UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Machine Learning Report

Student Number: 220639927

Module Code: ST3189

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

Table of Contents

Unsupervised Learning

Introduction

Unsupervised Learning is a machine learning technique that involves discovering insights and patterns in **unlabeled data** without human intervention (Google Cloud, 2025). Popular unsupervised learning techniques include clustering techniques such as K-Means and Hierarchical Clustering, which classify data into clusters based on similar characteristics, and dimensionality reduction methods like PCA and T-SNE, which reduces the number of features while preserving key information. The 'Palmer Penguins' dataset from the UCI Machine Learning Repository will be used to showcase the workflow unsupervised learning. Principal Component Analysis and K-Means clustering will be the unsupervised learning techniques utilized for this particular task. The uncovered insights and the findings will be presented and compared against a similar research conducted.

Existing Literature

Gentoo Penguins are the heaviest with longer flippers and shorter culmen depths, and are found only on Biscoe Islands. Adelie Penguins are found on all islands, and have shorter culmen lengths (under 40mm) and longer depths. Chinstraps have longer culmen measurements and are exclusive to Dream Islands. Over time, climate change has altered habitats, influencing penguin populations and plant life (Jadhav, 2022).

Research Questions

1. Can we accurately classify penguin species (Adelie, Gentoo, Chinstrap) based on the physical features bill length, bill depth, flipper length, and body mass?
2. Is it possible to distinguish male and female penguins from physical characteristics such as bill length, bill depth, flipper length, and body mass?
3. Do physical attributes of penguins vary across different islands (Biscoe, Dream, Torgersen) in the Palmer Archipelago?
4. Are there any unique characteristics in the principal components and the clusters created?

Exploratory Data Analysis

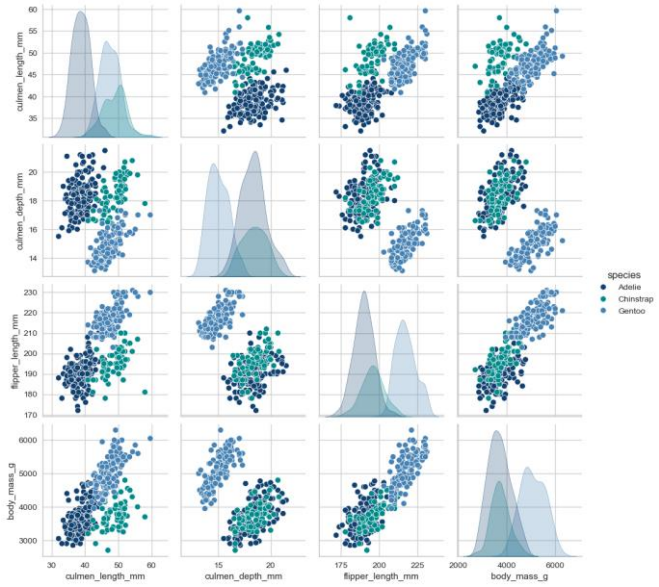


Figure 1

The scatterplot matrix (Figure 1) indicates that Gentoo penguins display comparatively higher body masses, flipper lengths, culmen lengths, but lowers culmen depths, while Adelie and Chinstrap showcase low to moderate body masses, flipper lengths, and higher culmen depths across species. The key distinguisher among Adelie and Chinstrap penguins is the higher culmen lengths observed in Chinstrap penguins. The boxplot grid (Figure 2) shows that male penguins have comparatively higher values for all physical traits but there's no substantial disparity among the genders. The violinplot grid (Figure 3) displays that Biscoe islands stands out for having penguins with

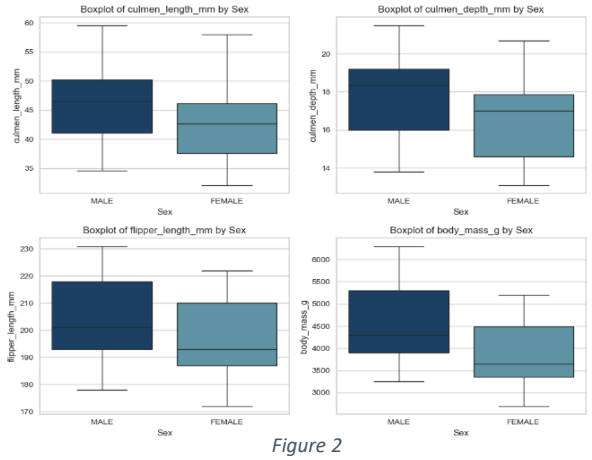


Figure 2

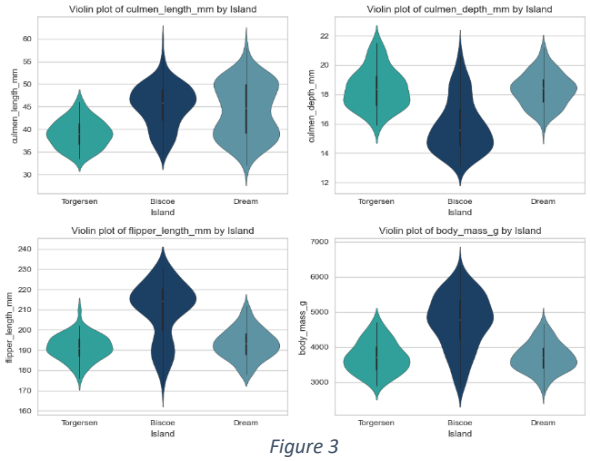


Figure 3

higher body masses, flipper lengths, and culmen lengths, but lower culmen depths. Apart from the higher culmen lengths observed in penguins from Dream island, there is no significant separation observed in the physical features of penguins from Torgersen and Dream islands. Additionally, the correlation heatmap indicated moderate to strong positive correlations across the features body mass, flipper length, and culmen length. Weak to moderate negative correlations were observed across culmen depth, culmen length, and flipper length.

