Heart Disease Prediction Using Machine Learning Techniques

Samarth R. Parekh[1], Yogendra L Patel[2]
[1]*Student, Dept. of IT, Dharmsinh Desai University, Nadiad, Gujarat, India*
[2]*Assistant Professor, Dept. of IT, Dharmsinh Desai University, Nadiad, Gujarat*

***Abstract*—Being a significant global health challenge, early detection is thus still considered vital in improving patient care with heart disease. This project aims to develop a user- friendly, home-based questionnaire that predicts heart disease, allowing individuals to check their heart status from a distance. We applied diverse machine learning models with more than 400,000 adults from the 2022 CDC survey. The models applied include logistic regression, decision trees, K-Nearest Neighbors (KNN), naive Bayes, random forest, and ensemble techniques, including bagging, boosting, AdaBoost, XGBoost, and a voting classifier, for classification, where performance is checked using classification metrics like precision, recall, and F1 score. The voting classifier that combined the KNN, naive Bayes, AdaBoost, and XGBoost models scored the highest level of performance with a weighted precision of 89.14%, a recall of 85.51%, and an F1 score of 87.01%. These results illustrate the usefulness of the machine learning model in creating an accessible aid for health use to help users identify heart disease quickly through an easily administered survey, thereby enabling timely intervention and better health outcomes.**

***Index Terms*—Machine Learning (ML), Ensemble Techniques, Voting Classifier, K-Nearest Neighbors (KNN), AdaBoost, XGBoost, Random Forest, Precision, Recall, F1 Score.**

I. INTRODUCTION

Heart disease is one of the leading causes of death around the world, which is why detection before things get worse would be very prudent. Conventional methods of diagnosing heart diseases are, more or less, specialized in nature, hence not accessible to several populations, mainly those in far-flung or neglected regions. This, therefore, calls for other alternatives that make the diagnosis of heart disease more accessible and affordable.

Machine learning has emerged as a very potent tool in the health sector, able to scan large data to identify patterns useful in earlier diagnosis. Techniques available in ML allow for the building of predictive models to assess the risk of heart disease, making quicker, cost-effective, and accessible diagnosis without the equipment or an expert's involvement.

We apply multiple machine learning techniques to predict the risk of having heart disease in this study. Our data contains over 400,000 adult health records drawn from the CDC's 2022 health survey with 17 features as shown in Fig 1. We will analyze a multiple model with logistic regression, decision trees, K-Nearest Neighbors, naive Bayes, random forest, and advanced ensemble methods such as bagging, boosting, AdaBoost, XGBoost, and a voting classifier. The models were tested for the sake of classification performance using precision, recall, and F1 score to determine their efficiency in predicting heart disease presence. So essentially, the main objective here would be to create a user-friendly home-based survey tool that incorporates these models and thus enables an individual to assess heart health independently; early intervention would eventually lead to better health outcomes.

#	Column	Non-Null Count	Dtype
0	HeartDisease	319795 non-null	object
1	BMI	319795 non-null	float64
2	Smoking	319795 non-null	object
3	AlcoholDrinking	319795 non-null	object
4	Stroke	319795 non-null	object
5	PhysicalHealth	319795 non-null	int64
6	MentalHealth	319795 non-null	int64
7	DiffWalking	319795 non-null	object
8	Sex	319795 non-null	object
9	AgeCategory	319795 non-null	object
10	Race	319795 non-null	object
11	Diabetic	319795 non-null	object
12	PhysicalActivity	319795 non-null	object
13	GenHealth	319795 non-null	object
14	SleepTime	319795 non-null	int64
15	Asthma	319795 non-null	object
16	KidneyDisease	319795 non-null	object
17	SkinCancer	319795 non-null	object

dtypes: float64(1), int64(3), object(14)

Fig. 1. Dataset Description

II. LITERATURE REVIEW

Heart disease is still among the leading causes of morbidity and mortality in almost every part of the world. Early diagnosis is imperative since ease and accuracy in diagnosing the disease may lead to better patient outcomes. Towards this, machine learning becomes an emerging approach to predict heart disease by making use of vast datasets with advanced algorithms in enhancing diagnostic accuracy. Literature has discussed several ML algorithms, like Decision Trees and Random Forests, which state that the ensemble technique is a necessity for enhancing predictive accuracy in the field of clinical diagnostics. The findings agree with the general conclusion that in most cases, a set of multiple models performs better than a single model, and hence it is very feasible to use ensemble methods in clinical settings.

