Prediction of Heart Failure Using Machine Learning

Table of Contents

Overview This project aims to develop machine learning models for predicting heart failure risk using patient clinical data, then choose the one with the best performance metrics. The candidate algorithms to develop these models are: K-Nearest Neighbors , Random Forest , Logistic Regression and Neural Network. Such predictions can assist pharmaceutical companies in optimizing clinical trial participant selection and improving treatment effectiveness.

Business Case: Pharmaceutical Industry

In the pharmaceutical industry, **identifying target patients** is very important for the success of clinical trials and drug treatments. The machine learning model in this project provides several benefits:

1. Improves candidate selection for clinical trials.
2. Reduces trial time and costs by selecting the right participants.
3. Enhances treatment effectiveness.

Project Structure:

```
Prediction_Of_Heart_Failure/
├── data/
│   ├── processed/            # Folder for processed data
│   ├── raw/                  # Folder for raw data
│   └── SQL/                  # Folder for tables
├── experiments/              # Jupyter notebooks for experiments and EDA
├── models/                   # Trained models
├── reports/                  # Visualizations, evaluation reports from EDA and model performance
├── README.md                 # Project overview and instructions (this file)
├── requirements.txt          # List of dependencies
└── src/                      # Python scripts for preprocessing, training, and evaluation
```

Dataset: The dataset is sourced from Kaggle and contains 918 samples with 11 clinical features, including:
Numerical Features:

- Age
- Resting Blood Pressure

- Serum Cholesterol
- Maximum Heart Rate Achieved
- ST Depression (OldPeak)
- FastingBS (Fasting Blood Sugar)

Categorical Features:

- Sex
- Chest Pain Type
- Fasting Blood Sugar
- Resting Electrocardiogram Results
- Exercise-Induced Angina
- Slope of ST Segment

Target Variable:

- HeartDisease (binary: 0 = No heart disease, 1 = Heart disease)

Data Exploration and Preprocessing

Steps in Preprocessing:

1. **Inspection and Cleaning:**

i. Checked for duplicates and missing values (none found).
ii. Summarized numerical features with descriptive statistics and visualizations.
iii. Conducted statistical tests (e.g., Chi-Square and T-tests) to identify feature significance.

2. **Feature Engineering:**

i. Separated features into numerical and categorical groups.
ii. Encoded categorical variables using one-hot encoding.
iii. Standardized numerical variables for model training.

3. **Visualizations:**

i. Target variable distribution: Proportion of heart disease cases.
ii. Histograms and KDE plots for numerical features.
iii. Bar charts for categorical features grouped by heart disease status.

Model Building and Comparison

Models Tested:

1. **Random Forest**

i. Best Parameters: max_depth=3, min_samples_leaf=20, n_estimators=100
ii. Performance: Train Accuracy: 86.38%, Test Accuracy: 92.39%, AUC-ROC: 0.9755

2.
3. **Logistic Regression**

i. Best Parameters: C=0.19, penalty='l2', solver='liblinear'
ii. Performance: Train Accuracy: 86.38%, Test Accuracy: 90.76%, AUC-ROC: 0.9697

4. **K-Nearest Neighbors (KNN)**

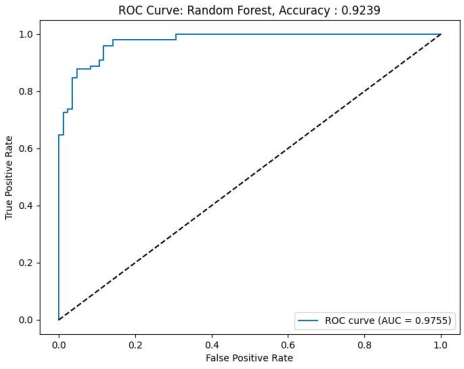
i. Best Parameters: n_neighbors=20
ii. Performance: Train Accuracy: 86.38%, Test Accuracy: 90.76%, AUC-ROC: 0.9700

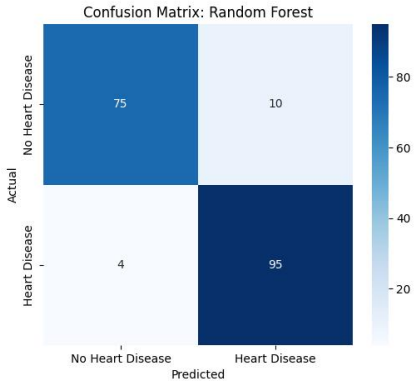
5. **Neural Network**

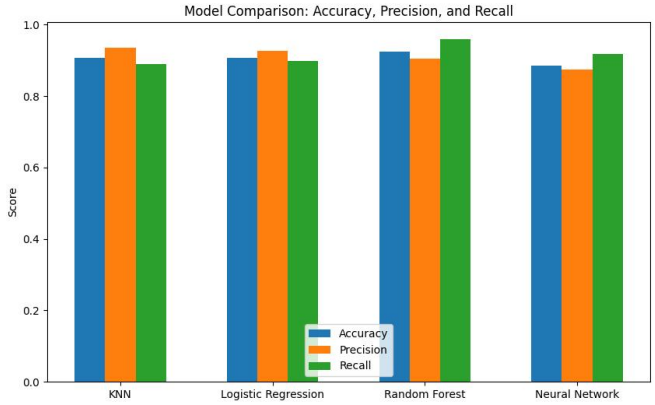
i. Best Parameters: batch_size=32, epochs=100
ii. Performance: Train Accuracy: 84.88%, Test Accuracy: 88.59%, AUC-ROC: 0.9626

Results

The Random Forest model outperformed others with the highest test accuracy and AUC-ROC score, indicating superior predictive ability without overfitting.







Conclusion

The Random Forest classifier was identified as the best-performing model, achieving a 92.39% test accuracy and an AUC-ROC of 0.9755. This model provides a reliable tool for predicting heart failure risk, which can benefit pharmaceutical companies in improving clinical trial participant selection and optimizing treatment strategies.