JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION AND RISK ANALYSIS USING MACHINE LEARNING TECHNIQUES

Rimsy Dua[1],Dr.Santosh Singh[2],Shubhada Bhatkhande[3],Jyoti Prajapati[4]
[1]Assistant Professor, Department of IT,Thakur College of Science and Commerce
[2] HOD Department of IT,Thakur College of Science and Commerce
[3,4]PG Student, , Department of IT,Thakur College of Science and Commerce
Thakur Village, Kandivali (East), Mumbai-401107, Maharashtra, India

ABSTRACT:-

Heart diseases remain one of the leading causes of death globally. It, therefore, calls for early detection and proper diagnosis of heart diseases. The paper uses ML classification for heart diseases, which includes no disease, coronary artery disease, arrhythmia, and cardiomyopathy. The project is analyzed in a variety of risk factors-to include age, sex, type of chest pain, body mass index, alcohol consumption, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, smoking status, history of stroke, diabetes, ST segment depression, and slope-and does its best to make inferences towards improved classification accuracy of risk in the case of heart disease. Our approach was training and validating multiple models of ML to understand their classification ability regarding disease categories using the described risk factors for possible prediction. The classification, as such, would give a profile of the patient and allow clinicians to identify those at high risk for definite preventive interventions. The results revealed that good clinical decision making can indeed be supported by ML algorithms and could be a highly valuable tool in reducing the burden of the disease from the earliest accurate predictions.

KEYWORDS:-

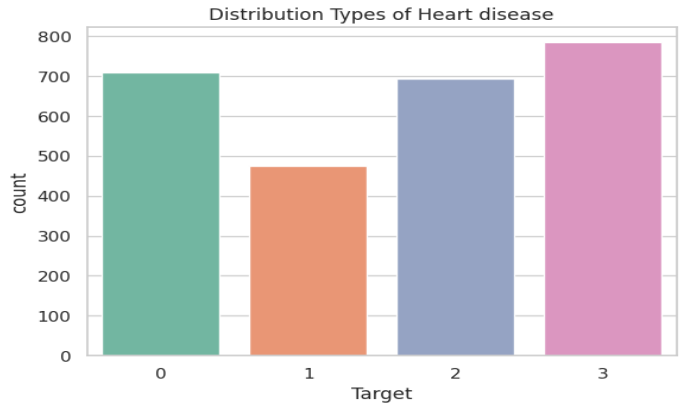
Heart Disease Prediction, Machine Learning, Naive Bayes, XGBoost, Decision Tree, Risk Analysis, Predictive Modeling.

INTRODUCTION:-

Heart disease is still one of the leading causes of death in all parts of the world, hence early diagnosis and assessment of risks can help improve patients' outcomes. Generally, traditional methods of diagnosis rely on clinical observation and subjective interpretation by practitioners that are time-consuming, and their precision levels sometimes do not suffice to sustain the growing population's needs. This requires further integration of heart disease prediction using more machine learning techniques to better improve the diagnostic efficiency and accuracy. There are many machine learning algorithms, such as XGBoost, Naive Bayes, and Decision Trees, applied in medical research for pattern discovery in large data sets regarding heart disease. For instance, XGBoost is said to handle complicated structures of data despite obtaining a high level of predictive accuracy. It, therefore, suits the case where performance is to be favored above the rest. Instead, Naive Bayes bases its premise on simplicity and speed much more significantly based on the independence of features that can save some time fast in several cases. Although less efficient than the other two algorithms, the interpretability provided by decision trees is unparalleled.