Dimensionality Reduction – Principal Component Analysis

Principal component analysis (PCA) is a popular unsupervised learning technique that reduces the dimensionality of large and complex datasets, while preserving key insights and patterns. PCA creates uncorrelated variables called principal components, that are linear combinations or mixtures of the initial variables in the dataset (Built In, 2024).

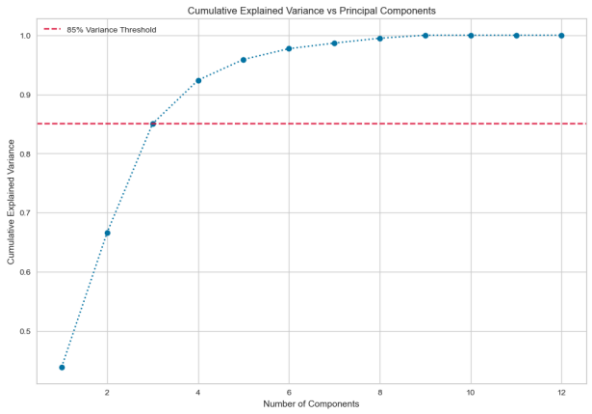


Figure 4

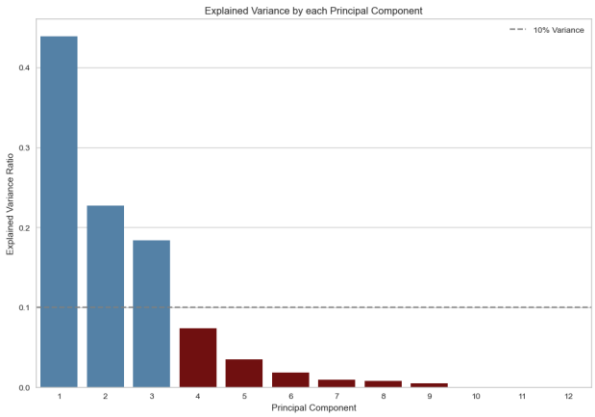
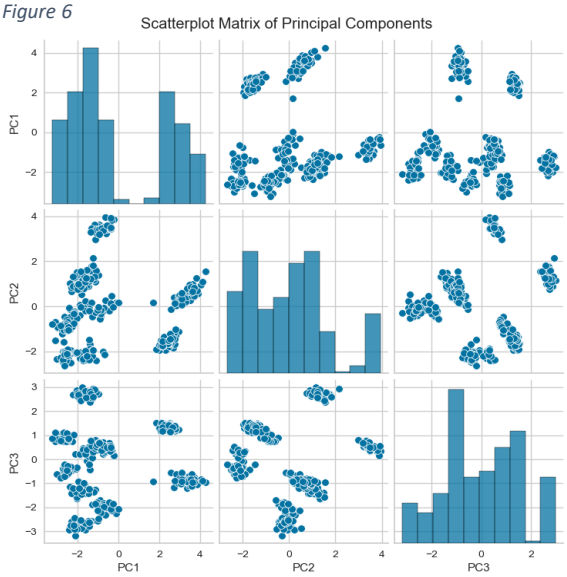


Figure 5

The data was standardized to ensure that the initial features equally contribute towards the principal component analysis. To determine the optimal number of principal components, a trade-off value between the number of components corresponding to the maximum cumulative explained variance (captures the most amount of information) and the number of components individually explaining more than 10% of the variance was obtained. For this particular task, **3 principal components were chosen** since 85% of the variance (cumulatively) was explained by the 3 components and only the first 3 components explained more than 10% of the variance individually (Figure 4 and Figure 5).

Figure 6



The scatterplot matrix (Figure 6) and the 3D projection of the principal components data show clear separation among tightly grouped data points that indicates distinct characteristics to be further explored using unsupervised techniques such as K-Means Clustering. The widely spread points depict the significant variance captured by the principal components from the original dataset while the random scattering indicates the weak correlations among these principal components.

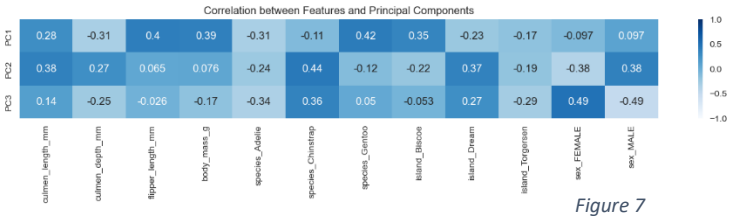


Figure 7

The correlation matrix (Figure 7) indicates how each feature in the original dataset contributes to each principal component. **PC1** shows moderate positive correlations with flipper length (0.40), culmen length (0.28), body mass (0.39), Gentoo penguins (0.42), and Biscoe islands (0.35), and weak to moderate negative correlations with Adelie penguins (-0.31), Dream islands (-0.23), and culmen depth (-0.31). **PC2** displays moderate positive correlations with culmen length (0.38), culmen depth (0.23), Chinstrap penguins (0.44), Dream islands (0.37), and male penguins (0.38), and weak to moderate negative correlations are observed with female sex (-0.38), Adelie species (-0.24), and Biscoe islands (-0.23). **PC3** showcases significant positive correlations with Chinstrap species (0.36), female penguins (0.27), and Dream islands (0.49), and significant negative correlations with culmen depth (-0.25), Adelie species (-0.34), Torgersen islands (-0.29), and male penguins (-0.49) are also observed.

