IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 11, Issue 6, November-December 2024

Impact Factor: 7.394

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

A Machine Learning-Based Web Application for Cardiac Disease Prediction

Prabhavathi K[1*], Mareeswari V[2]

Research Scholar, AMC Engineering College, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[1]

Assistant Professor, Department of CSE, RV Institute of Technology and Management, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[1]

Professor and Head of Department, Department of CSE, AMC Engineering College, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[2]

ABSTRACT: This study used predictive modelling approaches in machine learning (ML) to predict cardiac disease using a dataset from the US Centres for Disease Control and Prevention (CDC). The dataset was pre-processed and then used to train five machine learning models: random forest, support vector machine, logistic regression, extreme gradient boosting, and light gradient boosting. The goal was to employ the best performing model to create a web application that could consistently forecast heart disease based on user-supplied data. The extreme gradient boosting classifier produced the most reliable findings, with precision, recall, and F1-scores of 97%, 72%, and 83% for Class 0 (no heart disease) and 21%, 81%, and 34% for Class 1 (heart disease). The model was then deployed as

I. INTRODUCTION

Significant advancements in computing have recently been made in hardware and software, including the development of reliable artificial intelligence (AI) tools and systems. Artificial intelligence (AI) systems produce predictions and improve the prediction performance by applying different types of techniques, particularly machine learning (ML), to gain knowledge through data and/or experience. Because of its "ability," artificial intelligence (AI) is being used more widely in various fields, organisations, and sectors, including the health care sector. ML has shown to be helpful in providing the medical industry with an enormous platform that allows health care concerns to be resolved quickly and easily. [1] [2] [3] [4].

According to recent studies in the literature, decision trees (DT), logistic regression (LR), random forest (RF), support vector machines (SVM), and gradient boosting (GB) are among the models frequently employed for illness prediction. Few people have utilised other models such naïve bayes, bayesian networks, and neural networks (deep learning). In addition to the traditional models, other researchers, such as [9], have also used hybrid models, which combine many models with or without new methods. Predicting diabetes, as in [10] [11] [12], depression, as in [16] [17], hypertension, as in [18] [19], anxiety, as in [10] [20] [21] [22], and heart disease, as in [16] [17] [22] were the main topics of these investigations. Heart disease is a common health problem worldwide, and it's important to estimate your risk of developing heart disease.

The current study considered the previously cited works, particularly those from [10]. [10] employed a two-stage machine learning (ML) model (logistic regression and Evimp functions) to forecast the co-occurrence of cardiovascular disease (CVD) and diabetes mellitus (DM). In order to be tested and have their data collected, 2000 individuals who were over 40 years old had to fulfil certain prerequisitesTheir research produced outcomes that demonstrated the ML model's 94.09% prediction accuracy, 93.5% sensitivity, and 95.8% specificity in detecting the co-occurrence of DM and CVD. [22] employed two deep learning models (multilayer perceptrons and convolutional neural networks) and five machine learning (ML) models (LR, k-nearest neighbours—k-NN, RF, and Extreme Gradient Boosting) to predict eight significant chronic ailments, including cardiovascular disease.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

II. METHODOLOGY

1.2.1 Data Source

The Centres for CDC [34] in the US provided the initial dataset for this research. The Behavioural Risk Factor Surveillance System (BRFSS) is used by the CDC to gather data. The CDC, together with all of the states and participating U.S. territories, are working together on the BRFSS initiative. When the BRFSS was first launched in 1984, Monthly telephone interviews were conducted in 15 states to gather surveillance data on risky behaviours. The number of states taking part in the poll grew over time. These days, the District of Columbia, participating U.S. territories, and all 50 states participate in the BRFSS data collection process. Every state in the US, the District of Columbia, Guam, and Puerto Rico gathered BRFSS data in 2020.

With over 400,000 adult interviews completed annually, BRFSS is the world's largest continually operating health survey system. Heart disease is one of the leading causes of death in the United States for individuals of most races, according to the CDC. Smoking, high BP, and high cholesterol are the three biggest risk factors for heart disease that over half of all Americans (47%) have at least one of. Other important markers are having diabetes, being obese, not exercising enough, or consuming excessive amounts of alcohol. In the realm of medicine, it is vital to identify and avoid the factors that contribute most significantly to heart disease. As a result, developments in computers allow for the application of machine learning algorithms to identify "patterns" in data that suggest a patient's prognosis.

1.2.2 Data Preprocessing

The original dataset comprises 279 variables and 401,958 rows in SAS Transport Format. Python 3 was used to preprocess the data using the matplotlib, pandas, numpy, and sklearn modules. Handling missing or NaN values, feature selection, categorical data encoding, data standardisation, data splitting into train and test sets, and class weighting were all involved in the preparation.

Missing values (NaNs): Using the dataset's size to guarantee data integrity, rows with incomplete entries or missing values (NaNs) were eliminated throughout the data cleansing process.

Feature selection: A significant portion of the values in the majority of characteristics (columns) were missing, necessitating their complete removal. Important heart disease risk factors were incorporated in the chosen features.

Categorical data encoding: As Table 1 illustrates, most of the collected data consists of categorical variables. Label encoding was used to translate non-numeric categories into matching numerical representations in order to aid ML model interoperability. Binary categorical values, like "Yes" and "No," for example, were converted to 1 and 0, respectively.

Data standardisation: Many machine learning estimators have as one of their common requirements the standardisation of a dataset; otherwise, the estimators may perform poorly if the individual features do not roughly resemble standard normally distributed data (e.g., Gaussian with 0 mean and unit variance). In statistics and mathematics, the idea of standardisation and the corresponding z-score formula ([35], p. 880) have been extensively utilised. Equation (1) can be obtained and used for data standardisation by substituting the arithmetic mean [36] and the standard deviation [37] formulas into the z-score formula.

$$z = \frac{x - \frac{1}{N}\sum_{i=1}^{N} x_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)}} \quad (1)$$

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

Heart Disease	BMI Category	Smoking	Alcohol Drinking	Stroke	Physical Health	Mental Health	Diff Walking
No	Underweight (BMI < 18.5)	Yes	No	No	3.0	30.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	Yes	0.0	0.0	No
No	Overweight (25.0 ≤ BMI < 30.0)	Yes	No	No	20.0	30.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	No	0.0	0.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	No	28.0	0.0	Yes

Table 1.1. The first five rows of the data with the selected 17 features and the target variable (Heart Disease). Features 1 to 8 and in the top panel and 9 to 17 are in the bottom panel.

Class weighing: To reduce model bias towards the majority class and the poor prediction performance of the minority class, class weighting is crucial for imbalanced datasets. The study's dataset was unbalanced, with just 27,373 individuals (8.56% of the total) actually having heart disease out of the 319,795 individuals that were chosen. Equation (2) [38] provides the inverse-frequency class weights, which are a useful default way for computing weights.

$$w_i = \frac{N}{K\sum_{n=1}^{N} t_{ni}} \qquad (2)$$

where: N = total number of samples, K = number of classes, w_i = weight for the class i, and t_{ni} = indicator that the n^{th} sample belongs to the i^{th} class

1.2.3 Machine Learning Models and Training

The dataset was utilised to train five (5) machine learning models, including logistic regression, light gradient boosting (LightGBM), random forest (RF), support vector machine (SVM), and extreme gradient boosting. Python 3's sklearn, lightgbm, and xgboost modules was employed to implement them. Two situations were utilised to train the models: Case 1 trained the models before class weighting, while Case 2 trained them after class weighting. The goal of Case 1 was to see how the models would behave if they were applied to all classes equally. In the event of imbalanced data, this would be especially helpful in determining any bias towards the majority class and comprehending the baseline performance.

Gradient Boosting: Extreme (XGBoost) and Light (LightGBM)

Radiant boosting improves the efficiency, accuracy, and interpretability of the Model by utilising a group of weak learners, usually decision trees. Ensembles are built using decision tree models.To address the prediction mistakes from previous models, more trees are fitted to the ensemble. Two well-liked techniques based on gradient-boosted decision trees (GBDT) are extreme gradient boosting (XGBoost) and light gradient boosting (LightGBM), each with unique advantages. In tasks involving regression, classification, and ranking, XGBoost and LightGBM are preferred.

