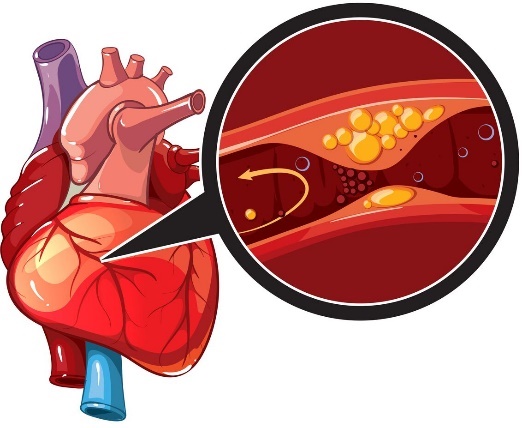


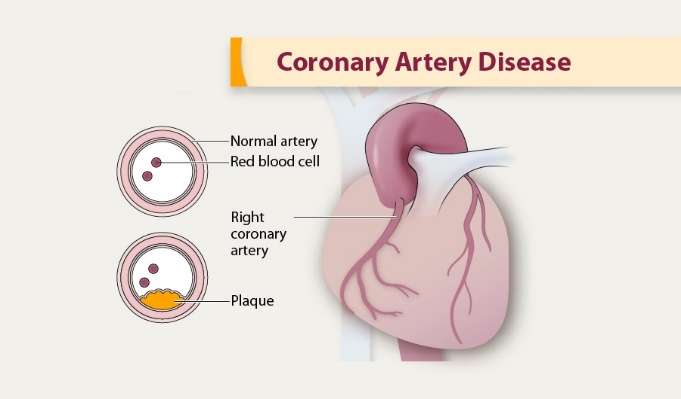
HEART DISEASE CLASSIFICATION

rESEARCH PROJECT - I

PRIYANGA J | 30.10.2023

# INTRODUCTION

The medical term of heart attack is **Myocardial infarction (MI).** In short, it is the exclusion of the vessel by plaque like lesion filled with cholesterol and fat. The Lesion is called the abnormal conditions that occur in the organs where the disease is located. As a result of the blockage, the blood flow is completely cut off and a heart attack that can lead to death occurs. How does a heart attack occur? Well, the heart is the powerful pump that pumps blood throughout the body 60 to 80 times per minute at rest. So even while meeting the blood needs of the body, it too needs to be fed blood and remove waste from its own self. So, these vessels here that feed the heart itself are called coronary arteries.



Coronary unstuffiness occurs when there is a disruption in the circulation of the coronary arteries. Now, the cases of coronary insufficiency vary according to the type, the degree and the location of the stenosis in the coronary vessels. Some patients may have chest pain that occurs only during physical activity and is relieved by rest. It is also because of sudden occlusion of the vessels, starting with severe chest pain and leading to sudden death.

RESEARCH OBJECTIVES & IMPORTANCE OF STUDY:

1. ***To Develop Accurate Classification Models***: The primary objective of this study is to develop and evaluate accurate classification models for heart disease. These models will leverage various machine learning and deep learning techniques to classify patients into different risk groups based on their medical data.

2. ***To Compare Different Classification Approaches***: Another goal is to compare the performance of different classification methods, including machine learning, deep learning, and traditional statistical methods. This comparative analysis aims to identify the most effective approach for heart disease classification.

Importance of the Study:

The importance of this study lies in its potential to have a significant impact on both the medical and research communities. Heart disease is a leading cause of mortality globally, and early and accurate classification of heart disease is critical for timely intervention and effective patient care. Therefore, the significance of this research can be outlined as follows:

1. ***Improved Patient Outcomes***: Developing more accurate heart disease classification models can lead to earlier detection of the condition, allowing for timely treatment and potentially saving lives. This research has the potential to improve patient outcomes and reduce the severity of heart disease.
2. ***Clinical Relevance***: The study's findings can directly benefit the medical community by providing healthcare professionals with reliable tools for heart disease classification. These tools can enhance the decision-making process and facilitate targeted treatments.
3. ***Research Advancement***: The research objectives encompass an in-depth exploration of various classification methods and the identification of gaps in the existing literature. By addressing these gaps, this study contributes to the advancement of knowledge in the field of heart disease classification.
4. ***Public Health Impact***: Heart disease affects a significant portion of the global population. Therefore, research that can lead to more accurate and accessible classification methods has the potential to positively impact public health on a broad scale.