K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that groups data points with similar characteristics into clusters by minimizing the mean distance between geometric points. This technique operates by identifying a specific number of centroids (arithmetic mean of every data point in a particular cluster) in the dataset, and assigning each data point to the nearest cluster while keeping the clusters as small and differentiated as possible (Nvidia, 2025).

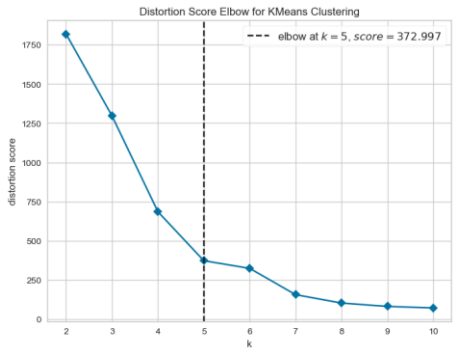


Figure 8

The K-Means Elbow method (Figure 8) is utilized to obtain a value for the optimal number of clusters to consider by assessing the distortion score, which is the sum of squared differences between each observation and centroid divided by the number of observations in the cluster. The silhouette score (Figure 9) explains how well-defined each cluster is, by considering cohesion (how close the data points are within the same cluster) and separation (distinctness of each cluster).

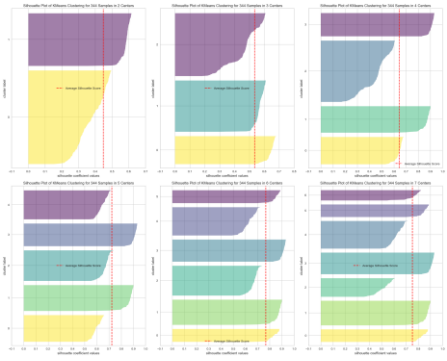
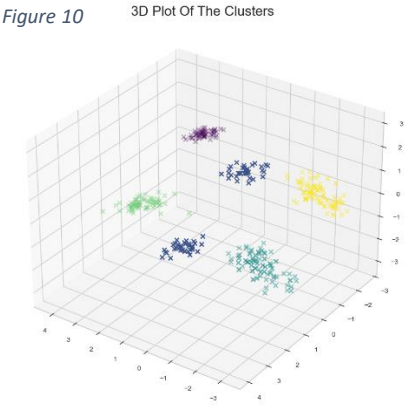


Figure 9

Figure 10



K-Means Clustering was carried out with **5 clusters** chosen as the optimal value based on a trade-off point from both graphs since it corresponded to the elbow point and a **silhouette score of 0.72**. This indicates a good clustering of the data with significant cohesion and separation of the clusters. The distribution of clusters across the 3 dimensions of the principal components (Figure 10) and the cluster counts (Figure 11) are shown. It is evident that cluster 1 has the highest count while cluster 5 has the least.

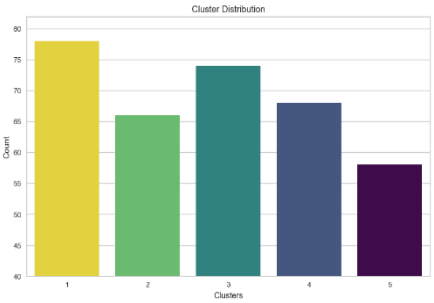


Figure 11

Cluster Analysis and Profiling

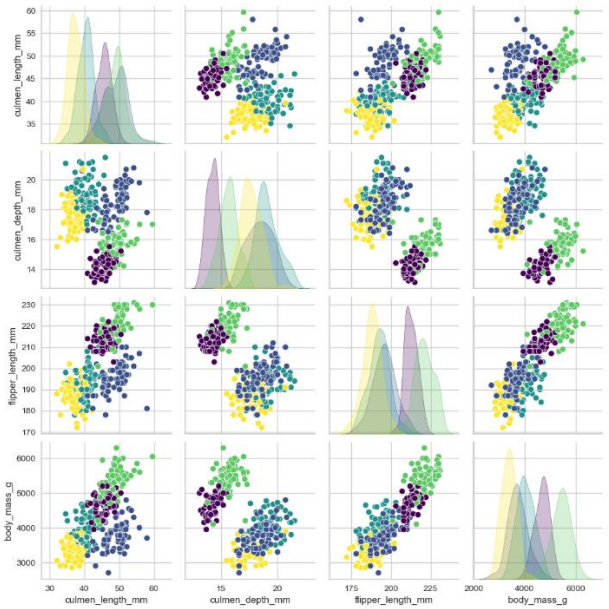


Figure 12

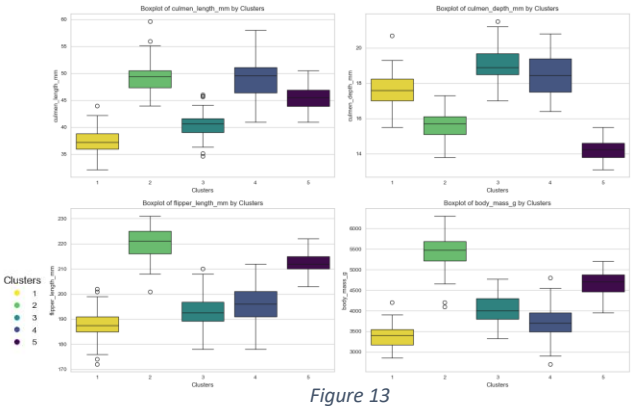


Figure 13

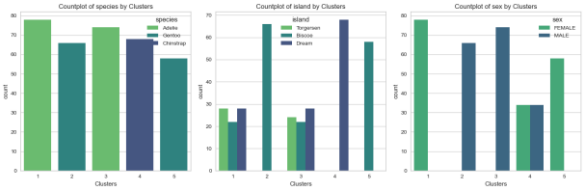


Figure 14

The scatterplot matrix (Figure 12) visualizes feature intercorrelations within the clusters. The boxplots (Figure 13) reveal the numerical feature variation among clusters, while the bar plots (Figure 14) show categorical feature distributions. These visualizations collectively reveal detailed insights into the individual characteristics that differentiate each cluster, while uncovering the hidden patterns and natural groupings in the data.

The table below provides a summary of these findings, highlighting the defining attributes of each cluster.

Cluster	Species	Sex	Island	Culmen Length	Culmen Depth	Body Mass	Flipper Length
1	*Adelie*	*Female*	*All islands*	*Shortest*	*Moderate*	*Lowest*	*Shortest*
2	*Gentoo*	*Male*	*Biscoe*	*Longest*	*Low*	*Highest*	*Longest*
3	*Adelie*	*Male*	*All islands*	*Moderate*	*Largest*	*Moderate*	*Moderate*
4	*Chinstrap*	*Both*	*Dream*	*Moderate*	*Moderate*	*Low*	*Moderate*
5	*Gentoo*	*Female*	*Biscoe*	*Moderate*	*Lowest*	*Moderate*	*Moderate*

Supervised Learning

Supervised learning is a subset of machine learning that uses labeled data to train algorithms in order to predict outcomes and identify patterns (Google Cloud, 2025). The use of labeled data allows the model to determine relationships between input and output data, identifying patterns that eventually enables these predictive models to forecast the outputs. The variable we predict is known as the dependent/target variable, while the independent input variables used to predict the outcome are known as features.

Regression

Regression is a supervised machine learning technique that predicts a continuous numerical outcome (y) based on one or more independent variables (**x**) (BuiltIn, 2024).The dataset chosen for this task will be the 'Medical Cost Personal' dataset used as a practical in the book 'Machine Learning with R by Brett Lanz'. Obtained from Kaggle, the inspiration behind this data will be accurately predicting the medical insurance costs of individuals based on features such as age, BMI, and more.

Existing Literature

A study investigating healthcare costs using data analytics and machine learning found that smoking status had the most significant impact on medical expenses. It was also observed that demographic factors such as age and BMI also played role medical costs incurred. Furthermore, several machine learning models were used for medical cost prediction, where Gradient Boosting achieved the highest accuracy at 92%, outperforming other models such as Random Forest and Linear Regression, which achieved accuracy rates of 83.44% and 76.59%, respectively (Rana, 2023).

Research Questions

1. What are the key features that determine medical insurance charges and how do they interact with other features?
2. Are medical insurance charges affected by geographical and demographical factors?
3. What is the best regression model to predict medical insurance charges?

Exploratory Data Analysis

The most influential feature determining medical insurance charges was the smoking status of individuals, as observed in the scatterplot matrix (Figure 15) and the boxplots (Figure 17). Additionally, the age of the individual showed a positive linear relationship with medical insurance charges. Furthermore, it was observed that the BMI was significantly higher in smokers. The higher medical insurance charges observed among individuals who smoked regardless of their age was another distinguishable characteristic identified during EDA. The southeast region corresponded towards higher medical insurance charges. There was no relationship observed between age and BMI. A sample discrepancy was observed in the 'smoker'

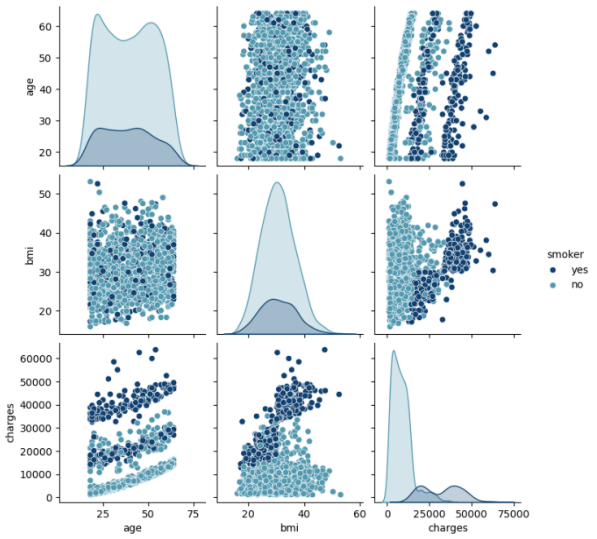


Figure 15

column, where the number of smokers in the dataset was significantly lower than the number of non-smokers in the dataset. Majority of the individuals in the sample had no children and a downward trend was observed as there were fewer individuals with more children.

