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

# Introduction

Our goal for the project is to approach heart disease and compare different classification algorithms on a Heart Failure Prediction Dataset 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.

# 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. 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.

# 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.

## 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.

## 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.

## 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:

A picture containing black, darkness

Description automatically generated

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.

## Outline of the Proposed Work

A diagram of data processing

Description automatically generated with low confidence

**Fig -1**: Outline of the Proposed Work.

# 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*

.

# Evaluation Metrics

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

## 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%).

## 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 actually 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%).

## 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%).

## 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.

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.

## Principle Component Analysis (PCA)

E. **Principal Component Analysis (PCA):**

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

# Conclusion

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>.
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**