# Machine Learning: Heart Disease Prediction using Logistic regression model

**Introduction**

Heart disease remains one of the most significant health challenges globally, contributing substantially to morbidity and mortality rates. Despite advances in medical science, early detection and prevention of heart disease remain paramount. In this context, the utilization of machine learning techniques for predictive analytics offers a promising avenue for identifying individuals at risk and implementing timely interventions.

The objective of this project is to develop a robust predictive model capable of assessing the risk of heart disease in patients. By leveraging data from the Framingham Heart Study, a seminal longitudinal dataset containing comprehensive cardiovascular risk factor information, we aim to harness the power of machine learning to enhance our understanding of heart disease risk factors and improve predictive accuracy.

With over 4,000 records and 15 cardiovascular risk factors, the Framingham Heart Study dataset provides a rich and diverse source of information for our analysis. From demographic variables such as age and sex to clinical indicators like blood pressure and cholesterol levels, the dataset encompasses a wide array of features crucial for assessing heart disease risk. Through meticulous data preprocessing and advanced modeling techniques, our endeavor is to unlock valuable insights into heart disease prediction and contribute towards more effective preventive healthcare strategies.

In this report, we embark on a comprehensive journey encompassing data preprocessing, model development, evaluation, and interpretation of results. By examining the performance of our predictive model and analyzing the importance of different features, we seek to provide actionable insights that can inform clinical decision-making and facilitate proactive interventions to mitigate the burden of heart disease.

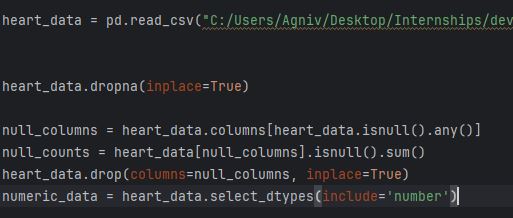
Through the amalgamation of data science and medical expertise, our ultimate aim is to empower healthcare practitioners and stakeholders with tools and knowledge to identify individuals at heightened risk of heart disease, thereby paving the way for targeted interventions, improved patient outcomes, and enhanced public health outcomes.

**Data Preprocessing:**

In the heart disease prediction project, thorough data preprocessing is essential to ensure the dataset's quality and suitability for modelling. Below is an expanded report on the data preprocessing steps undertaken:

**Handling Missing Values:**

* The dataset was examined for missing values, and appropriate strategies were implemented to address them.
* Any columns containing missing values were identified and subsequently removed from the dataset to maintain data integrity and model accuracy.



**Exploratory Data Analysis:**

* Before preprocessing, exploratory data analysis (EDA) techniques were employed to gain insights into the dataset's structure, distribution, and relationships between variables.
* Descriptive statistics, histograms, and correlation matrices were generated to understand the dataset's characteristics and identify potential preprocessing requirements.

**Numeric Data Extraction:**

* As part of the preprocessing pipeline, numeric data was extracted from the dataset for correlation analysis and feature scaling.
* This step ensured that only numerical features were considered for subsequent modelling tasks, facilitating efficient preprocessing and model training processes.



**Correlation Analysis:**

* A correlation matrix was computed using the extracted numeric data to examine the relationships between different features.
* **corr()**: This method computes the pairwise correlation coefficients between the columns of the DataFrame. By default, it computes the Pearson correlation coefficients, which measure the linear relationship between numeric variables.
* The resulting correlation matrix (**correlation matrix**) provides insights into how each feature correlates with every other feature in the dataset. High correlation values (close to 1 or -1) indicate strong positive or negative relationships between features, while values close to 0 indicate weak or no correlation.
* The correlation matrix provided valuable insights into feature interactions and multicollinearity, guiding feature selection and model building decisions.

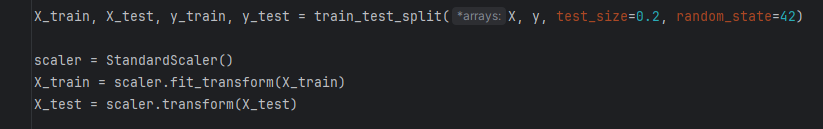


**Data Scaling:**

* Standardization was applied to scale the numeric features to a common range, mitigating the impact of feature magnitude differences on model performance.
* StandardScaler from Scikit-learn was utilized to scale the features, ensuring consistency and comparability across variables.