XGBoost: A distributed GBDT machine learning library with parallel tree boosting capabilities, XGBoost was created in 2016 as a member of the Distributed Machine Learning Community (DMLC) by [44]. It is expandable. A weighted total of all the tree forecasts is the final prediction made by GBDTs, which train an ensemble of shallow decision trees iteratively. Each iteration fits the next model using the error residuals from the prior one. Using random bootstrap samples of the data set, Random Forest builds entire decision trees in parallel using bagging; the majority vote (classification) finds the last forecast. The computational speed and model performance are the major considerations in the design of XGBoost's boosted tree approaches.

LightGBM: First presented by [45] in 2017, LightGBM aims to increase scalability and training efficiency. LightGBM is distinguished by two major innovations: exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS). GOSS greatly reduces computing complexity by focussing on instances with bigger gradients, which optimises the learning process. However, EFB improves automatic feature selection by bundling sparse mutually incompatible characteristics like one-hot encoded categorical variables. With all of these improvements together, training times can be accelerated by up to 20 times, which makes LightGBM an effective GBDT implementation.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

III. PERFORMANCE EVALUATION

The models compute evaluation metrics. These are useful for assessing how well the trained machine learning models perform. A number of variables, including as the problem's nature, the distribution of the data, domain-specific considerations, and many more, might affect which evaluation is best. Important evaluation matrices include F1-score, Accuracy, Precision, and Recall. The relationship between recall and precision was discovered by [46], which helped to construct the suitable formulas that are given in Equations (4)–(7) and supplied by [47].

In this instance, our goal is to forecast a class label of 1 or 0, which denotes the presence of heart disease or its absence. We refer to this as a deterministic classifier. We must select a probability threshold t in order for our probabilistic classifiers to give a label prediction. If the anticipated likelihood is more than t=50, label 1 (heart disease) is predicted by default. This default is used implicitly by all the metrics that follow.

- Samples that were misclassified are known as false positives (FP) and false negatives (FN).
- The samples that were successfully identified are known as true positives (TP) and true negatives (TN).
- Accuracy -The percentage of samples successfully classified is called accuracy
- The percentage of real positives that were appropriately classified is called recall.
- Precision-The fraction of accurately categorised expected positives is known as precision.
- Recall-The percentage of real positives that were appropriately classified is called recall.
- The F1-Score is a harmonic mean that combines recall and precision into a single score. It offers a harmony between recall and precision.

IV. RESULTS AND DISCUSSION

This section presents the model findings along with an assessment of their performances. The models were trained under two main scenarios: Case 1, or the base case, comprised training the models without assigning weights to the classes, and Case 2, or the weighted class training, entailed training the models. The findings are condensed and shown appropriately.

The models' performance for Case 1 (unweighted classes) is displayed in Table 2, and the performance for Case 2 (weighted classes) is displayed in Table 3. While the Precision, Recall, and F1-Score show how each class (Class 0 and Class 1) of the models performed, the Accuracy indicates the total accuracy of each model.

		CLASS 0			CLASS 1		
Algorithm	Accuracy	Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	91.33	91	100	95	75	1	2
SVM	91.26	91	100	95	0	0	0
LR	91.28	92	99	95	51	9	15
XGBoost	91.32	91	100	95	73	1	2
LightGBM	91.37	92	99	95	54	9	15

Table 1.2. Summary of performance metrics for unweighted classes (Case 1).

		CLASS 0			CLASS 1		
Algorithm	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
RF	70.81	98	70	81	21	82	33
SVM	73.12	97	73	83	21	78	34
LG	74.25	97	74	84	22	76	34
XGBoost	72.51	97	72	83	21	81	34
LightGBM	74.66	97	74	84	22	78	35

Table 1.3 Summary of performance metrics for weighted classes (Case 2).

