Heart Disease Prediction

1st Showmick Kar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
showmick.kar@g.bracu.ac.bd
ID:20301177

2nd Sajidul Islam Khandaker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
sajidul.islam.khandaker@g.bracu.ac.bd
ID:20301190

3rd Tahmina Talukdar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
tahmina.talukdar@g.bracu.ac.bd
ID:20301412

4th Md. Shakil Anawar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
md.shakil.anawar@g.bracu.ac.bd
ID:20301162

5thHumaion Kabir Mehedi
Dept. of CSE
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

6th Md. Farhadul Islam
Dept. of CSE
BRAC University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

7th Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Heart disease, or cardiovascular disease (CVD), is a leading cause of death globally, resulting in millions of fatalities each year. Atherosclerosis, characterized by the accumulation of fatty deposits in the arteries, is a major contributor to CVD. This paper explores the application of machine learning, a subfield of computer science, in predicting cardiac disease by leveraging an individual's medical history and behavioral features. By analyzing past data and training predictive models, accurate predictions can be made regarding the presence of cardiovascular problems in individuals. This research presents a significant advancement in medical science, offering the potential for remote diagnostics and early identification of individuals at risk for heart disease. By identifying high-risk individuals, interventions can be implemented to save lives and improve patient outcomes.**

***Index Terms*—Sentiment Analysis, BERTweet, Hate Speech**

I. Introduction

Cardiovascular disease (CVD), commonly known as heart disease, is a pervasive health issue that disrupts the normal functioning of the heart and poses significant risks to overall well-being. According to the World Health Organization (WHO), CVD accounts for a staggering 17.9 million deaths annually, making it the leading cause of mortality worldwide. This complex array of conditions encompasses disorders affecting both the heart and the blood vessels, with atherosclerosis standing out as a major contributor. Atherosclerosis involves the accumulation of fatty deposits within the arteries, increasing the likelihood of blood clots and compromising blood flow to vital organs. The consequences of CVD manifest in various forms, with coronary heart disease, stroke, peripheral artery disease, and aortic disease being the most prevalent types recognized by England's National Health Service (NHS).

Despite extensive research efforts directed towards developing effective treatments and preventive measures, heart disease continues to exert a significant toll on global health. Traditional approaches to diagnosis and risk assessment involve invasive procedures, costly diagnostic tests, and reliance on subjective clinical judgment. However, recent advancements in the field of computer science, particularly in machine learning and artificial intelligence (AI), offer promising avenues for transforming cardiac care and enhancing patient outcomes.

Machine learning, a subfield of computer science fueled by AI, focuses on leveraging data and algorithms to model the acquisition of knowledge, aiming for ever-improving levels of accuracy. By analyzing historical data, machine learning algorithms can learn patterns, detect trends, and generate predictions for future outcomes. This ability holds tremendous potential for accurately predicting cardiac diseases based on an individual's medical history and behavioral characteristics.

One compelling application of machine learning in the medical sciences is the prediction of cardiac disease in individuals, enabling the identification of high-risk individuals for timely intervention and preventive measures. By analyzing comprehensive datasets encompassing medical histories, physiological parameters, lifestyle factors, and genetic predispositions, machine learning algorithms can extract intricate patterns and create predictive models. These models can then be used to assess an individual's risk of developing heart disease, potentially revolutionizing diagnostics by enabling remote and non-invasive prediction methods.

Remote diagnostics, facilitated by machine learning algorithms, offer an innovative approach to cardiac care. Instead of relying solely on real-time medical testing, remote diagnostics

empower healthcare professionals to utilize readily available data and predictive models to evaluate an individual's risk of cardiovascular problems. By comparing an individual's medical history with those of previously diagnosed heart disease patients, the machine learning models can identify subtle risk factors and facilitate early intervention. This transformative shift towards proactive and personalized healthcare has the potential to save lives, optimize resource allocation, and improve patient outcomes.

In this paper, we delve into the realm of machine learning applications in predicting cardiac disease, exploring the intricacies of leveraging historical medical data and behavioral features for accurate risk assessment. By shedding light on the potential of remote diagnostics in cardiac care, we aim to contribute to the ongoing efforts towards preventing, diagnosing, and managing heart disease. Through advancements in machine learning and AI, we strive to improve patient outcomes, reduce healthcare costs, and enhance the overall quality of cardiac care on a global scale.

