TERM PROJECT: Forecasting Heart Disease with Machine Learning Algorithms

Sereiwathna Ros

Abstract—**The goal of this research project is to develop a machine learning-based artificial intelligence system for the identification of heart disease. We demonstrate how machine learning may be used to forecast if someone would get heart disease. A python-based application is created for healthcare research in this study since there are many available libraries to work with machine learning. Our primary objective this time is to determine if a person will be impacted by a Savior heart conditions or not. Data mining technology is the technique of extracting information from a large amount of data. Today, data mining is widely used in many facets of human existence. Data mining has many different and varied applications. A significant use of data mining is in the field of health care. We will be examining the heart disease dataset from which we will gain numerous insights that assist us know the weightage of each factor and how they are associated to one another.**

I. Introduction

THE Cardiovascular disorders are frequently substituted for heart diseases. These illnesses are generally associated with blood vessel blockages or narrowing, which can lead to heart attacks, strokes, and angina. Heart disorders come in a variety of forms, including those that damage the heart's rhythm, valve, or muscle. Machine learning, on the other hand, is essential for evaluating whether anybody has had cardiac disease. In either event, if they are anticipated in advance, physicians would find it much simpler to gather vital information for diagnosing and treating patients. The most mistaken sign of coronary artery disease is heart disease.

Our applications will be mainly based on Python Programming Language since there are many available libraries such Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn and XGBoost which have been used in this research project.

Several hospitals and patients are currently providing information to the health care industry. Doctors may quickly predict improved treatment procedures and improve the entire delivery system of the healthcare sectors by making the best use of this data . The Python framework's ability to assist make sense of and stimulate computational facilities in extracting vital insights from data related to the health care sectors is one of its most significant uses. Additionally, Python is regarded as one of the most well-known programming languages in the world.

A. Research Aim

The purpose of this research project is to use the machine learning algorithms to forecast whether a patient is having a heart disease or not.

B. Research Objectives

To conduct a rigorous analysis of how heart disease is detected using machine learning methods.

To critically analyze the earlier actions and use a proper methodological strategy for superscribing the found issue.

To effectively use the python library with data interpretation techniques for forecasting health issues.

II. Literature Review

(H. Jindal et al., 2020) trained some machine learning models including **Decision Tree**, **Support Vector Machine**, **K-Nearest Neighbors**, **Random Forest**, and **Logistic Regression** for heart disease detection where among those models the ones which perform best are **K-Nearest Neighbor** and **Logistic Regression** with the accuracy of **88.5%** respectively with the dataset of UCI Machine Learning Repository[1] with the dataset of UCI Machine Learning Repository. (M. Pal & S. Parija, 2019) conducted the research about heart disease by using **Random Forest** algorithm in which the model reaches an accuracy of **86.9%** [2] with the dataset of UCI Machine Learning Repository. (V. et al., 2022) made many machine learning models for heart diseases detection with algorithms including **K-Nearest Neighbor** with an accuracy of **87%**, **Decision Tree** with an accuracy of **79%**, and **Support Vector Machine** with an accuracy of **83%** **Random Forest** with an accuracy of **84%** [3] with the dataset of UCI Machine Learning Repository. (N. Kumar et al., 2021) researched about machine learning models for healthcare with estimating a person's risk of COVID-19, heart disease, and diabetes by asking them a few questions about their travel history, age, gender, and blood pressure. They proposed a machine learning model which outperforms other models with the accuracy of **96%** (approximately) for diabetes dataset, **95&** (approximately) for COVID-19 dataset, and **88%** (approximately) for heart diseases dataset [4]. A hybrid breast cancer detection system was created by (M.S. Uzer, O. Inan and N. Yilmaz, 2013). The classification technique employed was the artificial neural network. Cancer is a condition when the body's cells stop functioning and begin to grow and divide uncontrolled. Cancer is a condition in which body cells stop functioning normally and start to grow uncontrolled. The leading cause of mortality worldwide is cancer, which is based on malignant tumors. The hybrid function selection technique, which the authors used, is based on **Sequential Forward Search**, **Sequential Backward Selection**, and **Principal Component Analysis** (PCA). A ten-fold cross validation methodology was also used to obtain the findings, which had an accuracy rate of **98.57%** [5]. The study of (M.R. Senapati, A.K. Mohanty, S. Dash, and

