CSE422: Artificial Intelligence

SUMMER 2022

Group - 3

Section- 08

Report Writing

Member 1 : Apu Islam
ID : 20101430

Member 2 : G. M. Refatul Islam
ID : 20101482

Member 3 : Faiza Bushra
ID : 20101554

Member 4 : Tahsin Alam
ID : 19301171

Table of Contents

Introduction

Heart disease is a fatal condition, globally impacting millions of lives every year. From Congenital heart defects, to Heart Failure, the term "heart disease" covers a bigger spectrum of conditions. Reports by World Health Organization (WHO) suggest cardiovascular conditions are indeed the leading cause of death worldwide, with an estimation of 17.9 million deaths per year. Heart controls a fundamental functionality of the body that is the blood it pumps, has to suffice the need of other organs, for those to function properly. Any failure to this process can cause other organs to start dysfunctioning over time. The common blockages causing that are high cholesterol, diabetes, hypertension, obesity, increase in triglycerides levels, unhealthy diet etc. As symptoms vary from condition to condition for this multifaceted disease, the correct diagnosis of every one of those is crucial for it to not cause another death. Conventional invasive-based methods for diagnosing heart diseases are usually based on the patient's medical history, physical test results, and examining related symptoms.There are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg daily. A patient experiencing fatigue and irregular heartbeat can get diagnosed with both neurological conditions and a heart disease. Critically assessing all the indications can minimize the possible outcomes, resulting in an accurate diagnosis. Since the accuracy is vital, implementing intelligent learning-based computational techniques can make the process for effective heart disease diagnosis faster. There are many open sources for accessing the patient's records and research can be conducted so that various computer technologies could be used for doing the correct diagnosis of the patients and detect this disease to stop it from becoming fatal. Nowadays it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. We can use different machine learning and deep learning models to diagnose the disease and classify or predict the results. Applying Machine Learning algorithms for the detection of heart disease can be proven revolutionary for the healthcare industry because trained ML would advance faster over time. In our study, we are using different machine learning models to assess and predict the possibility of an individual having a heart condition, outcome scaling from 0 to 1. A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge of pandemic predictions and also medical records can be transformed and analyzed more deeply for better predictions. As early detection of heart diseases has higher chances of faster recovery and lower mortality rate, this can be a step in the right direction for resolving the global phenomena.

Methodology

Dataset Description:

In this study we use the publicly available heart disease prediction dataset from the UCI machine learning Repository which contains a large collection of datasets that have been widely used by the machine learning community. The datasets are related to the various types of heart disease. These datasets can be viewed as classification or regression tasks. The classes are ordered. The heart disease dataset contains 1100 instances. Input features are based on the report of different cases and output variables based on the sensory data are scaled in 4 quality classes.

Pre-processing Techniques:

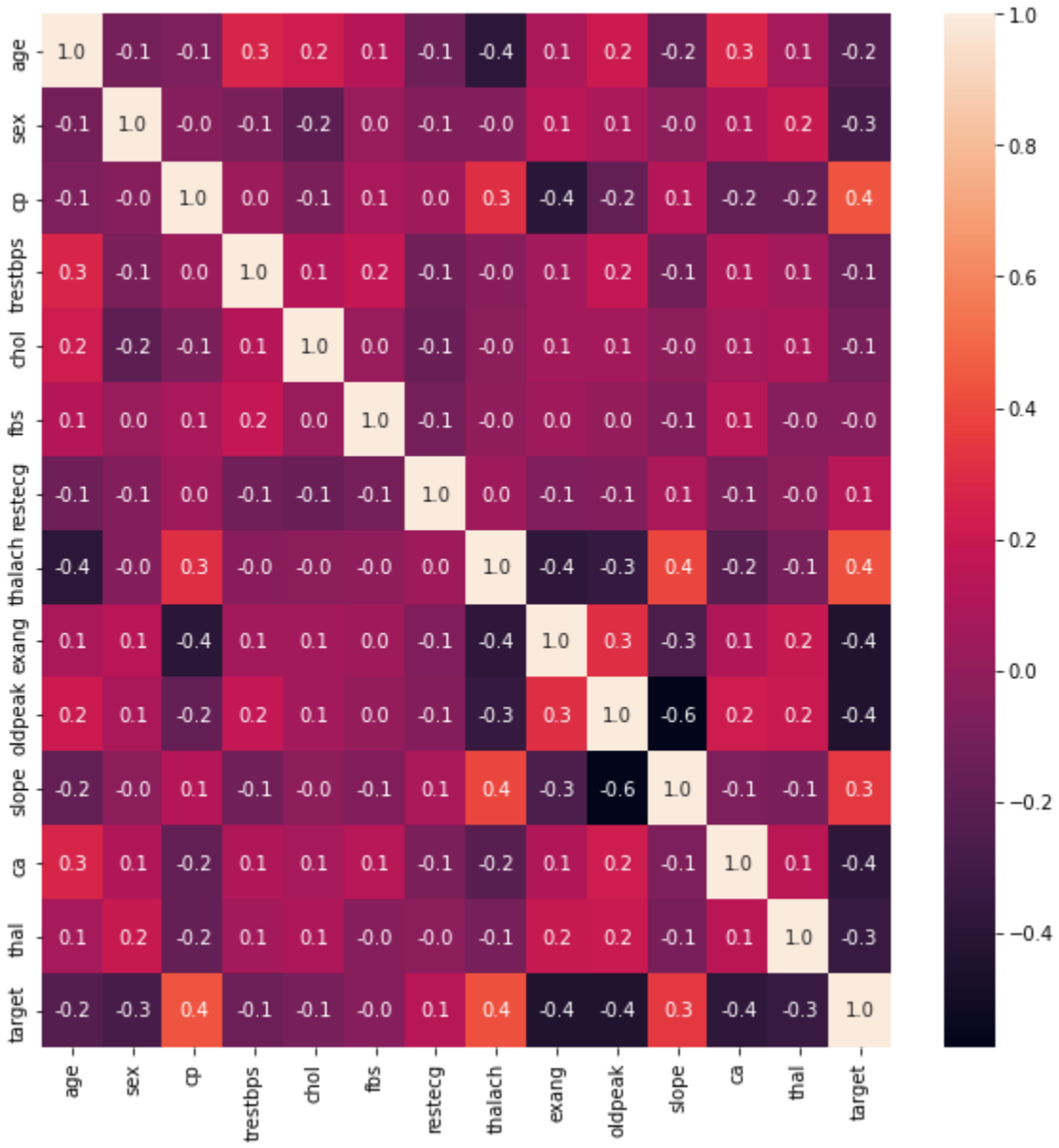
Handling Missing values:

As the ML algorithm processes datasets for improving accuracy, faulty or incomplete data requires fixing before analyzing it. This is called data cleaning which handles missing values before testing. It is a much required process because dataset retrieved for regular scenarios may have them in abundance. If not troubleshooted, most ML models will give an error if null values are pushed through the system.One of the ways to troubleshoot that is to replace the missing value with the mean value of that particular feature. In context of the dataset we are using, having null values for features like resting blood pressure (“trestbps”) , serum cholesterol (“chol”) or other features with non-binary inputs will be resorted by by its (feature) mean value.

Feature Selection:

Only a small portion of the dataset's variables may be used to create a machine learning model; the others are either redundant or useless. The overall performance and accuracy of the model may decrease if all these redundant and pointless features are included in the dataset. In order to remove the unnecessary or less significant features from the data, it is crucial to discover and choose the most appropriate features from the data, which is accomplished with the aid of feature selection in machine learning. Feature selection is the method of selection of the best subset of features that will be used for classification. In this study, for a better understanding of the features and to examine the correlation between the features, the Pearson correlation coefficient is calculated for each feature which shows the pairwise Pearson correlation coefficient. The range of the correlation coefficient from -1 to1. Point 1 value implies a linear equation describing the correlation between X and Y strong positive, which is all data points are lying on a line for Y

increases as X increases. Point -1 value indicates strong negative correlations between data points. All data points lie on a line in which Y decreases as X increases and point 0 indicates that there is an absence of correlation between the points. Here is our correlation heat map looks like:



