Capstone Project

Heart Disease Prediction Final Report

1. Define the Problem Statement

The primary objective of this project is to develop a robust machine learning model that can accurately predict the presence of heart disease in a patient based on a set of key clinical and demographic attributes.

Goals:

- To create a predictive tool that can serve as a decision-support system for healthcare professionals, enabling earlier and more accurate diagnosis.
- To identify the most significant predictors of heart disease from the available features, providing valuable clinical insights.
- To compare multiple machine learning algorithms to determine the most effective and reliable model for this specific diagnostic task.

2. Model Outcomes or Predictions

- **Type of Learning:** This is a Classification problem. The model's task is to assign one of two class labels to each patient: 1 (Heart Disease) or 0 (No Heart Disease).
- **Expected Output:** The selected model will take a patient's 11 feature values (e.g., Age, Sex, Cholesterol, etc.) as input and will give a binary prediction (1 or 0) to assess if they may have heart disease or not. Additionally, the model can output the probability of the patient belonging to each class, which is crucial for assessing the model's confidence in its prediction.
- **Supervised vs. Unsupervised:** In this Capstone Project, Supervised learning algorithms were used. The reason is that the training data (heart.csv) contains a labeled target variable, 'HeartDisease', which the models learn from to make future predictions. Unsupervised learning was not deemed useful or appropriate.

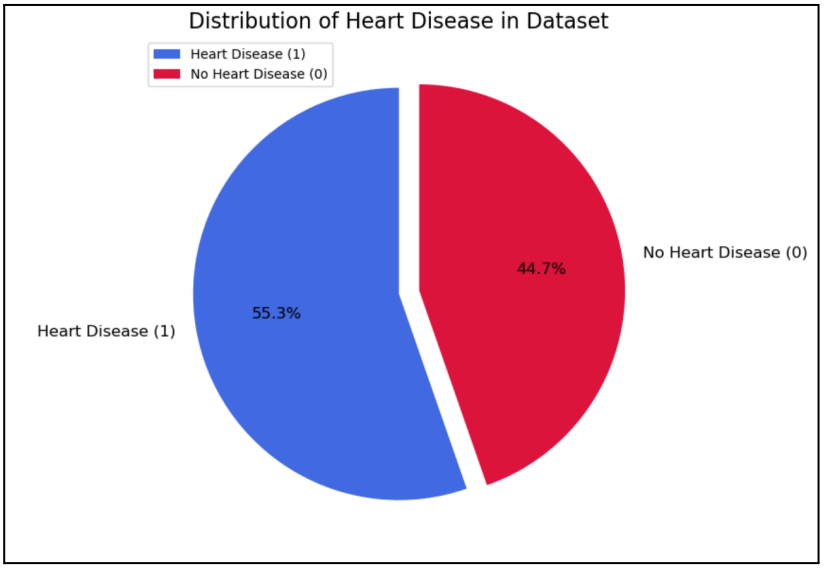
3. Data Acquisition

- **Data Source and Analysis:** The data used for this model was acquired from a single source: the 'heart.csv' file located on Kaggle's website (url: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction).
    - This dataset contains 12 columns, including 11 feature columns and one target variable (HeartDisease).
    - The features are a well-documented mix of demographic data (Age, Sex), clinical measurements (RestingBP, Cholesterol, MaxHR), and categorical observations (ChestPainType, RestingECG, ExerciseAngina, ST_Slope).
    - The data was deemed suitable for this problem because it directly contains the necessary features and target labels required for a supervised classification task.

Column Descriptions

Column	Description
Age	Patient's age in years. Older age is associated with higher cardiovascular risk.
Sex	Biological sex (Male/Female). Males have higher risk at younger ages; gap narrows after menopause.
ChestPainType	Chest pain type: typical angina, atypical angina, non-anginal pain, or asymptomatic.
RestingBP	Resting blood pressure (mmHg). High values indicate hypertension and increased risk.
Cholesterol	Serum cholesterol (mg/dL). High levels contribute to atherosclerosis.
FastingBS	Fasting blood sugar (1 if >120 mg/dL, else 0). Indicates diabetes/prediabetes risks.
RestingECG	Resting ECG results. Can detect arrhythmias and heart damage.
MaxHR	Maximum heart rate during exercise. Lower values may indicate poor fitness or heart problems.
ExerciseAngina	Chest pain during exercise (Yes/No). Suggests coronary artery disease.
Oldpeak	ST depression on ECG during exercise (in mm). Indicates inadequate blood flow.
ST_Slope	Slope of peak exercise ST segment (upsloping, flat, downsloping). Downsloping/flat are more concerning.

- A key initial assessment was to check the balance of the target variable, which is crucial for training an unbiased classifier. A pie chart was created in the notebook to show the distribution of data in the target variable where 55.3% of the data was for patients with heart disease and remaining 44.7% for patients who did not have heart disease. This helped ensure there was no bias and the data was balanced.



4. Data Preprocessing/Preparation

Techniques used to ensure your data was free of missing values, and inconsistencies:

- **Checking for Null Values:** The heart.info() function was used to confirm that there were no missing (null) values in any of the columns.
- **Handling Inconsistent values:** Through descriptive statistics (heart.describe()), it was identified that the RestingBP and Cholesterol columns contained zero values, which are technically not possible for any living human. These were treated as data errors. This inconsistency was corrected without losing data by replacing

the zero values with the median of the non-zero values for each respective column. The median was chosen as it is robust to outliers.

Training and Test sets data split:

- The data was split into features (X) and the target variable (y).
- The train_test_split function from Scikit-learn was used to partition the data.
- A 70/30 split was implemented, where 70% of the data was used for training the models and 30% was reserved as an unseen data set for final evaluation.
- A random_state=42 was set to ensure that the split is reproducible, which is required for consistent results.

Encoding and Transformation:

A ColumnTransformer pipeline was created to apply the correct preprocessing to each type of feature. This dataset had both Numerical and Categorical features:

- **Categorical Feature Encoding:** The categorical columns (Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope) were transformed using One-Hot Encoding. This technique is used to convert each category into a new binary (0/1) column, this allows the machine learning models to process the non-numeric data.
- **Numerical Feature Scaling:** The numerical columns (Age, RestingBP, Cholesterol, etc.) were scaled using StandardScaler. This process standardizes the features to have a mean of 0 and a standard deviation of 1. This is done to ensure features with larger ranges do not influence the models training process incorrectly.

5. Modeling

A comparative approach was taken by selecting four machine learning algorithms suitable for this classification problem. This strategy allows for a robust comparison across different model types to determine which model is best suited. Below are the four models trained:

- **Logistic Regression:** A linear model that is highly interpretable and serves as a strong baseline.
- **Random Forest Classifier:** An ensemble model based on decision trees (bagging). It is known for its high accuracy and robustness against overfitting.
- **Bagging Classifier:** A general ensemble meta-algorithm (bagging) that fits multiple base classifiers on random subsets of the data and aggregates their predictions. It's effective at reducing variance.
- **XGBoost (Extreme Gradient Boosting):** An advanced and highly efficient implementation of gradient boosting, which is another ensemble technique.

  For each of these models, GridSearchCV was used with 5-fold cross-validation to perform hyperparameter tuning. This ensures that the best parameters of each model were found and evaluated.

6. Model Evaluation

This was a classification problem, so four different classification models were trained and evaluated, namely: Logistic Regression, Random Forest, Bagging, and XGBoost.

- **Evaluation Metrics:** A suite of metrics was used to provide a holistic view of model performance:

- **Accuracy**: The overall percentage of correct predictions.
- **Precision**: Of all the patients the model predicted had heart disease, how many actually did have a heart disease? (Measures the cost of false positives).
- **Recall** (Sensitivity): Of all the patients who actually have heart disease, how many did the model correctly identify? (Measures the cost of false negatives).
- **F1-Score**: The harmonic mean of Precision and Recall, providing a single score that balances both.
- **AUC-ROC Curve**: A plot of the model's true positive rate against its false positive rate. The Area Under the Curve (AUC) represents the model's ability to distinguish between the two classes. A value closer to 1.0 indicates a better model.

7. Conclusion:

The Random Forest Classifier was determined to be the most optimal model:

- **Recall:** In a medical context, it is far more dangerous to miss a positive case (a false negative) than to incorrectly flag a healthy person (a false positive). Therefore, Recall for the "Heart Disease" class was the most important metric.
- **Random Forest's Performance:** As shown in the comparison table, the Random Forest model achieved the highest Recall (0.908), meaning it was the best at identifying patients with the disease.
- **Overall Strength:** While the Logistic Regression model had slightly higher accuracy and precision, the Random Forest's excellent Recall, combined with a very strong F1-score (0.894) and AUC (0.945 from the ROC plot), makes it the most reliable and safest choice for this application.
- **Top Predictors (features)**
  - ST_Slope: Slope of the ST segment during exercise.
  - Oldpeak: ST depression induced by exercise.
  - ChestPainType_ASY: Asymptomatic chest pain.
  - Age: Patient's age.
  - ExerciseAngina: Whether angina was induced by exercise.

========== COMPARISON OF MODEL PERFORMANCE FOR HEART DISEASE PREDICTION ==========

	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	0.876812	0.922078	0.865854	0.893082
Random Forest	0.873188	0.881657	0.908537	0.894895
Bagging	0.829710	0.897959	0.804878	0.848875
XGBoost	0.855072	0.907895	0.841463	0.873418