P.K. Dash, 2015) the usage of a local linear neural wavelet network to identify breast cancer by refining the network's parameters using the Recursive Least Squares (RLS) method to enhance each component's performance. The findings showed that the suggested technique is quite effective and offers a good classification and were estimated and compared with other studies [6]. To estimate the risk of type II diabetes, (M.A. Sapon et al., 2011) examined three algorithms and created prediction models. When the pancreas is unable to create enough insulin, type II diabetes develops, which results in abnormal blood sugar levels. **Artificial neural networks** (ANN), **Naïve Bayes**, and **K-nearest Neighbor** (KNN) were the three techniques employed. The outcomes demonstrated that the **neural network**, with a prediction accuracy of **96%**, outperformed the **Naïve Bayes**, with a prediction accuracy of **95%**, and the **K-Nearest Neighbor**, with a prediction accuracy of **91%** [7]. (A. Charleonnan et al., 2016) trained four machine learning techniques to predict chronic kidney disease, including the **Support Vector Machine**, **Decision Tree**, **Logistic Regression**, and **K-Nearest Neighbors**. In order to select the best classifier, the effectiveness of these models for predicting chronic kidney disease was compared to one another. According on the experimental findings, the **SVM** classifier has a maximum accuracy of **98.3%**. **SVM** also has the maximum sensitivity following training and testing utilizing the suggested methodology. It can therefore be inferredthe **SVM** classifier is appropriate for the forecast of chronic kidney disease [8]. (H. Elshazly et al., 2013) has provided a hybrid method for lymphatic illness prognosis that combines the genetic algorithm with the random forest. Fluids are transported throughout the body through the lymphatic system, which is a component of the immune system.Fluids are transported throughout the body through the lymphatic system, which is a component of the immune system. The lymphoma dataset size was reduced using a genetic approach, and the **Random Forest** method was applied as a classifier. Performance of the proposed system was contrasted with existing feature selection techniques combined with an **Random Forest** classifier, such as **Principal Component Analysis** (PCA) and etc. According to the study's findings, the model was able to classify objects with an accuracy of roughly **92.20%** [9]. By using algorithms for machine learning. (C. Boukhatem et al., 2022) introduces a number of machine learning techniques for heart disease prediction that make use of patient data on key health indicators. In order to construct the prediction models, the study exhibited four classification techniques: **Multilayer Perceptron** (MLP), **Support Vector Machine** (SVM), **Random Forest** (RF), and **Naïve Bayes** (NB). On the basis of accuracy, precision, recall, and F1-score, the models were assessed. The **SVM** model has the highest accuracy, at **91.67%** [10].

III. Methodology

In order to implement the heart disease detection, we use methods such as **Standard Score** [11] to scale our data and for machine learning algorithms such as **Logistic Regression** [12], **Naïve Bayes** [13], **Decision Tree** [14], and **Gradient Boosting** [15] to predict the heart disease of patients. In order to evaluate the models prediction, **Precision**, **Recall**, **F1-score**, and **Accuracy** are used.

Standard Score A raw score x is transformed into a standard score z by if the population mean and standard deviation are known:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where

μ is the mean of the population,
σ is the standard deviation of the population.

Logistic Regression is estimating the parameters of a logistic model in which it is given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

where

$z = \sum_{i=0}^{d} w_i z_i$ is the linear combination of the features
w_i is the weight parameter,
z_i is the value of the feature,
d is the number of features (dimensions).

Naïve Bayes is a technique for applying the Bayes' theorem in which it makes the "naïve" assumption that every pair of features is conditionally independent given the value of the class variable. Given the class variable y and the dependent feature vectors x_1 through x_n, the Bayes' theorem establishes the following relationship:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{3}$$

where

$P(y)$ is the probability of the class y,
$P(x_i|y)$ is the probability of the feature x_i given y

Decision Tree is learning straightforward decision rules derived from the data characteristics in which it aims to develop a model that forecasts the value of a target variable.

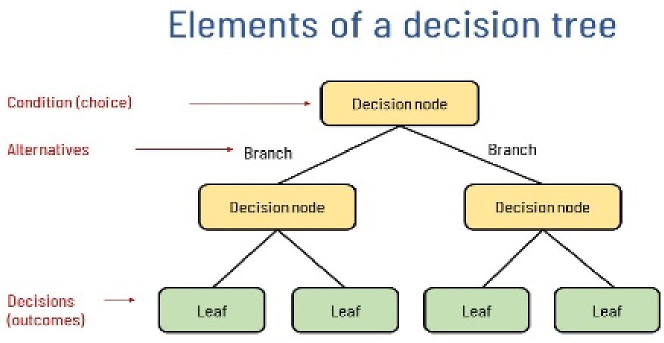

Fig. 1. Diagram of Decision Tree

Gradient Boosting is the model which has been provided by several models in the form of an ensemble of weak prediction models, most often decision trees.

