iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52922

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Interpreting Cardiovascular Disease using Random Forest and Explainable AI

Aishwarya Dabir[1], Pratiksha Khedkar[2], Laxmi Panch[3], Tejal Thakare[4], Dr M A Pradhan[5]
[1, 2, 3, 4]*Student,* [5]*Associate Professor, Department of Computer Engineering, AISSMS College of Engineering, Kennedy Road, Pune-411001, India*

Abstract: ***These days, artificial intelligence and machine learning in trendy have proven terrific performances in lots of obligations, from image processing to natural language processing, specifically with the advent of machine learning in conjunction with studies development, they've encroached upon many specific fields and disciplines. a number of them require excessive degree of duty and as a result transparency, as an instance the clinical region studies into Explainable Artificial Intelligence (XAI) has been increasing in current years as a response to the need for extended transparency and believe in AI. that is especially crucial as AI is utilized in sensitive domain names with societal, moral, and safety implications reasons for system choices and predictions are as a consequence had to justify their reliability. This requires extra interpretability, which frequently approach we need to understand the mechanism underlying the algorithms. by means of applying the same categorization to interpretability in clinical research, it is hoped that (1) clinicians and practitioners can in the end method those strategies with caution, (2) insights into interpretability could be born with greater issues for scientific practices, and (3) initiatives to push ahead statistics-based totally, mathematically- and technically-grounded scientific schooling is recommended.***

I. INTRODUCTION

Artificial Intelligence (AI) has become more and more popular within the healthcare industry for sickness diagnosis, analysis, and remedy. however, the "black box" nature of a few AI algorithms has raised concerns regarding the interpretability and transparency of the results produced[1]. This loss of interpretability is specifically tricky within the case of complicated sicknesses together with cardiovascular ailment (CVD), that's the leading cause of loss of life globally. To address this problem, Explainable AI (XAI) has emerged as a new paradigm in AI studies that specializes in growing fashions that could offer a clean and comprehensible clarification in their decision-making manner.

Explainable AI (XAI) refers to the development of artificial intelligence systems that may provide clean and understandable causes for their decision-making processes.

The purpose of XAI is to make AI systems greater transparent and responsible, so that human beings can consider them and make informed decisions primarily based on their output. XAI objectives to enhance the transparency and interpretability of AI algorithms, making them more accessible and useful for clinicians and patients. XAI is an active area of research, and there are numerous exclusive procedures to developing explainable AI structures.

A few strategies include building AI structures that use simpler models or are trained on smaller, more comprehensible datasets, or creating visualization equipment that permit people to look the internal workings of an AI system. other strategies contain developing algorithms that can generate natural language motives for his or her output or that could spotlight the most relevant capabilities in the facts that contributed to a selected selection.

In general, XAI is a vital area inside the development of accountable and ethical AI systems, and it is likely to play an increasingly more vital position as AI becomes more incorporated into our everyday lives.

The Explainable AI (XAI) program goals to create a collection of system gaining knowledge of techniques that:

1) Produce more explainable models, while maintaining a high stage of gaining knowledge of performance (prediction accuracy)
2) Enable human users to understand, appropriately trust, and efficaciously manage the emerging generation of artificially intelligent partners.

©IJRASET: All Rights are Reserved

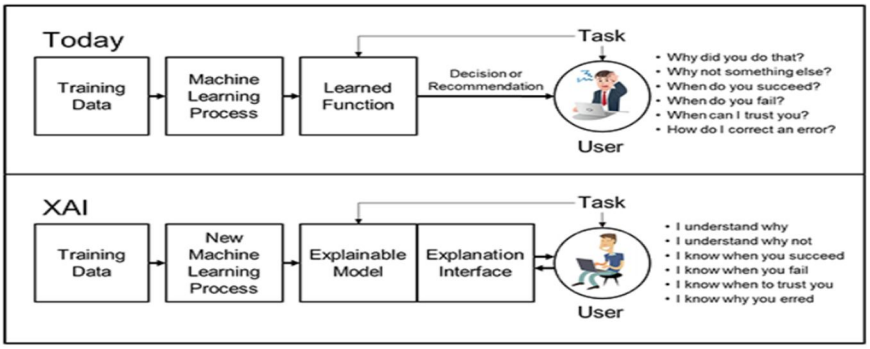


Fig 1. Introduction to Explainable AI

Explainable AI (XAI) is an emerging field that goals to make artificial intelligence (AI) more transparent and understandable to human beings. within the context of cardiovascular disease evaluation[1], XAI can assist clinicians and researchers understand how AI models arrive at their predictions, that can result in better decision-making and patient outcomes. a number of the goals of XAI for cardiovascular disease analysis encompass:

a) *Transparency:* XAI can assist uncover the black box of AI models, permitting clinicians and researchers to peer the reasoning behind the model's predictions.

