# Major Project

**Submitted by : Karri Deepak Seshu Reddy**

**Code link :** https://colab.research.google.com/drive/1sl8aGo82Z2-bwgKHMWiGGlvL9g19QBw0

**Project Report: Heart Disease Prediction**

**Introduction**

Heart disease is one of the leading causes of death globally. Early prediction and diagnosis are crucial for effective treatment and management. This project aims to build machine learning models that can predict the presence of heart disease based on various health metrics.

**Objectives**

- To analyze and visualize a heart disease dataset.
- To build predictive models using machine learning algorithms.
- To evaluate model performance and compare their accuracies.

**Dataset**

The dataset used for this project is a CSV file containing various health attributes associated with heart disease. It consists of features such as age, cholesterol level, maximum heart rate, and more, with a target variable indicating the presence of heart disease (1: Present, 0: Not Present).

**Dataset Exploration**

1. **Shape**: The dataset contains **303 rows and 14 columns**.
2. **Missing Values**: No missing values were detected in the dataset.
3. **Data Types**: The dataset contains numerical features and a target variable.

## Exploratory Data Analysis (EDA)

### Visualizations

1. **Distribution of Target Variable**: A count plot shows the distribution of heart disease presence in the dataset.
2. **Feature Correlation Heatmap**: A heatmap illustrating the correlation between features, helping identify relationships.
3. **Histograms**: Histograms for each feature indicate their distributions.
4. **Pairplot**: Pairplots show relationships between selected features with respect to the target variable.
5. **Boxplot**: A boxplot visualizing the age distribution across the target variable categories.

### Data Preprocessing

- **Feature Selection**: The target variable was separated from the feature set.
- **Train-Test Split**: The dataset was divided into training (80%) and testing (20%) sets.
- **Feature Scaling**: StandardScaler was used to standardize the features for better model performance.

### Models Implemented

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVM)**

### Results

| Model | Accuracy |
|---|---|
| Logistic Regression | 85 |
| Random Forest | 84 |
| Support Vector Machine | 87 |

**Model Training and Evaluation**

Each model was trained on the training set, and predictions were made on the testing set. The accuracy and classification report were generated for each model.

**Results**

1. **Logistic Regression**
    - **Accuracy**: 0.85
    - **Classification Report**:

```
Logistic Regression Accuracy: 0.85
              precision    recall  f1-score   support

           0       0.83      0.86      0.85        29
           1       0.87      0.84      0.86        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```

2. **Random Forest**
    - **Accuracy**: 0.84
    - **Classification Report**:

```
Random Forest Accuracy: 0.84
              precision    recall  f1-score   support

           0       0.83      0.83      0.83        29
           1       0.84      0.84      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

### 3. SVM

- **Accuracy**: 0.87
- **Classification Report**:

```
SVM Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        29
           1       0.88      0.88      0.88        32

    accuracy                           0.87        61
   macro avg       0.87      0.87      0.87        61
weighted avg       0.87      0.87      0.87        61
```

### 4. Decision Tree

- **Accuracy**: 0.75
- **Classification Report**:

```
Decision Tree Accuracy: 0.75
              precision    recall  f1-score   support

           0       0.69      0.86      0.77        29
           1       0.84      0.66      0.74        32

    accuracy                           0.75        61
   macro avg       0.77      0.76      0.75        61
weighted avg       0.77      0.75      0.75        61
```

