## ****Title Page****

**Project Title:**  
Heart Disease Prediction Using Machine Learning

**Student Name:**

SAI SRI LAKSHMI CHIRIKI

**Roll Number:**  
2366250374

**Institution:**  
SRI ADITYA DEGREE COLLEGE BHIMAVARAM

**Course/Department Name:**  
BCA [COMPUTER APPLICATIONS]

**Submission Date:**  
04-FEB-2025

## ****Abstract****

This project focuses on predicting the presence of heart disease using machine learning techniques. The main objective was to create a predictive model that can accurately identify whether a person has heart disease or not, based on various health attributes. The dataset used in the project contains features such as age, sex, cholesterol levels, blood pressure, and exercise-induced angina, among others. Various machine learning algorithms were explored, with the Random Forest Classifier providing the best results in terms of accuracy and prediction. The model was evaluated using accuracy score, confusion matrix, and classification report, which provided insights into precision, recall, and F1-score. The findings show that machine learning can be effectively applied to predict heart disease, with potential real-world applications in healthcare for early diagnosis. The project highlights the importance of data preprocessing, model selection, and evaluation in building a robust predictive system.

**Table of Contents**

1. **Introduction** ........................................................... 1
2. **Frameworks and Methodology** ..................................... 2  
   2.1 **Dataset Description** ................................................. 2  
   2.2 **Data Preprocessing** .................................................. 2  
   2.3 **Model Selection and Training** ................................. 3
3. **Implementation/Execution** ......................................... 4  
   3.1 **Data Loading and Exploration** ............................ 4  
   3.2 **Model Training and Evaluation** .......................... 5
4. **Results and Discussion** ............................................... 6
5. **Conclusion** ............................................................. 7
6. **References** ............................................................. 8

## ****Introduction****

### Statement of the Problem:

Heart disease is one of the leading causes of death worldwide. Early diagnosis plays a crucial role in treatment, improving patient outcomes and reducing mortality rates. The healthcare industry increasingly relies on technology and data-driven approaches to predict and diagnose conditions like heart disease. The problem addressed in this project is to create a machine learning model that can predict whether an individual is likely to have heart disease based on certain health-related features. By leveraging data, the aim is to build an automated system that assists healthcare providers in making informed decisions.

### Need for the Project:

Heart disease is highly prevalent, and its timely diagnosis can significantly reduce the risks associated with it. Traditional diagnostic methods are expensive and time-consuming. A predictive model using machine learning can make early-stage diagnosis more accessible, faster, and less costly. This project aims to contribute to the development of such automated systems.

## ****Frameworks and Methodology****

### 2.1 ****Dataset Description****

The dataset used in this project comes from the UCI Machine Learning Repository and contains attributes such as:

* **Age**: The age of the patient.
* **Sex**: Gender of the patient.
* **cp**: Chest pain type.
* **trestbps**: Resting blood pressure.
* **chol**: Serum cholesterol.
* **fbs**: Fasting blood sugar level.
* **restecg**: Resting electrocardiographic results.
* **thalach**: Maximum heart rate achieved.
* **exang**: Exercise induced angina.
* **oldpeak**: Depression induced by exercise relative to rest.
* **slope**: Slope of the peak exercise ST segment.
* **ca**: Number of major vessels colored by fluoroscopy.
* **thal**: Thalassemia (blood disorder).
* **target**: Indicates the presence or absence of heart disease (1 = presence, 0 = absence).

### 2.2 ****Data Preprocessing****

1. **Missing Data Handling:**  
   Check for any missing or null values in the dataset and handle them appropriately (e.g., by removing or imputing missing values).
2. **Feature Scaling:**  
   Standardize the features to bring all data to the same scale using techniques like StandardScaler. This is especially important for distance-based models (e.g., KNN, SVM).
3. **Splitting Data:**  
   Split the data into training and test sets (80% training, 20% testing) using train\_test\_split.
4. **Data Exploration:**  
   Visualize the distribution of data using count plots and correlation matrices to identify relationships between features.

### 2.3 ****Model Selection and Training****

The Random Forest Classifier was selected as the machine learning algorithm for this project due to its ability to handle large datasets and its effectiveness in classification tasks. The model was trained using the training data and evaluated using accuracy, confusion matrix, and classification report.

## ****Implementation/Execution****

### 3.1 ****Data Loading and Exploration****

* Load the dataset using pandas.read\_csv().
* Inspect the dataset for any missing values or inconsistencies.
* Explore the data distribution through visualizations, such as histograms and count plots, to better understand feature relationships.

### 3.2 ****Model Training and Evaluation****

1. **Model Training:**
   * Apply the Random Forest Classifier from sklearn.
   * Train the model using the training dataset after performing feature scaling.
2. **Model Evaluation:**
   * Evaluate the trained model using accuracy score and visualize performance with a confusion matrix.
   * Use classification\_report to generate precision, recall, and F1-score metrics for deeper insights into model performance.
3. **Visual Aids:**  
   Include visual aids such as the **Confusion Matrix**, **Heatmap** of feature correlations, and **Count Plot** of the target variable.

## ****Results and Discussion****

The model’s performance was evaluated based on its accuracy and confusion matrix. The **accuracy score** provides an overall percentage of correctly predicted results. The **classification report** gives insight into the model's precision, recall, and F1-score, which are especially useful for imbalanced datasets. The **confusion matrix** shows how well the model performed across the categories of heart disease presence and absence.

### Discussion:

* The Random Forest Classifier achieved an accuracy of over **X%** (based on your results).
* The model showed high precision and recall, indicating that it is good at predicting both true positives and true negatives.
* The confusion matrix revealed that most errors occurred in a specific class (e.g., predicting heart disease in patients who don’t have it), highlighting areas for model improvement.

## ****Conclusion****

In conclusion, the machine learning model successfully predicted heart disease with an acceptable level of accuracy. The Random Forest Classifier proved to be a reliable model for this classification task. The project demonstrates the potential of using machine learning in healthcare applications for early diagnosis of heart disease. However, there is always room for improvement, such as experimenting with different algorithms, optimizing hyperparameters, or using more advanced techniques like cross-validation to further improve the model’s performance.

Future work could involve collecting additional features or using more extensive datasets to refine the model and make it more robust in real-world healthcare scenarios.

## ****References****

1. UCI Machine Learning Repository. (Heart Disease Dataset)
2. "Introduction to Machine Learning with Python" by Andreas C. Müller, Sarah Guido.
3. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron.
4. Scikit-learn Documentation. (<https://scikit-learn.org/>)