II. MOTIVATION

Accurately predicting cardiovascular disease has long been a sought-after goal in both the fields of contemporary medicine and computer engineering. The inherent challenges associated with cardiac disease diagnosis, coupled with individuals' tendency to overlook critical risk factors, have made it difficult to achieve accurate and timely assessments. Moreover, the elusive nature of cardiac disease, often progressing silently and leading to severe consequences or even death without apparent symptoms, further compounds the challenge.

The pressing need to reduce avoidable deaths caused by cardiovascular disease fuels the search for an effective model or algorithm to address this formidable problem. Machine learning, with its ability to analyze data from previously diagnosed cardiac patients, presents an opportunity to make precise predictions about future individuals at risk. Leveraging the power of machine learning algorithms to draw meaningful insights and learn from past research has the potential to revolutionize the medical field. However, achieving the highest levels of precision necessitates the careful selection of appropriate modeling approaches and input data. Identifying patterns and similarities between healthy individuals and those with heart disease could unlock highly efficient solutions that have far-reaching implications for the global population, significantly improving countless lives.

Early prediction of heart disease, coupled with proactive interventions and the promotion of healthy lifestyles among individuals identified as high-risk, holds immense value for the medical community and the global population at large. Machine learning, through its capacity for early detection of cardiovascular diseases, can empower patients and their families with the opportunity to adopt preventive measures and make informed decisions regarding their health.

By harnessing the potential of machine learning in cardiovascular disease prediction, we can pave the way for enhanced healthcare outcomes and a higher quality of life for individuals worldwide. This research endeavor aims to contribute to the advancement of predictive models and algorithms that have the potential to transform cardiac care, reducing mortality rates, and enabling more proactive and personalized approaches to managing cardiovascular diseases.

III. DATASET PREPROCESSING

During the preprocessing phase, several issues were identified in the dataset that required attention and resolution to ensure reliable and accurate analysis.

Faults:

NULL Values: The dataset was found to contain null or missing values, which could potentially impact the quality and validity of the data.

Categorical Values: The dataset includes 14 columns with categorical values, which need to be appropriately handled to enable the utilization of machine learning algorithms.

Imbalanced/Biasness: A noticeable imbalance was observed in the distribution of the "HeartDisease" label, indicating a significant disparity between the number of heart-healthy and heart-unhealthy patients. This imbalance can introduce bias into the analysis and affect the performance of the machine learning models.

Solutions:

Dealing with NULL Values: To address the issue of null values, a common approach is to remove the rows or columns containing these missing values. In this preprocessing phase, we opted to drop the rows with null values, ensuring that only complete and reliable data points were retained for further analysis.

Encoding Categorical Values: To handle the categorical values present in the dataset, an encoding technique is applied to convert them into numerical representations. In our specific case, we utilized the Ordinal Encoder, which assigns numeric labels to the categorical variables while preserving the ordinal relationship between the categories. This transformation allows the machine learning models to process and interpret the categorical data effectively.

Balancing the Dataset: Addressing the issue of imbalanced data is crucial to prevent biased predictions. In our dataset, we employed a technique called random oversampling to address the class imbalance. Random oversampling involves replicating instances from the minority class to create a more balanced distribution between heart-healthy and heart-unhealthy patients. By increasing the representation of the minority class, the models can better learn patterns and make accurate predictions for both classes.

By implementing these preprocessing steps, including handling null values, encoding categorical variables, and balancing the dataset, we aimed to ensure the dataset's quality and optimize its suitability for subsequent analysis and model training. The resulting preprocessed dataset serves as a reliable foundation for developing robust machine learning models capable of predicting heart disease with enhanced accuracy and generalizability.