This paper seeks to experiment these three algorithms within the scope of the heart disease predictor in order to assess their performances regarding raising the particularly sensitivity for the diagnostic clues. This work aims to contribute to insights that could help in more accurate prognosis for individuals at risk of coronary heart disease by probing the capabilities of machine learning applied in this domain. A careful comparison of predictive performance will allow the study to conclude which algorithm can provide a good balance between accuracy, efficiency, and interpretability for use in a clinical environment.

```
Class distribution:
 Target
3    784
0    708
2    693
1    474
Name: count, dtype: int64
```



LITERATURE REVIEW:-

C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai,United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
This paper discusses heart disease prediction using major health-related data through the use of machine learning techniques. Four classification techniques used in this paper are MLP, SVM, RF, and NB. The models were tested on accuracy, precision, recall, and F1-score after executing data

cleaning and feature selection. Among them, the best-performing model of SVM was 91.67%, and it can be used to predict heart disease.[1]

Pal, M., et al. "Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers." Open Med (Wars), vol. 17, no. 1, June 2022, pp. 1100–13, https://doi.org/10.1515/med-2022-0508.
This study presents the power of machine learning methods, that is, MLP and K-NN, in the identification of CVD. The removal of outliers and handling of missing values improve the models greatly so that the MLP has achieved an accuracy of 82.47% and an AUC value of 86.41% compared to K-NN. Therefore, this MLP model may be employed for the automatic detection of CVD and potentially for many other diseases also. Further research should be conducted on further datasets to evaluate more performance of the model as the next step for further validation of the model.[2]

A., Lakshmanarao., Thotakura, Venkata, Sai, Krishna., Tummala, Srinivasa, Ravi, Kiran., Chinta, Venkata, Murali, krishna., Samsani, Ushanag., Nandikolla, Supriya. "Heart disease prediction using ML through enhanced feature engineering with association and correlation analysis." Indonesian Journal of Electrical Engineering and Computer Science, null (2024). doi: 10.11591/ijeecs.v34.i2.pp1122-1130
The authors attempted to emphasize the relevance of machine learning in the health domain, particularly concerning the ability to predict several diseases. The review has fully achieved its purpose by discussing the following varieties of machine learning classifiers: including decision trees, k-nearest neighbors, random forests, XG-Boost, and support vector machines-all of which have been shown to be so successful in classification tasks.[3]

Nagavelli, U., et al. "Machine Learning Technology-Based Heart Disease Detection Models." *J Healthc Eng*, vol. 2022, Feb. 2022, p. 7351061, doi:10.1155/2022/7351061.
This article assesses the performance of four approaches for heart disease detection by machine learning: Naïve Bayes with weighted prediction, two SVM models using XGBoost, an improved SVM (ISVM) with duality optimization, and a standalone XGBoost model. This paper suggests current ML-based heart disease detection methods and future enhancements, such as an expanded dataset attribute, developing a mobile application, and integrating the system with the hospital's databases for ease of use and simplicity.[4]

Chicco, D., and G. Jurman. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone." *BMC Med Inform Decis Mak*, vol. 20, no. 16, 2020, https://doi.org/10.1186/s12911-020-1023-5. This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not. Indeed, medical doctors aiming at understanding if a patient will survive after heart failure may focus mainly on serum creatinine and ejection fraction.[5]

HIMANSHI, HIMANSHI., Srinibas, Pattanaik., Krishna, S., Nayak. "Heart Diseases Prediction Using machine learning and Deep learning Models." null (2024).doi: 10.1109/ccict62777.2024.00063 This literature survey reflects the evolution of heart disease prediction methodologies, highlighting the transition from traditional statistical approaches to

advanced machine learning and deep learning techniques, while also addressing ethical considerations in the field.[6]

Qianjun, Zheng. "Machine Learning Analysis in the Field of Heart Disease." (2024). doi: 10.61173/qz08vs80 . The manuscript deals with the significant influence of machine learning in cardiovascular disease research, bringing new ideas on prevention, diagnosis, and therapeutic interventions, but underlines the growing role of technological innovations within healthcare. Such applications have many relevant algorithms in ML that can help improve the prediction of diseases, as well as the stratification of risk, considered important factors in planning the patient's optimal care.[7]

METHODOLOGY:-

1.Data Collection:

Here, it is a collection of dataset containing numbers of demographic variables and health indicators relevant to heart diseases. The main features of the dataset include age, gender, BMI, alcohol consumption, type of chest pain, resting blood pressure, cholesterol level, fasting blood sugar, results from a resting electrocardiogram, maximum heart rate, symptoms produced by exercise, smoking, history of stroke, diabetes status, ST-Depression, slope of the ST-Segment, and the target variable is presence or absence of heart disease.

