HEARTFAILUREPREDECTIONUSINGMACHINELEARNING

A TERM PROJECT REPORT SUBMITTED

by

Nukapeyyi. Latha Kumari
322103282071

Sathi. Sanvitha Reddy
322103282088

Galla. Jyothsna
322103282114

Karakavalasa. Likitha
322103282122

Department of Computer Science and Engineering (AI & ML)
GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN
(Affiliated to Andhra Pradesh University, Visakhapatnam) 2022-26

GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN

Department of Computer Science and Engineering (AI &ML)

CERTIFICATE

This is to certify that the term project report titled << Project Title >> is a bonafide work of following III B.Tech Ist Semester students in the Department of Computer Science and Engineering(AI&ML), Gayatri Vidya Parishad College of Engineering for Women affiliated to ANDHRA UNIVERSITY, Visakhapatnam during the academic year 2024-25.

Ms (Nukapeyyi Latha Kumari) Ms (Sathi Sanvitha Reddy)

Ms (Galla Jyothsna) Ms (Karakavalasa Likitha)

Project Coordinator

TABLE OF CONTENTS

1.Abstract

1.1 Background:

Heart failure is one of the leading causes of morbidity and mortality in the developed world. Traditionally diagnosis and prediction of heart failure are mainly based on a patient's clinical symptoms and history as well as on any physical examinations that might prove subjective and time-consuming to obtain. Recently, more and more attention has come to machine learning in order to enhance diagnostic accuracy when it comes to predicting a heart failure onset and evolution. ML algorithms, when analyzing large, complex datasets, can reveal clinical, demographic, and medical patterns that are hard to be detected by humans. These models provide the possibility of more accurate and earlier identification of patients at risk of heart failure, thus allowing timely interventions. ML techniques may also improve risk assessment, help guide personalized treatment plans, and reduce hospital re-admissions, thus improving patient outcomes.

1.2 Dataset Characteristics:

The "Heart Failure Prediction" dataset has a variety of attributes in terms of patient health conditions and clinical measurements that may be used to predict heart failure. The principal attributes of the dataset are:
Age- The age of the patient.

- Gender: Male or female.
- Ejection Fraction: Measures the amount of blood pumped from the heart with every contraction, thereby indicating how efficient the heart is at the time of contraction.
- Serum Creatinine: Measured creatinine in the serum to assess renal function.
- Blood Pressure: Systolic and diastolic BP obtained from the patient
- Platelets: The count of platelets in the blood, which is equivalent to the coagulability of blood.
- Smoking History: Whether the patient has ever smoked (Yes/No)
- Diabetes: Whether the patient has diabetes mellitus (Yes/No).
- Heart Disease: Whether the patient has some heart disease (Yes/No).
- Target Variable: A binary variable that identifies whether the patient ever had a history of heart failure(1=Yes,0=No).
  This will enable us to predict the probability of heart failure occurrence in a patient using a Support Vector Machine-based machine learning model on this dataset. In other words, the early detection will enable the facilitation of preventive measures, as well as more effective decision-making on the part of healthcare providers, all of which will help in reducing the burden of heart failure both on patients and the health system.

2.Methodology

The methodology to forecast heart failure is divided into steps from the collection of the data through preprocessing, actual model building, and finally validation. Given below is a detailed description in building and deploying a machine learning approach for heart failure forecasting:

1. Data Collection:

The first step involves gathering data on heart failure. Here, we use the "Heart Failure Prediction" dataset that contains clinical attributes along with the history of a patient. The essential features in the dataset would include the age and gender of the patient; medical history such as smoking, diabetes, and heart disease; and clinical measures, for example, ejection fraction, blood pressure, serum creatinine. The target variable to be used will be binary classification for heart failure-1 for having heart failure and 0 for no heart failure.

2. Data Preprocessing:

- Preprocessing ensures that a data set is ready and clean for running on machine learning algorithms. Dealing with Missing Data: All missing or incomplete values in the data set are dealt with. This may include imputation-that is, filling in missing values with mean, median, or mode-or deleting rows/columns that have too many missing data.
- Normalization and Scaling: If blood pressure, serum creatinine, and age are continuous variables, then it is done such that all feature ranges are comparable. A technique most commonly used in this area includes either Min-Max scaling or Standardization.
- Encoding categorical variables. In this case, one-hot encoding or label encoding is applied to convert categorical variables such as gender, smoking history, and heart disease history into numerical format. Feature engineering based on domain knowledge: new features consist of creating new binary features in the risk categories or aggregating existing features.

3. Exploratory Data Analysis (EDA)

- EDA gives insights about the structure and relationship between variables in a data set.
- Descriptive Statistics: Summary statistics like mean, median, variance, and standard deviation for numerical attributes.
- Visualization: Histograms, scatter plots, and correlation matrices will help identify patterns, outliers, and correlations between features.
- Class Imbalance Check: The distribution of the target variable, which is heart failure vs. no heart failure, is checked. If the dataset is imbalanced, one may apply oversampling (SMOTE) or under sampling techniques to balance the classes.

