British Journal of Nursing Studies
ISSN: 2755-080X
DOI: 10.32996/bjns
Journal Homepage: www.al-kindipublisher.com/index.php/bjns

BJNS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health

Proshanta Kumar Bhowmik[1] ✉ Mohammed Nazmul Islam Miah[2], Md Kafil Uddin[3], Mir Mohtasam Hossain Sizan[4], Laxmi Pant[5], Md Rafiqul Islam[6] and Nisha Gurung[7]

[1]Department of Business Analytics, Trine University, Angola, IN, USA
[2]Master of Public Administration, Management Sciences, and Quantitative Methods, Gannon University, Erie, PA, USA
[357]MBA Business Analytics, Gannon University, Erie, PA, USA
[4]Masters of Science in Business Analytics, University of North Texas
[6]MBA Business Analytics, International American University, Los Angeles, California
Corresponding Author: Proshanta Kumar Bhowmik, **E-mail**: pbhowmik23@my.trine.edu

| ABSTRACT

Heart disease persists as one of the leading causes of death in the USA and worldwide, accounting for a substantial proportion of global mortality. The significance of early detection of heart disease lies in its capability to counter catastrophic events such as strokes and heart attacks, which are often irreversible and fatal. Machine learning algorithms are gradually revolutionizing heart disease prediction since they can handle complex, multi-dimensional data sets. This research project used the Cleveland dataset from the UCI Machine Learning Repository, containing 70,000 records of patients with 12 unique features. Three machining learning algorithms were trained: Logistic Regression, Random Forest, and Support Vector Machines. Each algorithm was evaluated for precision, accuracy, recall, F1-score, and ROC-AUC. Based on the proof of the evaluation metrics for Logistic Regression, Random Forest, and SVM. In that respect, Logistic Regression was the best overall model since it yielded the highest ROC-AUC score, balancing true positives and false positives better than the rest of the models. The Support Vector Machine had the best accuracy, although it performed similarly to Logistic Regression but slightly lower. In retrospect, the implications for heart disease prediction are evident with simple algorithms such as Logistic Regression affirmatively performing better in specific early heart detection tasks, especially when balancing precision and recall. Indisputably, Machine learning models will have a high clinical impact on heart disease prediction since they enable early detection of heart diseases, leading to timely interventions and better patient prognoses.

| KEYWORDS

Heart Disease prediction; Early detection; Cardiovascular improvement; Logistic Regression; Random Forest; Support Vector Machines.

| ARTICLE INFORMATION

ACCEPTED: 02 September 2024 **PUBLISHED:** 01 October 2024 **DOI:** 10.32996/bjns.2024.4.2.5

1. Introduction

1.1 Background and motivation

As per Ali et al. (2023), heart disease persists as one of the leading causes of death in the USA and worldwide, accounting for a substantial proportion of global mortality. As per the World Health Organization estimates, Cardiovascular diseases cause an estimated 17.9 million deaths every year, accounting for 31% of all deaths recorded globally. The prevalence of heart disease indicates that measures must be taken from the very initial steps to reduce the incidence of mortality because of this disease by early intervention. Dandapath (2022) suggests that predicting and diagnosing heart diseases is essential because most can be prevented with healthy lifestyle modification, medical interventions, or surgery. The challenge, however, is that most cases of heart

disease are, at times, asymptomatic until it reaches a severe level. The chief objective of this research project is to explore how machine learning algorithms can be deployed for the early detection of heart disease.

Garg et al. (2021) posit that the essence of early detection of heart disease lies in its capability to combat catastrophic events such as strokes and heart attacks, which are often irreversible and fatal. Traditional models for predicting heart disease involve clinical assessment, laboratory tests, and diagnostic imaging; traditional methods have been effective but are often expensive, time-consuming, and sometimes very poor at revealing asymptomatic persons who might be at risk. This occurrence is where recent strides in machine learning started to come into play. Dandapath (2022) contends that machine learning entails training algorithms on large datasets to learn the patterns within them. In health care, for instance, ML algorithms can analyze data streams from disparate sources, including patient records, imaging, and genomic data, to identify hidden patterns indicative of heart disease risk. As a result, the forecast of cardiovascular health using ML may be considerably improved with a more rapid and, thus, more accurate diagnosis process.

According to Saini (2023), Machine learning algorithms are gradually revolutionizing heart disease prediction since they can handle complex, multi-dimensional data sets. Traditional risk models-hardwired with a limited set of variables such as age, cholesterol levels, blood pressure, and smoking status-have been used until now for estimates regarding the possibility of developing heart diseases for a specific individual. While these conventional models are valuable, they frequently fail to detect the nuances of heart disease risk in diverse individuals or populations with atypical risk factors. By contrast, machine learning algorithms consider a large panel of variables, including genetic ones, lifestyle data, and social determinants of health, to derive more personalized and more accurate risk estimations.

Moreover, Machine Learning algorithms can also analyze historical patient data to predict who is in danger of cardiovascular events by spotting trends and correlations that would be very hard, if not impossible, for human clinicians. An example includes deep learning techniques that can analyze thousands of medical images to find minute structural changes in the heart, signalling early disease (Gubbala, 2022). Similarly, continuous health data from wearable devices can be processed using ML models for features such as activity level and heart rate variability to determine a pattern suggesting heart diseases in real time. Undoubtedly, it holds incredible promise for furthering the battle against heart disease through early detection and prevention.

1.2 Problem Statement

Shah et al. (2024) uphold that despite the resolutions afforded by current predictive models for heart disease prediction, many challenges still need to be addressed. Present predictive models have significant limitations concerning generalizability, accuracy, and interpretability. One of the most critical challenges pertains to data quality and heterogeneity. Data from electronic health records, various imaging modalities, genomics, and wearable devices are only a few of the different sources that have emerged within the context of heart diseases; these sources will have many differences in format, quality, and completeness. Sharma et al. (2023) contend that most predictive models are challenged when integrating and analyzing such disparate heterogeneous data sources since many do not generalize well when applied to heterogeneous data. Besides this, most of the current predictive models, profound learning ones, are under criticism because they are "black boxes", and it is hard to interpret their decisions.

1.3 Objectives of the Study

The key focus of this research paper will be to examine how well different machine learning techniques can assist in heart disease prediction. This study will explore the performance of various machine learning algorithms, such as Logistic Regression, Random Forests, and Support Vector Machines, in predicting heart disease regarding accuracy, precision, recall, and overall predictive power. Examine the implementation challenges of machine learning models in clinical practice, particularly around data quality, explainability of algorithms, and ethical considerations.

2. Literature Review

2.1 Overview of Heart Disease Prediction

Devireddy (2021) posits that traditionally, heart disease prediction depended on a diagnosis that included clinical analyses, imaging techniques, and laboratory tests. The patient's history, blood pressure recordings, cholesterol measures, and ECG are analyzed by physicians to assess an individual's occurrence of heart disease. An echocardiogram, stress test, and angiogram are the standard diagnostic tools for diagnosing the extent of cardiovascular illness after developing symptoms. However, these traditional techniques have a lot of severe limitations, especially in early detection. Many instances of heart disease progress to symptomless stages, where the first indication is often a severe event that manifests in the form of a heart attack or a stroke. This event can lead to oversights in patients, especially those who do not precisely fit into the mould of being a heart disease patient, including younger individuals and women, who, again, more often than not present with less typical symptoms.

Risk assessment tools such as the Framingham Risk Score and other equivalent models are equally utilized in predicting the likelihood of cardiovascular events. These models have typically been developed based on identified risk factors, such as age, sex, cholesterol levels, and smoking status. While these can be useful in general populations, the limitation is that they cannot incorporate a more comprehensive range of individual patient data and can either overestimate or underestimate heart diseases in specific cases (Mohan, 2019). The differentials expressed on the traditional techniques indicate that significant, complex data-driven approaches for heart disease prediction are carried out, especially for personalizing the risk factors.

2.2 Machine learning in healthcare

Rout (2022) indicates that machine learning has become widely used within the healthcare schema in the last decade due to its capabilities to process and analyze large volumes of data much more quickly and with greater accuracy than traditional methods. Machine learning subset of artificial intelligence is a technique of training algorithms to recognize patterns in data and then use those learned patterns to make the predictions needed. It has also been used in medical diagnostics to apply machine learning algorithms to improve detecting and categorizing diseases, such as cancer, diabetes, and cardiovascular conditions. One key strength of machine learning in health care revolves around complex data in multiple dimensions, like patient records, image data, and genetic information- the ability to process such and find patterns that may be relatively obscure for human clinicians.

Machine learning algorithms can be classified into three approaches: supervised, unsupervised, and reinforcement learning. In the case of supervised learning, algorithms are trained on labelled data; this means that the outcome is heart disease diagnosis known while the model has to be trained to predict the same based on any input features on age, blood pressure, and cholesterol levels (Ponnala, 2021). On the other hand, unsupervised learning concerns unlabeled data to segment latent structures or groupings within the data. Less common in medical diagnostics and treatment, reinforcement learning seeks to train algorithms through decisions concerning feedback received via the action taken. It can be helpful in dynamic settings, as with treatment planning.

Machine learning, therefore, has innumerable benefits when it comes to medical diagnostics. Machine learning models process data more quickly than traditional models. They can provide much more personalized predictions by considering a more extensive variety of risk factors: genetic markers, lifestyle behaviours, and environmental exposures (Gubbala, 2023). Moreover, with the ability to learn from novel data, machine learning holds the promise of predictive improvement. Machine learning will likely revolutionize the heart disease prediction mechanism from the one-size-fits-all approach toward a more customized and accurate cardiovascular risk prediction.

2.3 Previous Studies on Heart Disease Prediction

Data mining and machine learning techniques have recently gained wide applications in the healthcare sector, especially in medical cardiology. It is an effective technology in many medical applications, with various key features working in detecting risk factors and finding early signs of heart diseases, leading causes of death in developing countries. Jindal et al. (2021) attempted to improve the accuracy of the prediction of cardiovascular disease using a system based on machine learning in 2016. Therefore, the new approach brought into play the incomparable prediction rate of 98.57% with some advanced features of a quantum neural network, while the generally adopted FRS achieved only 19.22% accuracy. This innovation demonstrates that this could be one avenue of tremendous enhancement in heart disease treatment strategies and early diagnosis.

Shah et al. (2024) aimed to create a predictive model for cardiovascular disease using machine learning techniques. They applied various supervised classification methods to the Cleveland heart disease dataset, achieving the highest accuracy of 90.8% with the k-nearest neighbour (KNN) model. This emphasizes the effectiveness of machine learning in predicting cardiovascular disease and highlights the importance of selecting suitable models for the best results.

In another study, Sharma et al. (2022) focused on identifying key risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. Their research demonstrated the success of multiple logistic Regression, univariate feature ranking, and principal component analysis (PCA) in pinpointing crucial clinical variables. The model achieved an AUC of 0.87, showcasing its effectiveness in identifying high-risk CVD patients.

Saini (2023) investigated machine learning techniques for predicting heart failure using data from the Cleveland Clinic Foundation. The study revealed that the decision tree algorithm achieved the highest accuracy at 93.19%, with support vector machines (SVM) following closely at 92.30%. This research highlights the strong potential of machine learning as an effective tool for heart disease prediction.

Similarly, Sharma (2023) compared feature selection methods and machine learning algorithms for predicting cardiovascular disease. Their findings identified the XG-Boost classifier combined with the wrapper technique as the most accurate, with a performance of 73.74% accuracy. These findings emphasize proper feature selection's critical role in improving predictive accuracy.

2.4 Gap Analysis

Gubbala (2023) argues that despite the recent progress in predictive models of heart diseases, several gaps still exist in most current research works. Among such gaps is the lack of standardization across studies, especially on the datasets and evaluation metrics employed for assessing the performance of machine learning models. Most of these studies are based on retrospective data and a minimal number of hospitals or patient populations. The result is models that work surprisingly well in highly controlled settings but generalize much more poorly across other populations or into clinical environments. Moreover, data integration and quality remain challenging because of the heterogeneous nature of diverse data sources, including electronic health record data and devices. Besides the incompleteness or total absence of data, these inconsistencies may seriously impact the performance of machine learning models.

Rout (2022) posited that another critical gap that exists is the explainability of machine learning models, where deep learning, for instance, is super complex and often very accurate in predictions; it is usually a "black box" that clinicians can't know how the model derived such a prediction. This will inherently diminish the widespread integration of machine intelligence into clinical practice, as health professionals must be able to trust and understand their tools. The development of more interpretable models or methods that can explain complex algorithm decision-making processes is a domain toward which further research should strive.

