Heart Disease Predictions Using Machine Learning

Harsh Bolakani
College of Computing and Informatics
Drexel University
New Jersey, USA
hvb36@drexel.com

Greg Morgan
College of Computing and Informatics
Drexel University
Philadelphia, USA
gm655@drexel.com

Trevor Pawlewicz
College of Computing and Informatics
Drexel University
San Diego, USA
tmp365@drexel.com

***Abstract*— Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Classification algorithms based on supervised learning, a type of machine learning, can make diagnoses of cardiovascular diseases easy. Three supervised machine learning algorithms are used in this paper which are Logistical Regression, Naive Bayes, Decision Trees. These algorithms can be used to classify people who have a heart disease from people who do not although Some risk factors for heart disease cannot be controlled, such as your age or family history.**

Keywords— Machine learning, Heart Disease, Heart Disease Dataset, Classification, Logistical Regression, Naive Bayes, Decision Trees

I. Introduction

Our goal for the project is to approach heart disease and compare different classification algorithms on a Heart Failure Prediction Dataset [1] to predict and compare which one does well. We will also compare our algorithms with a neural network to see if does any better. Our predictions will have a substantial impact on detecting early identification of heart disease. This important because it can provide treatment, increase life expectancy, prevent disability and costly hospitalizations while improving the quality of life. Loss of heart function is irreversible. Once the damage has been done, a person cannot return to having a fully functional heart. Early detection for a heart condition is key to finding a possible remedy.

II. Background

Cardiovascular disease (CVD) is a term used to describe a class of diseases that affect the heart and blood vessels. It is a broad term that encompasses various conditions, including coronary artery disease, heart failure, stroke, and peripheral artery disease, among others. CVD is a leading cause of death globally, accounting for a significant proportion of mortality in many countries, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide [2]. Four out of 5 CVD deaths are due to heart attacks or strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

III. Methodology

The manufacturing process models are done with the following steps: data collecting, pre-processing, model building, comparison of models, and evaluation. The algorithms we will be evaluating to train a model are Logistic Regression, Naive Bayes, and Decision Trees. Additional analysis will involve utilizing a Neural Network comparison.

A. Logistic Regression

One of the simplest and best ML classification algorithm is Logistic Regression. It is a supervised ML binary classification algorithm widely used in most applications. It works on categorical dependent variables and the result can be discrete or binary categorical variable 0 or 1.

The logistic regression model is based on the logistic function, also known as the sigmoid function. The sigmoid function maps any real-valued number to a value between 0 and 1. In logistic regression, this function is used to transform a linear combination of predictor variables into a probability value. The goal of logistic regression is to estimate the probability of the "success" outcome given the values of the independent variables.

B. Naive Bayes

Naive Bayes classifier is a statistical based classifier which is based on Bayes Theory. The "naïve" assumption in Naïve Bayes refers to the assumption of independence among the features. This classifier is based on probabilities. Given two events A and B, P (A) is prior probability and P (A|B) is posterior probability, then according to Bayes theorem.

P (A|B) = P (B/A) P (A)/P (B) and
P (B|A) is computed as **P (A ∩ B) = P (A)**

These Bayesian probabilities are used to determine the most likely next event for the given instance given all the training data. Conditional probabilities are determined from the training data. The Naive Bayes model is based on the conditional independence model of each predictor give the target class. This classifier yields optimal prediction (given the assumptions). It can also handle discrete or numeric attribute values.

C. Decision Trees

The decision tree algorithm is a supervised learning algorithm that can be used in both classification and regression

analysis. Unlike linear algorithms, decision trees algorithms are capable of handling nonlinear relationships between variables in the data. The information gained in the decision tree can be defined as the amount of information improved in the nodes before splitting them for making further decisions.
To measure the information gain we use the entropy. Which is a quantified measurement of the amount of uncertainty because of any process or any given random variable. Mathematically the formula for entropy is:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

Decision trees other advantages include their interpretability and the ability to handle both numerical and categorical features. They can handle large datasets and are relatively insensitive to outliers.