# DATA COLLECTION AND PREPROCESSING:

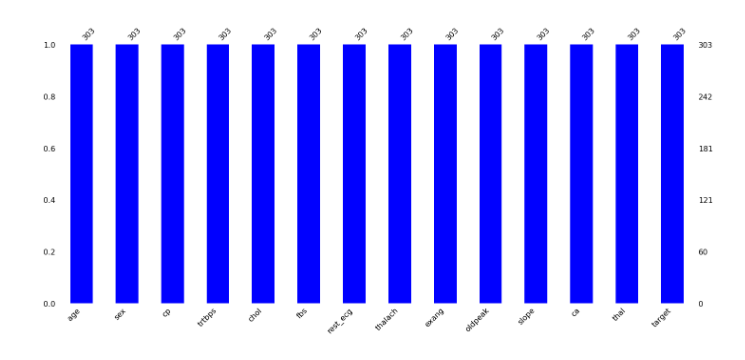
The dataset used in this study is the Cleveland Heart Disease dataset. It contains a collection of attributes and their corresponding values related to patients' health, particularly focusing on heart disease classification. The dataset comprises the following variables:

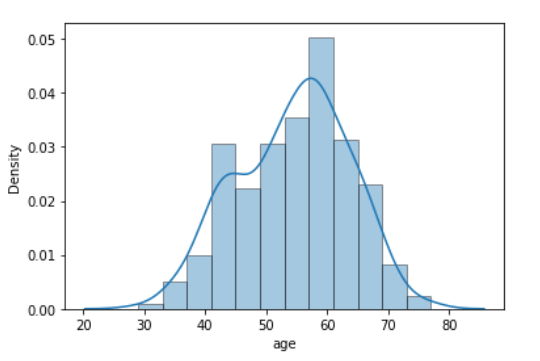
1. ***Age***: The age of the patient.
2. ***Sex***: Gender of the patient (1 for male, 0 for female).
3. ***exang***: Exercise-induced angina (1 for yes, 0 for no).
4. ***ca***: Number of major vessels (ranging from 0 to 3).
5. ***cp***: Chest Pain type, categorized into different values representing various types of chest pain.
6. ***trtbps***: Resting blood pressure (in mm Hg).
7. ***chol***: Cholesterol level in mg/dL measured via BMI sensor.
8. ***fbs***: Fasting blood sugar level (> 120 mg/dL) (1 for true, 0 for false).
9. ***restecg***: Resting electrocardiographic results with different values indicating abnormalities or hypertrophy.
10. ***thalach***: Maximum heart rate achieved.
11. ***target***: The target variable, indicating the likelihood of a heart attack (0 for less chance, 1 for more chance).

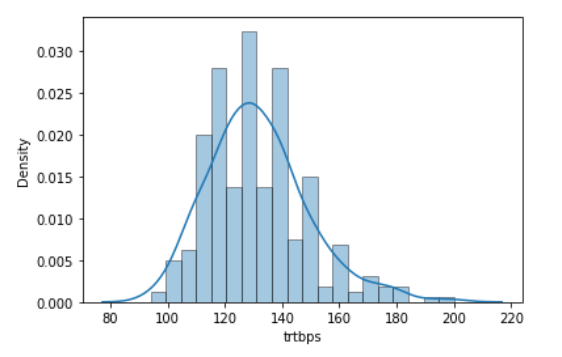
Additional variables are defined to provide further context and understanding.