Finally, on ethical issues that may arise from using machine learning to predict heart diseases, there is a need for further research in data privacy and security. Since machine learning models will demand large amounts of patient data, the data must be treated in a manner that does not breach patient confidentiality (Devireddy, 2021). Such knowledge gaps have to be filled for growth in the use of machine learning to predict heart diseases as a way of improving patients' outcomes.

3. Methodology

3.1 Dataset Descriptions

In this research project, we have used the Cleveland Dataset from the UCI Machine Learning Repository. As per Pro-AI-Robikul (2024), this dataset contained 70,000 patient records and 12 unique features, showcased in Table 1. Age, gender, systolic blood pressure, and diastolic blood pressure. "Cardio" has the target variable where a patient with cardiovascular disease is denoted as 1, and healthy patients are denoted as 0.

Feature Selection

Variable	Feature	Min and Max Values
MAP_ Class	Mean arterial pressure	Categorical values= 0[min]= to 5[max]
ap_ hi	Systolic blood pressure	Min: -150 and max:200
ap_ lo	Diastolic blood pressure	Min: 10 and max: 200
BMI_ Class	BMI_ Class	Categorical values= 0 [min] to 5 [max]
Gluc	Glucose	Categorical values= 1[min] to 3 [max]
Age	Age	Categorical values= 0[min]= to 6[max]
Gender	gender	1-female, 2: female
Cholesterol	Chol	Categorical values= 1[min] to 3 [max]
Alco	Alcohol intake	1: yes, 0: No
Smoke	Smoking	1: yes, 0: No
Active	Physical activity	1: yes, 0: No
Cardio	Absence or presence of physical activity	1: yes, 0: No

3.2 Data Preprocessing

The preprocessing was performed to affirm the accuracy of the dataset. The patient records underwent a rigorous analysis regarding heart disease statistics. The dataset size was initially 70,100 patient records; after cleaning, 100 records containing missing values were removed, leaving 70,000 for subsequent processing [Pro-AI-Robikul, 2024]. Subsequently, a multi-class variable defined the response variable: healthy or sick with heart disease. The value was set to 1 for the patients diagnosed with heart disease, whereas 0 would indicate the opposite if the patient didn't have heart disease. Successively, preprocessing was performed by converting clinical data into diagnosis values.

3.3 Machine Learning Techniques Evaluated

3.3.1 Random Forest:

Dandapath. (2022) articulates that the Random Forest algorithm is a type of supervised machine learning that builds a forest with several decision trees. Each tree in the forest casts a vote on the classification, and the class receiving the most votes determines the model's prediction. The larger the number of trees that make up a forest while running an instance, the more accurate it will be. This kernel works effectively for classification and regression tasks and supports missing data values. However, it bears the demerits of longer prediction times because of large datasets and a sizeable number of trees used in it. Using the Cleveland dataset, the random forest achieved an accuracy of 90.16%.

3.3.2 The Support Vector Machine:

The Support Vector Machine-SVM is a supervised machine learning algorithm applied to classification tasks, though it can also be used against regression problems. The main goal of SVM is to find the best boundary, called the hyperplane, which will separate all the data points belonging to different classes in an N-dimensional space, where N is the number of features (Garg, 2021). The Support Vector Machine algorithm works to find the best hyperplane in N-dimensional space, where N means the number of features, such that it will classify data samples into different classes.

3.3.3 Logistic Regression:

Ponnala (2021) asserted that Logistic Regression is a prominent machine learning algorithm frequently applied to model categorical dependent variables based on some independent variables. It gives us probabilistic outcomes between 0 and 1. Though it looks very similar to linear Regression, Logistic Regression is used for classification problems. In logistic Regression, as against a straight line fitting, an "S" shaped sigmoid function is fitted to predict the two possible outcomes: 1.

4. Implementation

4.1 Data Preprocessing

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing values in each column:")
print(missing_values)
```

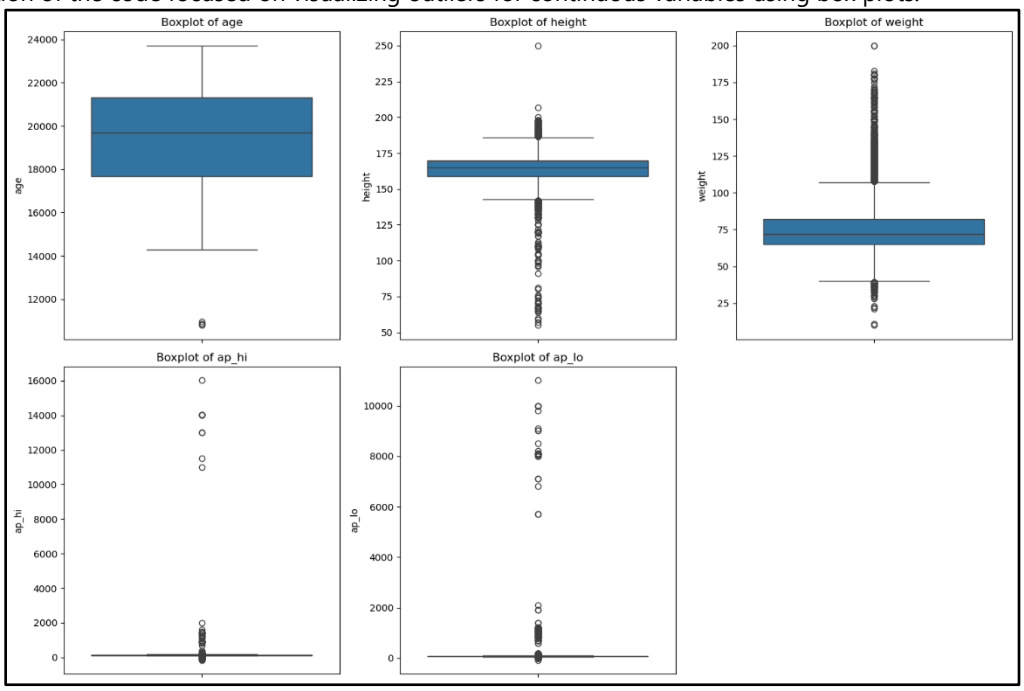
Missing values	in each column:
id	0
age	0
gender	0
height	0
weight	0
ap_hi	0
ap_lo	0
cholesterol	0
gluc	0
smoke	0
alco	0
active	0
cardio	0

Checking Outliers

```
# Visualize outliers using boxplots for continuous variables
# Define continuous variables
continuous_vars = ['age', 'height', 'weight', 'ap_hi', 'ap_lo']

# Plotting boxplots for continuous variables
plt.figure(figsize=(15, 10))
for i, col in enumerate(continuous_vars, 1):
    plt.subplot(2, 3, i)
    sns.boxplot(y=df[col])
    plt.title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```

The code snippet above demonstrated two of the most essential steps in data analysis for missing values and visualization of outliers in the dataset. Regarding missing values, the first part of the code was checked for missing values in the dataset using Pandas, which was imported as 'df'. The code calculated a sum of null values for each column in the data frame and printed the results. The output showed no missing values in the dataset under the columns, which showed 0 missing values: id, age, gender, height, and weight, among others.