IV. FEATURE SCALING:

Feature scaling is a crucial step in the preprocessing pipeline that aims to normalize the range of feature values. It ensures that all features contribute equally to the machine learning algorithms, preventing biases and improving the models' performance. In this project, we addressed the need for feature scaling to facilitate effective training and accurate predictions. Why Feature Scaling is required: Different Magnitudes: The features in the dataset may have varying magnitudes and scales. Some features might have values in the range of thousands, while others could be in decimals or fractions. Such discrepancies in magnitudes can adversely impact certain algorithms that rely on distance-based calculations, such as K-nearest neighbors (KNN) or support vector machines (SVM). Convergence Speed: Gradient-based optimization algorithms, like gradient descent, tend to converge faster on datasets where features are on a similar scale. Scaling the features aids in achieving quicker convergence and more efficient model training. Solutions: To address the issues related to varying feature magnitudes, we applied feature scaling techniques to bring all features to a similar scale. Two commonly used methods for feature scaling are: Standardization (Z-score normalization): This technique transforms the features to have a mean of zero and a standard deviation of one. It achieves this by subtracting the mean of each feature and dividing by its standard deviation. Standardization ensures that the features follow a Gaussian distribution, which can be beneficial for certain algorithms like linear regression or logistic regression. Min-Max Scaling: This method scales the features to a specific range, often between 0 and 1. It achieves this by subtracting the minimum value and dividing by the range (maximum minus minimum). Min-max scaling is useful when maintaining the original distribution and preserving the relationships between data points is important. Both techniques help mitigate the impact of varying feature magnitudes and promote fair comparisons and meaningful interpretations of the data. In our project, we employed the Standardization (Z-score normalization) technique to scale the features. By standardizing the features, we ensured that they have zero mean and unit variance, making them more amenable to various machine learning algorithms. By applying feature scaling, we improved the convergence speed of optimization algorithms and created a level playing field for all features. This step enhances the reliability and accuracy of our machine learning models, enabling them to effectively learn patterns and make robust predictions based on the scaled features.

V. DATASET SPLITTING

The goal of a successful machine learning project is to predict accurate data by testing it after providing suitable data and training the algorithm. For this, we need to split our preprocessed data into train and test sets. Two types of splitting can be performed - stratified and random splitting. Here, we use random splitting to divide the data into train and test sets.

In our case, we have randomly split our preprocessed and encoded data into a train:test ratio of 80:20. This means that 80% of the data has been used for training our model, while 20% of the data has been reserved for testing the performance of our model.

VI. MODEL TRAINING

For our project, we employed three classification models to train our dataset. The models selected for this task are:

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm that classifies data points based on their proximity to neighboring data points. It assigns a label to a new data point based on the majority class among its k nearest neighbors. KNN is intuitive and easy to understand, making it a popular choice for classification tasks.

Logistic Regression: Logistic regression is a widely used statistical model that predicts the probability of binary outcomes. It models the relationship between the input variables and the binary response using the logistic function. Logistic regression is interpretable and well-suited for datasets with linearly separable classes.

Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm that finds an optimal hyperplane to separate data points into different classes. It aims to maximize the margin between classes, leading to robust and effective classification. SVM can handle complex decision boundaries and is particularly useful when dealing with high-dimensional data.

By training our dataset with these three models, we aim to explore their individual strengths and weaknesses in predicting heart disease based on the provided features. Each model brings its own set of assumptions and algorithms, allowing us to evaluate their performance and identify the most suitable approach for our specific problem.

Through extensive training and evaluation, we will assess the accuracy, precision, recall, and other relevant metrics of each model. This analysis will enable us to make informed decisions about the model selection and ultimately deploy the most effective and reliable solution for predicting heart disease.

.

VII. RESULT ANALYSIS

Accuracy Score: The accuracy score indicates the overall correctness of the predictions made by the models. In our evaluation, the SVM model achieved the highest accuracy score, while the KNN model had the lowest accuracy among the three models.

Precision Score: Precision measures the proportion of correctly predicted positive instances out of all predicted positive instances. In our analysis, the SVM model demonstrated the highest precision, indicating a lower rate of false positive predictions. On the other hand, the KNN model had the lowest precision score among the three models.

Recall Score: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances

correctly identified by the models. Interestingly, the KNN model achieved the highest recall score, indicating a lower rate of false negative predictions. Conversely, the SVM model had the lowest recall score.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation of a model's performance. In our assessment, the KNN model obtained the highest F1 score, reflecting a better balance between precision and recall. Meanwhile, the SVM model had the lowest F1 score.