4. Model Selection

A few machine learning algorithms are chosen for heart failure prediction.

- Logistic Regression: A Very Common Model When Binary Classification Applied.
- Decision Trees: A model, tree-based model that splits the data based on the values of the features and this makes it very interpretable.
- Random Forest: Ensemble learning that combines a number of decision trees to improve the accuracy and reduce overfitting.
- Gradient Boosting Machines (GBM): Techniques like Boost or Light that improve model performance by sequentially correcting errors made by previous trees.
- Neural Networks: These are deep learning models that can approximate complex, nonlinear relationships in the data.

5. Training Models

Data Splitting : The split between the training and testing is usually made of the most frequently used dataset split of 80 / 20 or 70/30. Training set for models picked is usually performed with cross-validation, including

K-fold cross validation to assess the generalisation of a model and to not overfit. Derivation of hyperparameter search by means of grid or random search will also help in getting optimal settings of a model.

6. Model Evaluation

The performance of the trained models is measured against the following metrics:
Accuracy: The ratio of instances that were correctly predicted by the model.
Precision: The ratio of true positives in all the positive predictions, especially crucial for preventing false positives.
Recall (Sensitivity): The ratio of true positives in all actual positive instances, especially crucial for identifying heart failure patients.
F1 Score: It will be the harmonic mean of both precision and recall that give a good balance on how good the model is performing
AUC-ROC: Measures how good a model would be in terms of discrimination of classes since the greater value that would be obtained, then the better would the model be.
Confusion Matrix: These summaries are related to the result of the predicted results and also encompass false positives, false negatives, true positives, and true negatives.

7. Model Comparison and Selection

Here, various models are compared through the evaluation metrics and the model that is best performing at that time is selected for further application. If required, then merit of different models is combined with the help of techniques such as voting classifiers.

8. Model Deployment

This selected model further applies it to the clinical or decision support environment. It gets plugged into the health care system with actual live predictions and decisions that make it a reality in action. It develops an interface user-friendly, captures the input of the patient data provided by health care providers from where heart failure risk prediction will be generated.

9. Continuous Monitoring and Upgrades

Its accuracy is maintained for long periods because this model is always monitored and checked. It is constantly retrained and updated whenever it discovers new data or changing practices in healthcare.

Outcomes:

The SVM model showed better performance with 89.33% accuracy on the test dataset and hence was found to be the best fit algorithm to predict heart failure.
The application developed here gives the following features:
•It offers a simple and interactive interface to input data by loan officers and applicants.
•It also gives visualizations that help understand the distribution of the features.
•Accurate prediction of likelihood.
• SVM. Vector Machine was used as it is powerful for handling complex and nonlinear interactions in data. It is trained using an RBF kernel, which enables it to classify the patients with respect to the pattern in the high-dimensional feature space.
• Hyperparameter Tuning: Main parameters such as C (penalty parameter) And gamma (kernel parameter) were optimized with respect to best performance possible using grid search

3. INTRODUCTION

Heart failure is a chronic condition that impairs the pumping ability of the heart, leading to inadequate oxygen supply to organs in the body. Forecasts of heart failure are an important step towards improving patient outcomes since early interventions prevent severe complications and lead to improvement in the quality of life. Considerable progress in machine learning and AI has dramatically improved our capability to predict heart failure much more accurately.
This paper shall talk about the evolution of ML models, especially using the SVM algorithm, in forecasting a heart failure with accuracy and the use of clean processed data along with advanced techniques proven to be used by the ML for getting the best efficiency with a reduction in biases,

3.1 Objectives

Develop a Prediction Model that uses Machine Learning:

It is a prediction model built based on machine learning with the SVM algorithm for heart failure prediction. The model would be evaluated with proper metrics like accuracy, precision, recall, F1 score, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve), ensuring that the hyperparameters of the model are fine-tuned so that the prediction becomes as accurate as possible, and finally deployed in real-time prediction.
You can elaborate further with this approach by trying other algorithms such as Random Forest, Boost, or Neural Networks, and experimenting with other techniques for feature engineering to fine-tune performance even better.

- Hyperparameter Tuning: Optimize the models to increase accuracy while minimizing false negatives, meaning it should not reject eligible applications. Design an User-Friendly Interface.
- Web Application: Create an interactive web application using Streamlit an highly versatile open-source framework for machine learning applications.
- Input Features: Collect user's loan application information including the applicant's income, co-applicant's income, credit history, loan amount, and property area.
- Real Time Predictions: Offer the loan approval predictions in real-time to users in a friendly manner.
- Data Visualizations: Let us not forget the intuitive visualizations to make users understand trends in applicant incomes and the percentage of loan approvals.

3.2 Scope of the Project

Using Publicly Available Datasets

This project uses the dataset heart.csv that contains 918 records of loan applications. This will have features like age, Gender, Pistolong history ejection fraction, Diabetes, target value(0=No)(1=Yes) to train and test all of these models.
Four Machine Learning Model Implementation
Four different machine learning models are used to predict the outcomes of loan approvals and employ multiple approaches to prediction:
a. Logistic Regression: A simple statistical model that could be applied for binary classification to determine the relationship between features and loan approval.
b. Random Forest: Ensemble learning technique where multiple Accuracy:83.45
c. Decision Tree: The Accurcy:84.72%
d. Support Vector Machine (SVM): This is a robust algorithm that separates all the data points using hyperplanes in high-dimensional feature spaces. Accuracy: 89.36
The model computes accuracy, recall, and precision on each model to compare and establish the best algorithm with loan approval predictions.