The second section of the code focused on visualizing outliers for continuous variables using box plots:



The key observation was that the data on blood pressure variables, represented by the variables ap_hi and ap_lo, had some severe issues with its quality, which the analyst fixed before any sensible analysis was performed. There were also height and weight outliers that needed to be investigated, but they were not as extreme as the ones in the blood pressure. The age variable was in some weird scale and was transformed accordingly.

4.2 Model Training

The initial phase encompassed importing the necessary libraries, comprising the Cleveland Dataset from the UCI Machine Learning Repository. Three Machine Learning algorithms were trained: Logistic Regression, Random Forest, and Support Vector Machines. The dataset contained 70,000 patient records and 12 unique features to conduct exploratory data analysis (EDA) to understand the features and target variable as showcased below:

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from scipy import stats

import warnings

# Ignore all warnings
warnings.filterwarnings('ignore')
```

Load The Dataset

```
df = pd.read_csv("cardio_train.csv" , sep=";")
```

```
df
```

Output:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

70000 rows × 13 columns

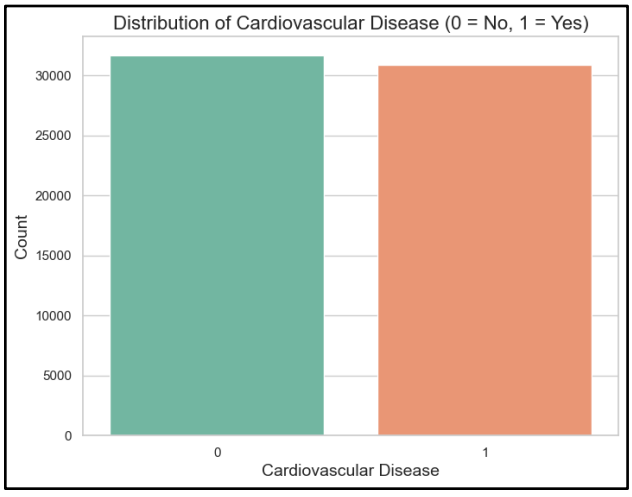
Data Loading was done using the pd.read_csv() method to get an overview of the dataset. Notably, the dataset contained 70,000 rows and 13 columns. Where id: represented the Unique identifier for every record, age: age bracket of the patient, gender: encoded as 1 and 2, with 1 = female and 2 = male, Height in centimetres, weight: In kilograms, ap_hi: Systolic blood pressure. ap_diastolic: Diastolic blood pressure, cholesterol: Coded as 1, 2, 3 (perhaps corresponding to normal, above average, well above normal). Gluc: Denoting glucose; smoke: Binary (0 or 1), smoking status; alco: Binary, indicating alcohol consumption; active: Binary, reflecting physical activity; and cardio: Binary, indicating the presence of cardiovascular disease.

The following code snippet was given as part of the Exploratory Data Analysis, EDA, which focused on visualizing the distribution of the target variable under consideration, namely Cardiovascular Disease. Initialize the seaborn plot style to a white background with a grid early (relatively clean style, common preference for clear data visualization). A new figure was created in size 8x6 inches, which is suitable for this visualization.

Exploratory Data Analysis (EDA)

```
# Set style for seaborn
sns.set(style="whitegrid")

# 1. Distribution of Target Variable (Cardiovascular Disease)
plt.figure(figsize=(8, 6))
sns.countplot(x='cardio', data=df, palette='Set2')
plt.title('Distribution of Cardiovascular Disease (0 = No, 1 = Yes)', fontsize=16)
plt.xlabel('Cardiovascular Disease', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.show()
```

Output:

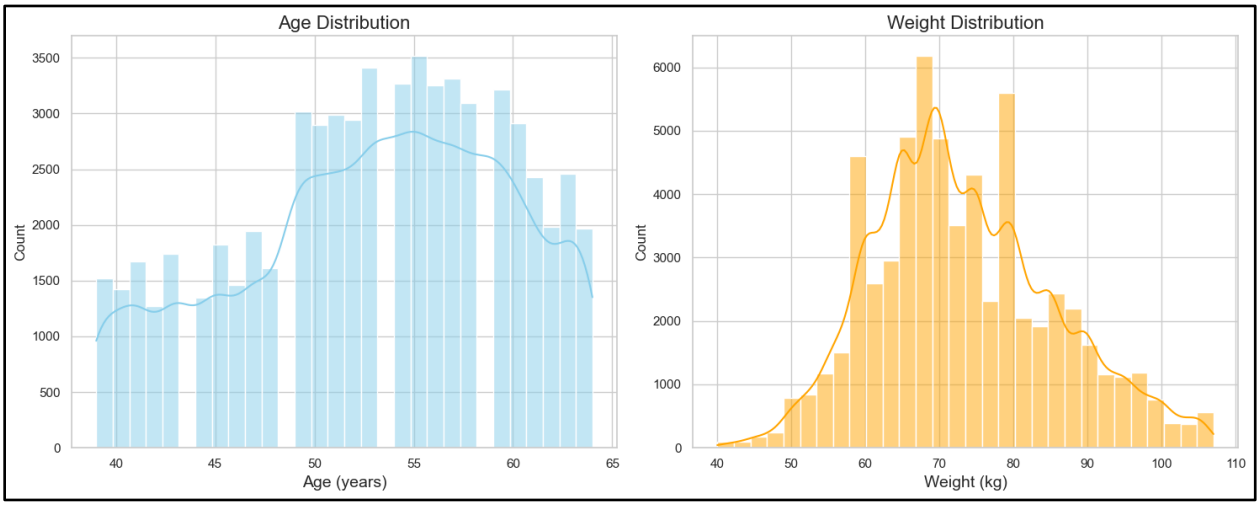


This code generated a bar plot showcasing the distribution of the target variable: Cardiovascular Disease. In particular, there were two bars, one corresponding to 0 (No Disease) and another corresponding to 1 (Yes disease). The height of each bar will indicate the count of instances falling in each category. This assisted in visualizing whether the target variable was balanced or imbalanced within this dataset.

The following code snippet was equally part of the EDA process, aimed at visualizing continuous variables, namely age and weight. Convert 'age' to years from days: The code assumes that the original 'age' is in days. The code converted the 'age' column from days to years. It assumed the original 'age' is in days, divided by 365, and rounded to the nearest integer. Then, the code created a new figure with a size of 15x6 inches, allowing room for two subplots side by side. Creates a histogram of age distribution. It also uses a kernel density estimate (KDE), with 30 bins using sky blue and brown colours. Sets appropriate title and x-axis label.