b) *Interpretability:* XAI can help give an explanation for how AI model arrive at their predictions, permitting clinicians and researchers to understand the factors that contribute to the model's output.

c) *Trust:* by way of making AI greater transparent and interpretable, XAI can assist build consider among clinicians and AI models, that is crucial for the adoption of AI in healthcare.

d) *Scientific selection-making:* XAI can help clinicians make greater knowledgeable decisions by means of supplying them with the reason behind the AI model's predictions.

e) *Patient Outcomes:* By means of allowing better decision-making, XAI can in the long run result in advanced patient consequences within the prognosis, remedy, and control of cardiovascular ailment.

A. Dataset

For our system we have used heart disease dataset, which is collected from UCI repository. All information about the dataset is given in following table.

Dataset	Instance	Attributes
Heart Disease Dataset	918	12

Table1.Total Instances

Attribute	Values
1. id (Unique id for each patient)	Integer
2. age (Age of the patient in years)	Integer
3. sex	1 = male; 0 = female
4. cp: chest pain type	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
5. trestbps (resting blood pressure)	in mm Hg on admission to the hospital
6. chol	serum cholesterol in mg/dl
7. fbs (if fasting blood sugar > 120 mg/dl)	1 = true; 0 = false
8. estecg: resting electrocardiographic results	Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
9. thalach: maximum heart rate achieved	
10. exang: exercise-induced angina	True/ False (1 = yes; 0 = no)
11. oldpeak: ST depression induced by exercise relative to rest	Float
12. slope: the slope of the peak exercise ST segment	Value 1: upsloping Value 2: flat Value 3: downsloping
13. ca: number of major vessels	0-3
14. thal	3 = normal; 6 = fixed defect; 7 = reversable defect
15. num: the predicted attribute simply attempting to distinguish presence	values (1,2,3,4) from absence (value 0)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

II. FEATURE-BASED TECHNIQUES

This segment gives function-based totally model explainability strategies, which de-notice how much the input features make a contribution to model's output. there are numerous function-primarily based methods available consisting of Local Inter-pretable Model-agnostic Explanations (LIME) and Shapley Additive Explana-tions (SHAP).

A. LIME

LIME (Local Interpretable Model-agnostic Explanations) is an explainable AI technique used to interpret the predictions made by a machine learning model at the local level. It provides an explanation of how a particular prediction is made by the model by highlighting the important features that contributed to the prediction. Here we are using random forest as our model to generate explanations

LIME creates a surrogate model that is trained on a small subset of the original data, in the vicinity of the prediction to be explained. The surrogate model is usually a linear model or a decision tree that is easier to interpret. The important features are identified by perturbing the input features and observing the impact on the prediction. The features that have the greatest impact on the prediction are considered the most important.

LIME can be applied to a wide range of machine learning models, including complex models such as neural networks. It is model-agnostic, which means it can be used with any machine learning algorithm, including black box models.

The output of LIME is a set of local feature importance scores that indicate the importance of each input feature in the prediction for a particular instance. These scores can be visualized using various techniques such as bar charts or heat maps, to help users understand how the model makes predictions[2].

LIME is useful for explaining individual predictions and gaining insights into how a machine learning model works. It can help increase the transparency and trust in the model by providing interpretable explanations for its predictions.

Mathematical formula

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Steps of LIME for cardiovascular ailment evaluation:

1) Train a black-box machine learning model: A machine getting to know model is trained on a huge dataset of patient heath records to predict the risk of cardiovascular ailment.
2) Select a patient for explanation: A affected person record is selected for which the machine learning model has made a prediction of cardiovascular disease.
3) