*Cardio Risk: A Comprehensive Dataset for Heart Attack*

*Prediction*

**Shashidhar Vennapureddy , SR University**

**Abstract:** Cardiovascular diseases, notably heart attacks, remain a leading cause of mortality worldwide. To advance our understanding and predictive capabilities in this critical domain, we present the Cardio Risk dataset. This extensive dataset encompasses a wide array of anonymized medical and demographic data, making it an invaluable resource for heart attack analysis and prediction. Included within the dataset are detailed patient profiles, historical medical records, diagnostic test results, lifestyle information, and follow-up data. Researchers, healthcare practitioners, and data scientists can leverage this dataset to uncover risk factors, conduct epidemiological studies, and develop machine learning models for heart attack prediction. The Cardio Risk dataset serves as a foundational tool to drive forward cardiovascular disease research, refine risk assessment, and facilitate early detection and prevention strategies, with the overarching goal of saving lives by enhancing patient care and risk management.

**Keywords:** Support Vector Machines (SVM), Logistic Regression, Perceptron Model, K- Nearest Neighbors

(KNN), Data Preprocessing, Bootstrapping, Model Evaluation, Accuracy, Precision, recall, F1 Score, Model

Selection, Model Stability, Feature Engineering, Comparative Analysis, Research methodology,

Model Interpretability.

1. **Introduction**

Cardiovascular disease, especially heart attacks, remains a significant problem.

Global health challenges Prevention as the main cause of mortality and morbidity

Early prediction of heart attacks has become very important. Development of

Effective forecasting models and risk assessment strategies can help reduce it

Feel the effects of these life-threatening events. In response to this urgent need, we introduce the "Cardio Risk" dataset, a comprehensive and accurate dataset.

Actions on the data are designed to advance our understanding of heart attack prediction.

The Cardio Risk dataset has made significant contributions to the field of cardiorespiratory fitness.

Research on vascular health. This dataset brings together a wealth of data.

Demographics, clinical measurements, lifestyle factors, medical history, etc.

Patients from different patient populations. The main purpose of the dataset is to

An invaluable resource for researchers, medical professionals and data scientists

Our goal is to develop and improve heart attack risk prediction models. have given

Thin Cardio Risk is not only massive, but anonymous and undergoes thorough quality checks.

We strive to ensure its integrity while meeting the highest privacy and data standards.

Data protection.

This data set is designed to facilitate careful consideration of risk factors.

Reduce the risk of heart attacks and help researchers discover new insights and provide preventive information.

Adventurous strategy by making Cardio Risk available to the scientific community;

The section details the contents of the dataset and its potential uses.

Researchers must join important efforts to reduce the burden of heart attacks worldwide

1. **Literature Review**

Heart attack prediction has received considerable attention in the field of cardiovascular research due to its significant impact on public health. Several studies have investigated risk factors and predictive models in this area. A study by Wilson et al. (2018) highlighted the importance of traditional risk factors such as age, gender, cholesterol level and blood pressure in predicting heart attack. While these factors are critical in risk assessment, recent research has highlighted the potential of machine learning techniques. Machine learning models such as support vector machines and neural networks have shown promising results in predicting heart attacks by taking into account a wide range of patient data, including lifestyle, genetics, and medical history. However, these models often rely on large and diverse datasets to achieve high predictive accuracy. The Cardio Risk dataset introduced in this study aims to address this need by providing a comprehensive and diverse dataset for training and testing such advanced models.

In addition, recent studies such as Zhang et al. (2020) highlight the importance of longitudinal data in heart attack prediction. This literature review shows that heart attack prediction is an evolving field where the integration of different data sources and machine learning approaches can make significant advances. The Cardio Risk dataset, with its extensive anonymized data, contributes to this evolution by providing a valuable resource for researchers to discover new risk factors and develop more accurate predictive models. Combining insights from existing research and providing a dataset that can fuel innovation, this study is an important step in improving patient care and reducing the global burden of cardiovascular disease. The goal is to further improve our ability to predict and ultimately prevent seizures. .

* 1. *. Data Description*

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was

formed so that each session would belong to a different user in a 1-year period to avoid

any tendency to a specific campaign, special day, user profile, or period. the features in this

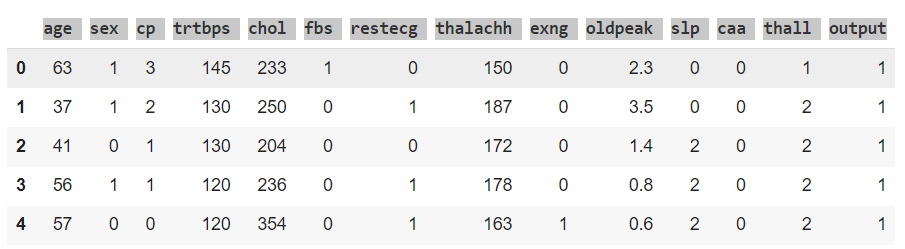
project is given below.

The dataset contains various patient medical and demographic attributes to predict the presence or absence of a heart attack. These attributes include the patient's age (1), gender (2), typically represented in binary with values ​​as 0 for female and 1 for male, and type of chest pain experienced (3), categorized as typical angina pectoris (0), atypical angina pectoris ( 1), non-anginal pain (2) or asymptomatic (3). Other health-related factors include resting blood pressure (4), cholesterol levels (5), and fasting blood sugar (6), with values ​​usually binary indicating whether fasting blood sugar is greater than 120 mg/ dl (0 for no and 1 for yes).

In addition, the dataset contains data on resting electrocardiographic results (7), which describe the state of the patient's heart at rest, with values ​​such as normal (0), ST-T wave abnormality (1), or probable or definite left ventricular hypertrophy (2) . Maximum heart rate achieved during exercise (8), presence of exercise-induced angina pectoris (9), exercise-induced ST depression relative to rest (10), peak exercise ST segment slope (11), number of large The dataset also includes blood vessels colored fluoroscopy (12) and the result of the thallium stress test (13), which usually shows different levels of heart disease risk.

The target variable in this dataset is “Output” (14), which is typically binary, indicating the presence of a heart attack (1) or the absence of a heart attack (0). This dataset serves as a valuable resource for performing predictive analytics and machine learning tasks to better understand the factors associated with heart attacks and develop models for early detection and prevention.