The technique adopted for heart disease diagnosis involves applying data mining techniques- especially Naive Bayes and Decision Trees applied by Soni et al. [2]. This study probes into the strengths and weaknesses of these algorithms in clinical settings, pointing to an effective choice of algorithms that best suit the requirements of clinical needs. The comparative analysis provided valuable insight into what different algorithms are doing better in order to make better decisions in clinical practice.

The research of Parvin et al. focuses on the detection of heart disease based on ensemble machine learning algorithms, using models such as AdaBoost and Random Forest [3]. It is demonstrated that their benefits can increase classification performance to set new boundaries for diagnostics. The general result of this kind of analysis indicates that ensemble learning's application may effectively enhance the tasks of making predictions about heart diseases and provide a robust framework for using clinical diagnostics.

In the study of Velusamy and Elias, the data mining tools included algorithms like KNN and Decision Trees for diagnosis purposes related to heart disease. They comprehensively assessed the ability of these algorithms to predict something, so that whatever information was brought out in regard to such algorithms' performance in clinical scenarios can be gained. This work adds to a better understanding of the role played by different algorithms in reaching accurate predictions through the back and forth over which of the machine learning techniques will be the most efficacious.

Further progress in deep learning was reported by Ghorbani et al. [5], who explored deep models to interpret echocardiograms. Indeed, it was revealed that deep learning methods could be just as good as human experts in diagnosis. Therefore, this trend towards the application of advanced deep models represents the chance to improve conventional diagnostic methods to eventually boost the general landscape of diagnosis for heart diseases.

The literature discussed above reflects the dynamic emergence of modeling aspects behind the prediction of heart disease, starting from traditional algorithms like Decision Trees and Naive Bayes, advanced ensemble methods, and deep learning approaches. It recognizes that technology can play an important role in the accurate diagnosis of heart disease in patients and, hence, in a better patient outcome. Further research in this direction is expected to continue exploring the integration of such techniques into more robust, interpretable, and clinically applicable models for predicting heart disease

III. PROPOSED METHODOLOGY

The primary objective of this study is to develop an accessible, home-based tool for predicting the presence of heart disease using machine learning algorithms. The methodology involves multiple stages, including data preprocessing, model development, evaluation, and validation.

The following steps were performed: -

Step-1. Data Collection

Downloaded the "Indicators of Heart Disease” CDC 2022 survey from Kaggle with information regarding health status and associated factors from 400,000+ adults and 17 different features.

Step-2. Data Preprocessing

Handling Duplicates:

Removing duplicate rows using the drop duplicate () method in the panda’s library, as no. of duplicates is very insignificant.

Feature Encoding:

Convert the categorical variables into a numerical format by using one-hot encoding or ordinal encoding.

Feature Scaling:
Normalize or standardize features for uniform distribution of the data.

Class balancing:
SMOTEENN is used to balance the difference in the number of 'Yes' and 'No' classes of heart disease (target feature).

Step-3. Data Splitting
Now, splitting your data into two portions: the Training set and the Testing set, where the number of samples in the Training set is 70% while in the Testing set 30%.
Training the models on the training set and then testing them on the test set.

Step-4. Model Development
Different models were prepared to choose the right model for the project, such as:
Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, Ensemble Techniques (Bagging, Boosting, AdaBoost, XGBoost, Voting Classifier).

Step-5. Model Evaluation
These models were evaluated using various performance metrics. Typically, Precision, Recall, and F1-Score (classification report) are used in this case.

Step-6. Cross-Validation
Cross-validation is run in k-fold to ensure that the model performs on multiple subsets of data, thus avoiding overfitting.

Step-7. Final Ensemble Model
The final model will be a Voting Classifier, consisting of KNN, Naive Bayes, AdaBoost, and XGBoost. This is because this is the best obtained from the evaluation.

Step-8. Model Deployment
A pipeline of encoding and scaling transformers, followed by the model described above for predicting outcomes, is created. Then, a .pkl file is created from that pipeline that we can use in an API to predict the results given input values by users.

How does it work?
In this project, values are input via a web interface made using Flutter. These are then transmitted to a machine learning pipeline hosted on a Flask-based API. It first preprocesses values by encoding categorical features using Ordinal Encoding and One Hot Encoding. Numerical features are standardized to ensure homogeneity across the data by StandardAero. These inputs are then fed into the final stage of the machine learning model to predict the possibility of heart disease. The prediction result is sent back to the website and is displayed to the user.

User's Workflow
When the user accesses the heart disease predictor website, they are presented with a questionnaire consisting of 17 questions, each representing one of the 17 key features, as illustrated in Fig 1. After completing the questionnaire, the user clicks the "Predict" button, which sends the provided inputs to the hosted API. The API processes the data through the machine learning model to generate a prediction, which is then returned to the website and displayed to the user. The following is shown in Fig 2.

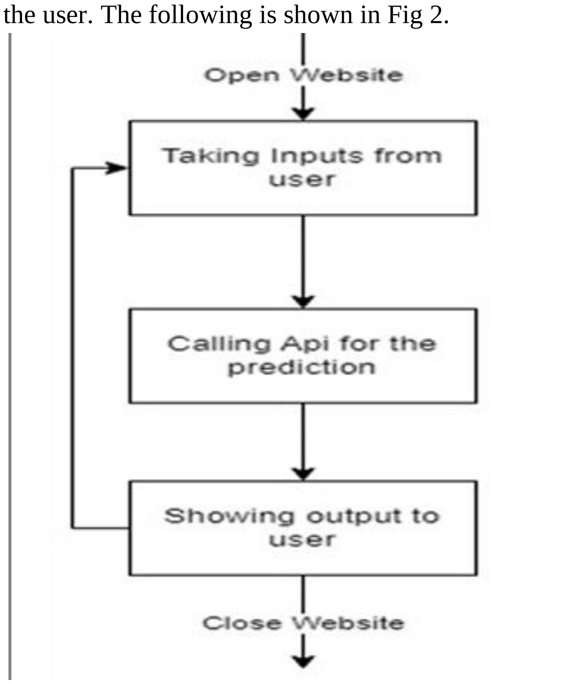

Fig. 2. User's Workflow Model's Workflow

The input values provided by users are first processed through the Ordinal Encoder and One Hot Encoder to transform categorical data into numerical form. This step ensures that the data can be used by machine learning models. Next, the encoded values are standardized using a Standard Scaler to normalize the numerical values. Once pre-processed, the data is passed to the final model, a Voting Classifier that combines Naïve Bayes, Random Forest, AdaBoost, and XGBoost. This model is selected based on evaluations of different models, and its performance is measured using metrics like precision, recall, and F1 score. The final prediction is returned to the user via the website. The following is shown in Fig 3.

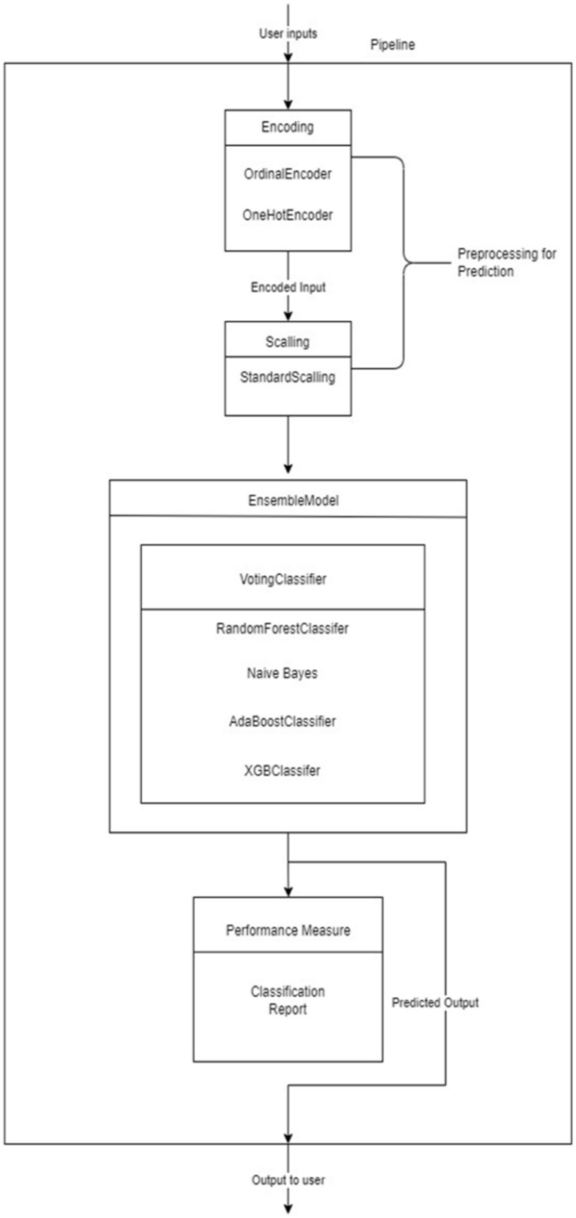

Fig. 3. Pipeline's Workflow

IV. RESULTS

The project evaluated multiple machine learning models for predicting heart disease, using a comprehensive dataset of over 400,000 adults from the CDC's "Indicators of Heart Disease" dataset. The primary models tested included Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and various ensemble techniques such as Bagging, Boosting, AdaBoost, XGBoost, and a Voting Classifier. The Voting Classifier produced the following results: Weighted Precision: 89.14%, Recall: 85.51%, F1 Score: 87.01%. The results are shown in Fig 4 & 5.