Figures 1-5 illustrate the results in confusion matrix for the models under the two cases (Case 1 & Case 2).
In the confusion matrix plots, it is clear that the predictions of Class 0 (majority class) in Case 1 were almost perfect (nearly 100% true negatives) while the predictions of the minority class (Class 1) were very poor (up to 99% false negatives).

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

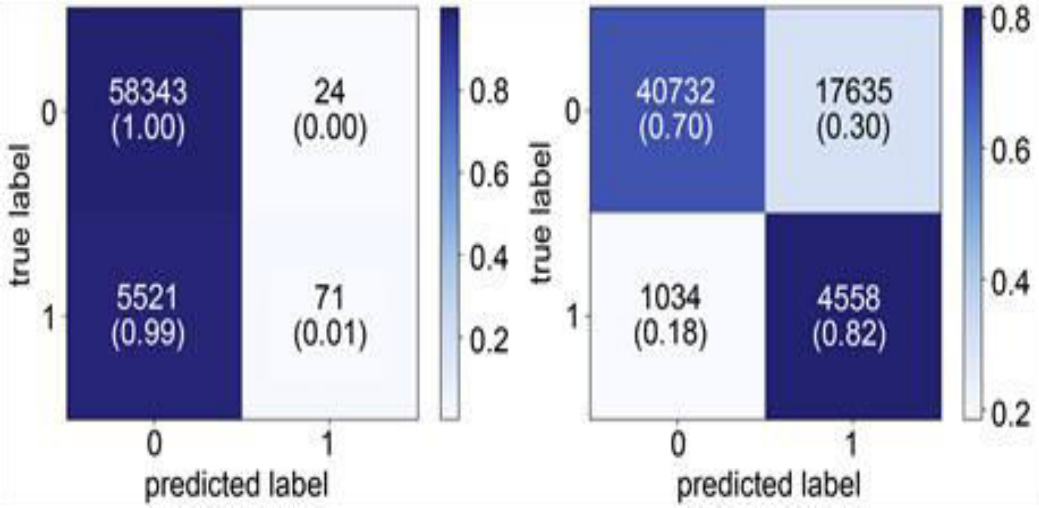

Figure 5.1 Confusion matrix of the RF classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

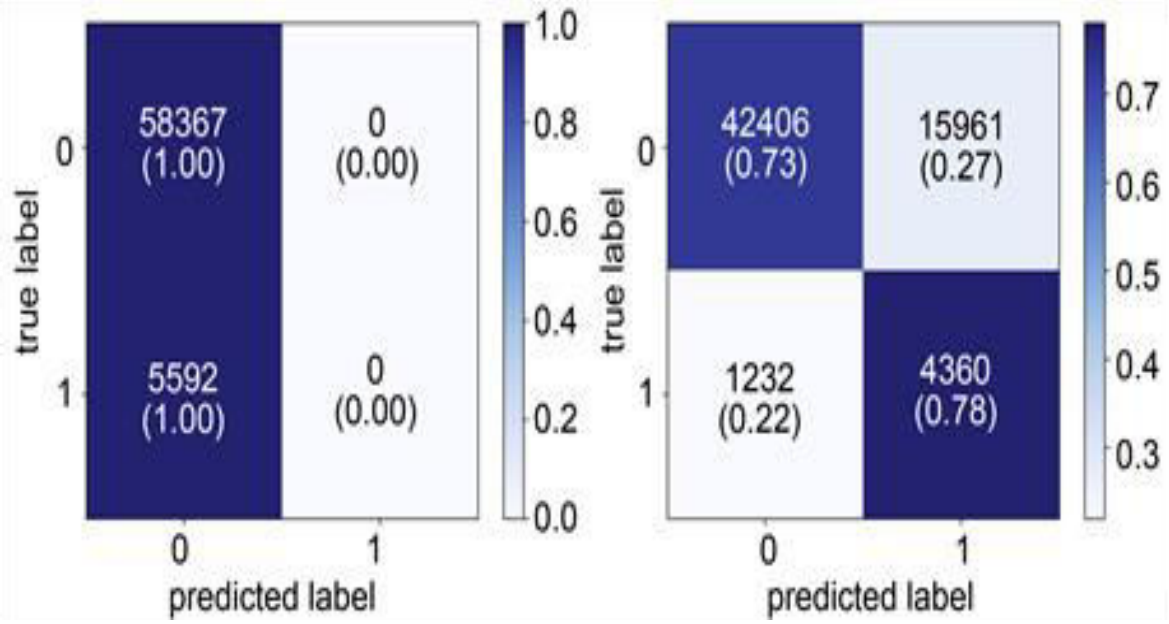

Figure 1.2. Confusion matrix of the SVM classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

1.4.1 Model Selection

Regarding the class-weighted scenario (scenario 2), the findings indicate that there aren't any significant variations in the models' overall performance between the two classes with respect to Accuracy, Precision, Recall, and F1-Score. Our primary goal is to effectively detect heart disease positive cases while lowering the amount of false positives. In both Class 0 and Class 1, the XGBoost model typically offers a superior trade-off in terms of Precision, F1-Score, and Accuracy, although the RF model produced the greatest Recall of 82%, followed by XGBoost at 81%. As a result, the XGBoost model has the best potential for heart disease prediction.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

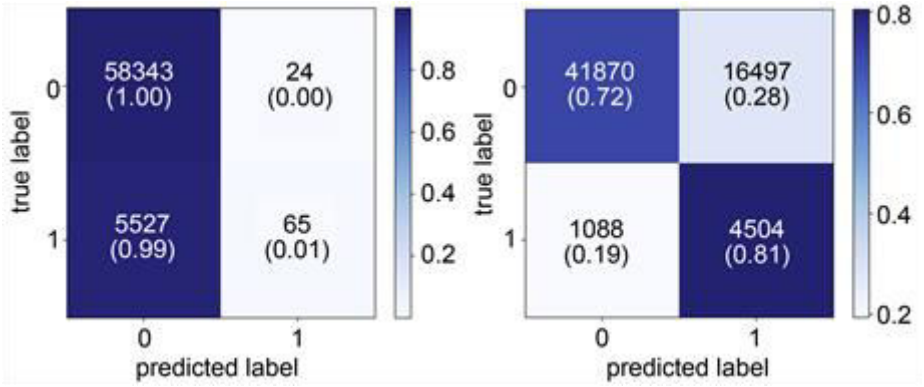

Figure 1.4. Confusion matrix of the XGBoost classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

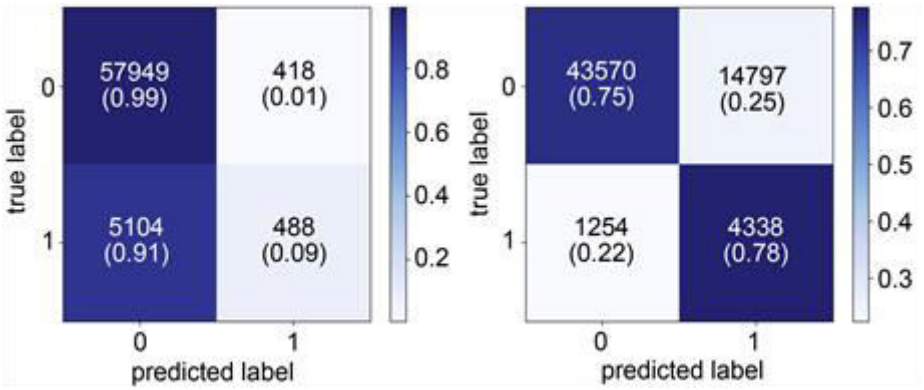

Figure 1.5. Confusion matrix of the LightGBM classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

The three widely used measures (types) of feature importance—gain, cover, and weight—are plotted in Figure 6 to demonstrate the feature's importance in the chosen model (trained XGBoost classifier). The average improvement in loss brought about by a feature is measured by the gain. Stated differently, it indicates the extent to which a characteristic contributes to the accuracy of our training data predictions. Gain was mostly attributed to the DiffWalking function. Additionally, the feature that contributed the most to cover was DiffWalking (Difficulty Walking The number of times a feature is applied to divide the data among all the trees are called weights. The feature that most affected weight was AgeCategory. Plotting generally indicates that the XGBoost model views walking difficulty (DiffWalking) as a critical predictor of a potential case of heart disease, whereas Alcohol consumption (Alcohol Drinking) is regarded as one of the least significant predictors of heart disease.