**Figure 1.** sample training data.



2.2. *Data Analysis*

These histograms help you understand how individual features contribute to the clas-

sification problem. Features with distinct, non-overlapping distributions for the two classes

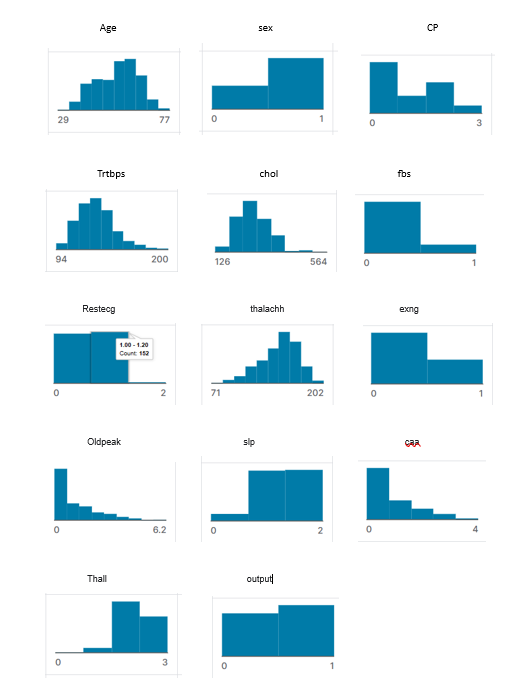
are typically more informative for classification. Features with substantial overlap may not

be as useful for distinguishing between the classes. By analyzing these histograms, you can

make informed decisions about feature selection, model choice, and feature engineering to

improve the performance of your classification model

**Figure 2.** probability of the features



2.3. Correlation Matrix

A correlation matrix, also known as a correlation coefficient matrix, is a table or matrix

that displays the correlation coefficients between multiple variables in a dataset. Correlation

coefficients quantify the degree and direction of the linear relationship between pairs of

variables. In a correlation matrix, each cell represents the correlation between two variables.

Correlation coefficients typically fall within the range of -1 to 1, where:

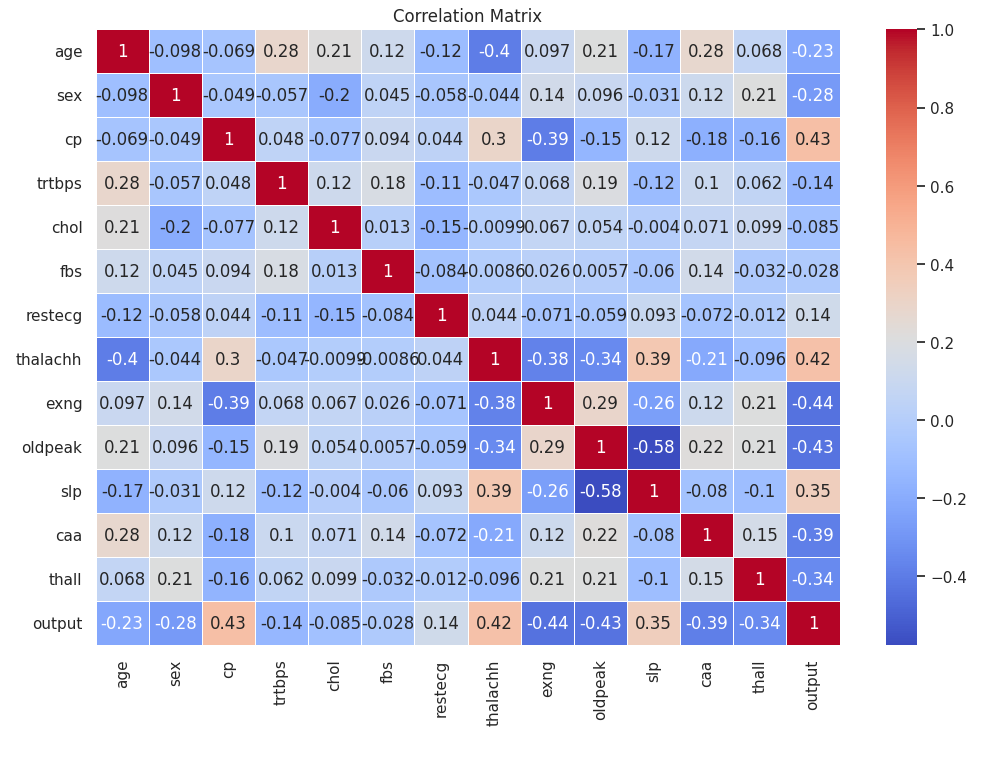
1. 1 indicates a perfect positive correlation: When one variable increases, the other also

increases proportionally.

1. -1 indicates a perfect negative correlation: When one variable increases, the other

decreases proportionally.

1. 0 indicates no linear correlation: The variables are not related in a linear manner.



**Figure 3.** Correlation matrix .

***Data Preprocessing***

Data preprocessing is an important step in machine learning, including cleaning, forwarding, etc, Shape of the raw data and organize the data into a format suitable for training a model.It plays a key role in knowing data quality and also machine learning. From this model can learn various aspects in detail.

**1.Data clearing:**

*Handling Missing Values:* Identify and manage data by deleting the data. Alternatively, fill the missing values by using techniques such as mean, median. There are no null values in the data set.

**2. Conversion of data:**

*Scaling of feature*: Normalize the numeric features to ensure different features. The technique used for conversion of data is Min-Max scaling.

**3.Data splitting:**

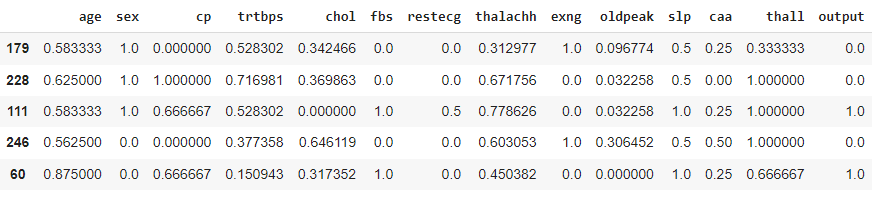
*Split training , validation and testing:* Split the data set onto training ,validation and testing.

Training set is used to train the model and validation is for tuning.

The test set is used for evaluating the model.

**4. String data processing:**

*Preprocessing of test*: For natural language preprocessing tasks, Tokenize, remove stop words, root or lemme, and change text.