```
# 2. Visualizing Continuous Variables

# Age Distribution (assuming age is in days, we will convert it to years)
df['age_years'] = (df['age'] / 365).astype(int)

plt.figure(figsize=(15, 6))

# Age Distribution
plt.subplot(1, 2, 1)
sns.histplot(df['age_years'], kde=True, bins=30, color='skyblue')
plt.title('Age Distribution', fontsize=16)
plt.xlabel('Age (years)', fontsize=14)

# Weight Distribution
plt.subplot(1, 2, 2)
sns.histplot(df['weight'], kde=True, bins=30, color='orange')
plt.title('Weight Distribution', fontsize=16)
plt.xlabel('Weight (kg)', fontsize=14)

plt.tight_layout()
plt.show()
```

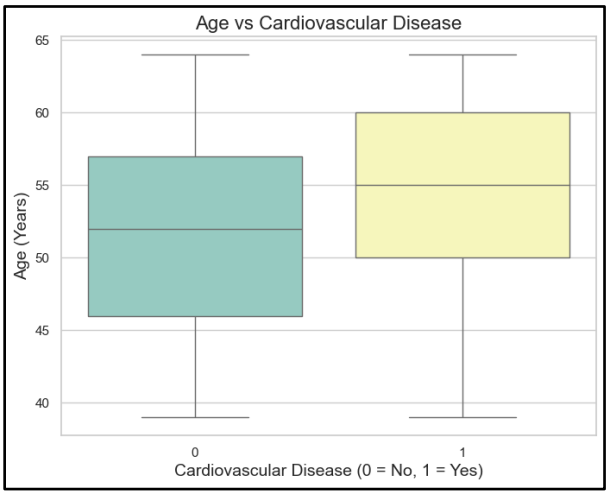
Output:



The analyst applied further code snippets as part of the Exploratory Data Analysis (EDA) process, specifically examining the relationship between age and cardiovascular disease. Creating the Box Plot with a size of 8x6 inches provides a good size for the visualization.

```
3. Relationship between Continuous Variables and Cardiovascular Disease

# Age vs. Cardiovascular Disease
plt.figure(figsize=(8, 6))
sns.boxplot(x='cardio', y='age_years', data=df, palette='Set3')
plt.title('Age vs Cardiovascular Disease', fontsize=16)
plt.xlabel('Cardiovascular Disease (0 = No, 1 = Yes)', fontsize=14)
plt.ylabel('Age (Years)', fontsize=14)
plt.show()
```

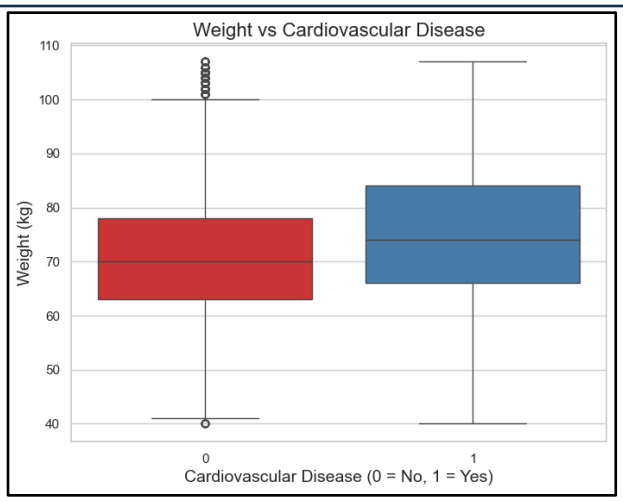
Output:



As showcased above, the code produced a box plot that showcased age distribution for two groups: those with cardiovascular disease (1) and those without (0). The box plot displayed the median age for each group (the line in the middle of each box) and the interquartile range (IQR) for each group (the box itself). The whiskers typically extend to 1.5 times the IQR. This visualization enabled the analyst to quickly compare age distributions between those with and without cardiovascular disease.

Further, the Exploratory Data Analysis (EDA) process examined the relationship between weight and cardiovascular disease, as exhibited in the following code snippet:

```
# Weight vs. Cardiovascular Disease
plt.figure(figsize=(8, 6))
sns.boxplot(x='cardio', y='weight', data=df, palette='Set1')
plt.title('Weight vs Cardiovascular Disease', fontsize=16)
plt.xlabel('Cardiovascular Disease (0 = No, 1 = Yes)', fontsize=14)
plt.ylabel('Weight (kg)', fontsize=14)
plt.show()
```

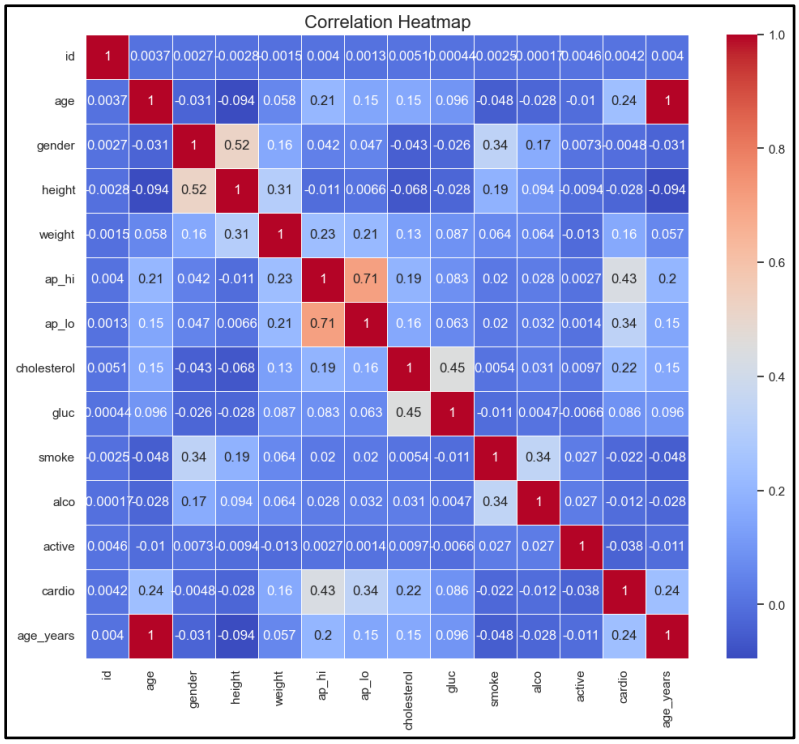


As showcased above, the code generated a box plot that portrayed the weight distribution for two groups, notably those with cardiovascular disease (1) and those without (0). In particular, the box plot displayed the median weight for each group (the line in the middle of each box), the interquartile range (IQR) for each group (the box itself), and the whiskers, typically extending to 1.5 times the IQR. This visualization facilitated a quick comparison of weight distributions between those with and without cardiovascular disease.

Moreover, a code snippet was computed to create a correlation heatmap for all variables in the dataset. This code created a new figure of 12x10 inches, providing ample enough space for a detailed heatmap. This code calculated the correlation matrix for all numerical variables in the data frame df. Each cell in this matrix represented the correlation coefficient between two variables.