D. *Outline of the Proposed Work*

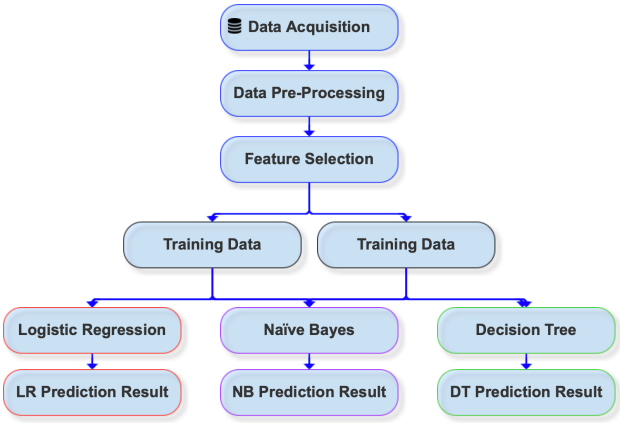

Fig -1: Outline of the Proposed Work.

IV. Dataset

Our source is the Heart Failure Prediction Dataset that includes 11 clinical features for predicting heart disease events. This dataset was taken from Kaggle.com created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations
Final dataset: 918 observations

Attribute Information

- *Age*: age of the patient [years]
- *Sex*: sex of the patient [M: Male, F: Female]
- *ChestPainType*: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- *RestingBP*: resting blood pressure [mm Hg]
- *Cholesterol*: serum cholesterol [mm/dl]
- *FastingBS*: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- *RestingECG*: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- *MaxHR*: maximum heart rate achieved [Numeric value between 60 and 202]
- *ExerciseAngina*: exercise-induced angina [Y: Yes, N: No]
- *Oldpeak*: oldpeak = ST [Numeric value measured in depression]
- *ST*_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- *HeartDisease*: output class [1: heart disease, 0: Normal]

V. Explatory Data Anaylsis

We will analyze our dataset and determine factors that contribute to heart failure and find correlation of various factors.
As part of preprocessing, clean up and EDA, we standardized the data, ensured that there is no class imbalance and converted categorical features to numerical features. As part of correlation analysis, we found that heart disease had maximum (negative) correlation with ST_Slope value as up and positive correlation with ExerciseAngina = Y and with ST_Slope value as flat.

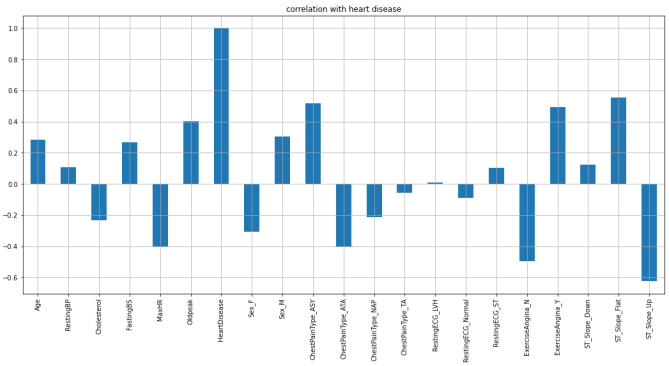

Fig - 2: Data Summary.

VI. Evaluation Metrics

We will quantify our results using accuracy, precision, recall, and F1-score. Accuracy is a proportional measure of the number of correct predictions over all predictions. Precision and recall are two numbers which are used together to evaluate the performance of a classification model. To fully evaluate the effectiveness of a model, you must examine both precision and recall. The precision of a model describes how

many detected items are truly relevant. Recall is a measure of how many relevant elements were detected. F1 score is the weighted average mean of Precision and Recall.

Based on the provided results for Logistic Regression, Naive Bayes, and Decision Tree classifiers, we can draw the following analysis:

A. Accuracy

- Logistic Regression achieved the highest accuracy of 0.86, indicating that it correctly classified 86% of the instances.
- Naive Bayes performed slightly lower with an accuracy of 0.855 (85.5%).
- Decision Tree had the lowest accuracy of 0.755 (75.5%).