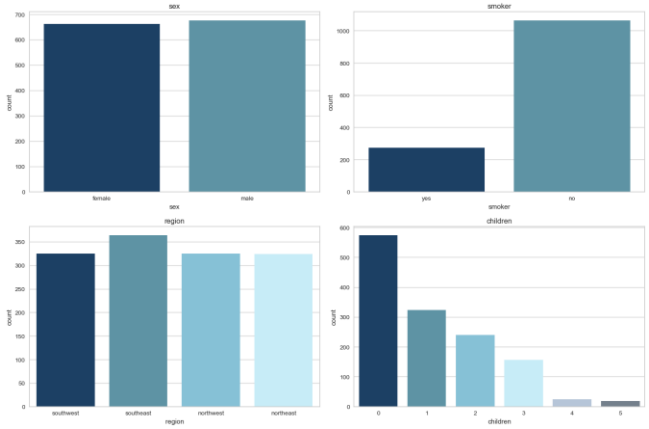


Figure 16

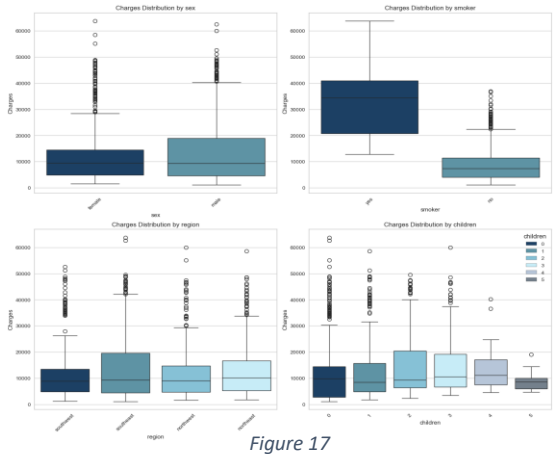


Figure 17

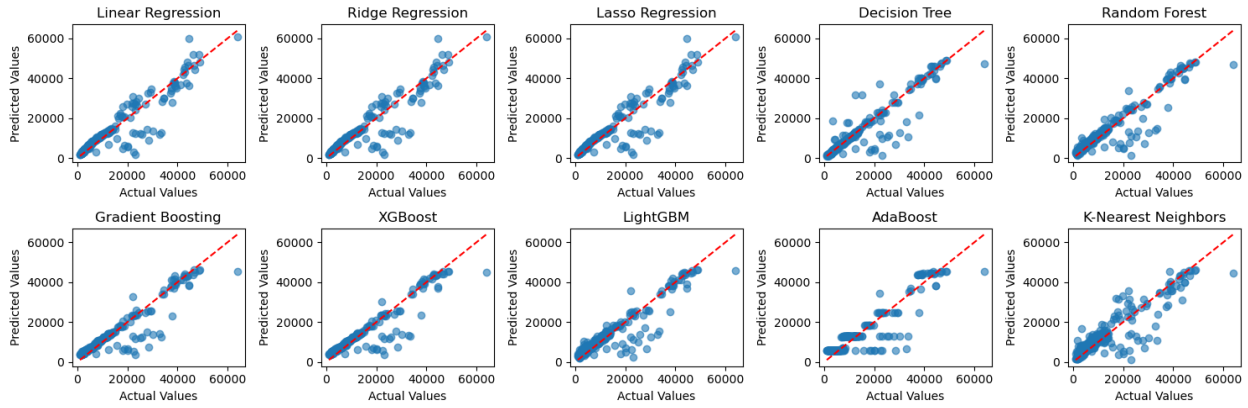
Machine Learning - Regression

The data was pre-processed initially by eliminating any null values and duplicate entries. The outliers were assessed and ultimately retained since they were all realistic values. Subsequently, the categorical features were appropriately encoded. No multicollinearity was observed among the independent variables, and hence no features were dropped from the dataset. The data was then split into train and test sets, and the features were standardized. The train set was used to train the machine learning model while the test data was used make predictions and evaluate the model.

Several regression models consisting of linear models (Linear Regression, Ridge Regression, Lasso Regression), tree-based models (Decision Tree Regressor), ensemble models (Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, and AdaBoost Regressor), and instance-based models (K-Nearest Neighbors Regressor) were trained on the data. Subsequently, hyperparameter tuning was conducted using GridSearchCV with 5-fold cross validation for each and every model (except for linear regression). After hyperparameter tuning, an overall increase in accuracy was observed for every model except for the K-Nearest Neighbors Regressor. It was also observed that the boosting ensembles performed the best while the linear and instance-based models performed the worst, indicating the complexity and the non-linear nature of the data. The results obtained are depicted below:

Regression Model	R^2 Score	MSE	RMSE	MAE
Linear Regression	0.81	35478020.68	5956.34	4177.05
Ridge Regression	0.81	35829011.84	5985.73	4202.53
Lasso Regression	0.81	35552990.01	5962.63	4178.73
Decision Tree Regressor	0.88	21217632.33	4606.26	1920.20
Random Forest Regressor	0.90	18883089.84	4345.47	2470.06
Gradient Boosting Regressor	0.90	18867550.47	4343.68	2621.70
XGBoost Regressor	0.90	19151862.16	4376.28	2664.02
LightGBM Regressor	0.89	20017335.23	4474.07	2642.99
AdaBoost Regressor	0.90	19104732.63	4370.90	2775.10
K-Nearest Neighbors Regressor	0.82	32171868.70	5672.03	3456.21

Figure 18



The scatterplots (Figure 18) depict the performance of each model by comparing the predicted values and actual values. Each blue dot represents a data point while the red dashed line shows the perfect prediction line (y=x). Points lying further away from the line indicates larger prediction errors (the further the point, the larger the error).