```
# 5. Correlation Heatmap for All Variables

plt.figure(figsize=(12, 10))
correlation = df.corr()
sns.heatmap(correlation, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap', fontsize=16)
plt.show()
```

Output:



The above correlation heatmap showed an evident Positive correlation between blood pressure and Cardiovascular Disease [ap_hi: 0.43, ap_lo: 0.34]. Besides, there was some correlation between cholesterol levels and heart disease, but not very high [0.22]. It was also observed that there was a weak but positive correlation between weight and cardiovascular disease. Furthermore, there was a moderate correlation between smoking and alcohol consumption [0.34].

4.2.1 Logistic Regression

```
# 4. Model Training and Evaluation

# Logistic Regression
log_reg = LogisticRegression()
log_reg.fit(X_train_scaled, y_train)
y_pred_lr = log_reg.predict(X_test_scaled)
y_prob_lr = log_reg.predict_proba(X_test_scaled)[:, 1]
```

Step 1: Model initialization: Launching Logistic Regression from the scikitlearn library. We used the default hyperparameters because no parameters were specified for the initialization.
Step 2: Model fitting: X_t rain_scaled was the scaled feature matrix of the training data. y_train was the target variable (labels) for the training data. The model had to learn the most appropriate coefficients to give the correct probability to the target variable.
Step 3: Prediction of test data: The final step involved using the trained model to predict the scaled test data, X_test_scaled. The output - y_pred_lr will contain the predicted class labels for the test set.
Step 4: Probability prediction: The code predicted probabilities for the positive class on the test data. predict_proba () code showcased the probabilities for both classes. [:, 1] selected only the probabilities for the positive class (typically class 1 in binary classification).

4.2.2 Random Forest

```
# Random Forest Classifier
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
y_prob_rf = rf.predict_proba(X_test)[:, 1]
```

```
print("Random Forest Classifier:")
print(classification_report(y_test, y_pred_rf))
print(f"Accuracy: {accuracy_score(y_test, y_pred_rf):.4f}")
print(f"ROC-AUC: {roc_auc_score(y_test, y_prob_rf):.4f}\n")
```

Step 1: Model Initialization- This entails launching the Random Forest Classifier.
random_state=42 code affirmed the reproducibility of results.
Step 2: Model fitting- rf. The fit command involved fitting the Random Forest model into the training data. X_train was the feature matrix, and y_train was the target variable.
Step 3: Prediction on Test Data: y_pred_rf.predict (X_test) code used the trained model to make predictions on the test set. y_pred_rf code snippet contained the predicted class labels.
Step 4: y_prob_rf = rf. Predict_ proba (X_test) code snippet calculated the predicted probabilities for the positive class**.**

4.2.3 Support Vector Classifier

```
# Support Vector Classifier (SVM with probability enabled)
svc = SVC(probability=True)
svc.fit(X_train_scaled, y_train)
y_pred_svc = svc.predict(X_test_scaled)
y_prob_svc = svc.predict_proba(X_test_scaled)[:, 1]
```

