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:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

Numerical Features:

These numerical features are used to assess various aspects of heart health, helping in identifying risk factors and predicting heart disease outcomes:

1. Age: Represents the patient's age in years. Age is an important factor in heart disease risk, with older individuals generally having a higher likelihood of developing heart disease.

2. Resting Blood Pressure: The blood pressure measured while the patient is at rest, typically taken when the person is seated and relaxed.

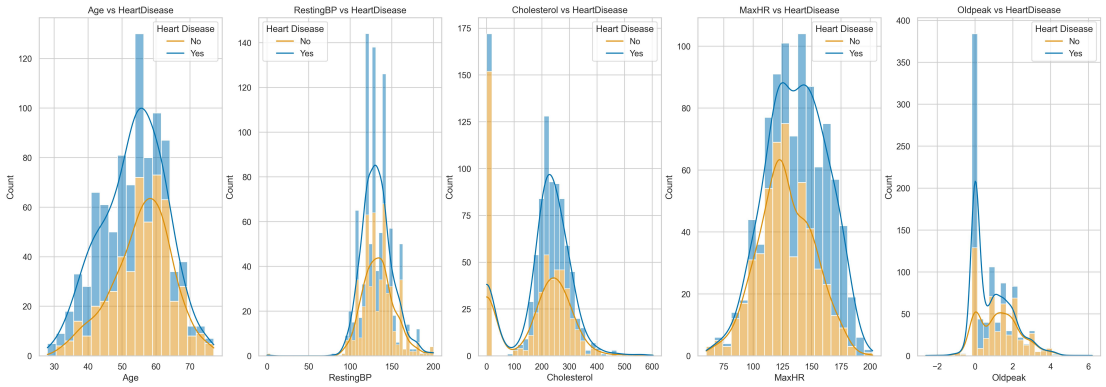
3. Serum Cholesterol: The total amount of cholesterol in the blood, which includes LDL (low-density lipoprotein, often called "bad" cholesterol), HDL (high-density lipoprotein, or "good" cholesterol), and triglycerides. Elevated cholesterol levels, especially high LDL and low HDL, can lead to plaque buildup in the arteries, increasing the risk of heart disease.

4. Maximum Heart Rate Achieved: The highest heart rate the patient achieves during exercise or stress tests. This measure helps evaluate the heart's capacity for physical exertion. A lower-than-normal maximum heart rate may indicate underlying heart issues.

5. ST Depression (OldPeak): Refers to a depression in the ST segment of the electrocardiogram (ECG) during exercise, which is used to assess the heart's response to physical activity. ST depression is often associated with insufficient blood flow to the heart and can be indicative of ischemia (lack of oxygen) or coronary artery disease.

6. Fasting Blood Sugar (FastingBS): - The level of glucose (sugar) in the blood after fasting for at least 8 hours. Elevated fasting blood sugar levels can be an indication of diabetes or pre-diabetes, both of which are risk factors for developing heart disease.

Analysis of Numerical Features Related to Heart Disease is shown the figure below:



Categorical Features:

Each of these categorical features helps provide important insights into the patient's heart health and aids in diagnosing or predicting heart disease

1. Sex: This feature represents the gender of the individual. It is typically encoded as a binary categorical variable, where "1" represents male and "0" represents female.

2. Chest Pain Type (ChestPainType): This feature categorizes the type of chest pain the individual has experienced. It is classified into several types, such as typical angina, atypical angina, non-anginal pain, or asymptomatic.

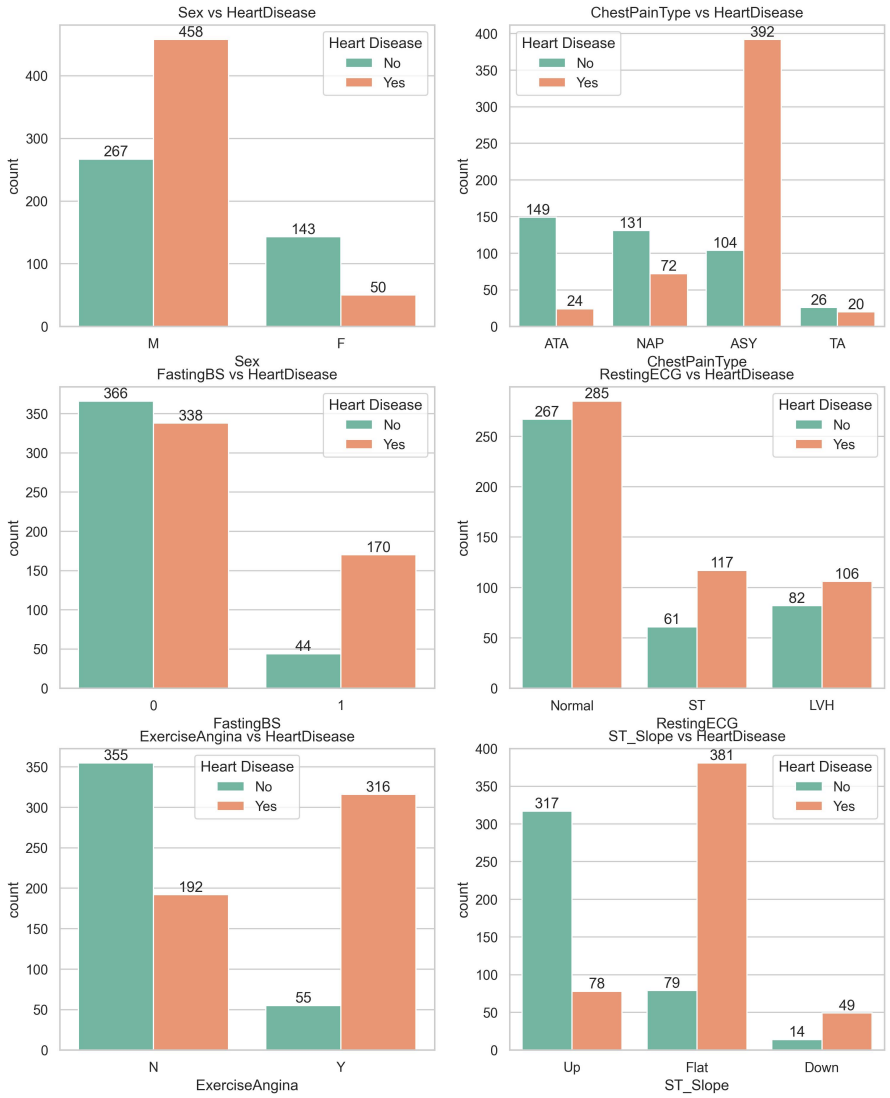
3. Fasting Blood Sugar (FastingBS): This variable represents whether the individual's fasting blood sugar is greater than 120 mg/dl. It is a binary categorical feature, where "1" indicates that the fasting blood sugar level is above 120 mg/dl (indicating potential risk for diabetes), and "0" indicates it is below or equal to 120 mg/dl.

4. Resting Electrocardiogram Results (RestingECG): This feature refers to the results of the resting electrocardiogram (ECG), which is used to measure the electrical activity of the heart. It is typically classified into categories like normal, ST-T wave abnormality, or left ventricular hypertrophy.

5. Exercise-Induced Angina (ExerciseAngina): This variable indicates whether the individual experiences angina (chest pain) during exercise. It is a binary categorical feature, where "1" means that the person experiences angina during exercise, and "0" means they do not.

6. Slope of ST Segment (ST_Slope): This feature refers to the slope of the ST segment in the ECG during exercise testing. The ST segment is a part of the ECG that represents the period when the heart is electrically neutral. The slope can be classified into three types: up-sloping, flat, or down-sloping. The type of slope can provide insights into the presence of heart disease or abnormal heart function during exercise.

Analysis of Categorical Features Related to Heart Disease is shown the figure below:



Target Variable:

HeartDisease: This is the binary target variable in the dataset, representing the presence or absence of heart disease in the individual. It is a classification variable with two possible outcomes:

0 (No heart disease): This indicates that the individual does not have heart disease.
1 (Heart disease): This indicates that the individual has been diagnosed with 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 and the results:

Model	Test Accuracy	Precision	Recall	AUC Score_test
Random Forest	**0.9239**	**0.9048**	**0.9596**	**0.9755**
Logistic Regression	**0.9076**	**0.9271**	**0.8990**	**0.9697**
KNN	**0.9076**	**0.9362**	**0.8889**	**0.9700**
Neural Network	**0.9185**	**0.9375**	**0.9091**	**0.9471**

1.Random Forest : Best model saved as models/'best_rf_model.pkl'

2.Logistic Regression : model saved as 'models/best_logreg_model.pkl'

3.K-Nearest Neighbors (KNN) : model saved as 'models/best_knn_model.pkl'

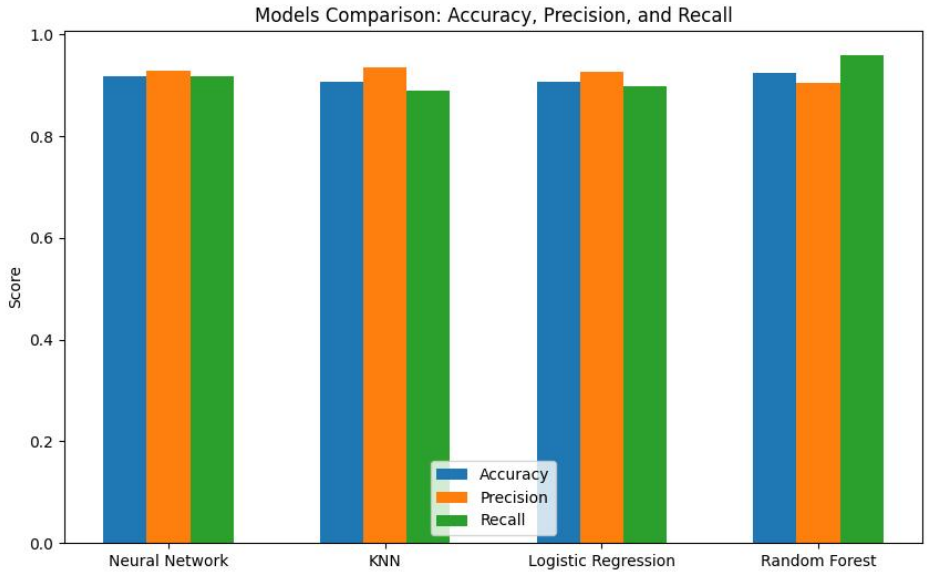
4.Neural Network : Best model saved as 'models/best_nn_model.pkl'

Conclusion:

The chart below represents the comparison of four models-Neural Network, KNN, Logistic Regression, and Random Forest-based on three performance metrics: Accuracy, Precision, and Recall.

The performance scores for each model across all three metrics are very similar. The bars for each model show high values, typically close to 1, indicating strong performance. Neural Network, KNN,

Logistic Regression, and Random Forest all have almost identical scores across the three metrics, with each metric showing strong performance for all models.



The Random Forest classifier was identified as the best-performing model, achieving a **92.39%** test accuracy and an **AUC-ROC of 0.9755,** see the figure below. 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.