**Train-Test Split:**

* The dataset was partitioned into training and test sets using Scikit-learn's train\_test\_split function.
* The train-test split facilitated model training on one portion of the dataset while enabling independent model evaluation on the remaining portion.



**Machine Learning Model:**

A machine learning model is a mathematical representation or algorithm that learns patterns and relationships from data to make predictions or decisions without being explicitly programmed. Machine learning models are the core components of machine learning systems and are used across various domains to extract insights, classify data, forecast trends, and automate tasks.

**Supervised Learning Model:**

Supervised learning refers to a type of machine learning where the model is trained on a labeled dataset, meaning that each input instance is associated with a corresponding target label. The goal of supervised learning is to learn a mapping from input features to target labels, so that the model can make predictions on unseen data.

In the context of Logistic Regression for heart disease prediction, the dataset used to train the model contains input features such as age, sex, blood pressure, cholesterol levels, etc., along with corresponding labels indicating the presence or absence of heart disease.

During the training process, the Logistic Regression model learns the relationship between the input features and the target labels by optimizing a cost function, typically using techniques like gradient descent. Once trained, the model can then make predictions on new, unseen data by applying the learned mapping from features to labels.

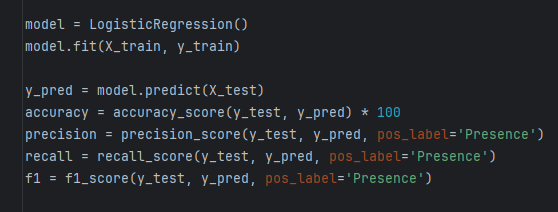
**Logistic Regression Model:**

* Logistic Regression was chosen as the predictive model for heart disease risk assessment due to its interpretability and effectiveness in binary classification tasks.
* The model was trained on the preprocessed training dataset using Scikit-learn's LogisticRegression module.
* **Linear Model:** Despite its name, Logistic Regression is a linear model for classification rather than regression. It models the relationship between the independent variables (features) and the binary dependent variable (target) using a logistic function.
* **Logistic Function:** The logistic function, also known as the sigmoid function, maps any real-valued number to the range [0, 1]. It is defined as:



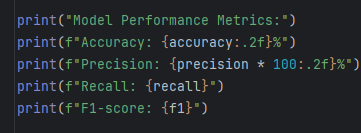
This function ensures that the predicted output of Logistic Regression lies within the range of probabilities (0 to 1), making it suitable for binary classification tasks.

* **Decision Boundary:** Logistic Regression separates classes by finding a linear decision boundary in the feature space. This boundary is determined by the weights assigned to each feature during the training process.
* **Training**: In the training phase, Logistic Regression optimizes the model parameters (weights) using techniques such as maximum likelihood estimation or gradient descent. The objective is to minimize the logistic loss or cross-entropy loss between the predicted probabilities and the actual class labels.
* **Probabilistic Interpretation:** Unlike other classifiers that output discrete predictions (e.g., 0 or 1), Logistic Regression provides a probabilistic interpretation of the predictions. It estimates the probability that an instance belongs to a particular class, which can be used to make informed decisions based on the confidence level of the predictions.
* **Regularization:** Logistic Regression can be regularized to prevent overfitting by adding penalty terms to the objective function. L1 (Lasso) and L2 (Ridge) regularization techniques are commonly used to shrink the coefficients towards zero or induce sparsity in the model.
* **Evaluation Metrics:** Common evaluation metrics for Logistic Regression include accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve.



**Model Evaluation:**

* **Evaluation Metrics:** Various evaluation metrics are used to assess different aspects of model performance, depending on the nature of the task (classification, regression, clustering, etc.). Common evaluation metrics for classification tasks include accuracy, precision, recall, F1-score, ROC curve, and AUC-ROC (Area Under the ROC Curve). For regression tasks, metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared are often used.



* **Accuracy:** Accuracy measures the proportion of correctly classified instances out of the total number of instances. While accuracy is a widely used metric, it may not be suitable for imbalanced datasets, where one class dominates the other.