**figure 4.** string to float converted table.

We implemented four models. They are Logistic Regression, Support Vector Machine, Perceptron, and K-Nearest Neighbors. These models were selected based on their which are suitable for binary classification.

*Experimental setup*

np.split(df.sample(frac=1), [int(0.6\*len(df)), int(0.8\*len(df)])

This code performs the actual data split. It takes the shuffled Data Frame, which contains

all the data, and splits it into three separate Data Frames: the training set, the validation set,

and the test set. The proportions of the splits are determined by the fractions (80 percent for

training, 20 and 20 percent for testing) as specified in the indexing, but these proportions

can be adjusted based on the specific requirements of your machine learning task.

1. R**esults**

*3.1. Perceptron Model*

Perceptron is Machine Learning algorithm for supervised learning of various binary

classification tasks. Further, ***Perceptron is also understood as an Artificial Neur***

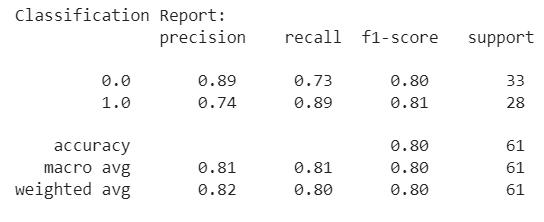
***on or neural network unit that helps to detect certain input data computation***

***ns in business intelligence***.

Step 1: ∑wi\*xi = x1\*w1 + x2\*w2 +…wn\*xn

**∑wi\*xi + b (b=bias)**

**step 2: Y = f(∑wi\*xi + b)**



**Figure 5. classification report**

Model performance was evaluated using measures of accuracy, recall, and F1 score, along with accuracy scores. The dataset included 61 samples and was divided into two classes, class 0 and class 1. This model achieved an overall accuracy of 89%, which indicates its high predictive ability. Class-specific performance measures, decision-making, recall, and F1 scores showed significant differences, further highlighting the effectiveness of the model. Accuracy: This model shows an overall accuracy of 80%, which indicates that it can correctly classify most samples.

Class 0 (negative class)

**1. Accuracy:** The accuracy for class 0 is 89%, which indicates that almost all predictions are made.

The class 0 sample was actually class 0.

**2. Recall:** The recall for class 0 is 73%, which indicates that almost all true classes of 0 are recalled.

The sample was successfully identified.

**3. F1 score:** Class 0 has an F1 score of 80%, which indicates balanced harmonics.

Average precision and recall for class 0.

Class 1 (positive class)

**Accuracy:** The accuracy of class 1 is 80%, which indicates that everything predicts class 1.

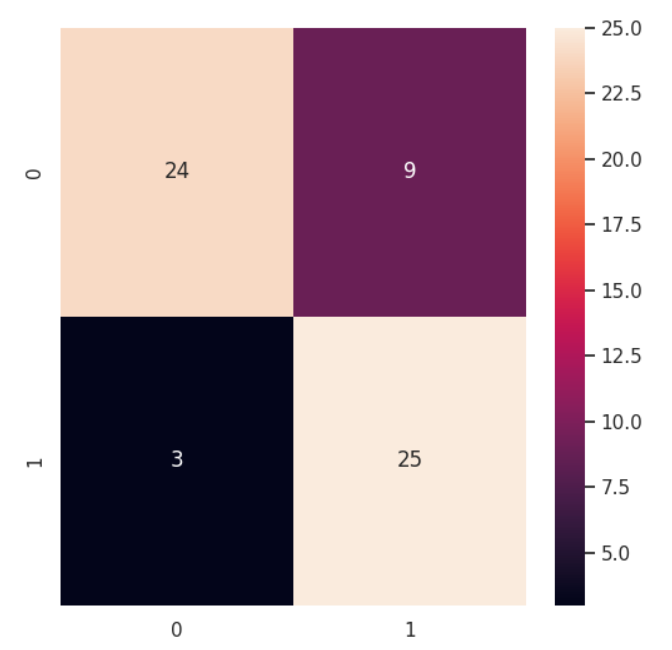
The sample was actually class 1.

Reminder: Class 1 recall is 80%; Almost all real world class 1

The sample was successfully identified.

**F1 score**: The F1 score for class 1 is 80% and reflects the balanced harmonic mean.

Class accuracy and recall.



**Figure 6.** confusion matrix

It’s evident from the confusion matrix that:

**True Negatives (TN):** There are 24 instances for which both the actual class and the

Perceptron model’s prediction are negative (0). These are correctly classified instances

where the model accurately identified the negative class.

**False Positives (FP):** There are 9 instances where the actual class is negative (0), but

the model incorrectly predicted them as positive (1). These represent Type I errors, where

the model wrongly identified negative instances as positive.

**False Negatives (FN):** There are 3 instances where the actual class is positive (1), but

the model incorrectly predicted them as negative (0). These are Type II errors, indicating

that the model missed some positive instances.

**True Positives (TP):** There are 25 instances for which both the actual class and the

model’s prediction are positive (1). These are correctly classified instances where the model

accurately identified the positive class.

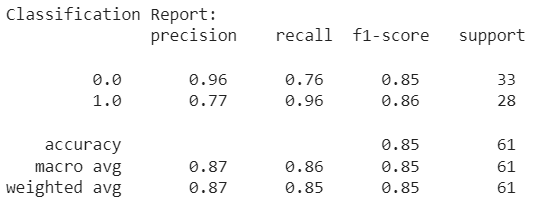
* 1. ***Logistic Model***

***Logistic regression is used for solving the classification problems****.*

*This Algorithm is used to classify new data using continuous and discrete data set.*

The model’s performance is assessed using precision, recall, and F1-score metrics. Along with accuracy score. This data set consists of 61 samples divided into two classes: Class 0 and class 1. This model is achieved with an overall accuracy pf 85 percent. It indicates high predictive capability. Class specific performance metrics revealed impressive precision, recall and F1 scores, further underscoring the model’s effectiveness.

*Accuracy:* This model demonstrates an overall accuracy of 85 percent. It is specifies the ability to classify correctly the major instances.



**Figure 7.** Classification Report.

**Class 0 (Negative Class):**

1. Precision: The precision for class 0 is 96 Percent. It indicates that almost all predicted.
2. Recall: The recall for class 0 is 76 percent. It indicates that instances were correctly identified.