2. Data Preprocessing:

Data Collection Collect the data set regarding heart disease by searching information on multiple medical websites and other internet resources. Then observe the structure of the data set as to what feature other than response variable are included. Missing Value Handling: Identifies missing values and performs relevant operations, such as filling up with statistics like mean, median, and mode or delete rows and columns that have a very high missing data count. Encoding categorical variable: The categorical attribute is converted into a numerical attribute using techniques such as one-hot encoding or label encoding. Scaling of Features: The features are scaled since algorithms like XGBoost and Decision Tree are feature-scale sensitive.

3. Data Exploration and Analysis(EDA)

Correlation Analysis: Determine the associations that exist between all features and the target variable by using heatmaps or correlation matrices.
Feature Distributions: Distribution Analysis Conclusion Finally, analyze the feature distributions and determine skewness and whether transforms, such as log transforms, are required.
Representations using plots: Create histograms, box plots, and pair plots to help understand the interrelationships within the data.

4. Feature Selection

RFE eliminates the recursion process of features systematically until, at the end, the optimal subset of features is reached.**Regularization methodologies,** such as Lasso (L1) or Ridge (L2) regression, serve to diminish dimensionality by imposing penalties on features deemed less significant.

5. Model selection and training

Data partitioning Divide the dataset into training and testing subsets, typically in a 70:30 split or 80:20 ratio.

Algorithm 1: XGBoost

In summary, fit the XGBoost classifier using learner rate and maximum depth as hyperparameters set, along with the number of estimators, and regularised parameters alpha and lambda as shown.

Algorithm 2 Naive Bayes

A Naive Bayes classifier uses conditional probabilities in order to classify either categorical or Gaussian-distributed data.

Algorithm 3 Classification Using Decision Trees

The decision tree classifier is to be used in its optimized parameters of maximized depth and minimized number of samples required for splits along with pruning techniques.

6. Model Evaluation

Performance Metrics To analyze the performance of the models, several metrics were used such as accuracy, precision, recall, F1-score, and the AUC-ROC curve.

Cross-validation: Apply k-fold cross-validation to check how well the respective models work on the yet unseen data. This can be achieved using the confusion matrix. It will then analyze the kinds of mistakes, and their accuracy in predicting heart disease regarding each model. Risk Analysis

7. Probabilistic Evaluations:

Include probabilistic assessments for each case: the probability of developing heart disease. Threshold Adjustment: Adjust decision thresholds so that sensitivity and specificity are in good balance with each other, given that medical diagnoses depend on the costs of false positives as much as false negatives. Interpretability: Features attributes related to feature significance contribute to considering these as predictors to be important and strengthen interpretive support given to physicians.

8. Deployment of the Implementation and Observation Model:

The best model has to be run on such frameworks as Flask or FastAPI in order to provide real-time predictions. Continuous Monitoring: The performance of the model is continuously checked in terms of drift detection or any other metric for ensuring that a model works well for long periods.

RESULTS AND OBSERVATION:-

Data Collection Gather the heart disease dataset with the help of several medical and online sources. Dataset Attributes Number of Columns: 17 Rows: 2692 The heart disease class is highly imbalanced, which is why we are going to apply SMOTE because our model needs to learn well for both classes. Distribution: Training set- 60%, validation set-20%, test set-20%. The standardized features were fed into the process in order to upgrade the efficiency of the model prior to the training process. The usage of standardized variables improved the strength of the training model built for training classifiers XGBoost, Decision Tree, and Naïve Bayes on the

training dataset besides developing many models for making predictions from the validation and test datasets. Lastly, the ability of the model is measured using the metric accuracy, and the results are also shown in the confusion-matrix.