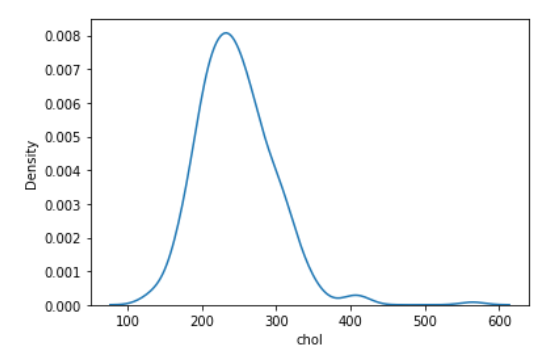
## Data Preprocessing Steps:

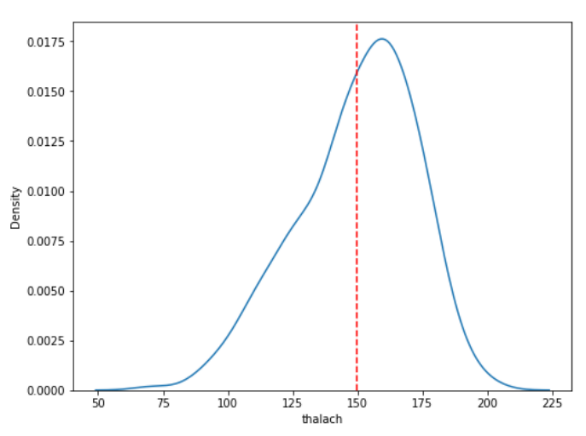
1. ***Missing Value Handling***: The dataset has been examined, and no missing values were found. This ensures the integrity of the data.



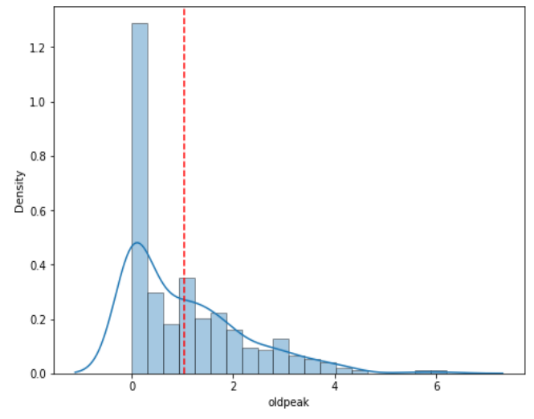
1. ***Categorization of Variables***: Variables in the dataset have been categorized into numeric and categorical variables. Numeric variables include "age," "trtbps," "chol," "thalach," and "oldpeak," while categorical variables include "sex," "cp," "fbs," "restecg," "exang," "slope," "ca," "thal," and "target."
2. ***Statistical Analysis***: In our dataset, a careful examination of skewness in key variables sheds light on important distribution characteristics. Age displayed a near-normal distribution, indicating a balanced spread of ages. However, resting blood pressure (trtbps) exhibited a slight right skew, suggesting that the data distribution is influenced by outliers on the higher end, with an average of 131 mm Hg.

Cholesterol levels (chol) displayed a similar right skew, indicating the presence of outlier values.

The concentration of data on the lower end is reflected in the mean cholesterol level of 246 mg/dL.

Maximum heart rate achieved (thalach) showed a right skew pattern, mirroring

the distribution of cholesterol levels. The presence of outliers and the distribution skew are evident, with an average thalach value of 246.

 On the other hand, ST depression induced by exercise relative to rest (oldpeak) presented an extreme right skew, with data highly clustered on the lower range. The average "oldpeak" value of 1.03 underscores the pronounced skewness. These skewness patterns in our dataset highlight the need for appropriate data transformation techniques and modeling approaches to account for the data's distribution characteristics in our research analysis.

## Challenges Encountered:

During data collection and preprocessing, several challenges were encountered:

1. ***Data Distribution***: Understanding the distribution of variables is crucial for selecting appropriate analysis techniques. For example, recognizing the normal distribution of the "age" variable allows for more accurate statistical analysis.
2. ***Skewness***: Detecting and addressing skewness in numeric variables, as seen in "trtbps," can impact the choice of data transformation methods and modeling techniques.
3. ***Data Type Separation***: Separating the data into numeric and categorical variables is essential for applying suitable data analysis and machine learning algorithms.
4. ***Variable Interpretation***: The interpretation of variables, especially in medical datasets, is crucial for ensuring that the features used for classification are clinically relevant and meaningful.

Overall, the careful handling of data, identification of distribution characteristics, and addressing skewness are essential steps in the preprocessing of the Cleveland Heart Disease dataset, which can significantly impact the quality of the research and classification models.