The **R^2** metric, is a statistical measure that shows how well the data fits a regression model by calculating the proportion of variance in the target variable explained by the features. The **Mean-Squared Error (MSE)** evaluates the average squared difference between the predicted values and the actual values, while the **Root Mean Squared-Error (RMSE)** is the square root of the MSE, which provides the metric in the same units as the original data. The **Mean-Absolute Error (MAE)** quantifies the average size of the errors in a prediction set, disregarding the direction. The **Random Forest Regressor** emerged as the best model among all models, showcasing the highest R^2 value, the lowest Mean-Squared Error, the second lowest Root Mean-Squared Error, and the lowest Mean Absolute Error.

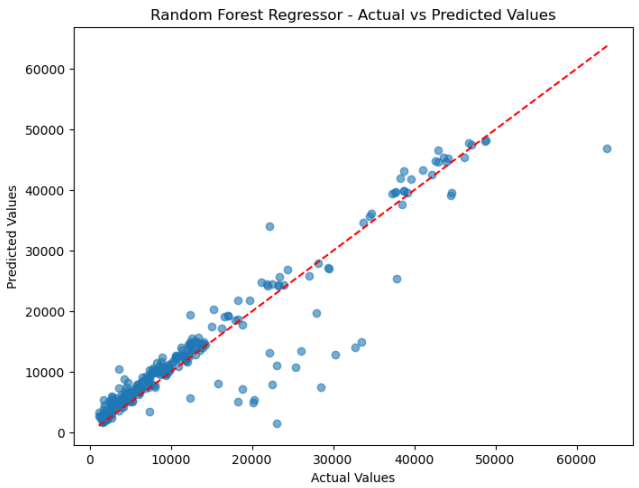


Figure 19

Conclusion

It was observed that smoking status had the greatest impact on the medical insurance charges of an individual, which corresponds to the insights obtained from the assessed existing literature. The age of the individual was also found to be significantly impactful towards medical insurance costs, while the BMI was identified to be higher in people who smoked. The south-east region portrayed higher medical insurance costs in comparison to other regions.

A variety of machine learning models were trained and tested for this case. The best model to predict medical insurance charges was identified to be the Random Forest Regressor model, reinforced by the R^2 value of 0.90, with MSE, RMSE, and MAE scores of 18883089.84, 4345.47, and 2470.06. These results solidified the Random Forest Regressor's position as the optimal model and the best overall performer for this particular task.

Classification

Classification is a supervised machine learning technique that assigns or predicts a class for the target variable based on the independent variables (features). These algorithms use labeled data and the output is a discrete variable (BuiltIn, 2023). The dataset used for this task is the 'Heart Failure Prediction Dataset' from Kaggle, which is a combination of 5 heart datasets from the UCI Machine Learning Repository. The dataset will have a total of 11 features and the aim of this task will be to accurate classify individuals with heart disease and without heart disease based on these features.

Existing Literature

A research published on the National Library of Medicine, 'Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel' used the same dataset, and created a machine learning metamodel (created based on Random Forest Classifier, Gaussian Naive Bayes, Decision Tree models, and k-Nearest Neighbor) showcasing an accuracy of 87% for predicting heart failure (Mahmud, 2023). No prior Exploratory Data Analysis was published in this research.

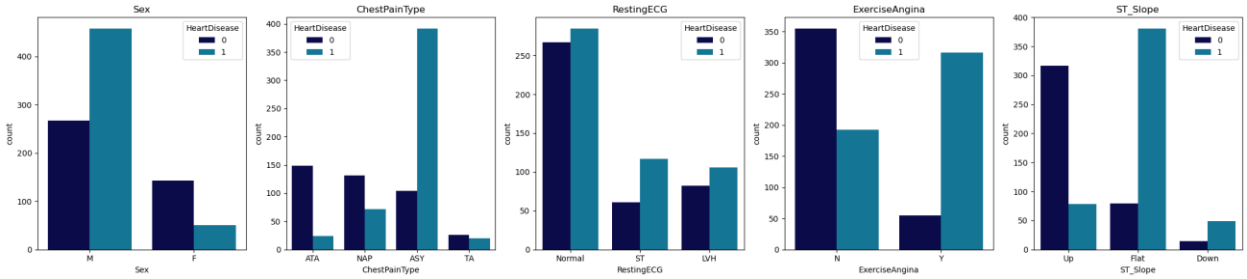
Research Questions

1. What are the most significant health indicators that correlate with heart failure?
2. What are the most crucial demographic factors that affect heart failure?
3. What is the best classification model to predict heart failure in individuals?