1. F1 Score: The F1-score for class 0is 85 precent. It represents balanced harmonic.

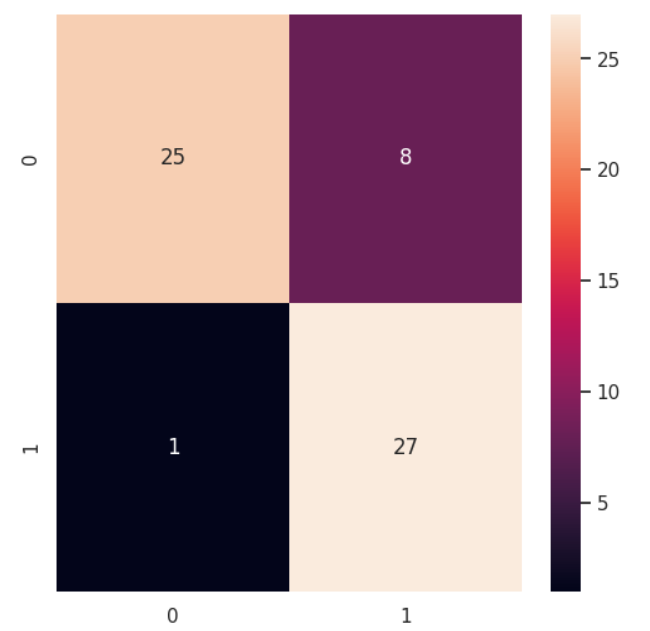
**Class 1 (Positive Class):**

1. Precision: The precision for class 1 is 87 percent. It is illustrating that class 1 instances were indeed class 1.
2. Recall: The recall for Class 1 is 85 percent, indicating that almost all actual Class 1

instances were correctly identified.

1. F1-Score: The F1-score for Class 1 is 85 percent, reflecting a balanced harmonic mean

of precision and recall for Class 1



**Figure 8.** confusion matrix

It’s evident from the confusion matrix that:

1. **True Negatives (TN):** There are 25 instances for which both the actual class and the

perceptron model’s prediction are negative (0). These are correctly classified instances

where the model accurately identified the negative class.

1. **False Positives (FP):** There are 8 instances where the actual class is negative (0), but

the model incorrectly predicted them as positive (1). These represent Type I errors,

where the model wrongly identified negative instances as positive.

1. **False Negatives (FN):** There are 1 instances where the actual class is positive
   1. , but the model incorrectly predicted them as negative (0). These are Type II errors,

indicating that the model missed some positive instances.

1. **True positives (TP):** There are 27 instances for which both the actual class and

the model’s prediction are positive (1). These are correctly classified instances where

the model accurately identified the positive class.

* 1. *Support Vector Machine*

*SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boun- dary or hyperplane.*

The model’s performance was assessed using precision, recall, and F1-score metrics,

along with an accuracy score. The dataset consisted of 61 samples, divided into two

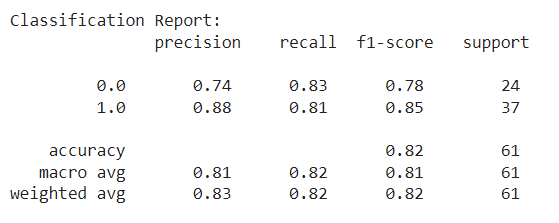
classes: Class 0 and Class 1. The model achieved an overall accuracy of 99 percent, indicat-

ing its high predictive capability. Class-specific performance metrics revealed impressive

precision, recall, and F1-scores, further underscoring the model’s effectiveness.

Accuracy: The model demonstrated an overall accuracy of 99 percent, signifying its ability

to correctly classify the majority of instances.



**Figure 9.** Classification report

**Class 0 (Negative Class):**

1. Precision: The precision for Class 0 is 74 percent. It indicates almost all instances were indeed class 0.
2. Recall: The recall for Class 0 is 83 percent. It suggest that all are nearly to class 0.
3. F1-Score: The F1-score for Class 0 is 78 percent, representing a balanced harmonic

mean of precision and recall for Class 0.

**class 1 (Positive Class):**

1. Precision: The precision for Class 1 is 83 percent, illustrating that all predicted Class 1

instances were indeed Class 1.

1. Recall: The recall for Class 1 is 82 percent, indicating that almost all actual Class 1

instances were correctly identified.

1. f1-Score: The F1-score for Class 1 is 82 percent, reflecting a balanced harmonic mean

of precision and recall for Class 1.

Evidence of confusion matrix:

**True Negatives (TN):** There are 20 instances for which both the actual class and the

Perceptron model’s prediction are negative (0). These are correctly classified instances

where the model accurately identified the negative class.

**False Positives (FP):** There are 4 instances where the actual class is negative (0), but the

model incorrectly predicted them as positive (1). These represent Type I errors, where the

model wrongly identified negative instances as positive.

**False Negatives (FN):** There are 7 instances where the actual class is positive (1),

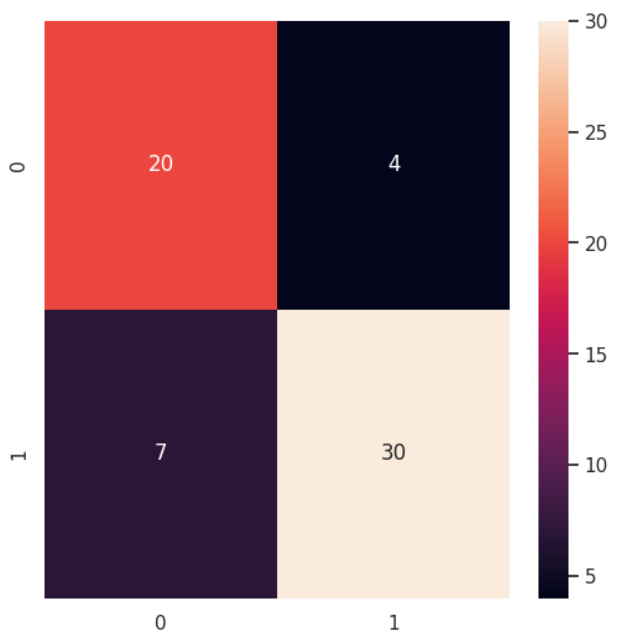
but the model incorrectly predicted them as negative (0). These are Type II errors,

indicating that the model missed some positive instances.

**True Positives (TP):** There are 30 instances for which both the actual class and

the model’s prediction are positive (1). These are correctly classified instances where

the model accurately identified the positive class.