B. Precision

- Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive.
- Logistic Regression achieved a precision of 0.851, indicating that 85.1% of the predicted positive instances were positive.
- Naive Bayes had a precision of 0.838 (83.8%), slightly lower than Logistic Regression.
- Decision Tree had the lowest precision of 0.748 (74.8%).

C. Recall

- Recall measures the proportion of correctly predicted positive instances out of all actual positive instances.
- Naive Bayes achieved the highest recall of 0.894 (89.4%), indicating that it identified a high percentage of the actual positive instances.
- Logistic Regression had a recall of 0.885 (88.5%), slightly lower than Naive Bayes.
- Decision Tree had the lowest recall of 0.798 (79.8%).

D. F1-Score

- The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance.
- Logistic Regression achieved an F1-Score of 0.868, indicating a good balance between precision and recall.
- Naive Bayes had an F1-Score of 0.865, slightly lower than Logistic Regression.
- Decision Tree had the lowest F1-Score of 0.772.

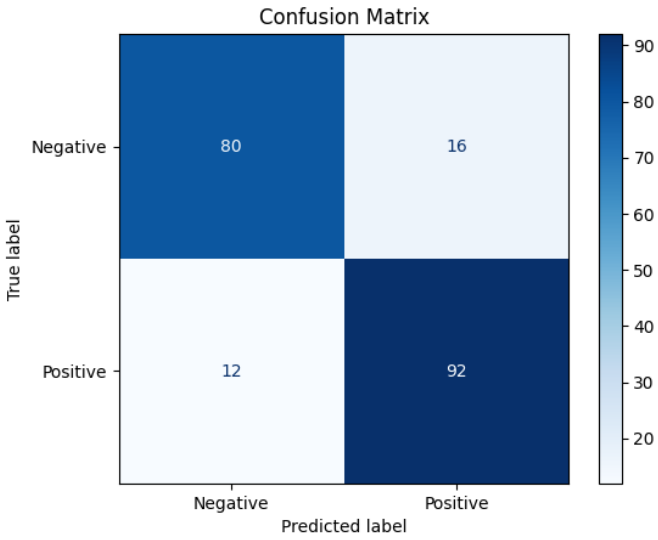

Fig - 3: Logistic Regression Confusion Matrix.

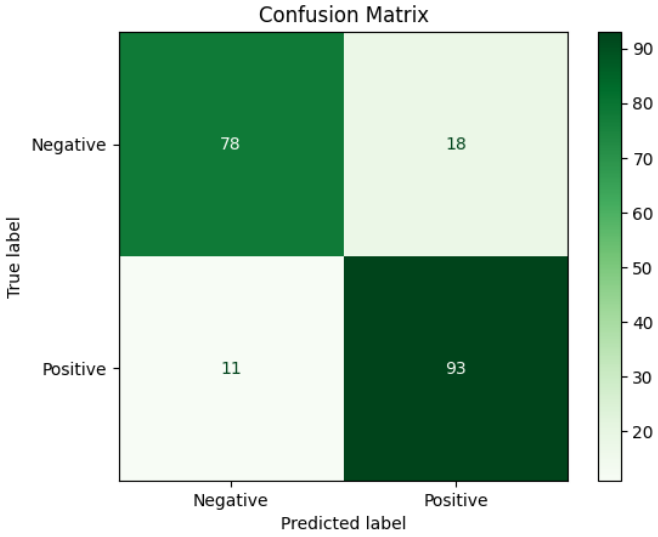

Fig - 4 Naïve Bayes Confusion Matrix.

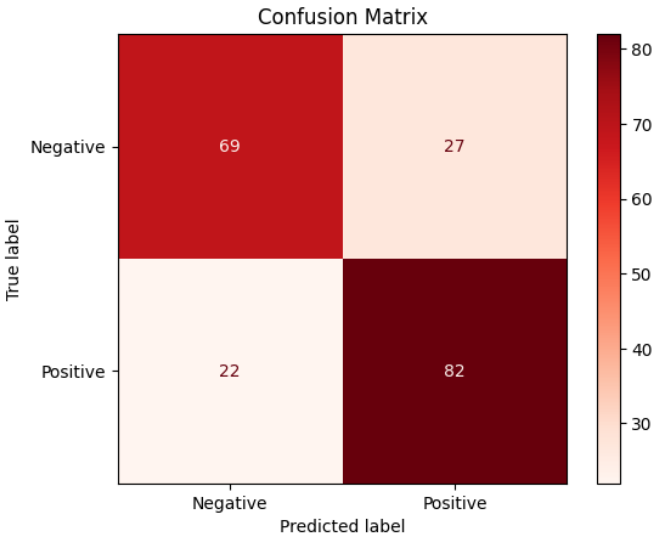

Fig - 5: Decision Tree Confusion Matrix.