Table 1. Precision,Recall,F1-Score accuracy between various models.

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	**97.40%**	**0.97**	**0.97**	**0.97**
Decision Tree	**93.88%**	**0.89**	**0.88**	**0.89**
Naive Bayes	**91.65%**	**0.91**	**0.91**	**0.91**

Table1:Comparative Efficiency Analysis of Model-XGBoost outperforms all the metrics studied and thus undoubtedly remains the best fit model for this specific classification task. Improved precision and recall can classify positive and negative cases quite well. As for the Decision Tree model, although it executed pretty well, its efficiency is relatively inferior to that of the XGBoost model. However, the increased precision related to the Decision Tree entails that despite sometimes mistaking many instances as positive, it is more prolific about false positives than the XGBoost model is. Naive Bayes presents high efficiency, especially concerning recall, that is detection of positive cases; however, it has been proved to have in general significantly lower accuracy compared to the other models tested in this paper.

The bar chart in Fig. 1 presents the accuracy obtained by three machine learning models on a classification task, most probably heart disease prediction given the previous charts. The graphs are XGBoost: Accuracy: 96.43%. The graph is represented by a green bar. Naive Bayes: Accuracy: 94.74%.Graphically, the orange bar represents it. Decision Tree: Accuracy: 91.73%.Graphically represented by the blue bar. From the above graph, the y-axis refers to the accuracy in percentage and the x-axis refers to the names of the models. From the diagram, it is evident that XGBoost was the best of the three, followed by Naive Bayes, where the decision tree itself shows that it has the least accuracy among the three models. The space between the bars shows the spread of the variations of the models, and at the top of every bar, there is the corresponding percentage inscribed.

Figure 2. This is a representation of a stacked bar chart illustrating the distribution of risk scores in the predicted categories for cardiovascular disease. The Y-axis or vertical scale displays the percent of distribution for the risk score from 0% to 80%. Along the x-axis, a timeline is plotted labeled with the integers 0, 1, 2, and 3, showing the types of heart disease that are forecasted to occur. Each integer stands for one different kind of expected heart disease. Each group has a categorization of high-risk distinguished with red, the low-risk levels distinguished in blue, and the moderate risk factor as green. The red is historic coding and therefore high-risk, the blue is low-risk levels, and the green is a moderate risk factor. This means a large percentage of this population is at a higher risk for Type 1 and Type 3. The lower-risk group is represented by the color blue, which indicates the lowest prevalence of all types; however, this prevalence is a bit higher in Type 3 with a prevalence rate of 18%. In particular, this category of medium-risk that is represented by green represents 27%. The following graph outlines the relationship of all these types of expected heart disease with their respective risk factors, where high-risk factors are more predominant in the predictions for Type-1 and Type-3. Fig.

Figure 3, Actual vs Predicted Outcome for the Classification of Four Types of Heart Disease Obtained by Confusion Matrix: No Heart Disease It correctly predicts 2 patients. Type 1 Coronary Artery Disease:2 correct as "Having Heart Disease", 1 incorrect as "No Heart Disease". Type 2 Arrhythmia Heart Disease:2 correct as "Having Heart Disease", 1 incorrect as "Type 3 Cardiomyopathy". Type 3 Cardiomyopathy correct, 1 incorrect placed as "Type 1 Coronary Artery Disease". Off-diagonal cells signify incorrect classifications. The cells on the diagonal are not incorrectly classified. The darker the shades, the higher the frequency required for classification.
Heart Disease Prediction is about 97.18%.