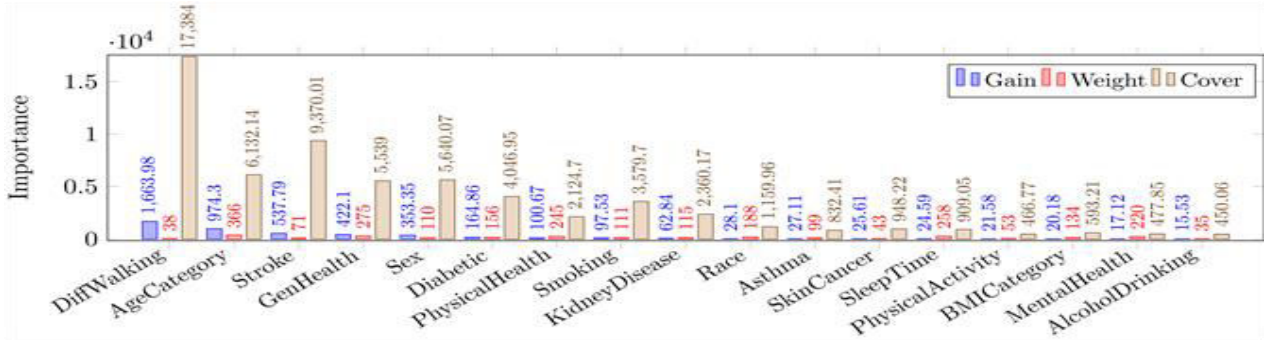

Figure 1.6. Feature importance in the XGBoost model in terms of the gain, weight and cover for the 17 features.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

V. MODEL DEPLOYMENT: ONLINE WEB APPLICATION

The chosen model, the XGBoost classifier, was made available online and is publicly accessible at https://drxgboost.streamlit.app/. Anybody can use and access the software by clicking the link. The web user interface, depicted in Figure 7, provides a brief summary of how to use the application. In order to identify the prognosis for heart disease, the user must enter all 17 inputs (features) in a style that best depicts their experiences, living circumstances, and overall health. After entering the data, the user can select the "predict" icon at the end to receive a result in a matter of seconds.

A percentage probability (1% - 100%) of the user having heart disease shall be a part of the outcome that is given. The returning probabilities had been separated even more into four groups for the aim of showing the findings: 0 - 25% is the "green zone," 25% - 50% is the "yellow zone," 50% - 75% is the "orange zone," and 75% - 100% is the "red zone." Additionally, because of the model's imperfect accuracy, the web application has stated that the results do not equate to a medical diagnosis. Nonetheless, the approach can assist in offering helpful insights for early professional medical consultation.