### 5. KNN

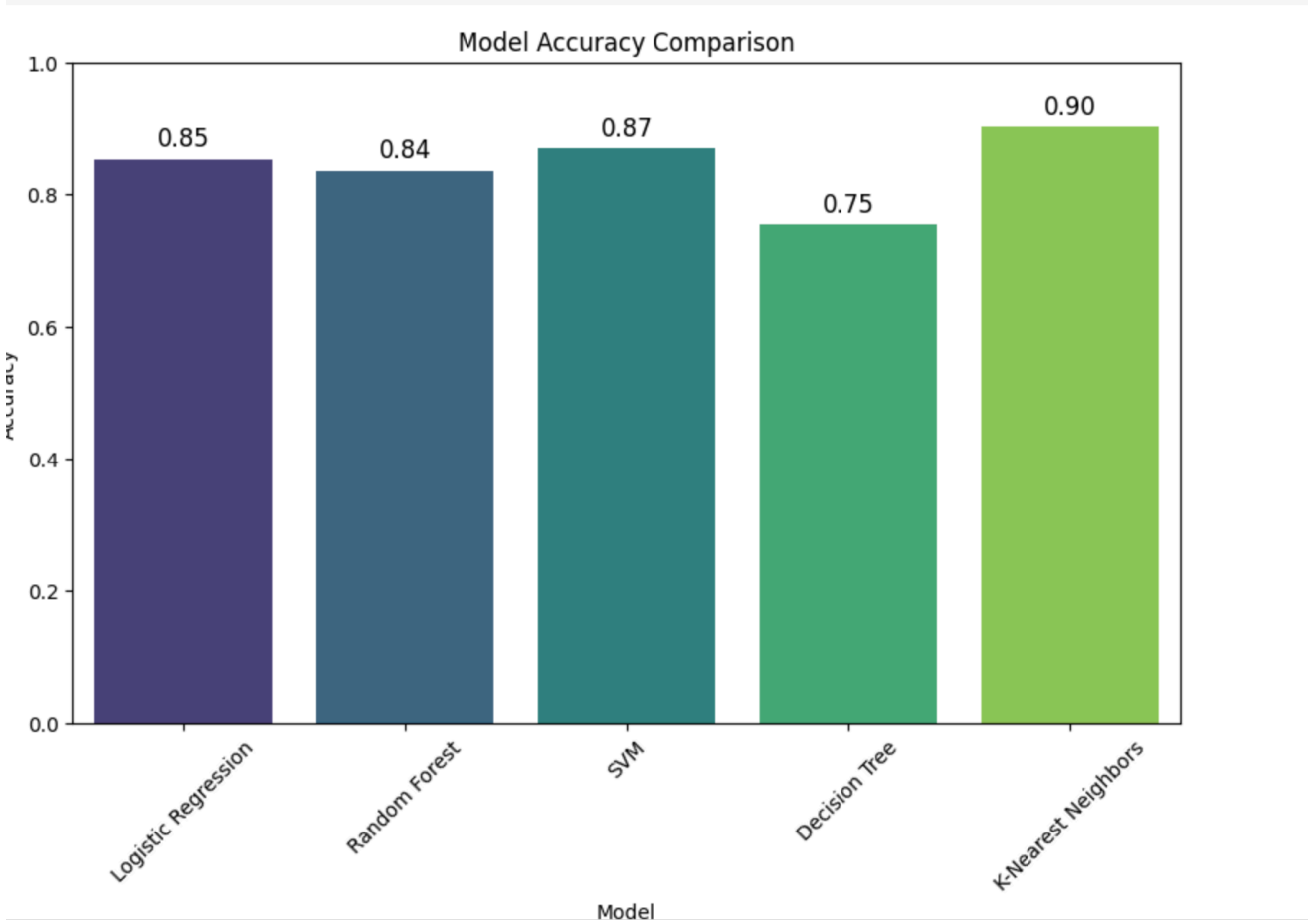
- **Accuracy**: 0.90
- **Classification Report**:

K-Nearest Neighbors Accuracy: 0.90

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.93 | 0.90 | 29 |
| 1 | 0.93 | 0.88 | 0.90 | 32 |
| accuracy | | | 0.90 | 61 |
| macro avg | 0.90 | 0.90 | 0.90 | 61 |
| weighted avg | 0.90 | 0.90 | 0.90 | 61 |

Confusion matrices for each model were also plotted for a better understanding of model performance.

**Model Accuracy Comparison**



A bar chart was created to visually compare the accuracies of the models.

**Conclusion**

The project successfully developed multiple machine learning models for heart disease prediction. Each model was evaluated, and their performances were compared. The KNN model exhibited the highest accuracy (0.90), indicating it may be the most effective for this prediction task. Future work could include tuning hyperparameters and experimenting with additional models to improve performance further.

**Future Work**

- Incorporating additional features or external datasets.
- Hyperparameter tuning for improved accuracy.
- Implementing cross-validation for more robust evaluation.