Fig.1:-Performance Comparison Model Accuracy

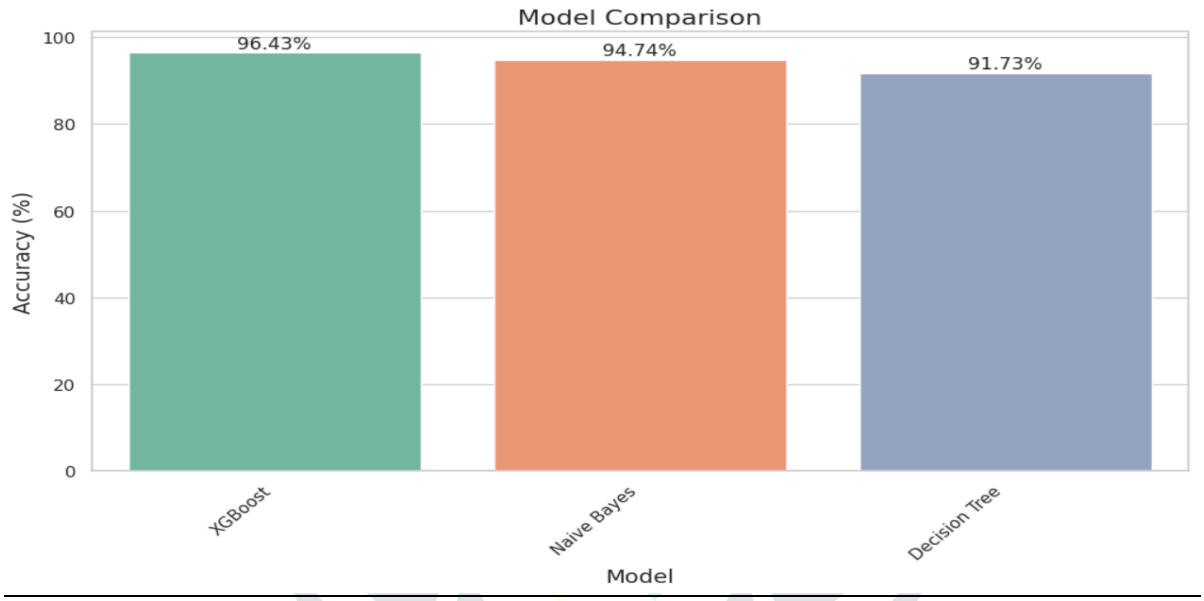


Fig.2:Stacked Bar Chart Of Predicted