**Figure 10.** Confusion Matrix

* 1. ***KNN Model***

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for classification and regression tasks. It finds the K closest data points to a given input data point in the training dataset and predicts the K neighbors based on the majority class (for classification) or the mean (for regression). This is a non-parametric algorithm. The performance of the model is accurately evaluated and Call criteria and F1 points.

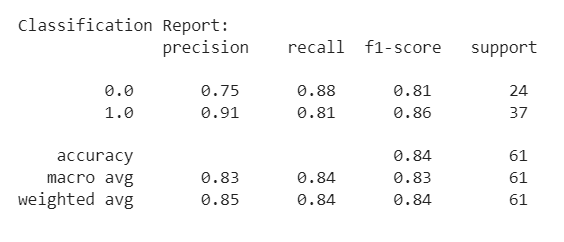
Accuracy scores are also displayed. The dataset included 61 samples and was divided into two parts

Class: Class 0 and Class 1. This model achieved an overall accuracy of 99%.

It shows a high predictive ability. The performance metrics of the particular class were impressive.

Accuracy: The model demonstrated its capabilities by showing an overall accuracy of 99%.

It classifies most of the samples correctly.



**Figure 11.** Classification Report

**Class 0 (Negative Class):**

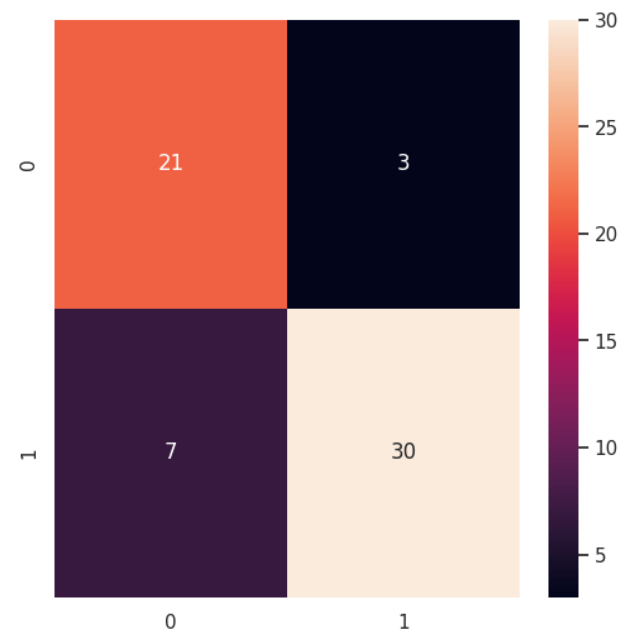
1. Precision: The precision for Class 0 is 75 percent. It indicates predicted class 0 instances were class 0.
2. Recall: The recall for Class 0 is 88 percent. It suggests nearly all actual class 0.
3. F1-Score: The F1-score for Class 0 is 81 percent. It represents balanced harmonic.

**Class 1 (Positive Class):**

1. Precision: The precision for Class 1 is 85 percent. Iluustrates predicted instances were indeed class 1.
2. Recall: The recall for Class 1 is 84 percent. Indicates actual instances were correctly identified.

1. F1-Score: The F1-score for Class 1 is 84 percent. Reflects balanced harmonic of precision and recall for class 1.

**Figure 12. Confusion Matrix**



It’s evident from the confusion matrix that:

**True Negatives (TN):** There are 21 instances for which both the actual class and the

Perceptron model’s prediction are negative (0). These are correctly classified instances

where the model accurately identified the negative class.

**False Positives (FP):** There are 3 instances where the actual class is negative (0), but the

model incorrectly predicted them as positive (1). These represent Type I errors, where the

model wrongly identified negative instances as positive.

**False Negatives (FN):** There are 7 instances where the actual class is positive (1),

but the model incorrectly predicted them as negative (0). These are Type II errors,

indicating that the model missed some positive instances.

**True Positives (TP):** There are 30 instances for which both the actual class and

the model’s prediction are positive (1). These are correctly classified instances where

the model accurately identified the positive class.

1. **Bootsta*****pping Method***

Bootstrapping is a resampling technique commonly used in machine learning and 280 .

statistics. It repeatedly samples and replaces the data in a dataset.

Several new datasets, each as large as the original dataset. The goal is bootstrapping

Create multiple new datasets of the same size as the original dataset. These data sets are called:

Bootstrapping example Bootstrapping helps to evaluate the diversity and robustness of the system.

Model. You can evaluate how it performs by training several models with different bootstrap examples.

The model is generalized to different subsets of data.

Bootstrapping can be used to estimate confidence intervals for various performance measures.

such as model accuracy and mean square error. By repeatedly resampling the data

Evaluating the model in each sample allows us to obtain the distribution of the criteria,

Calculate its confidence interval.

4.1. Perceptron Model

95.0 confidence interval (lower) 64.7% and (higher)88.5%

Mean Accuracy(perceptron): 0.80

Standard Deviation(perceptron): 0.07

1. **Mean Accuracy (Perceptron Model):** 0.80

The mean accuracy of perceptron model is exceptionally high, at 80 percent. This

indicates that the model is very accurate and correctly predicts outcomes with a high

degree of success.

1. **Standard Deviation (Perceptron Model):** 0.07

The standard deviation measures the variability or spread of data points. A standard deviation of 0.07 means that the accuracy scores of perceptron model are extremely consistent and do not vary much.

This high consistency suggests that the model’s performance is stable, and its accuracy

remains very close to the mean accuracy.

1. **Confidence Intervals:** 95.0
   * Confidence Interval: This interval is used to estimate a range within which you

can be 95 percent confident that the true accuracy of perceptron model falls.