Exploratory Data Analysis

Initially, Exploratory Data Analysis was conducted to find existing relationships between the features in the dataset and the target variable (heart disease). Bar plots were constructed in a subplot grid with individual counts to identify heart disease presence. For numerical features, a subplot grid of boxplots was created.

Figure 20



The analysis conducted for categorical features (Figure 20) indicated that males exhibited a higher prevalence of heart disease in comparison to females, with the proportion of men with heart disease being significantly greater than those without the condition. Additionally, Asymptomatic (ASY) chest pain shows a significantly higher association with heart disease presence in comparison to other types of chest pain. The resting ECG results indicated that individuals showing ST wave abnormality having a significant heart disease presence, as opposed to the approximately even distributions observed in the other types of Resting ECG. Exercise induced angina (Y) was identified to be a potentially strong predictor of heart disease, with a vast majority of individuals with exercise induced angina being diagnosed with heart disease. A 'Flat' slope of the peak exercise ST segment (ST_slope) showed a high correlation with heart disease presence, while an 'Up' slope was more common among unaffected individuals.

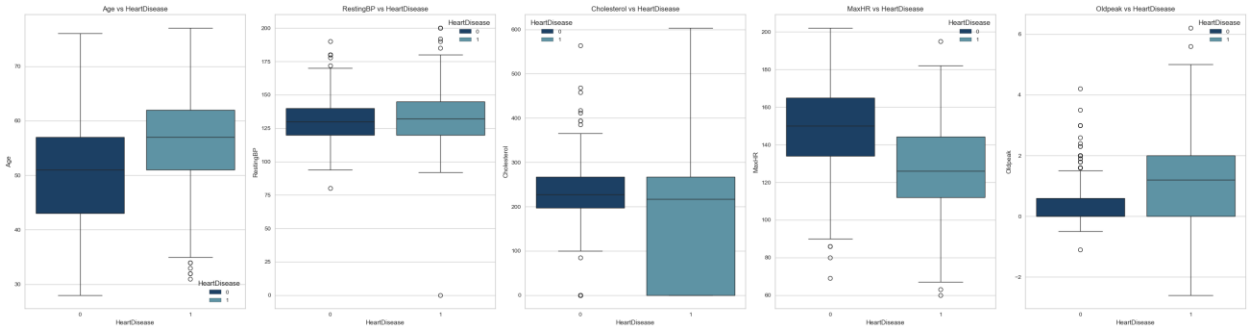


Figure 21

The analysis of numerical features (Figure 21) reveal that the distribution of age is marginally greater for individuals with heart disease. The Resting Blood Pressure levels did not reveal any significant disparity between the groups. The distribution of cholesterol levels for individuals with heart disease was lower, but this may have been largely influenced by the presence of 0 values for cholesterol levels in the dataset. Additionally, the distribution maximum heart rate was also identified to be marginally lower for individuals with heart disease. Furthermore, the distribution of the 'Oldpeak' metric (numeric value measured in depression), was identified to be significantly higher in individuals diagnosed for heart disease.

Machine Learning – Classification

Data pre-processing was conducted by examining null/missing values (no missing values were found), dealing with outliers and unrealistic values (unrealistic 0 values were removed and some were replaced with the mean), and encoding the categorical variables. The 'Cholesterol' column was dropped from the dataset since it contained a high proportion of 0 values (cannot be imputed with mean/median), which was an unrealistic observation in human beings. No multi-collinearity was observed among independent variables. The data was split into train and test sets and the features were standardized.

Various classification models consisting of Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, LightGBM Classifier, CatBoost Classifier, AdaBoost Classifier, and K-Nearest Neighbours Classifier were trained on the data, followed by hyperparameter tuning using GridSearchCV for each and every model. The results obtained after hyperparameter tuning are depicted below:

Classification Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.83	0.84	0.83	0.83	0.90
Decision Tree Classifier	0.77	0.79	0.77	0.77	0.89
Random Forest Classifier	0.86	0.86	0.86	0.86	0.93
XGBoost Classifier	0.87	0.87	0.87	0.87	0.93
LightGBM Classifier	0.85	0.86	0.85	0.85	0.94
CatBoost Classifier	0.88	0.88	0.88	0.88	0.94
AdaBoost Classifier	0.87	0.87	0.87	0.87	0.91
K-Nearest Neighbors Classifier	0.86	0.86	0.86	0.86	0.92

Key evaluation metrics that are specific for classification tasks include accuracy, precision, recall, f1-score, and ROC-AUC metrics. The confusion matrix provides a detailed breakdown of actual and predicted values, enabling the computation of these evaluation metrics (Labelf, 2022).

- **Accuracy:** The proportion of correct predictions across all predictions made by the model. *("How many true predictions out of all the predictions?")*
- **Precision:** The proportion of true positives across all positive predictions made by the model. *("How many true positive predictions out of all the positive predictions?")*
- **Recall:** The proportion of true positive among all actual positives. *("How many true positives out of all the actual positive values?")*
- **F1-score:** This provides the harmonic mean of precision and recall in a balanced single metric, particularly useful when dealing with imbalanced datasets.

The Receiver Operating Characteristic (ROC) curve is a graphical representation that visualizes the True Positive Rate (recall) against the False Positive Rate across different thresholds. The **AUC** (Area Under the Curve) value showcases the model's ability to distinguish between the classes.