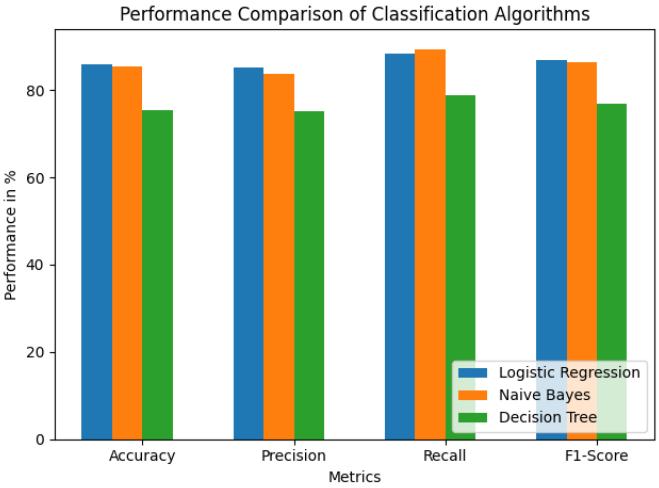

Fig - 6: Algorithm Comparison.

Based on these results, we can conclude that Logistic Regression performed relatively well overall, with high accuracy, precision, recall, and F1-Score. Naive Bayes also showed strong performance, especially in terms of recall. However, the Decision Tree classifier had lower performance compared to the other two algorithms, with lower accuracy, precision, recall, and F1-Score. These insights can help in selecting the most appropriate algorithm for the specific classification task and understanding the trade-offs between different evaluation metrics.

E. Principle Component Analysis (PCA)

PCA is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional space while retaining the most important information. PCA was used in simplifying our dataset and extracting the most important information, enabling more efficient and meaningful analysis.

- PC1 and PC2 collectively account for 31.7% of the total variance in the data, indicating that these two principal components capture a significant portion of the data's variability.
- The variance values of the principal components decrease as we move from PC1 to PC19, with PC1 having the highest variance and PC16 to PC19 having variances of 0.0, suggesting minimal contribution to the variability.

Performance Metrics with Increasing Principal Components:

- Accuracy, precision, recall, and F1-scores were evaluated for models trained using 1 to 19 principal components.
- The performance metrics exhibit stabilization or minimal fluctuations after considering a certain number of principal components.
- Notably, PC2 consistently yields the best results across all the performance metrics, indicating that it carries crucial information for the classification task.
- PC7 follows closely as the second-best principal component, reinforcing its significance in capturing important patterns in the data.

Implications of Achieving the Best Results with Just Two Principal Components:

- It is interesting and noteworthy that the best results are obtained using only PC1 and PC2, which collectively account for a relatively small portion of the total variance.
- This suggests that the data contains strong patterns or structures that are well-captured by these two principal components.
- The fact that such a small subset of the principal components achieves the best results implies that the remaining principal components may not provide significant additional discriminatory power for the classification task.
- It is possible that the data's intrinsic complexity is effectively represented in the first two principal components, implying a high level of dimensionality reduction potential and potential simplification of the classification model.

Overall, this analysis highlights the importance of PC1 and PC2 as the primary contributor to achieving the best results across the performance metrics. Additionally, the ability to achieve optimal results using just two principal components suggests that the data has a relatively low-dimensional structure with strong discriminatory patterns. This finding opens possibilities for more efficient modeling approaches, reduced computational complexity, and improved interpretability while maintaining high performance.