Encoding Categorical Features:

A process of converting categorical features into integer format for easier generalization and implementation with different models, is applied in machine learning. Since datasets can feature

data in formats that are written for better human readability, analyzing that data becomes more taxing with every usage. Even though many improved machine learning algorithms support categorical value without maneuvering that, there are algorithms that require its conversion into numeric form for ML to decide how the labels must operate.In our dataset, we used fasting blood sugar (fbs) feature as categorical that represents if the fasting blood sugar is greater than 120 mg/dl or not. If it is, the value would get converted into "1" , else it would get converted into a zero ("0").

Feature Scaling:

Feature scaling is a technique used in datasets, to standardize independent variables in a specific range. ML models map and traces variables for determining the output variable. The scaling and tracing outcome could differ from domain to domain, based on the use of different units and scales, which in return causes certain variables to dominate over the others, resulting in a flawed outcome. Minmax Scaler is used to normalize such data by putting the variables in the same range and scale. This normalized value would range from 0 to 1, making all the variables carry equal priority. The formula for Normalization (Minmax scaler) goes:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Applied Models

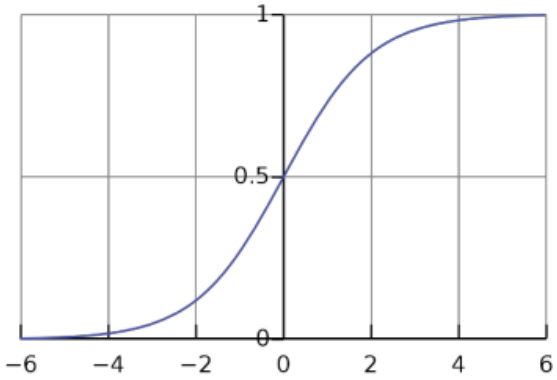
Logistic regression:

Logistic regression is a supervised learning classification algorithm which is used to predict the probability of a target variable and generally it indicates binary regression which targets binary target variables and a dependent variable always has two possible types either 0 or 1, true or false, success or failure. In multinomial, a dependent variable may have 3 or more unordered types, meaning they can represent type 1 or 2 or 3. Then again, the last one is ordinal logistic regression where a dependent variable can have 3 or more ordered types having quantitative significance for example, good, bad, excellent etc.

Logistic regression is very much similar to linear regression but how they are implemented or used in real life is where they differ. Linear regression is mainly used for regression problems, whereas logistic regression is used for classification problems.

Advantages of Logistic regression model are:

- It is easier to implement,
- It is very efficient to train,
- makes no assumptions about the distribution of classes in future space.
- It can easily extend to multiple classes



$$P(Y|X) = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Decision Trees:

Decision Trees are a type of Supervised Machine Learning algorithm where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

There are usually two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

There are two mainly types of decision trees:

- Classification trees
- Regression trees

In this report, we are using classification trees, where the outcome was a variable like ‘yes’ or ‘no’ and the decision variable is always categorical.

Entropy:

Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Attribute Selection Measures:

Attribute selection measure is used to select the best attribute for the root node and for sub-nodes. There are two popular techniques for ASM:

- Information Gain
- Gini Index

Information Gain = Entropy(S)- [(Weighted Avg) *Entropy (each feature)]

$$GINI\ Index = 1 - \sum(P(x = k))^2$$

Naive Bayes:

Naive Bayes algorithm is a supervised learning algorithm which takes the concept of Bayes Theorem and uses it to solve classification problems just like logistic regression. It is one of the simplest yet most effective classification algorithms that help build fast machine learning models that can predict fast predictions. The fundamental Naive Bayes assumption is that each feature makes an independent and equal contribution to the outcome. The assumptions made by Naive Bayes are not generally correct in real-world situations. In-fact, the independence assumption is never correct but often works well in practice.