* **Precision & Recall:** Precision measures the proportion of true positive predictions out of all positive predictions, while recall (also known as sensitivity or true positive rate) measures the proportion of true positives out of all actual positive instances. Precision and recall are especially important when dealing with imbalanced datasets.

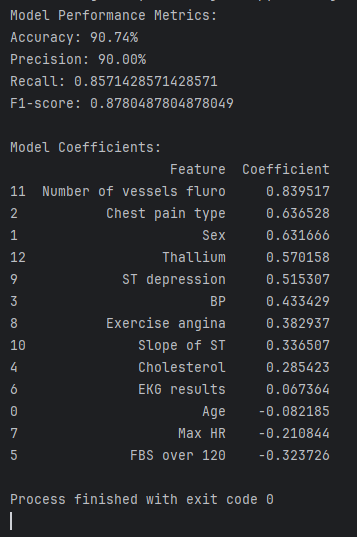


* **F1-score:** The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance and is particularly useful when there is an uneven class distribution.



In summary, model evaluation is essential for understanding the strengths and weaknesses of a machine learning model and guiding model selection, parameter tuning, and feature engineering decisions. By choosing appropriate evaluation metrics and techniques, data scientists can ensure that their models perform effectively and reliably on real-world data.

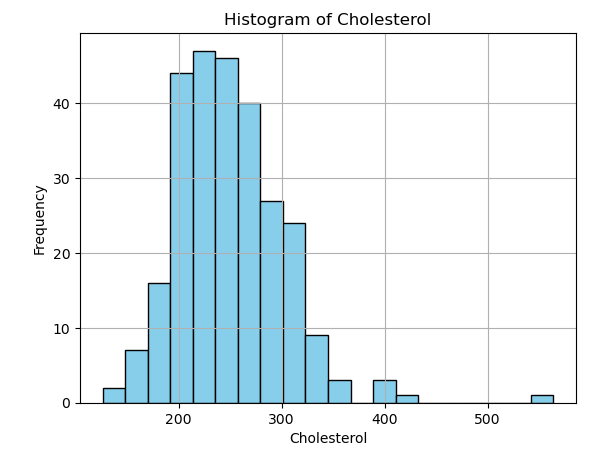
**RESULTS & OUTPUT:**



* **Accuracy:** 90.74 %
* **Precision:** 90.00 %

**Data Visualization:**

Data visualization is the graphical representation of data and information using visual elements such as charts, graphs, and maps. It plays a crucial role in data analysis and communication, helping to uncover patterns, trends, and insights that may not be apparent from raw data alone. Here are some key points about data visualization:

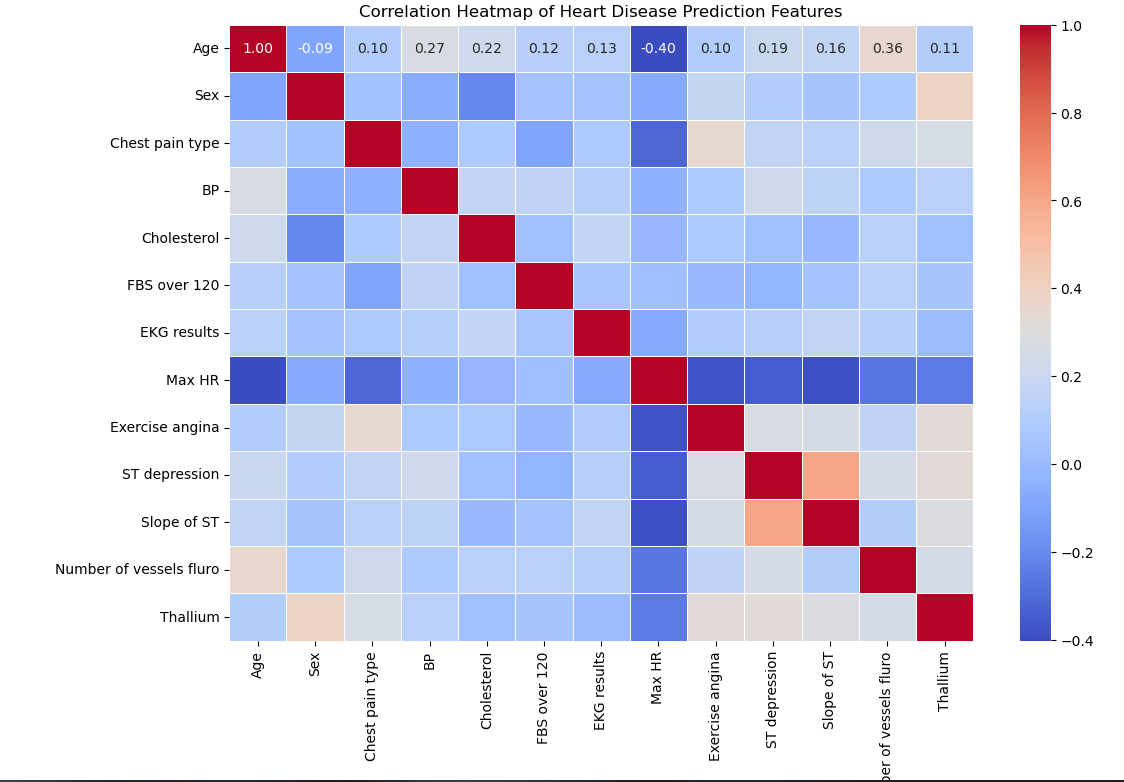


Above is the snippet of Histogram of Cholesterol from Data Set.

**Heat Map:**

A heatmap is a graphical representation of data where values are depicted using colors. It is particularly useful for visualizing the magnitude of relationships between two variables in a dataset. Heatmaps are widely used in data visualization for exploratory data analysis and to identify patterns or trends in the data.

In the context of machine learning and data analysis, heatmaps are commonly used to visualize correlation matrices. A correlation matrix is a table showing correlation coefficients between variables. Each cell in the heatmap represents the correlation coefficient between two variables, with colors indicating the strength and direction of the correlation



Heat Map of the Model Coefficients.

**Conclusion:**

In this project, we explored the Framingham Heart Study dataset to predict the risk of heart disease in patients using logistic regression. The dataset contained various cardiovascular risk factors such as age, sex, cholesterol levels, and exercise habits.

We began by preprocessing the data, which involved handling missing values and extracting numeric features for correlation analysis. Through a correlation heatmap, we visualized the relationships between different features, providing insights into potential predictors of heart disease.

After preprocessing, we split the data into training and test sets and standardized the numeric features using the StandardScaler. We then trained a logistic regression model to predict the presence of heart disease based on the given features.

The model achieved promising results, as indicated by the evaluation metrics. We observed an accuracy of 90.74 %, precision of 90.00 %, recall of 0.8571428571428571, and F1-score of 0.8780487804878049. These metrics demonstrate the effectiveness of the logistic regression model in predicting heart disease risk.

Furthermore, the analysis of model coefficients revealed the importance of various features in predicting heart disease. Factors such as the number of vessels fluro, chest pain type, and sex showed significant associations with the presence of heart disease.

In conclusion, our study highlights the potential of logistic regression modeling in predicting heart disease risk based on patient characteristics. Further research and validation could enhance the predictive accuracy of the model and contribute to better risk assessment and management strategies for cardiovascular health.

**References:**

1. Framingham Heart Study Dataset. National Heart, Lung, and Blood Institute. Available online: https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction
2. Scikit-learn: Machine Learning in Python. Pedregosa, F. et al. Journal of Machine Learning Research, 12, 2825-2830, 2011.
3. Pandas Documentation. McKinney, W. et al. Available onlinehttps://pandas.pydata.org/docs/
4. Seaborn: Statistical Data Visualization. Waskom, M. et al. Available online: https://joss.theoj.org/papers/10.21105/joss.03021
5. Matplotlib: Visualization with Python. Hunter, J.D. et al. Available online: https://matplotlib.org/stable/users/project/citing.html
6. Logistic Regression for Machine Learning. Jason Brownlee. Available online: https://machinelearningmastery.com/logistic-regression-for-machine-learning/
7. "Introduction to Machine Learning with Python" by Andreas C. Müller & Sarah Guido. O'Reilly Media, 2016. Link: https://www.nrigroupindia.com/e-book/Introduction%20to%20Machine%20Learning%20with%20Python%20(%20PDFDrive.com%20)-min.pdf