learndatasci.com/glossary/hierarchical-clustering/

Kothadia, S. (2022). Classification algorithms in Python – Heart Attack Prediction and Analysis. Retrieved march 30, 2024, from https://www.analyticsvidhya.com/blog/2021/05/classification-algorithms-in-python-heart-attack-prediction-and-analysis/

Kundu, R. (2022). F1 Score in Machine Learning: Intro & Calculation. Retrieved march 27, 2024, from https://www.v7labs.com/blog/f1-score-guide

Mason, D. (2019). Machine Learning Regression and Data Analysis with the Boston Housing Dataset in Python — Part 2. Retrieved march 20, 2024, from Machine Learning — Regression Models — Boston Housing Dataset | by Ravi | Feb, 2024 | Medium

MDPI. (2024). Multiclass Sentiment Prediction of Airport Service Online Reviews Using Aspect-Based Sentimental Analysis and Machine Learning. Retrieved march 21, 2024, from https://www.mdpi.com/2227-7390/12/5/781

Mehta, D. (2023). What Is Regression Analysis? Types, Importance, and Benefits. Retrieved march 20, 2024, from https://www.g2.com/articles/regression-analysis

Ozcan, M. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. Retrieved march 31, 2024, from https://www.sciencedirect.com/science/article/pii/S2772442522000703#b31

Paperspace. (n.d.). Gradient Boosting In Classification. Retrieved march 20, 2024, from https://blog.paperspace.com/gradient-boosting-for-classification/

Sharma, N. (2023). K-Means Clustering Explained. Retrieved march 30, 2024, from https://neptune.ai/blog/k-means-clustering

Singhal, M. (2020). Mall Customers Cluster Analysis. Retrieved march 29, 2024, from https://medium.com/analytics-vidhya/mall-customers-cluster-analysis-b2ece6effdaa

Taylor, S. (n.d.). R-Squared. Retrieved march 27, 2024, from https://corporatefinanceinstitute.com/resources/data-science/r-squared/