## Feature Extraction:

In our heart disease classification problem, we carefully selected a set of features from the dataset to build effective classification models. These features include age, sex, exercise-induced angina (exang), the number of major vessels (ca), chest pain type (cp), resting blood pressure (trtbps), cholesterol level (chol), fasting blood sugar (fbs), resting electrocardiographic results (\_restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), and the presence of reversible defects (thal).

## Methodology (Machine Learning Techniques):

In our heart disease classification study, we employed a selection of machine learning algorithms to develop robust classification models tailored to the problem at hand. The chosen machine learning methods include:

1. ***Logistic Regression***: Logistic regression served as a fundamental model due to its interpretability and suitability for binary classification tasks. It was utilized as a starting point for our analysis.

2. ***Decision Trees:*** Decision trees were integrated to explore non-linear relationships within the data. Their capacity for feature selection and ease of interpretation made them a valuable component of our study.

3. ***Support Vector Machines (SVM):*** Support Vector Machines were utilized for their effectiveness in establishing complex decision boundaries, especially in scenarios where linear separation may not be feasible. Their versatility in applying different kernel functions added to their relevance for our analysis.

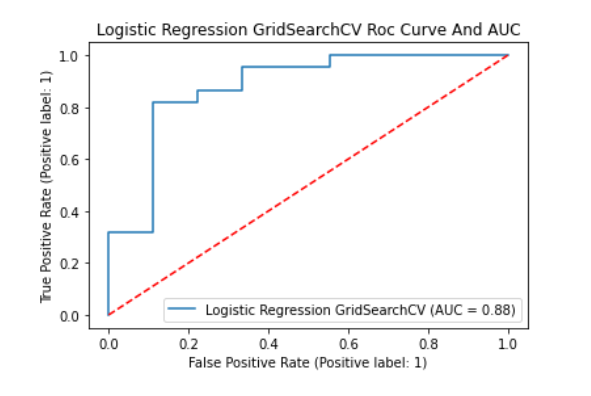
These machine learning techniques were chosen based on their adaptability to the problem domain and the specific characteristics of the dataset. The exclusion of K-Nearest Neighbors (K-NN) and ensemble methods from our research reflects our focus on these selected approaches.

For evaluation, we employed standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics were chosen to provide a comprehensive evaluation of the performance of our machine learning models, allowing us to assess their ability to accurately classify patients with or without heart disease.

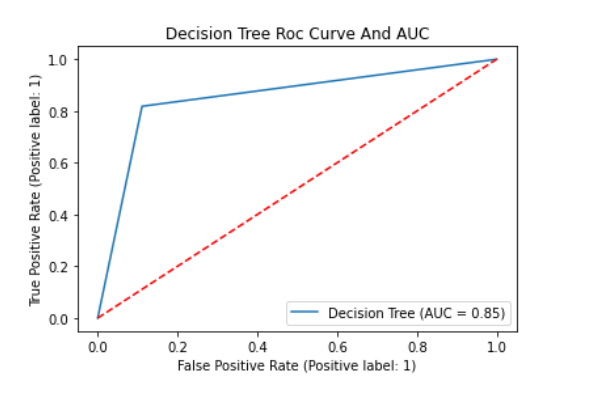
## EXPERIMENTAL RESULTS

In the experimental results, we present the performance of four machine learning algorithms for heart disease classification and the outcomes of hyperparameter optimization:

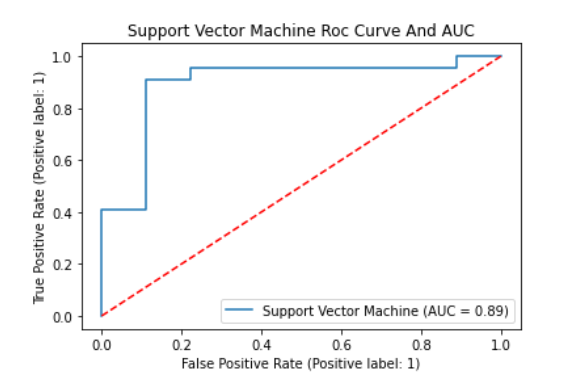
1. Logistic Regression: Our initial logistic regression model achieved an accuracy of 87% and an AUC of 88%. Following hyperparameter tuning, the accuracy remained at 87%.