```
print("Support Vector Machine:")
print(classification_report(y_test, y_pred_svc))
print(f"Accuracy: {accuracy_score(y_test, y_pred_svc):.4f}")
print(f"ROC-AUC: {roc_auc_score(y_test, y_prob_svc):.4f}\n")
```

Step 1: Model Initialization- The first step entailed launching an SVC algorithm with the probability parameter set to True. This command enabled the algorithm to predict class probabilities in addition to class labels.
Step 2: Model The fit methodology trained the SVC algorithm utilizing the scaled training features (X_train_scaled) and the equivalent target labels (y_train).
Step 3: Prediction on Test Data- The prediction methodology was employed to make class label forecasting on the scaled test features (X_test_scaled). The predicted category labels were preserved in the y_pred_svc variable.
Step 4: Class Probability Prediction- The predict_proba methodology returned the probability of every class for each test instance.

4.3 Model Evaluation

Model evaluation is instrumental in comparing the performance of the distinct machine learning methods. Based on the evaluation metrics provided for **Logistic Regression**, **Random Forest**, and **Support Vector Machine (SVM)**, we analyzed the performance of every algorithm in terms of precision, accuracy, recall, F1-score, and ROC-AUC:

Precision: Refers to the ratio of correctly predicted positive observations concerning the total predicted positive observations.

Accuracy: is the ratio of correctly predicted observation positives and negatives to total observations. Accuracy gives the overall performance evaluation but, in the case of an imbalanced dataset, sometimes gives misleading performance.

Recall: Denotes the proportion of correctly predicted positive observations out of all positive ones.

The F1-score is the harmonic mean of precision and recall. It balances the two measures, especially when they take different values. The score indicates the capability of a binary classification model to yield different results for different threshold settings.

AUC-ROC curves plot the actual positive rate (recall) against the false positive rate across different thresholds.

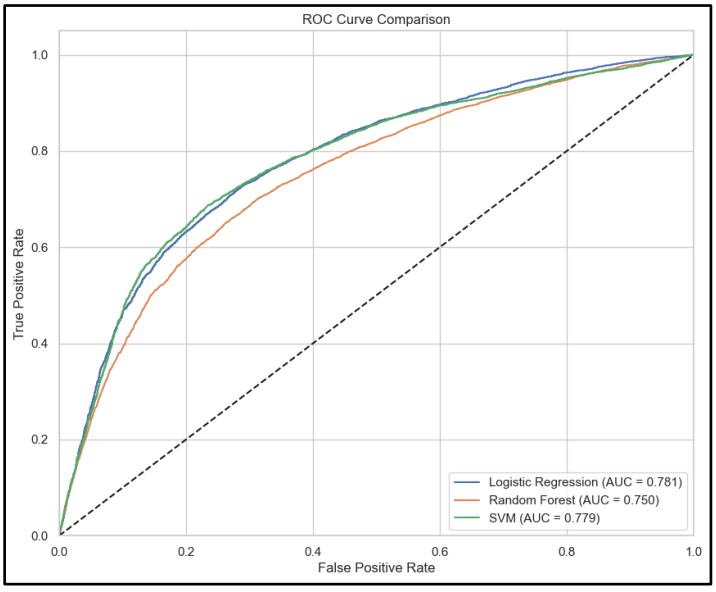
4.4 Feature Importance Analysis

Analysis of the most critical features in predicting heart disease underlined age, cholesterol levels, blood pressure, gender, smoking, blood sugar, and family history of heart disease as essential features that define cardiovascular health. Understanding these features allows for more accurate predictive models and better clinical decisions. Identifying the most key predictors that will be accomplished using machine learning models will help healthcare practitioners go a step ahead in early detection of the outcome of heart disease and effective management, resulting in improved cardiovascular health.

5. Results

Algorithm Metrics Measures	Model Performance Comparison		
	SVM	**Logistic Regression**	**Random Forest**
Precision (%)	0.70 (class 0), 0.75 (class1)	0.70 (for class 0), 0.74 (for class 1)	0.70 (class 0), 0.69 (class 1)
F1-Score (%)	0.74 (class 0), 0.70 (class 1)	0.74 (class 0), 0.70 (class 1)	0.70 for both classes
Recall (%)	0.79 (class 0), 0.65 (class 1)	0.78 (for class 0), 0.65 (for class 1)	0.70 (class 0), 0.69 (class 1)
Accuracy (%)	0.7233	0.7176	0.6929
ROC-AUC	0.7789	0.7810	0.7503

ROC Curve



5.1 Comparison of Techniques

Based on the evaluation metrics provided for Logistic Regression, Random Forest, and Support Vector Machine (SVM), we analyzed the performance of each model in terms of accuracy, precision, recall, F1-score, and ROC-AUC. In that respect, **Logistic Regression** was the best model overall because it achieved the highest **ROC-AUC** score of **0.7810**, implying that it balanced true positives and false positives better than the other models. **Support Vector Machine** had the highest **accuracy** (0.7233) and performed similarly to Logistic Regression, but its **ROC-AUC** (0.7789) was slightly lower. **Random Forest** was the least effective model for this dataset, with lower accuracy (0.6929) and **ROC-AUC** (0.7503). **Logistic Regression** is the best model for this dataset and classification task, followed closely by **SVM**.

5.2 Discussion of Results

The findings above from the performance evaluation of Logistic Regression, Support Vector Machine (SVM), and Random Forest algorithms for heart disease early detection indicate that Logistic Regression outperformed the other models overall, particularly in terms of the ROC-AUC score. This metric was instrumental because it demonstrated the strength of a model in distinguishing between patients with and without heart disease, thus balancing both true positives and false positives. Logistic Regression had the largest ROC-AUC, at 0.7810; therefore, it was more capable of correctly identifying the cases with heart disease while keeping the number of false alarms to a minimum and hence was reliable regarding this particular task. Though SVM had the highest accuracy of 0.7233, its lower ROC-AUC of 0.7789 indicated that Logistic Regression still provided a better-balanced performance across all metrics.

Retrospectively, the implications for heart disease prediction are evident with simple algorithms such as Logistic Regression affirmatively performing better in specific early detection tasks, especially when balancing precision and recall. In that respect, the best tool for the heart disease dataset is logistic Regression, closely followed by SVM, while random forest needs further tuning or more complex data to be more effective.

6. Discussion

6.1 Clinical Implications

Machine learning models may have a high clinical impact on heart disease prediction. The first advantage it offers is early detection of heart diseases, leading to timely interventions and better patient prognoses. Traditional diagnostic methods are usually effective but often invasive procedures or heavily reliant on clinician expertise, leading to variability in diagnosis. On the other hand, big datasets are analyzed consistently by machine learning models, and thus, such models can pick out subtle patterns that the naked eye could have easily missed. These provide automated data-driven insights to help the clinician arrive at better-informed decisions, reducing diagnosis errors and enhancing the precision of treatment plans.

Another clinical implication is the individualization of care. Machine learning models can analyze patient factors specific to medical history, genetics, and lifestyle to predict individual risk more precisely. This invention allows healthcare providers to tailor prospective treatment strategies or interventions for each patient and improve outcomes by customizing these care plans. For instance, if a model predicts that a patient is highly likely to have heart disease given certain risk factors, clinicians can recommend interventions to prevent the disease much earlier in the course of care. Also, integration into clinical practice can further facilitate workflows by automating portions of diagnosis processes, which in turn could enable clinicians to pay more attention to the more complex cases, hence improving overall healthcare delivery efficiencies.

6.2 Innovating the Prediction of Heart Diseases

Recent innovations in machine learning have enhanced heart disease prediction significantly. Among the most important ones is the application of ensemble learning methods, such as XG-Boost, where several models are combined to make one overall estimation that is usually better. Approaches like this work very well on complex datasets, in which most nonlinear relationships and missing data are standard features in cardiovascular analysis. These models can, for example, weigh variables such as the effect of age, level of cholesterol, blood pressure, and lifestyle factors to produce better predictions than single models.