4. Literature Review

4.1 Machine Learning Applied to Heart Failure Prediction

It has transformed the future of healthcare by predicting chronic diseases such as heart failure. Through large data sets, great computational powers, and advanced algorithms, machine learning may provide precise, patient-specific, and timely predictions regarding risk concerning heart failure. Here is how machine learning is applied to predict heart failure:

1. Early Detection of Heart Failure
• Predictive Onset: Perhaps, by learning machine, the onset could be predicted when heart failure initiates before the onset of symptoms. Thus, in case intervention occurs early enough, it would go quite a long way in presenting more favorable outcomes.
•Predictive Models: Algorithms, such as Random Forest, Support Vector Machine, and Logistic Regression can learn from historic patient data by providing a probability of heart failure that likely to happen in the near future based on the present existing medical conditions.

2. Risk Stratification and Prognosis
Identification of High-Risk Patients: The ML algorithms classify the patients based on their risk scores toward heart failure. The model predicts which patients have the higher chance of hospitalization or readmission or have worsened conditions of heart failure.
• Survival Estimation: Using data generated from ejection fraction, blood pressure, co morbid conditions, and laboratory findings, the ML algorithm estimates survival rates for individual patients and develops personalized recommendations based on the needs and requirements of the patients.

3. Enhanced Diagnostic Specificity
•Clinical Data Integration: ML can process many kinds of clinical data, including EHRs, echo-cardiogram images, ECG signals, and lab results to integrate to present an overall diagnosis. This reduces the errors and levels of subjectivity found with human diagnosis.
•Image Processing: The algorithms from ML, such as CNNs, process and explain medical imaging, for instance echo-cardiograms or CT scans, to identify structural heart abnormalities that are symptomatic of heart failure.

4. Tailored Treatment Plans
•Tailored Interventions: An ML model can provide a few tailor-made treatment plans by the detailed profile of the individual, like it could offer switching medicines, changing lifestyles, or observation plans that would prolong the progression of heart failure
•Medicine Management: An ML model can be fed a particular patient to predict an appropriate medicine for an individual based on their disease or gene and their associated histories regarding medical disorders that shall help improve.

5. Real-time monitoring and predictive alerts
•Continuous Monitoring: Through wearable devices and remote monitoring systems, the vital signs of patients like heart rate, blood pressure, and oxygen levels can be continuously monitored, and then predictive alerts based on ML algorithms will enable healthcare providers or patients to take appropriate measures to avoid emergencies in advance.
• Early Prediction of Decompensated Heart Failure: Based on real-time data, the ML model predicts early decompensated heart failure where the heart fails to pump adequate blood as per the demand, and alert for prompt intervention can be raised.

6. Data-Driven Insights
• Big Data Analysis: ML can analyze huge health data, including records from hospitals, patient surveys, sensor data, and genetic information. Using these patterns, ML can discover new risk factors, disease pathways, and interventions that have been unknown before.
This would give the healthcare practitioner the scope to use models of ML that can predict the new biomarkers or risk factors concerning heart failure or progression.

7. Health Care Cost Cuts
•Optimization of Resources: By knowing which patient is going to need intensive care or even a bed in a hospital, ML can manage health care resources in an optimized way such that the expense will be decreased for healthcare.
•Reducing Re-admissions: Machine learning can predict the probability of patient readmission and healthcare providers can make interventions to minimize these and reduce the overall cost of care.

8. Heart Failure Complications Forecasting
•Comorbidity Prediction: Heart failure is generally accompanied by other chronic conditions like diabetes, hypertension, or kidney disease. Using the ML model, comorbidities can be predicted, and thus preventive care strategies are assisted.
•Disease Progression Monitoring: Machine learning helps monitor heart failure disease progression. It tracks vital signs over time, thus alerting clinicians about a possible worsening of the condition.

9. Balancing Imbalanced Data
● Treatment of Infrequent Incidents: In almost all the data, class imbalance exists, in which number of patients are far below that of normal's. This kind of imbalance of the class and enhance its capacity for modeling could be increased so that one could have predictions of events like failure of heart. Through techniques such as SMOTE (Synthetic Minority Over-sampling Technique), Ensemble methods, and anomaly-based detection
10. NLP
● Information Extraction from Unstructured Data Information: All information about patients, in its entirety, of any health care system is fed into free text formats, including clinical notes and discharge summaries. The meaningful data that are going to be extracted from the text fields by the help of NLP techniques includes symptoms, history of previous treatments, results of tests, and more relating to the risk of heart failure.