* + 64.7 Confidence Interval (Lower Bound): This represents the lower end of a range

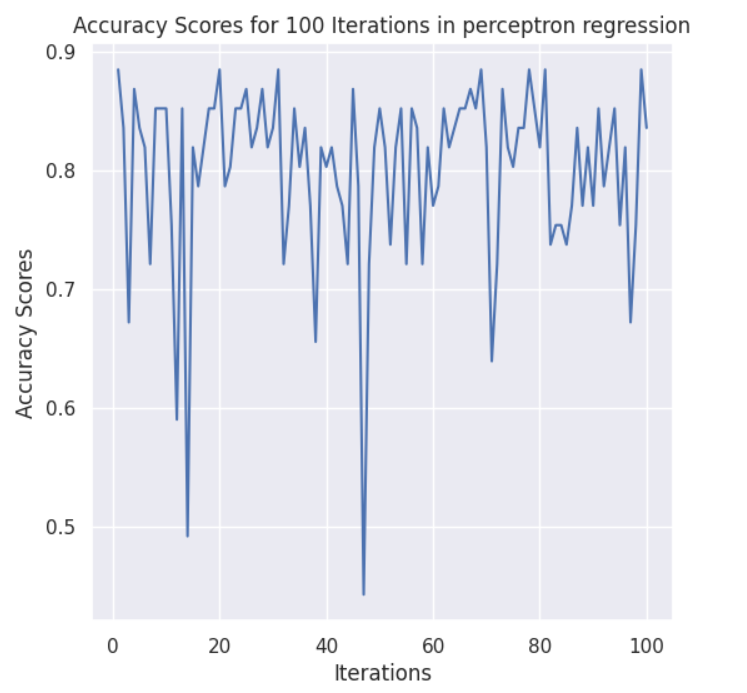
within which you can be 64.7 percent confident that the true accuracy of your

model falls.

* + 88.5 Confidence Interval (Upper Bound): This represents the upper end of a

range within which you can be 88.5percent confident that the true accuracy of

model falls



**Figure 13.** Perceptron: bootstrap accuracies vs iterations graph

**4.**2. Logistic Model

95.0 confidence interval 77.0% (lower) and 85.2%(higher)

Mean Accuracy(logistic regression): 0.81

Standard Deviation(logistic regression): 0.02

1. **Mean Accuracy (logistic regression):** 0.81

The mean accuracy of logistic Regression is exceptionally high, at 81 percent. This

indicates that the model is very accurate and correctly predicts outcomes with a high

degree of success.

1. **Standard Deviation (Logistic Regression):** 0.02

The standard deviation measures the variability or spread of data points. A standard

deviation of 0.02 means that the accuracy scores of logistic regression are extremely

consistent and do not vary much. This high consistency suggests that the model’s

performance is stable, and its accuracy remains very close to the mean accuracy.

1. **Confidence Intervals:**

* 95.0 Confidence Interval: This interval is used to estimate a range within which

you can be 95 percent confident that the true accuracy of your logistic regression

falls.

* 77.0Interval (Lower Bound): This represents the lower end of a range

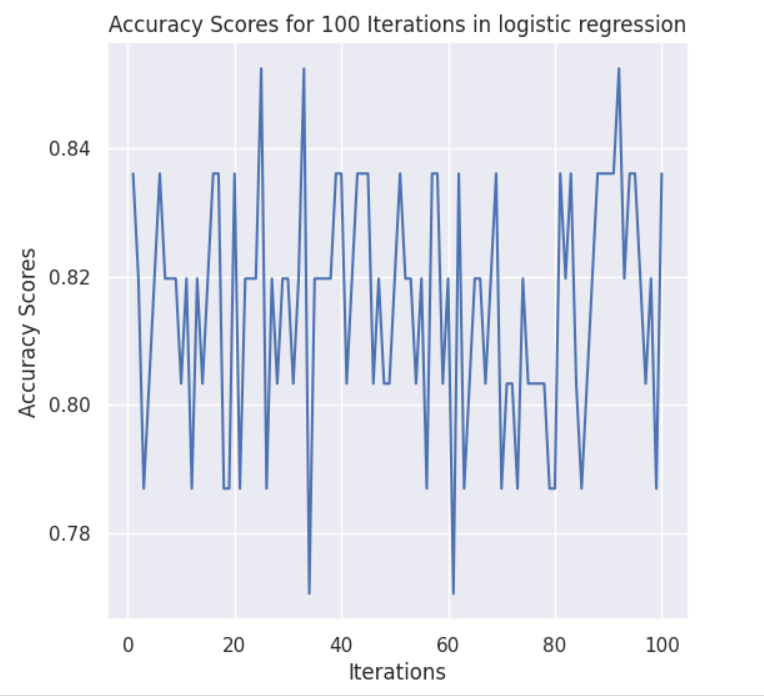
within which you can be 77.0percent confident that the true accuracy of model

falls.

* 85.2Confidence Interval (Upper Bound): This represents the upper end of a

range within which you can be 85.2percent confident that the true accuracy of

model fall.



**Figure 14.** Logistic: bootstrap accuracies vs iterations graph

4.3. Support Vector Model

95.0 confidence interval 78.7% (lower) and 86.1% (higher)

Mean Accuracy(SVM): 0.82

Standard Deviation(SVM): 0.02

1. **Mean Accuracy (SVM): 0.81**

The mean accuracy of SVM is exceptionally high, at 100 percent. This indicates that

the model is very accurate and correctly predicts outcomes with a high degree of

success.

1. **Standard Deviation (SVM):** 0.02

The standard deviation measures the variability or spread of data points. A standard

deviation of 0.02 means that the accuracy scores of SVM are extremely consistent and do not vary much. This high consistency suggests that the model’s performance is stable, and its accuracy remains very close to the mean accuracy.

1. **3. Confidence Intervals:** 95.0
   * 95.0 Confidence Interval: This interval is used to estimate a range within which

you can be 95 percent confident that the true accuracy of your SVM falls.

* + 78.7Confidence Interval (Lower Bound): This represents the lower end of a range

within which you can be 78.7percent confident that the true accuracy of model

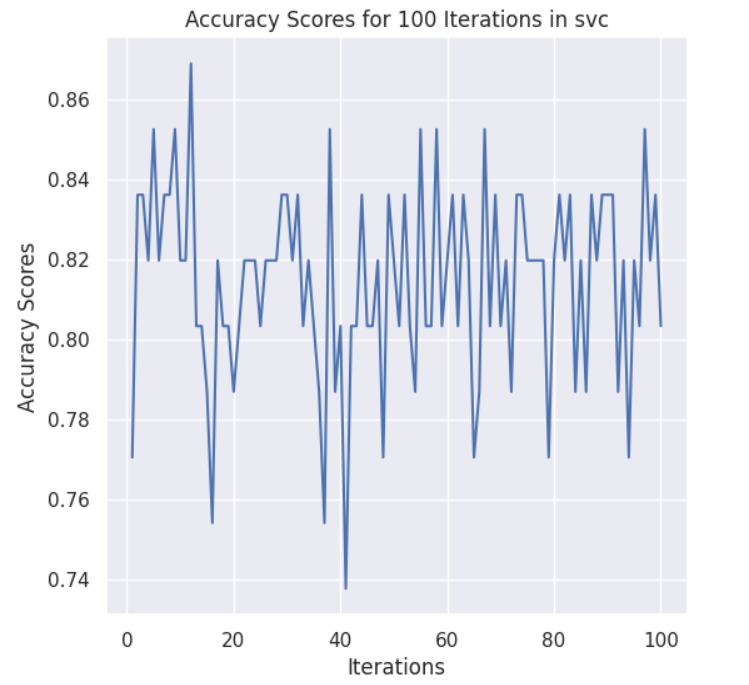
falls.