Based on these evaluation metrics, the **CatBoost Classifier** was identified to the best classification model for heart disease prediction. The CatBoost classifier is a gradient boosting technique that is known to handle categorical and numerical features seamlessly with minimal errors and better predictions. The confusion matrix (Figure 22), ROC curve (Figure 23), and the classification report (Figure 24) for the CatBoost Classifier is presented below:

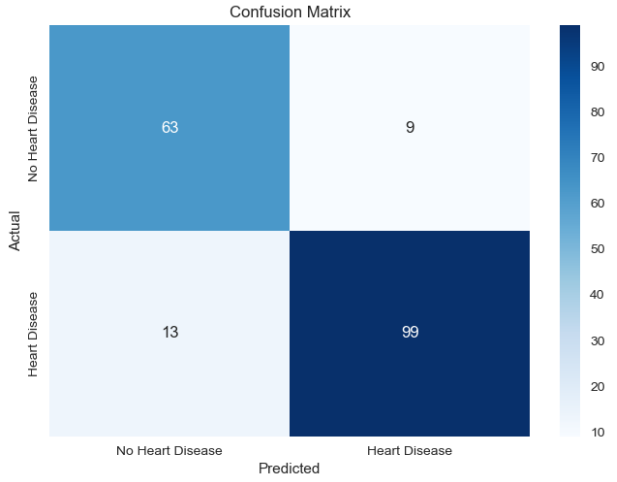


Figure 22

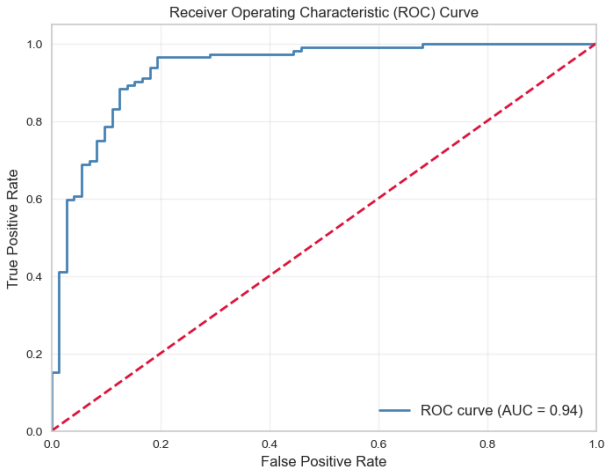


Figure 23

Conclusion

Exploratory Data Analysis revealed that male sex, ASY chest pain, flat ST slope, and exercise-induced angina were significantly present in individuals diagnosed with heart disease. In medical diagnosis, it is essential to ensure that we choose the model that has a better true positive prediction accuracy (predicts heart disease better), in order to mitigate any risks and consequences associated with late detection of the illness. The AUC value of 0.94 and F1-score of 0.90 for predicting heart disease, backed by the best overall accuracy, precision, recall, and true positive rate qualified the CatBoost Classifier as the best model for predicting the presence of heart disease.

Classification Report: CatBoost Classifier

	precision	recall	f1-score	support
No Heart Disease	0.83	0.88	0.85	72
Heart Disease	0.92	0.88	0.90	112
accuracy			0.88	184
macro avg	0.87	0.88	0.88	184
weighted avg	0.88	0.88	0.88	184

Figure 24

Bibliography

Built In. (2024, 2 23). *Principal Component Analysis (PCA): A Step-by-Step Explanation*. Retrieved 1 28, 2025, from BuiltIn.com: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

BuiltIn. (2023, 4 12). *5 Classification Algorithms for Machine Learning*. Retrieved 2 6, 2025, from builtin.com: https://builtin.com/data-science/supervised-machine-learning-classification

BuiltIn. (2024, 8 1). *Regression in Machine Learning: Definition and Examples of Different Models*. Retrieved 2 3, 2025, from builtin.com: https://builtin.com/data-science/regression-machine-learning

Google Cloud. (2025). *What is Supervised Learning?* Retrieved 2 3, 2025, from cloud.google.com: https://cloud.google.com/discover/what-is-supervised-learning

Google Cloud. (2025). *What is unsupervised learning?* Retrieved 1 25, 2025, from Google Cloud: https://cloud.google.com/discover/what-is-unsupervised-learning

Jadhav, D. R. (2022, 9 9). *Predictive analysis the study of different characteristics of Palmer penguins using R-programming*. Retrieved 9 9, 2025, from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4200205&download=yes

Labelf. (2022, 11 7). *What is Accuracy, Precision, Recall and F1 Score?* Retrieved 2 16, 2025, from www.labelf.ai: https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score

Mahmud, I. (2023, 7 31). *Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel*. Retrieved 2 6, 2025, from pmc.ncbi.nlm.nih.gov: https://pmc.ncbi.nlm.nih.gov/articles/PMC10417090/#sec2-diagnostics-13-02540

Nvidia. (2025). *K-Means Clustering Algorithm*. Retrieved 1 29, 2025, from Nvidia website: https://www.nvidia.com/en-us/glossary/k-means/

Rana, A. I. (2023, 11). *Healthcare Cost Patterns and Prediction: Investigating Personal Datasets using Data Analytics*. Retrieved 2 3, 2025, from Researchgate.net: https://www.researchgate.net/publication/375484166_Healthcare_Cost_Patterns_and_Prediction_Investigating_Personal_Datasets_using_Data_Analytics