Conclusion
Machine learning is transforming the landscape of heart failure prediction by allowing earlier detection, more accurate risk stratification, personalized treatments, and real-time monitoring. The potential to improve patient outcomes and
The burden of heart failure is huge as these technologies are embraced by health care systems. ML will enable clinicians to make decisions based on data, hence improving the quality of care and the quality of life for the patient

4.2 Related Work

Studies Applying Machine Learning to Predict heart failure
Several studies have used machine learning algorithms to predict heart failure. The studies are aimed at proving the ability of machine learning techniques to automate the loan evaluation process and enhance the accuracy of decisions. Some of the studies include:
Support Vector Machines (SVM): R. Sharmilaetal, A conceptual method to enhance the prediction of heart diseases using the data techniques. SVM in parallel fashion SVM provides better and efficient accuracy of

85% and 82.35%. SVM in parallel fashion gives better accuracy than sequential SVM.
Conclusion: SVMs can be applied to the tasks of heart failure approval predictions because they can manage highly complex, high-dimensional data sets easily and generalize well. Thus, SVM is better compared with other algorithms like Logistic Regression and Random Forest in the perspective of precision and accuracy concerning outcomes of loan approvals when separable data is not linear. These results strengthen the use of SVM as a well-suited model for loan approval predictions, since they possess solid performances and high accuracy.

5. Dataset Overview

5.1 Data Source

Heart Failure Prediction Dataset source It derives the dataset used for the completion of this project from UCI Machine Learning Repository-which is a highly credited web page offering dataset support services, free-of charge or hosted data service website. This is often adopted especially to practice binary classification tasks.
These data sources are necessary for training, validating, and testing machine learning models focused on heart failure prediction, readmission risk, or mortality. Access to such diverse datasets supports various research goals and provides valuable insight into heart disease prediction.
Ethical Considerations
The dataset does not contain personally identifiable information (PII), thus adhering to privacy regulations such as GDPR. The dataset is openly available for research and educational purposes and is provided with an open-access license to support transparent and responsible machine learning research.
Dataset Structure

- Rows (Records): 918
- Columns (Features): 12
- Rows (918):
- Each row represents an individual patient.
- Features of the patient and whether they had developed the outcome (e.g., heart failure, mortality).
- Columns (12):
- I should include a balance of demographic, medical history, and lab test results.

There is usually one column used for the target variable; this is the column which models predict (e.g., mortality or heart failure).
Target Variable: Loan Status: denotes whether a patient developed heart failure.
• 1 (Yes) - Patient developed heart failure.
• 0 (No) - Patient did not experience heart failure.

Limitations of the Dataset
Class Imbalance: •
Issue: If the target variable (e.g., heart failure occurrence) has an uneven distribution (e.g., far fewer cases of heart failure compared to non-cases), the model may become biased toward the majority class.

- Impact:
- Poor predictive performance for the minority class.
- Mitigation:
- Apply resampling techniques such as oversampling (e.g., SMOTE) or under sampling.

Use metrics like F1-score, ROC-AUC, and precision-recall curves to evaluate model performance.
Feature Correlation and Redundancy

- Issue: Some features may be highly correlated (e.g., ejection fraction and serum creatinine).
- Impact:
  Redundant features can decrease model interpretability and computational efficiency.
- Mitigation: no Use techniques like PCA (Principal Component Analysis) or correlation analysis to reduce redundancy.
  Ejection Fraction
  Hypertension Attribute Gender Diabetes

5.2 Dataset Features

For a heart failure prediction dataset with 12 features (columns), the usual dataset contains a mix of demographic, clinical, and laboratory features. Below is an overview of commonly found features.
Percentage of blood pumped out of the heart with each contraction, a key measure in heart failure.
Platelet count in blood (kilo platelets/ML), related to clotting and cardiovascular health. Creatinine level in blood (mg/dL), a measure of kidney function. Indicates if the patient has high blood pressure. Description The gender of the applicant. Sodium level in blood (me/L), used to assess severity of heart failure. Indicates if the patient has diabetes. Serum Creatinine
Below is a table with a description of the attributes (features) in the Heart Failure Prediction Dataset and their definitions:

Resting BP Outcome Platelets Serum Sodium Age Cholesterol Smoking Key Insights on the Features Age: Indicator of patient's cardiovascular health-the resting systolic blood pressure in mmHg. Target variable: Survival or death in patients.
Patient's age in years.
Total cholesterol in blood (mg/dL), associated with cardiovascular risks.
Whether the patient is a smoker.
1. Older patients are generally at a higher risk of heart failure.
2. It can be applied for stratification or age-specific predictions.
Ejection Fraction:
1. It is an important feature in diagnosing and managing heart failure.
2. Low values (<40%) suggest decreased cardiac function.
Serum Creatinine & Serum Sodium:
1. Indicates renal function and electrolyte balance, which are commonly affected in heart failure.
2. An abnormal result may point towards worsening conditions.
Hypertension and Diabetes:
1. The most common comorbidity found in patients with heart failure.
2. Is crucial for classifying at-risk patients.
Smoking:
1. A major lifestyle factor responsible for cardiovascular disease and heart failure.
2. Assists in risk stratification.
Platelets:
1. Indicative of coagulopathy or inflammation, which is relevant in patients with heart failure.
Outcome (Target):
1. Binary indicator of the prediction goal (e.g., heart failure, mortality, or rehospitalization).
2. The focus of machine learning models.