Advantages of Naive Bayes classifier includes the following:

- Fast and easy ML algorithms to predict class of datasets,
- Can be used for binary or multi class classification,
- Performs well in multiclass classification compared to other algorithms,
- Most popular for text classification concerns.

There are three types of Naive Bayes models:

- Gaussian
- Multinomial
- Bernoulli

In this study, we are using the Gaussian model, which follows a normal distribution meaning it predicts continuous values instead of discrete. The likelihood of the features is assumed to be Gaussian, hence conditional probability is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

Random Forest Classifier:

Random forest is a supervised machine learning algorithm commonly used in classification and regression problems. It builds decision trees on different samples taken from datasets and takes their majority vote for classification and take average. One of the most important feature of random forest algorithm is its ability to process datasets containing continuous variables as in regression and datasets containing categorical variables as in classification. Because a random forest combines multiple trees to predict classes in a dataset, it is possible that some decision trees predict the correct output and others do not. But together all the trees predict the correct output. Advantages of using Random forest algorithm is that:

- Takes less training time compared to other algorithms,
- It also predicts with higher accuracy even for the larger datasets which sometimes miss large proportions of data,
- It is capable of performing both classification and regression tasks,
- It can also handle datasets with higher volume containing high dimensionality.

In short, Random forest algorithm sort of takes the wisdom of the crowd and gets the average best possible voted result and gives it to the machine.

Results

Accuracy:

The proportion of accurately predicted observations to all observations is known as the accuracy ration. Accuracy can be determined by dividing the total number of predictions by the number of predictions by the number of right predictions. We can calculate Accuracy use the given formulation below,

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

Precision:

The proportion of accurately predicted positive observations to all expected positive observations is referred to as precision. We can find precision using the following formulation below,

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall:

The proportion of accurately predicted positive observations to all the actual class observations is known as recall. We can calculate Recall as,

$$Recall = \frac{True\ Positive}{True\ Positive + False\ negative}$$

F1 Score:

The F1 score is a weighted harmonic mean of precision and recall, such that the best score is 1.0 and the worst score is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. We can calculate F1 scores as,

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Support:

Support is the number of actual occurrences of the class in the specified dataset, Support helps to determine structural flaws in classifier reported scores. As the requirement for stratified sampling

or rebalancing can be indicated by unbalanced Support in the training data. Support remains constant across models.

Result from Logistic Regression:

Logistic Regression classifier had an accuracy of 80%

Target	Precision	Recall	F1 Score	Support
0	0.86	0.72	0.78	100
1	0.77	0.89	0.82	105

Result from Naive Bayes Theorem:

Naive Bayes classifier had an accuracy of 78%

Target	Precision	Recall	F1 Score	Support
0	0.79	0.75	0.77	100
1	0.77	0.81	0.79	105

Result from Decision Tree Classifier:

Decision Tree classifier had an accuracy of 80%

Target	Precision	Recall	F1 Score	Support
0	1.00	1.00	1.00	100
1	1.00	1.00	1.00	105

Result from Random Forest Classifier:

Random Forest classifier had an accuracy of 100%

Target	Precision	Recall	F1 Score	Support
0	1.00	1.00	1.00	100
1	1.00	1.00	1.00	105

It is observable that from Random Forest classifier and Decision Tree classifier, we have obtained 100% accuracy which is highest among all the classifiers. The accuracy of the Logistic Regression classifier was 80%. We have got least accuracy from Naive Bayes classifier which is 78%.

References:

1. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download

2. https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm#:~:text=Introduction%20Decision%20Trees%20are%20a,namely%20decision%20nodes%20and%20leaves.
3. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
4. https://www.javatpoint.com/logistic-regression-in-machine-learning
5. https://www.javatpoint.com/machine-learning-naive-bayes-classifier#:~:text=Na%C3%AFve%20Bayes%20Classifier%20is%20one,the%20probability%20of%20an%20object
6. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random%20forest%20is%20a%20Supervised,average%20in%20case%20of%20regression.
7. https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/