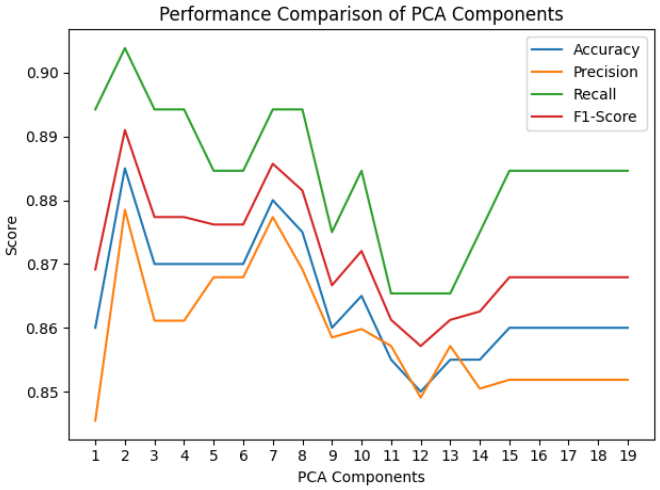

Fig - 5: PCA Comparison.

F. *Neural Network*

A neural network is a method that teaches computers to process data in a way that is inspired by the human brain that are designed to recognize patterns. We will analyze if a neural network can help cluster and classify our data.

a) Experiement

In our experiment, we chose the PyTorch library to build a neural network for this classification task. The motivation was to use an available deep learning library that provides various optimizers and loss functions and an easy way to tune hyperparameters so we can arrive at the most accurate network. In this experiment we found that a neural network with 2 layers performed best with our data set of heart data. we ran the experiment with 8 - 14 neurons in the first layer and 2 neurons in the final layer with the LeakyReLU activation to predict the chances of heart failure.

b) Observation

We were able to achieve a top accuracy of 86% over the test data with 12 neurons in the first layer and with optimizer set as RMSProp.

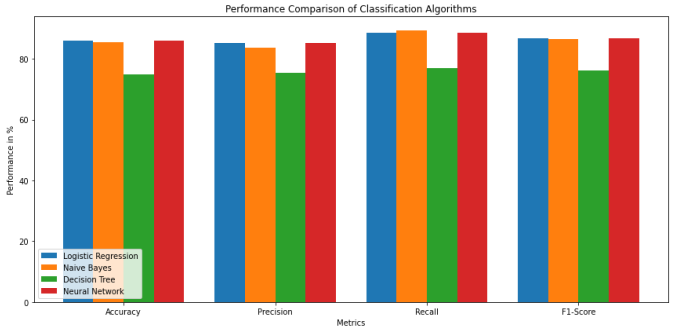

Fig - 7: Algorithm Comparison with Neural Network.

VII. Conclusion

Our investigation aimed to predict heart disease in patients using several different classification algorithms, including logistic regression, naïve bays, decision trees and neural networks. Through our analysis, we observed that logistic regression when projected onto the first two principal components obtained through PCA, yielded the best results in accuracy, precision, recall, and F1-score. Interestingly, logistic regression even slightly outperformed neural networks in this specific problem domain. This finding emphasizes the importance of carefully selecting appropriate algorithms, since more sophisticated models do not always yield better performance.

Furthermore, our model scored best in recall in the heart disease classification. The higher recall score indicates the model's ability to correctly identify positive instances of heart disease by minimizing false negatives. This ensures that individuals with the potential for heart disease are not overlooked. In this context, favoring recall over precision makes sense. Though our model already scored highly in recall, there could be future work to increase recall further at the expense of precision. This could be done by adjusting classification thresholds or incorporating additional data. Striking the right balance between recall and precision should be tailored to the specific objectives and requirements of the problem and our model does well at both.

Overall, our findings provide insights into the effectiveness of different algorithms, the significance of principal components, and the trade-offs between precision and recall in the classification of heart disease.

References

[1] Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[2] World Health Organization (June 2023) "Cardivascular Dieseases." Retrieved from https://www.who.int/health-topics/cardiovascular-diseases

[3] PyTorch (June 2023) “Tutorials” Retrieved from https://pytorch.org/tutorials/beginner/basics/quickstart_tutorial.html