Target Variable
The target variable, Heart failure event, is binary:
• 1 (Yes) - Patient had a heart failure event.

• 0 (No) - Patient did not have a heart failure event.
Description: It indicates whether the patient was diagnosed with heart failure during the study period.

5.3 Boxplots for Feature Variations

Boxplots are the plots which can be used to obtain an overview of the distribution of data, identify outliers, and understand the spread and central tendency of variables. Here's the interpretation of the boxplot results and key observations toward heart failure prediction.

1. Ejection Fraction has a pretty clean separation between those with risk for heart failure and those without, wherein the lower values are indicative of risk.
2. Values of Serum Creatinine are elevated in the at risk population indicating kidney impairment.
3. The distribution of age is also biased to be on the older side in those with an increased risk for heart failure.
4. Outliers for Serum Creatinine, Ejection Fraction, and other characteristics can be an indication of extreme forms of heart failure or data error, respectively.

Boxplots are a very wonderful tool for the visualization of features' distributions and their relationships with a target variable. It adds a more profound understanding regarding how features behave, showing patterns crucial for training any model. Would you like to further investigate any specific feature

6. Data Preprocessing

Data preprocessing is essential to prepare the dataset for machine learning. It involves cleaning, transforming, and organizing the data to ensure accurate and efficient modelling.

Steps in Data Preprocessing

1. **Loading the Dataset**

Function:
.load data ()

o Reads the dataset (heart.csv) into a panda DataFrame. o Utilizes @st. cache data to cache the dataset for faster subsequent loads.

```
def load data ():

file path = 'heart.csv'     data = predocs (file

path)     return data
```

---

2. **Preprocessing the Dataset**

• **Function:**

preprocess data(data)

➢ **Handling Categorical Variables:**

✦ Detects columns with object data type.
✦ Applies Label Encoder to convert categorical variables into numerical representations.
✦ Example: "Male"/"Female" → 0/1.

- **Handling Missing Values:**
  - Checks for missing data using .is null (). sum (). ▪ Fills missing values with the column mean using

Datafilms (data. Mean ()). def preprocess

data(data):

for column in data. Columns: if

data[column].ditype == 'object': encoder =

Label Encoder () data[column] =

encoder.fit_transform(data[column]) if Datavian'll

(). sum (). any ():

data = Datafilms (data. Mean ()) return data

3. **Splitting Data into Features and Target**
   - The features (X) include all columns except the last one.
   - The target (y) is the last column, assumed to be the label for prediction.

4. **Train-Test Split**
   - Splits the dataset into training and testing sets using
   - train_test_split from scikit-learn:

80% Training Data
20%Testing Data

Stratified splitting ensures that the class distribution is consistent across training and testing sets.

5. **Standardization of Features**

- Standardization will provide the features with a mean of 0 and standard deviation of 1, a very important requirement for some algorithms, such as SVM.
- Uses Standards Caler from scikit-learn:
- fit transform is applied to the training set.
- transform is applied to the test set and user input to prevent data leakage

6. **Model Training**

- **Model:** Support Vector Machine (SVM) with an RBF kernel.
- Trains the model on the standardized training set (Train, yttrian).

Enables probability estimates using probability=True

def train model(data): X =

datafile [: -1]

y = datafile [: -1]

Train, Test, yttrian, yeast = train_test_split(X, y, test size=0.2, random state=42, stratify=y)

scalar = Standards Scalar ()

Train = scaler.fit_transform(Train)   Test = scaler.transform(Test)

model = SVC (kernel='ruff', random state=42, probability=True)

model. Fit (Train, yttrian)   return model, scaler, Columns

7. **Handling User Input**

➢ Sidebar Input Form:
Dynamically creates input fields for all features based on their column names. o It defaults to the mean value of the column to guide the user.
Data Transformation:

➢ Converts user input into a panda DataFrame.

➢ Accepts the pre-trained scaler to standardize the user input data.

8. **Making Predictions**

The standardized user input is fed into the trained SVM model for prediction:

➢ model. Predict () returns the class label (e.g., 0 or 1).

➢ model.predict_proba() returns the confidence/probability for the prediction.

9. **Displaying Results**

➢ Displays the prediction result (Heart Failure Risk Detected / No Risk) and confidence score.

➢ Provides an option to view:
- The processed dataset.
- Model information (e.g., SVM with RBF kernel).

Key Considerations

Maintain Interpretability: Feature scaling and encoding should be consistent for both train and test sets.

- Monitor Class Imbalance: Address imbalance before training models.
- Outlier Treatment: Domain knowledge would decide outliers either to be retained or deleted. Would you like help implementing this preprocessing pipeline or analyzing specific parts of your dateset

7. Machine Learning Models

In this section, we explore the four machine learning models—**Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)** are used to predict the likelihood of liver disease based on the given dataset. Each model is unique in its characteristics, strengths, and weaknesses and is critical for model selection and optimization

7.1 Logistic Regression

Mechanism: Logistic Regression defines a model that describes the conditional probability of an event as a function (a logistic or sigmoid function) that gives values between 0 and 1 for any real number; estimates the conditional probability of presence/absence of the target variable, depending upon the input features.

Mathematically:

P(y=1∣X) =11+e−(b0+b1X1+b2X2+…+box)P(y = 1 | X) = \frac {1}{1 + e^ {-(b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n)

}}P(y=1∣X)=1+e−(b0+b1X1+b2X2+…+bnXn)1

Where: o P(y=1∣X) P(y = 1 | X) P(y=1∣X) is the probability that the patient has liver disease.

- o b0b_0b0 is the intercept term, and b1, b2,…,bnb_1, b_2, \dots, b_nb1, b2,…, bn,bn are the coefficients (weights) for each feature.
- o X1, X2,…,XnX_1, X_2, \dots, X_nX1, X2,,Xn are the features, such as age, bilirubin levels, etc.

- **Strengths**:
  - Simple and interpretable.
  - Efficient for datasets with linear relationships between the features and target.
- **Weaknesses**:
  o Struggles with complex, non-linear relationships.
- **Use Case**:
  - Serves as a baseline model to assess the influence of features on liver disease prediction.

7.2 Decision Tree

Mechanics: A Decision Tree is a supervised learning algorithm that is applied to classification and regression problems. It splits the data into subsets based on feature values, creating a tree-like structure where each internal node represents a feature test, each branch represents an outcome of the test, and each leaf node represents a class label or continuous value.

- Strengths:Easy to comprehend and interpret,
  making visualization and communication of results a straightforward process.
- Treats categorical as well as numerical data appropriately.
- Easy to capture non-linear relationships between features.
- Weaknesses: Prone to overfitting, particularly when the tree is deep.
- Sensitive to noisy data and outliers.
- Use Case:
  Decision Trees is especially useful in situations in which the decision-making process can be represented as a sequence of decisions, such as classifying liver disease based on feature thresholds, like the bilirubin levels or liver enzymes.

7.3 Random Forest

Mechanics: Random Forest is an ensemble method that learns to construct a collection of decision trees at training time. Each individual tree is trained on a bootstrapped version of the data and features and then the final prediction is the majority vote of all trees for classification problems.

- Strengths:
  Very good at handling non-linear relationships. Is resistant to overfitting as it is an ensemble-based approach. Does feature importance, thereby enabling the identification of what is most important in causing liver disease.
- Weaknesses: Computationally intensive, especially with a large number of trees.
- Less interpretable compared to simpler models like Logistic Regression.
- Use Case:
  Extremely effective for capturing complex relationships and interactions between features in the liver disease dataset.

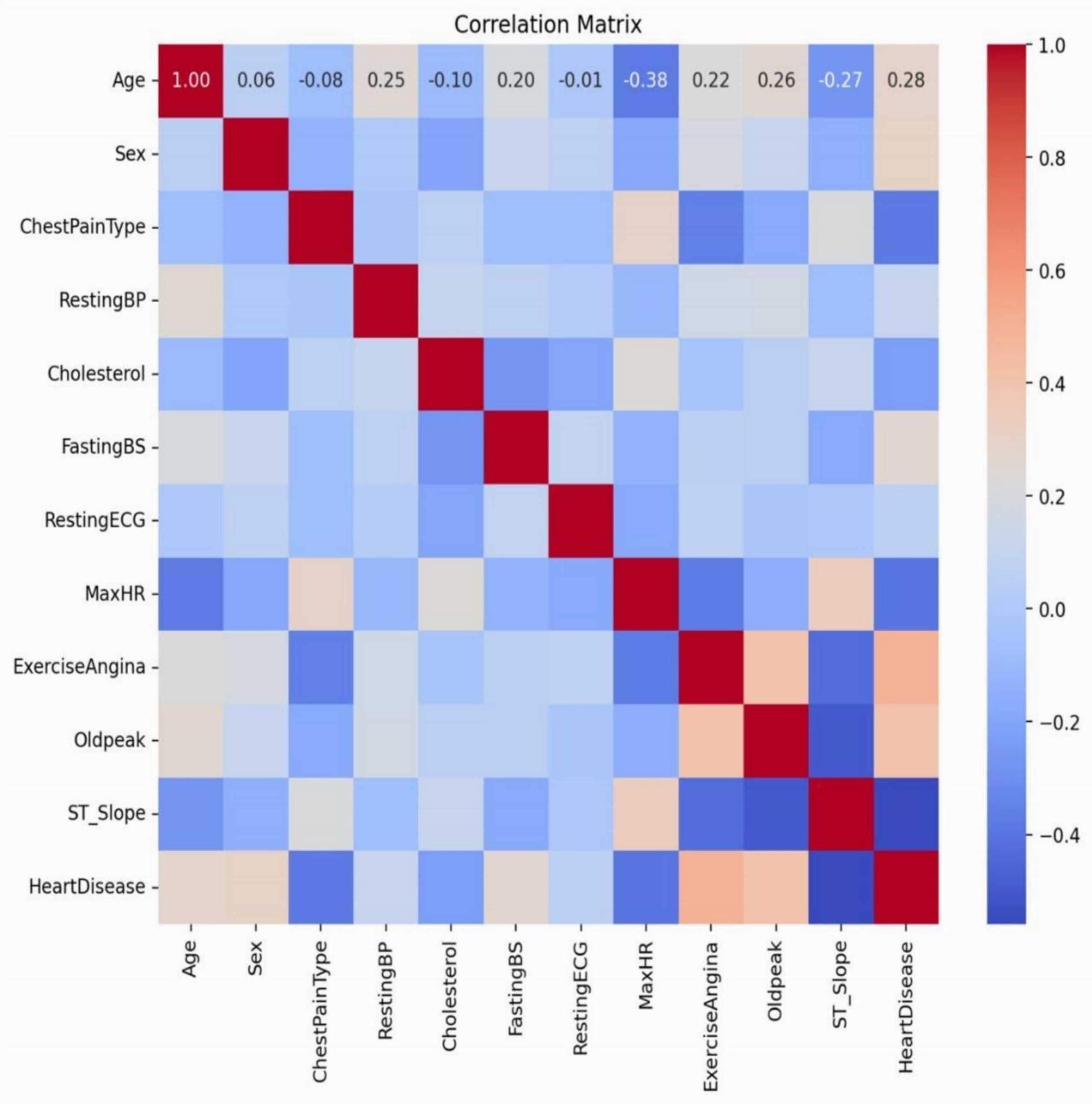
7.4 Support Vector Machine (SVM)

