**Heart Disease Prediction Project**

**Overview**

This project focuses on predicting heart disease using a dataset containing various medical attributes. The analysis involves Exploratory Data Analysis (EDA), data preprocessing, and model building using several machine learning algorithms. The dataset used in this project combines data from the Cleveland, Hungary, Switzerland, and Statlog heart disease datasets. It contains 1,190 entries with 12 attributes, including age, sex, cholesterol, and target (indicating the presence of heart disease).

**Author**

Saron Yaya

**Initial work**

* [Portfolio Projects](https://github.com/Saron222/PortfolioProjects)

**Released on**

* GitHub

**My professional profile on LinkedIn**

* [My LinkedIn Profile](https://www.linkedin.com/in/saron-yaya/)

**Showcase**

**EDA**

* **Data Overview**: Dataset with 1,190 entries and 12 attributes, fully complete with no missing values.
* **Age Distribution**: Age ranges from 28 to 77, with a mean of 53.72; outliers observed.
* **Heart Disease Insights**: 53% of entries indicate heart disease; correlation heatmap shows significant relationships between key variables.
* **Visualizations**: Histograms, countplots, and boxplots reveal age distribution, heart disease prevalence by age and sex, and important correlations for predictive modeling.

**Data Preprocessing**

1. **Handling Missing Values**: Confirmed there are no missing values.
2. **Outlier Detection**: Used boxplots to identify and handle outliers.
3. **Feature Scaling**: Standardized features using **StandardScaler**.

**Model Building**

1. **Splitting Data**: Data split into training (80%) and testing (20%) sets.
2. **Algorithms Used**:
   * Random Forest Classifier
   * Logistic Regression
   * Support Vector Machine (SVM)
   * Decision Tree Classifier
   * K-Nearest Neighbors (KNN) with GridSearchCV for hyperparameter tuning

**Evaluation**

Each model's performance was evaluated using accuracy, classification reports, and confusion matrices.

* **Random Forest**: Accuracy - 94.54%
* **Logistic Regression**: Accuracy - 86.13%
* **Support Vector Machine**: Accuracy - 84.45%
* **Decision Tree**: Accuracy - 90.34%
* **KNN**: Best parameters - **{'n\_neighbors': 1}**, Accuracy - 76.89%

**Visualization**

* **Correlation Heatmap**: Showcased the relationships between variables.
* **Confusion Matrix**: Visualized the performance of the KNN model.

**Technologies Used**

* Python
* Scikit-learn
* Pandas
* Matplotlib

**Installation**

1. Clone the repository: git clone [https:// https://github.com/Saron222/PortfolioProjects.git](https://github.com/saronyaya/DSC510.git)
2. Navigate to the project directory: **cd Heart Disease Prediction.ipynb**

**Contributing**

1. Fork the repository from [Saron222/PortfolioProjects](https://github.com/Saron222/PortfolioProjects/fork)
2. Create your feature branch: **git checkout -b feature/your-feature-name**
3. Commit your changes: **git commit -am 'Add your feature'**
4. Push to the branch: **git push origin feature/your-feature-name**
5. Create a new Pull Request in the original repository