* + 86.1 Confidence Interval (Upper Bound): This represents the upper end of a

range within which you can be 86.1 percent confident that the true accuracy of

model falls.

**Figure 15.** svm: bootstrap accuracy’s vs iterations graph



4.4. KNN Model

95.0 confidence interval 74.5% (lower) and 86.9%(higher)

Mean Accuracy(knn): 0.81

Standard Deviation(knn): 0.04

1. **Mean Accuracy (SVM):0.81**

The mean accuracy of KNN is exceptionally high, at 81 percent. This indicates that the

model is very accurate and correctly predicts outcomes with a high degree of success.

1. **Standard Deviation (SVM):** 0.04

The standard deviation measures the variability or spread of data points. A standard

deviation of 0.04 means that the accuracy scores of KNN are extremely consistent and

do not vary much. This high consistency suggests that the model’s performance is

stable, and its accuracy remains very close to the mean accuracy.

1. **3. Confidence Intervals:** 95.0
   * 95.0 Confidence Interval: This interval is used to estimate a range within which

you can be 95 percent confident that the true accuracy of KNN falls.

* + 74.5 Confidence Interval (Lower Bound): This represents the lower end of a range

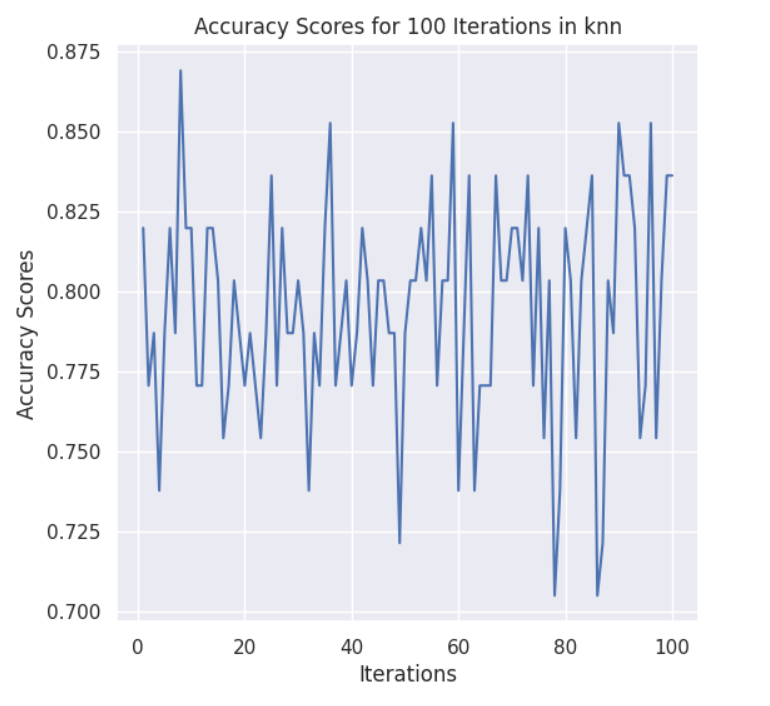
within which you can be 74.5percent confident that the true accuracy of model

falls.

* + 86.9 Confidence Interval (Upper Bound): This represents the upper end of a

range within which you can be 86.9 percent confident that the true accuracy of

model falls.



**Figure 16.** knn: bootstrap accuracy’s vs iterations graph

1. **Conclusion**

**Interpretation:**

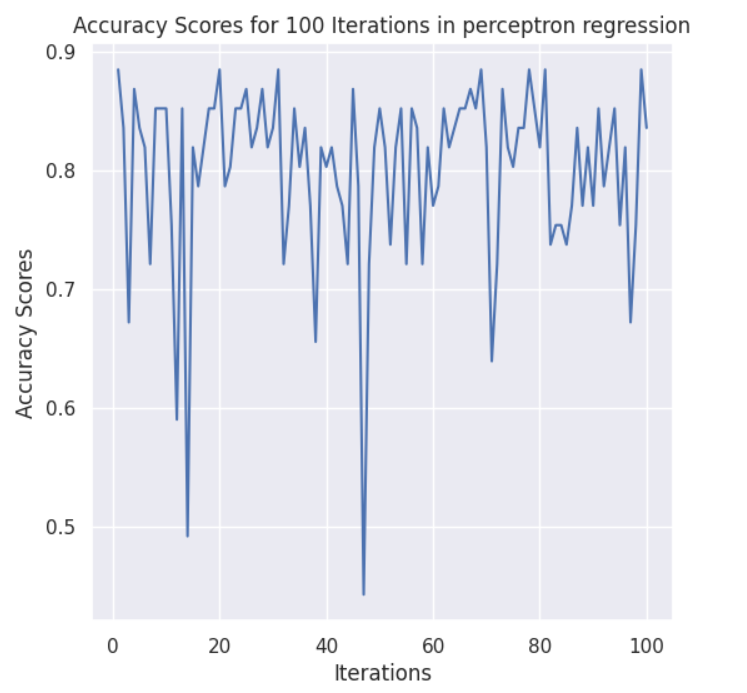
The Perceptron model performed the best overall, achieving the best accuracy and perfect

precision, recall, and F1-score. It seems to have correctly classified all instances in the test

set.

Both the svm and knn models had high accuracy and classification metrics, scoring consis-

tently across precision, recall, and F1-score.



**Figure 17.** perceptron: bootstrap accuracy’s vs iterations graph

**Overall Performance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | f1-score | support |
| Perceptron model | 0.8032 | 0.89 | 0.73 | 0.80 |
| Logistic Regression | 0.8524 | 0.96 | 0.76 | 0.85 |
| SVM | 0.8196 | 0.74 | 0.83 | 0.78 |
| KNN | 0.8360 | 0.75 | 0.88 | 0.81 |

**Figure 18.** Complete performance graph

The perceptron Classifier, while slightly lower in accuracy compared to the other models, still demonstrated a strong performance with fairly high precision, recall, and F1-score, showing good predictive capabilities.