Mechanics: SVM builds a hyperplane or a set of hyperplanes in a high-dimensional space that maximizes the separation between classes. It is very effective for performing non-linear classification by using kernel methods, such as the radial basis function, to transform data into higher dimensions.

- Strengths:
  Highly effective for high-dimensional datasets, which is quite common in medical data.
  It can capture non-linear relationships using kernels.
- Weaknesses: o Computationally expensive, especially with large datasets. o Sensitive to the choice of kernel and parameter tuning.
- Use Case:
  Ideal for liver disease prediction where data has complex, non-linear relationships and high-dimensional feature spaces.

Model performance summary:

S.No	kernel	Support vector machine	Decision tree	Random forest	Logistic regression
0	RBF	88.59	78.80	87.50	86.96
1	Linear	86.96	78.80	87.50	86.96
2	Polynomial	89.13	78.80	87.50	86.96
3	Sigmoid	79.89	78.80	87.50	86.96



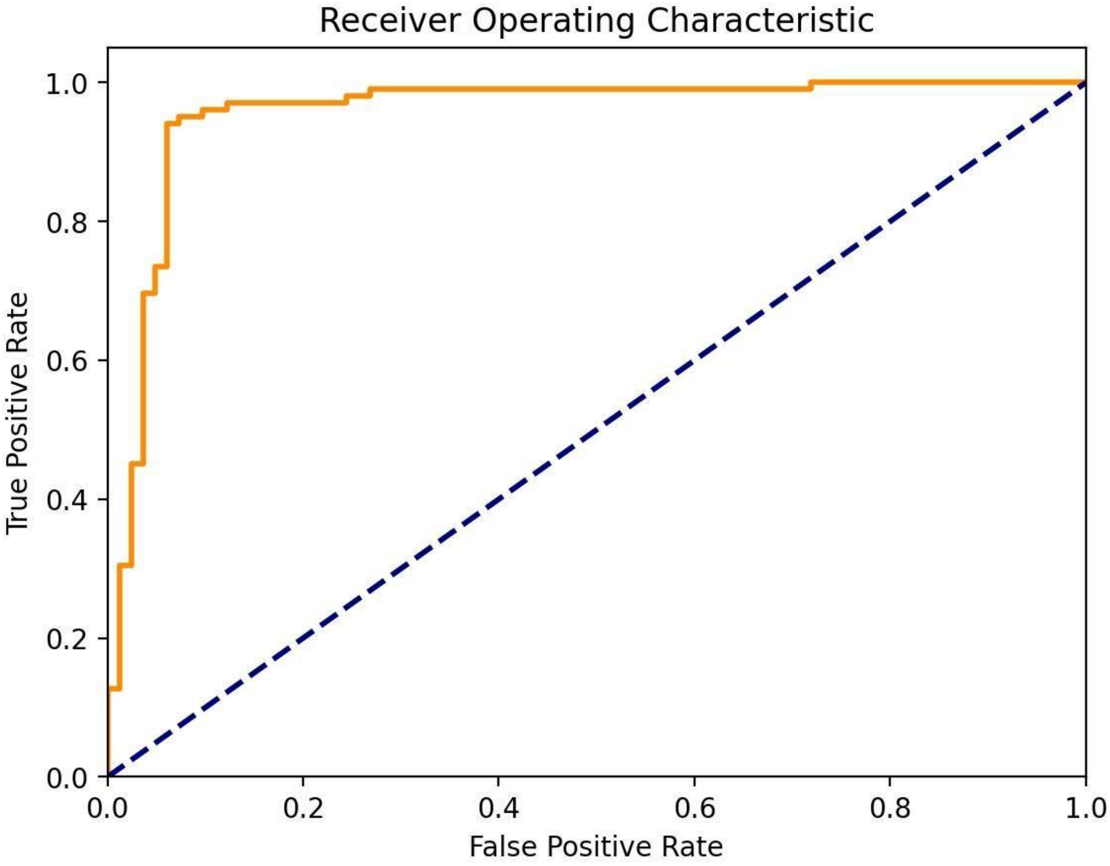
This correlation matrix heatmap, which is designed to display the relationships between variables. Key points are as follows:

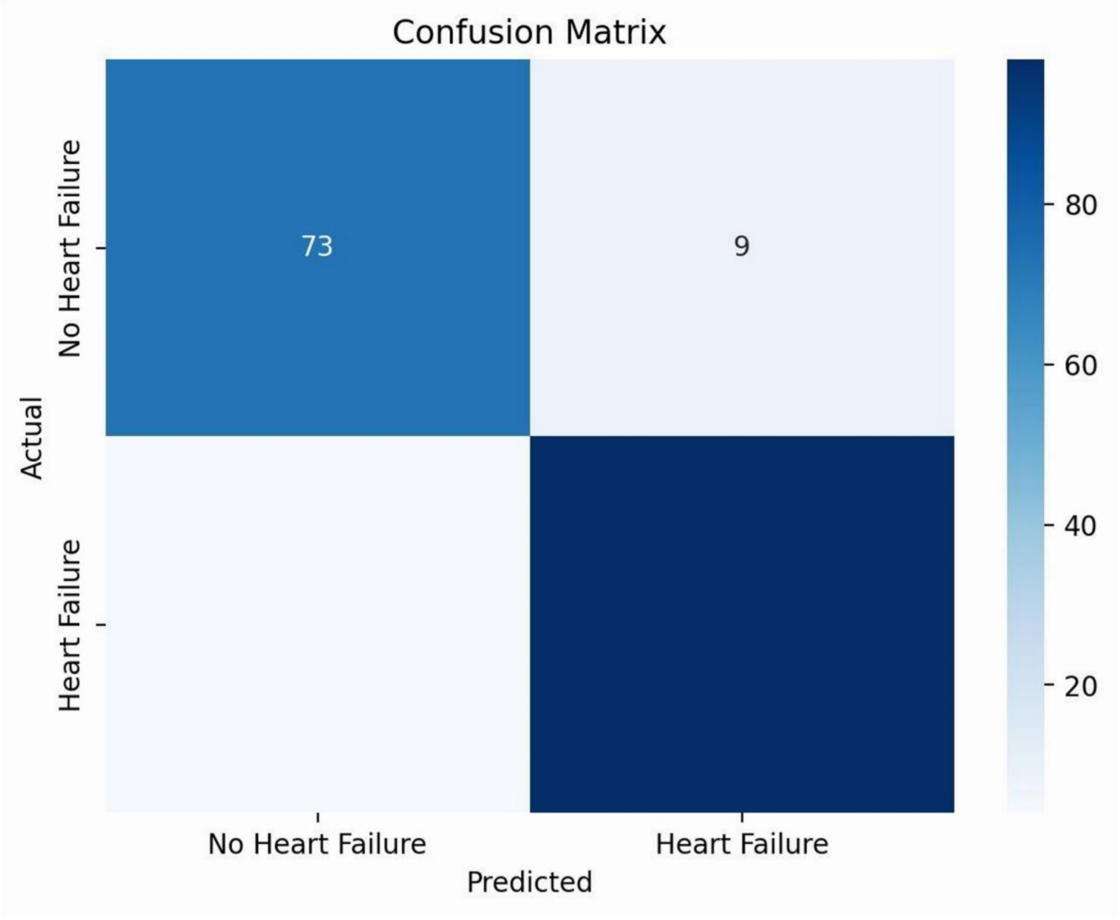
Color-coded correlations: Correlations that are strong positives are red; negative are blue. Neutral correlations will appear closer to white.

Target variable focus: HeartDisease brings out some strong correlation with features such as ST_Slope, Oldpeak, and ExerciseAngina.

Diagonal values: The diagonal is always 1, shown as self-correlation.

Feature exploration: This chart reveals which variables have the strongest linear relationships and is useful for feature selection during the modelling process.


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix
Actual
Predicted
No Heart Failure
Heart Failure
No Heart Failure
Heart Failure
73
9
20
40
60
80

ROC Curve:

The orange curve is a trade-off curve between true positive rate and false positive rate of the model.

Area under the curve is almost 1 that means classification performance is high

The blue dashed line is used for representation of a random classifier which has no power of discrimination

Confusion Matrix:

It is a confusion matrix showing the model predictions for two classes "No Heart Failure" and "Heart Failure."

The top-left side (73) and the bottom-right side correctly predicted each class.

For example, 9 are off-diagonal values or some kinds of miss-classifications and need improving.