Equally important is the development of deep learning algorithms, more so neural networks that can handle volumes of data in sensing designs and complex patterns associated with heart diseases. Large deep learning models, such as CNNs and RNNs, hold great promise for analyzing medical imaging coupled with physiological data from echocardiograms to CT scans for the early detection of cardiovascular issues. These models can be trained on millions of data and give superior results to traditional methods in recognizing subtle abnormalities that indicate heart disease. These advances open new frontiers in diagnostic imaging, where deep learning algorithms support the decisions of radiologists and cardiologists, increasing the speed and precision of clinical diagnosis.

Further, the innovations in Explainable AI [XAI] will undoubtedly stir up the world of transparency in machine learning models used for heart disease prediction. XAI is how developers can provide clinicians insight into how the model reached certain conclusions, for example, which features blood pressure and cholesterol, which mainly contributed to a specific prediction. This capability increases trust in the technology and enables healthcare professionals to integrate machine learning into their clinical decisions confidently. With continuous development in machine learning, techniques such as ensemble learning, deep learning, and XAI have been advancing the predictive accuracy of heart disease models and the interpretability and usability of developed tools in real-world clinical settings. These critical improvements are bound to revolutionize cardiovascular care and make predictions all the more precise and actionable for doctors and patients.

7. Case Studies:

At a relatively exponential rate, machine learning models are now applied to actual clinical settings to help with diagnosis, risk assessment, and sometimes even treatment planning for heart disease. Below are examples of how machine learning-based predictions transform healthcare practices in cardiology.

7.1 Mayo Clinic

One noteworthy application of machine learning in cardiology comes from the Mayo Clinic, where researchers developed a model to predict sudden cardiac arrest in patients with hypertrophic cardiomyopathy (HCM). The model at the Mayo Clinic used machine learning to study patient data, including ECG, genetic factors, and clinical history, to forecast sudden cardiac arrest more precisely than traditional methods. This model consolidated several risk factors to dynamically and dynamically estimate new patient risks. The impact on clinical practice was profound since it allowed doctors to be proactive with critical decisions, such as recommending ICDs that can prevent fatal outcomes.

7.2 Mount Sinai Health System:

AI-based heart attack prediction machine learning models have been deployed at Mount Sinai Health System in New York to predict the chances of a heart attack among patients with chest pain. Classic methods that could be employed to diagnose heart attacks include ECGs and blood tests, both of which, at times, can miss the subtlety of an imminent attack. To this effect, an AI-driven model proposed by researchers at Mount Sinai assessed a mix of biomarkers, ECG data, and patient demographics for the probability of a heart attack. Results of the real-world testing in clinical situations gave higher accurate predictions of heart attacks than their traditional diagnostic tools, enabling emergency room physicians to make swift, accurate decisions regarding patient care. By applying machine learning, cardiologists at Mount Sinai could take early interventions for patients who are at risk through medication administration or in preparation for emergency angioplasty, hence improving the patient outcome.

7.3 Cleveland Clinic

The Cleveland Clinic adopted a system of predicting heart failure early in patients. These were also diseases with symptoms that, at the time they were diagnosed, had often progressed very far, and treatment was no longer possible. Patient data formed the basis of the model used by the Cleveland Clinic to look for the pattern that would indicate heart failure at its earliest stages. This model was ensembled from supervised learning algorithms such as Random Forest and SVM, alerting doctors to subtle heart function changes before these become symptomatic. Physicians then recommended lifestyle modifications, medications, or close monitoring much before the disease advanced. Finally, this application of machine learning improved the precision of diagnosing heart failure. It allowed the clinic to provide personalized preventive care that reduced hospital admissions and ensured patients a better quality of life.

8. Conclusion

The chief objective of this research project is to explore how machine learning algorithms can be deployed for the early detection of heart disease. In this research project, we have used the Cleveland Dataset from the UCI Machine Learning Repository. This dataset contains 70,000 patient records and 12 unique features. Three Machine Learning algorithms were trained: Logistic Regression, Random Forest, and Support Vector Machines. Subsequently, we analyzed the performance of every algorithm in terms of precision, accuracy, recall, F1-score, and ROC-AUC. Based on the evaluation metrics provided for Logistic Regression, Random Forest, and Support Vector Machine (SVM), we analyzed the performance of each model in terms of accuracy, precision, recall, F1-score, and ROC-AUC. In that respect, Logistic Regression was the best model overall because it achieved the highest ROC-AUC score, implying that it balanced true positives and false positives better than the other models. Support Vector Machine had the highest accuracy and performed similarly to Logistic Regression but was slightly lower. Retrospectively, the implications for heart disease prediction are evident with simple algorithms such as Logistic Regression affirmatively performing better in specific early heart detection tasks, especially when balancing precision and recall.

References

[1] Ahmad, M., Ali, M. A., Hasan, M. R., Mobo, F. D., & Rai, S. I. (2024). Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins. In *Ethics, Machine Learning, and Python in Geospatial Analysis* (pp. 223-253). IGI Global.

[2] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2023). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.

[3] Dandapath, A. (2022). Heart Disease Prediction using Machine Learning Techniques. *www.academia.edu*.https://www.academia.edu/68528660/Heart_Disease_Prediction_using_Machine_Learning_Techniques

[4] Devireddy, S. (2021). Cardiovascular disease prediction using machine learning techniques. *www.academia.edu*. https://www.academia.edu/94882757/Cardiovascular_Disease_Prediction_Using_Machine_Learning_Techniques

[5] Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In *IOP Conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.

[6] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554.

[8] Nasiruddin, M., Dutta, S., Sikder, R., Islam, M. R., Mukaddim, A. A., & Hider, M. A. (2024). Predicting Heart Failure Survival with Machine Learning: Assessing My Risk. *Journal of Computer Science and Technology Studies*, *6*(3), 42-55.

[9] Ponnala, R. (2021). Heart Disease Prediction using Machine Learning Techniques. *Technoscienceacademy*. https://www.academia.edu/51348853/Heart_Disease_Prediction_using_Machine_Learning_Techniques

[10] Gubbala, S. (2023). Heart disease prediction using machine learning techniques. *Irjet*. https://www.academia.edu/89842470/Heart_Disease_Prediction_Using_Machine_Learning_Techniques

[11] Rout, D. (2022). Heart disease prediction using machine learning techniques. *Bput*. https://www.academia.edu/63306265/Heart_Disease_Prediction_Using_Machine_Learning_Techniques

[12] Saini, M. K. (2023). Machine learning techniques for precise heart disease prediction. *www.academia.edu*. https://www.academia.edu/122156116/Machine_Learning_Techniques_for_Precise_Heart_Disease_Prediction

[13] Shah, D., Patel, S., & Bharti, S. K. (2024). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.

[14] Sharma, V., Yadav, S., & Gupta, M. (2023, December). Heart disease prediction using machine learning techniques. In 2020, 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 177-181). IEEE.