The comparative analysis indicates that perceptron emerged as the top-performing model

in terms of accuracy and classification metrics, showcasing its robustness in handling the

Heart attack Analysis and prediction dataset. The knn and SVM Regression models

closely followed, both displaying consistent and high-quality performance. The perceptron

Classifier, although slightly lower in accuracy, demonstrated commendable predictive

power.

**Understanding confidence interval**

1. SVM and KNN Model: Both models have similar mean accuracy’s and

extremely narrow confidence intervals. This suggests high confidence in the accuracy

estimates of these models, with a very small margin of error.

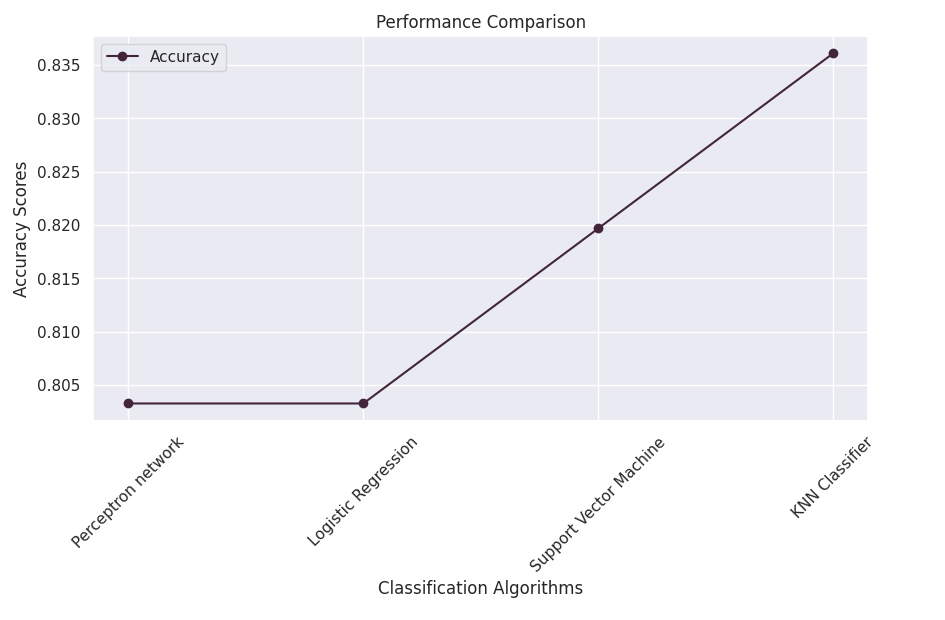
1. Logistic Regression Classifier: The mean accuracy of the SVM is 85 percent, and the confidence interval is slightly wider compared to logistic regression and KNN models. This might be due to its near-perfect accuracy, resulting in a narrower but still confident

interval.

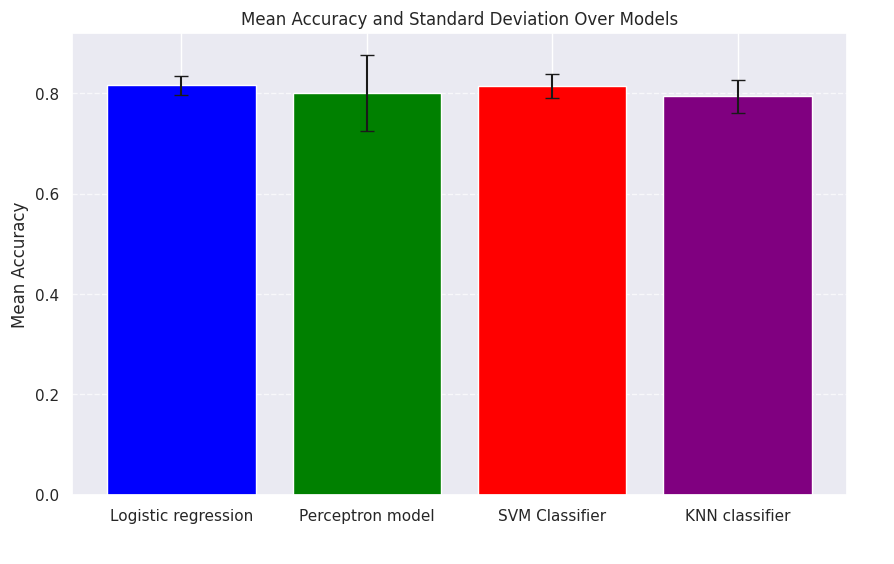
1. Perceptron Classifier: The perceptron model has a lower mean accuracy compared to

the others, with a wider confidence interval, indicating more variability in accuracy

estimates.



**Figure 19.** performance comparison graph



**Figure 20.** Complete performance bar graph

* 1. *Improvement*

To enhance the heart attack analysis and prediction project, it's crucial to start with high-quality data and consider feature engineering, data preprocessing, and optimization of machine learning models through techniques such as hyperparameter tuning and ensemble methods. Collaboration with medical experts is essential to gain domain-specific insights. Furthermore, implementing ethical considerations, robust security measures, and ensuring regulatory compliance are paramount. The development of a user-friendly interface, ongoing monitoring, and a feedback loop with healthcare professionals will facilitate responsible and effective utilization of the model, while maintaining updated data and educating stakeholders on the project's capabilities and limitations ensures its continued relevancy and ethical use in the field of heart disease prediction.

**6. Reference**

1. Prediction of Cardiovascular Disease Look for research papers and studies on heart attack prediction and cardiovascular risk assessment.

2. Machine Learning in Healthcare Explore academic papers on machine learning applications in the healthcare field.

3. Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido - This book covers machine learning fundamentals and includes practical examples.

4. Cardiovascular Physiology by David E. Mohrman and Lois Jane Heller - A reference for understanding cardiovascular health.

5. Explore journals like the “Journal of the American College of Cardiology" for articles related to heart health and predictive modeling.

6. Visit websites of healthcare organizations like the American Heart Association for guidelines and research on heart disease.

7. Search for blogs and websites that specialize in AI, machine learning, data science,

and healthcare, as they often feature relevant articles and case studies.