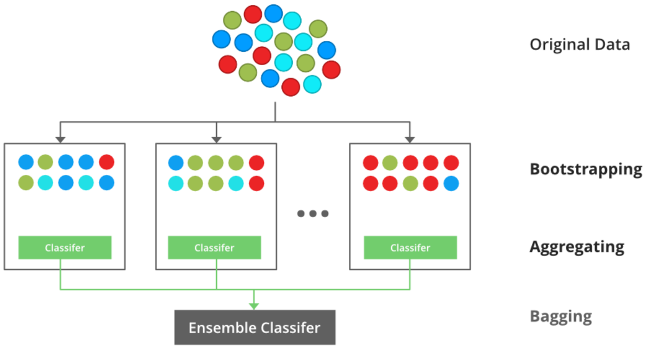

Fig. 2. Diagram of Gradient Boosting

Precision calculates the number of predictions from the positive class that are members of the positive class.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall calculates how many class predictions have been predicted that are positive based on all of the positive cases in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-score gives a single score that strikes a balance between recall and precision.

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6}$$

Accuracy quantifies the number of correct predictions divide by the all number of examples.

$$Accuracy = \frac{TP + TN}{TP + TP + FP + FN} \tag{7}$$

IV. PROPOSED SOLUTION

First, we take the dataset from UCI machine learning repository containing 303 data point where each of them consisting of 14 features. By splitting 85% for the training set and 15% for the test set, then we train and test them with the algorithms including **Logistic Regression**, **Naïve Bayes**, **Decision Tree**, and **Gradient Boosting**. Our evaluation metric results are:

TABLE I
PRECISION, RECALL, F1-SCORE, AND RESULT FOR THE TRAINING SET

Model	Precision	Recall	F1-Score	Accuracy
LR	85%	83%	84%	84%
NB	84%	84%	84%	84%
DT	100%	100%	100%	100%
GB	97%	97%	97%	97%

TABLE II
PRECISION, RECALL, F1-SCORE, AND RESULT FOR THE TEST SET

Model	Precision	Recall	F1-Score	Accuracy
LR	87%	87%	87%	87%
NB	80%	80%	80%	80%
DT	85%	85%	85%	85%
GB	80%	81%	80%	80%

V. CONCLUSION

In this challenge we have followed the normal workflow of a machine learning project. Moreover, during this work many techniques and 4 different classification models were applied, their advantages and disadvantages were also discussed. At the end, the best result that we obtained is by using **Logistic Regression** with **87%** accuracy score with the dataset from UCI Machine Learning Repository.

However, there are several Machine Learning techniques we would like to test here if we have more time. For example, Stacking, other Ensemble methods, can combine the strength of several tree family methods. We have test it with Decision Tree and Gradient Boosting which yields Accuracy score **85%** and **80%**. Furthermore, there are several tuning parameters which we haven't tried yet. We wish to try a lot of fine-tuning parameters and improve the further more in the future.

REFERENCES

[1] H. Jindal et al., 2020. Heart disease prediction using machine learning algorithms. IOP Conf. Series: Materials Science and Engineering.

[2] M. Pal S. Parija, 2020. Prediction of Heart Diseases using Random Forest. ICCIEA 2020. Journal of Physics: Conference Series.

[3] V. Chang et al., 2022. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analysics, Volume2.

[4] N. Kumar et al., 2021. Efficient Automated Disease Diagnosis Using Machine Learning Models. Journal of Healthcare Engineering Volume 2021, Article ID 9983652.

[5] M.S. Uzer, O. Inan and N. Yilmaz, 2012. A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS, and PCA. Springer Journal of Neural Computing and application, 2012.

[6] M.R. Senapati, A.K. Mohanty, S. Dash, and P.K. Dash, 2015. Local linear wavelet neural network for breast cancer recognition. Neural Computing and Applications,Volume 22, Issue 1, 2013, pp. 125-131.

[7] M.A. Sapon et al., 2011. Diabetes Prediction with Supervised Learning Algorithms of Artificial Neural Network. International Conference on Software and Computer Applications, Kathmandu, Nepal, 2011.

[8] A. Charleonnan et al., 2016. Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques, Proc. Management and Innovation Technology International Conference (MITiCON-2016), IEEE, Oct. 2016.

[9] H. Elshazly et al., 2013. Hybrid System for Lymphatic Diseases Diagnosis. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013.

[10] C. Boukhatem et al., 2022. An Analysis of Heart Disease Prediction using Machine Learning and Deep Learning Techniques. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp.1484-1491, 2022.

[11] E. Kreyszig, 1979. Advanced Engineering Mathematics (Fourth ed.). Wiley. p. 880, eq. 5. ISBN 0-471-02140-7.

[12] D. Jurafsky & J.H. Martin, 2022. Speech and Language Processing (Third ed.). p. 78.
[13] D. Jurafsky & J.H. Martin, 2022. Speech and Language Processing (Third ed.). p. 57.
[14] Breiman et al., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.
[15] A. Natekin & A. Knoll, 2013. Gradient boosting machines, a tutorial. Front. Neurorobot.