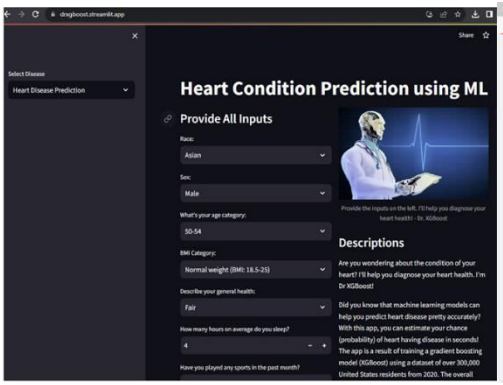

Figure 1.7 Home page

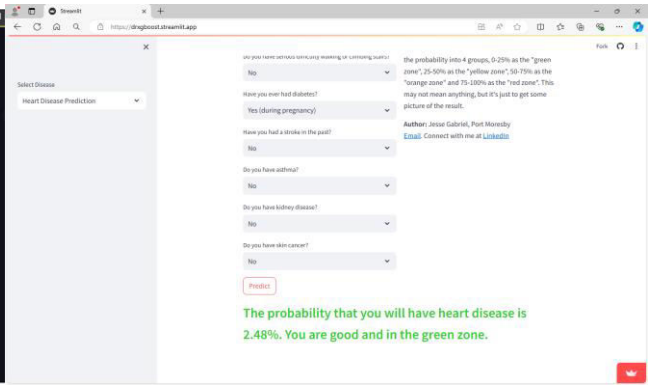

Figure 1.8 Result page

VI. CONCLUSION

Heart disease is among the primary causes of death globally. Early diagnosis and treatment of suspected cardiac disease can save many lives. A tough process requiring accuracy and efficacy in medicine is the identification of heart disease. Using machine learning for predictive modelling has proven to be quite useful in two areas: heart disease prediction and insight collection. In terms of heart disease prediction, the extreme gradient boosting model fared superior to the other five types of machine learning in this investigation. Sufficient data pre-treatment and model setup, including hyperparameter tweaking techniques, are key components in achieving accurate predictions from trained models.

For example, a number of issues (disease prediction, fraud detection, email spam detection, etc.) require careful data pre-processing approaches, such as class weighting, because the datasets in these situations have extremely imbalanced class labels. The web application that was utilised to train the model for this investigation is publicly available. The application's results may be able to offer helpful information for early professional medical consultations. Robust Machine learning approaches are expected as technology progresses, which will enhance AI-based applications in the healthcare industry.

REFERENCES

1. Chakraborty, C., Bhattacharya, M., Pal, S. and Lee, S. (2023) From Machine Learning to Deep Learning: An Advances of the Recent Data-Driven Paradigm Shift in Medicine and Healthcare. Current Research in Biotechnology, 7, Article ID: 100164. https://doi.org/10.1016/j.crbiot.2023.100164
2. Mbunge, E. and Batani, J. (2023) Application of Deep Learning and Machine Learning Models to Improve Healthcare in Sub-Saharan Africa: Emerging Opportunities, Trends, and Implications. Telematics and Informatics Reports, 11, Article ID: 100097.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106034

https://doi.org/10.1016/j.teler.2023.100097

3. Motwani, A., Shukla, P.K. and Pawar, M. (2022) Ubiquitous and Smart Healthcare Monitoring Frameworks Based on Machine Learning: A Comprehensive Review. Artificial Intelligence in Medicine, 134, Article ID: 102431. https://doi.org/10.1016/j.artmed.2022.102431
4. Rasheed, K., Qayyum, A., Ghaly, M., et al. (2022) Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey. Computers in Biology and Medicine, 149, Article ID: 106043. https://doi.org/10.1016/j.compbiomed.2022.106043
5. Liao, W., He, J., Luo, X., Wu, M., Shen, Y., Li, C. and Chen, N. (2022) Automatic Delineation of Gross Tumor Volume Based on Magnetic Resonance Imaging by Performing a Novel Semisupervised Learning Framework in Nasopharyngeal Carcinoma. International Journal of Radiation Oncology Biology Physics, 113, 893-902. https://doi.org/10.1016/j.ijrobp.2022.03.031
6. Pierre, K., Haneberg, A.G., Kwak, S., Peters, K.R., Hochhegger, B., Sananmuang, T., Tunlayadechanont, P., Tighe, P.J., Mancuso, A. and Forghani, R. (2023) Applications of Artificial Intelligence in the Radiology Roundtrip: Process Streamlining, Workflow Optimization, and Beyond. Seminars in Roentgenology, 58, 158-169. https://doi.org/10.1053/j.ro.2023.02.003
7. Zhai, K., Yousef, M.S., Mohammed, S., Al-Dewik, N.I. and Qoronfleh, M.W. (2023) Optimizing Clinical Workflow Using Precision Medicine and Advanced Data Analytics. Processes, 11, Article No. 939. https://doi.org/10.3390/pr11030939
8. Javaid, M., Haleem, A., Singh, R.P., Suman, R. and Rab, S. (2022) Significance of Machine Learning in Healthcare: Features, Pillars and Applications. International Journal of Intelligent Networks, 3, 58-73. https://doi.org/10.1016/j.ijin.2022.05.002

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 7.394

www.ijarety.in editor.ijarety@gmail.com