Heart Disease Types and Risk Scores

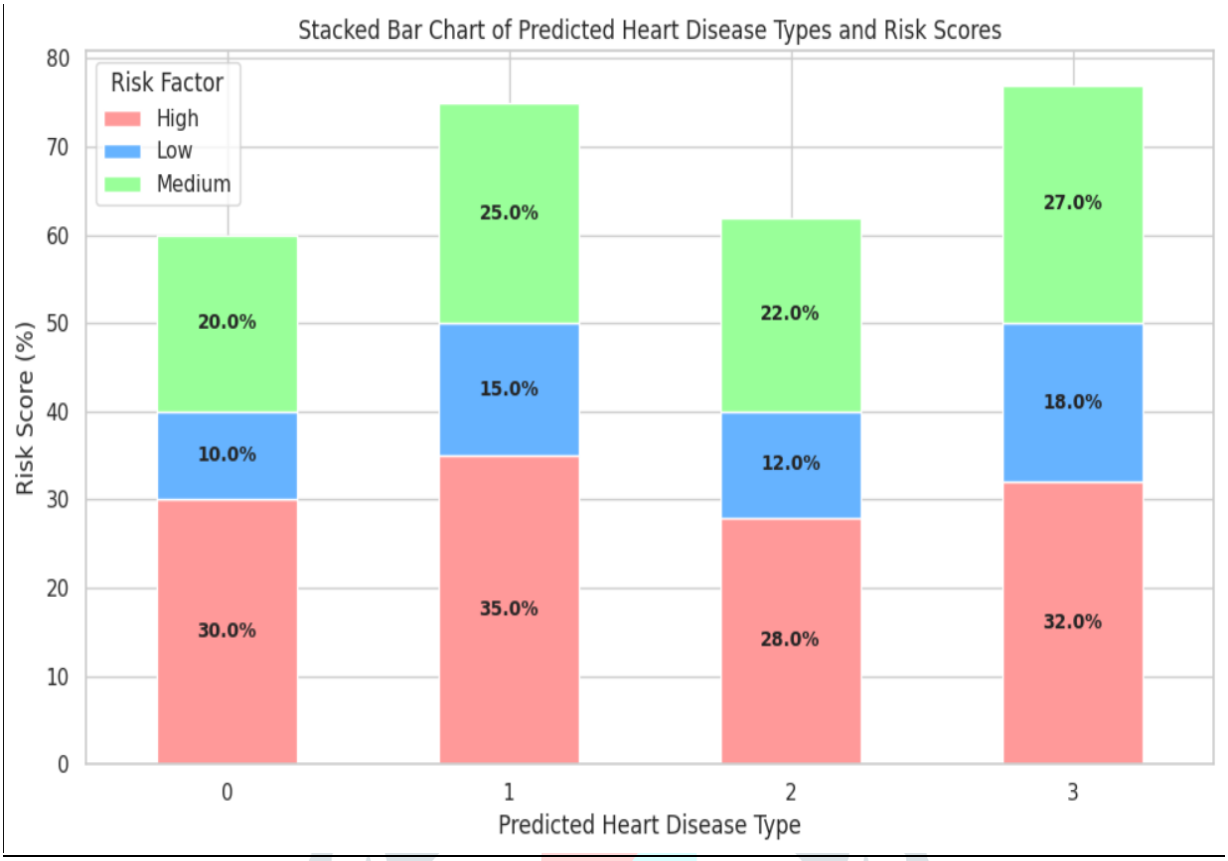
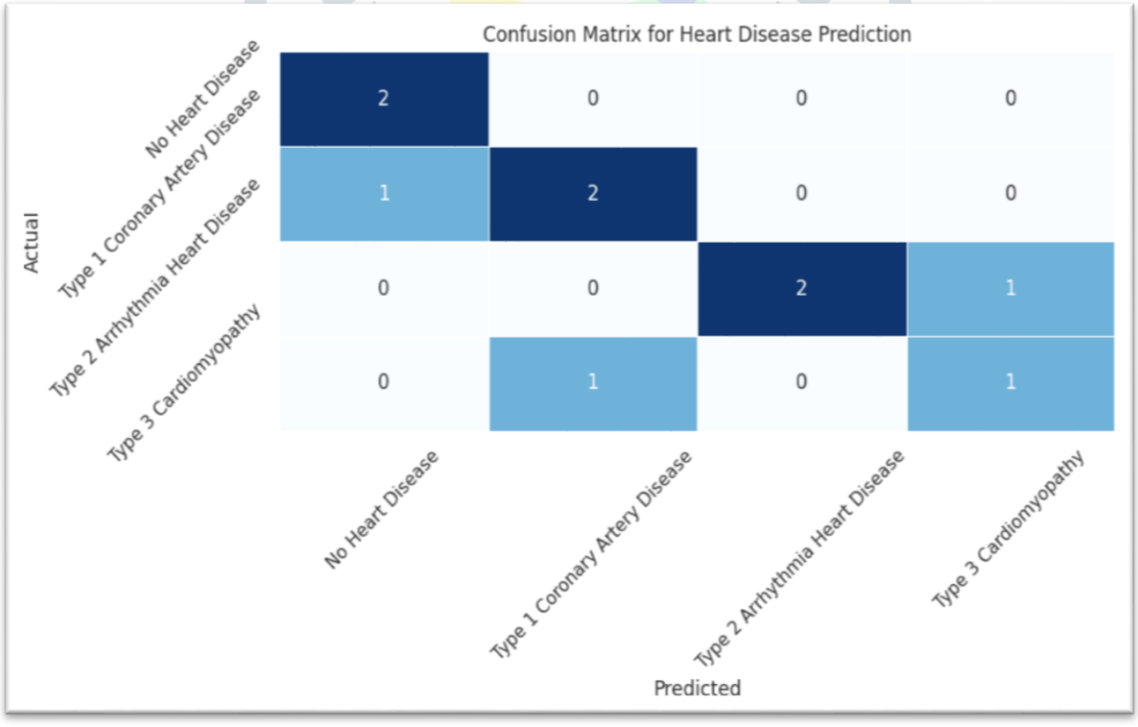