Confusion Matrix: The confusion matrix provides an overview of the predictions made by the models in a matrix format. It consists of four quadrants: True Negative (TN), False Negative (FN), False Positive (FP), and True Positive (TP). The values in these quadrants represent the percentages or counts of the corresponding predictions made by the models.

By analyzing the confusion matrix, we can observe the distribution of correct and incorrect predictions. The top right part of the matrix represents the TN values, indicating cases where the models correctly predict the absence of heart disease. The top left part represents the FN values, where the models incorrectly predict the absence of heart disease despite the actual presence of the condition. The bottom left part shows the FP values, signifying cases where the models wrongly predict the presence of heart disease in individuals without the condition. Finally, the bottom right part represents the TP values, indicating cases where the models correctly predict the presence of heart disease.

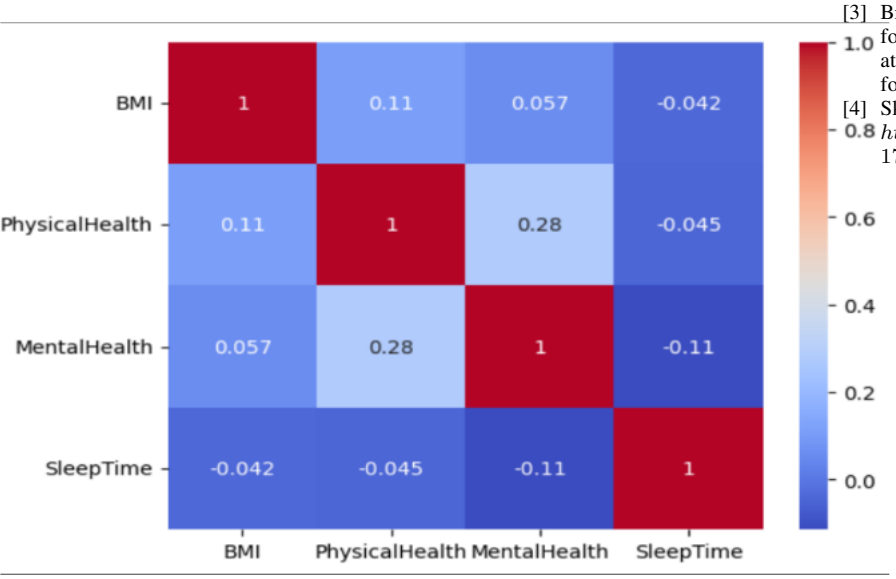

Fig. 1. Example of a figure caption.

Overall, the confusion matrix allows us to identify the type of errors made by the models and gain insights into their performance in classifying individuals with and without heart disease.

VIII. CONCLUSION

In this project, we employed three classification models, namely KNN, SVM, and Logistic Regression, to predict heart disease in a given dataset. After comparing the results based on various evaluation metrics such as accuracy, f1 score, error, recall, and precision, we observed that all models performed well, achieving an accuracy score of 80%.

Among the three models, SVM demonstrated the highest accuracy and precision, indicating its superior performance in making predictions about heart disease. Additionally, the Logistic Regression model also exhibited high precision values. These findings highlight the potential of these models as valuable tools for clinicians in anticipating and preventing cardiac disease, ultimately leading to improved patient health outcomes.

Overall, our analysis suggests that SVM stands out as the most effective model for predicting heart disease based on the given dataset. However, it is important to note that the performance of these models may vary depending on the specific dataset and context. Further research and validation with larger and diverse datasets would be beneficial to enhance the accuracy and reliability of these predictive models in real-world scenarios.

REFERENCES

[1] Documentation: Tutorial ¿ train models with scikit-learn (no date) Palantir. Available at: https://www.palantir.com/docs/foundry/model-integration/tutorial-train-model/ (Accessed: 17 May 2023).

[2] Brownlee, J. (2021) Random oversampling and undersampling for imbalanced classification, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (Accessed: 17 May 2023).

[3] Brownlee, J. (2020) How to calculate precision, recall, and F-measure for imbalanced classification, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ (Accessed: 17 May 2023).

[4] Sklearn.neural$_n etwork.MLPClassifier(nodate)scikit.Availableat$: $https://scikit-learn.org/stable/modules/generated/sklearn.neural_n etwo$ $17May2023$).