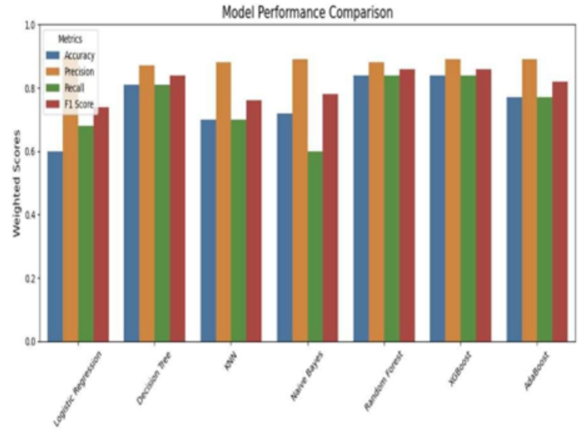

Fig 4 Different Model's evaluation metrics

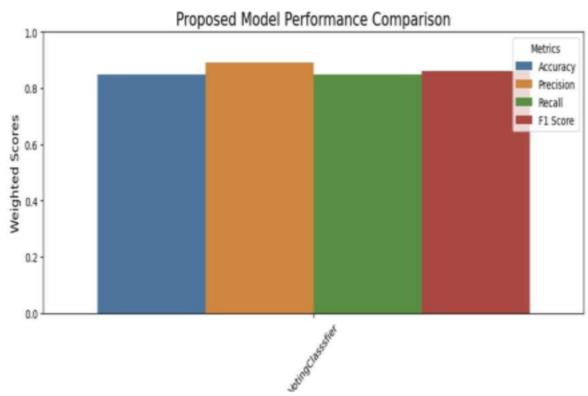

Fig 5 Proposed Model's evaluation metrics

V. CONCLUSION

The heart disease prediction project successfully demonstrated the efficacy of machine learning models in facilitating early detection and home-based diagnosis of heart disease. By leveraging a user-friendly interface and integrating various machine learning techniques—such as logistic regression, decision trees, k-nearest neighbors (KNN), naive Bayes, and ensemble methods like bagging, boosting, AdaBoost, and XGBoost—the study produced significant results. Among the models tested, the

Voting Classifier, which combines KNN, naive Bayes, AdaBoost, and XGBoost, emerged as the most effective, achieving a weighted precision of 89.14%, recall of 85.51%, and an F1 score of 87.01%.

These findings underscore the potential of machine learning as a complementary tool for healthcare professionals, allowing for initial risk assessment of heart disease in a non-clinical setting. The system's ability to deliver near-real-time predictions via a website-based platform highlights its accessibility and practical utility. The use of ensemble methods further showcases how combining different algorithms can enhance prediction accuracy and robustness.

With further optimization, including larger and more balanced datasets, such machine learning models can be refined for broader applications in preventive healthcare and remote monitoring. This approach represents a promising step toward integrating technology into healthcare workflows, particularly for early disease detection, thus improving patient outcomes and reducing the burden on healthcare systems and providing a seamless and visually appealing recommendation experience.

REFERENCES

[1] J. Patel, M. Tejal, and H. Keshwala, “Prediction of heart disease using machine learning algorithms,” 2022. [Online]. Available: https://www.researchgate.net/publication/365887966_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms

[2] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Data mining approach to detect heart diseases,” 2011. [Online]. Available: https://www.researchgate.net/publication/258374505_Data_Mining_Approach_to_Detect_Heart_Diseases

[3] F. Parvin, A. Paul, and P. C. Shill, “Ensemble machine learning algorithms for heart disease detection,” 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9398727/

[4] R. Velusamy and C. Elias, “heart disease diagnosis using data mining tools and techniques,” 2005. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d32ec14e005d9907603d2e46daa1b68a9b63d95b

[5] Ghorbani, D. Ouyang, and J. Zou, “Deep learning interpretation of echocardiograms,” 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31993508/