Fig3. Confusion Matrix For Heart Disease Prediction



CONCLUSION AND FUTURE WORK:-

In **conclusion,** the study shows that machine learning algorithms, such as XGBoost, Naive Bayes, and Decision Trees, can predict heart disease accurately and analyze the risk for heart disease.

Among the three models selected, XGBoost is found to perform well overall with better accuracy, precision, and recall in performing all the classification tasks. This strength in handling complex data structures and getting a high degree of accuracy in identifying patterns makes it a strong predictor of heart disease. Although the Naive Bayes algorithm is a little less accurate in this field, it has been quite effective in recall, thereby being useful in identifying true positive cases effectively. The Decision Tree model is less accurate, on average, but its interpretability does provide greater insights into the decision-making process-an aspect that can be helpful, especially in clinical settings where it's essential to provide transparent views of the decision-making process. The investigation also highlighted data imbalances and its handling, for which SMOTE application was used to let the model learn better among classes. Evaluation metrics such as accuracy, precision, recall, F1-score, visualization like confusion matrices, and distribution of risk helped draw a comprehensive comparison about the models' performances and their potential applicability into clinical diagnosis.

FUTURE WORK :- A few key enhancements in this area could focus on more implementation of algorithms, such as random forests, support vector machines, or even deep learning techniques like neural networks, or increasing the robustness of the system and its generalizability through higher diversity and comprehensiveness of the dataset. Additionally, exploring feature selection and dimensionality reduction techniques could help optimize model efficiency and reduce computational costs. Future research could also include the ensemble approach to learning, where the power of different algorithms can be merged to produce a more accurate and consistent model. Perhaps an EHR-integrated real-time prediction system would be a way of making the model accessible to healthcare providers and, therefore, liable to lead to faster, data-driven decisions in practice. Finally, introducing some XAI methods ensures that the models will not only give good predictions but also understandable and actionable insights; these are crucial for winning the confidence of medical professionals and patients. All these advances may put machine learning at the forefront in the early detection and management of heart disease, thus bringing better outcomes.

REFERENCES:-

[1] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai,United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.

[2] Pal, M., et al. “Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers.” Open Med (Wars), vol. 17, no. 1, June 2022, pp. 1100–13, https://doi.org/10.1515/med-2022-0508.

[3] A., Lakshmanarao., Thotakura, Venkata, Sai, Krishna., Tummala, Srinivasa, Ravi, Kiran., Chinta, Venkata, Murali, krishna., Samsani, Ushanag., Nandikolla, Supriya. "Heart disease prediction using ML through enhanced feature engineering with association and correlation analysis." Indonesian Journal of Electrical Engineering and Computer Science, null (2024). doi: 10.11591/ijeecs.v34.i2.pp1122-1130

[4] Nagavelli, U., et al. “Machine Learning Technology-Based Heart Disease Detection Models.” *J Healthc Eng*, vol. 2022, Feb. 2022, p. 7351061, doi:10.1155/2022/7351061.

[5] Chicco, D., and G. Jurman. “Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone.” *BMC Med Inform Decis Mak*, vol. 20, no. 16, 2020, https://doi.org/10.1186/s12911-020-1023-5.

[6] HIMANSHI, HIMANSHI., Srinibas, Pattanaik., Krishna, S., Nayak. "Heart Diseases Prediction Using machine learning and Deep learning Models." null (2024).doi: 10.1109/ccict62777.2024.00063

[7] Qianjun, Zheng. "Machine Learning Analysis in the Field of Heart Disease." (2024). doi: 10.61173/qz08vs80