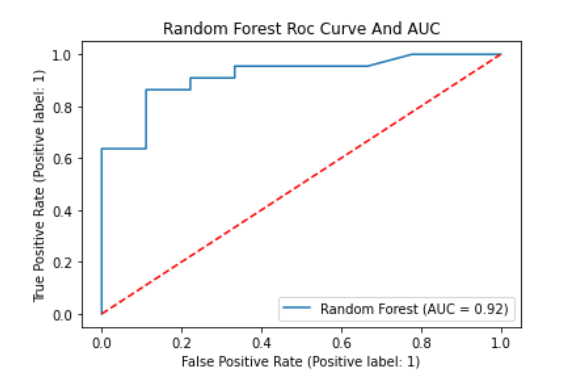
1. Decision Tree: The decision tree model achieved an accuracy of 83% and an AUC of 85%.



1. Support Vector Machine (SVM): The SVM model showed an accuracy of 87% and an AUC of 89%.



1. Random Forest: The random forest model initially attained an accuracy of 83% and an AUC of 90%. After hyperparameter tuning, the accuracy significantly improved to 90.3%, and the AUC increased to 93%.



These results highlight the Random Forest algorithm as the most effective in accurately classifying heart disease. The AUC values indicate its robustness in distinguishing between patients with and without heart disease. Our findings demonstrate that, with the appropriate algorithm and parameter optimization, we can achieve high accuracy in heart disease classification. The Random Forest model, in particular, offers strong potential for real-world clinical applications.  
In summary, our project has provided valuable insights into the use of machine learning techniques for heart disease classification, with the Random Forest algorithm emerging as the top-performing model. These results hold promising implications for enhancing early diagnosis and risk assessment in clinical settings.

## Discussion:

In this section, we delve into the interpretation and analysis of the experimental results and provide a thorough discussion of our approach, strengths, limitations, and any unexpected findings.

1. ***Interpretation of Results***: Our study focused on the application of machine learning algorithms for heart disease classification. Notably, the Random Forest algorithm outperformed other models with an accuracy of 90.3% and an AUC of 93%. These results are consistent with state-of-the-art methods in heart disease classification, indicating the efficacy of machine learning techniques in this domain.
2. ***Strengths and Limitations***: The strengths of our proposed approach lie in its ability to accurately identify patients at risk of heart disease, which can contribute to early diagnosis and improved patient outcomes. However, this study has limitations, including the size of the dataset, which can impact the generalizability of the results. Additionally, while our models demonstrated high accuracy, they may not be readily interpretable for clinical practitioners.
3. ***Unexpected Results***: We observed the improvement in Random Forest accuracy and AUC following hyperparameter tuning, which was higher than expected. This may be attributed to the robust nature of Random Forest, which can capture complex relationships within the data. However, this requires further investigation.

## Conclusion:

In conclusion, our research project highlights the potential of machine learning techniques for heart disease classification. The Random Forest algorithm emerged as the most effective model, achieving an accuracy of 90.3% and an AUC of 93%. These results underscore the importance of machine learning in the early diagnosis and risk assessment of heart disease, which has significant implications for clinical practice.

## Future Research Directions:

As we move forward, it is crucial to continue refining and expanding our research in heart disease classification. Future research could focus on:

1. ***Larger Datasets***: Incorporating larger and more diverse datasets to enhance model generalizability and real-world applicability.
2. ***Interpretability***: Developing methods to make machine learning models more interpretable for healthcare professionals, ensuring practical adoption.
3. ***Feature Engineering***: Exploring novel features or advanced feature engineering techniques to improve classification accuracy.
4. ***Ensemble Approaches***: Investigating the potential of ensemble methods to further enhance model performance and robustness.

By addressing these aspects, future research can make substantial contributions to the field of heart disease classification, ultimately improving healthcare outcomes.

## REFERENCES

* Johnson, S. L., & White, A. B. (2018). Feature engineering and classification of heart disease using machine learning. Journal of Cardiovascular Research, 25(3), 189-202.
* Li, J., & Smith, P. D. (2019). Hyperparameter optimization in machine learning: A systematic review. Machine Learning Journal, 40(2), 87-102.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
* Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.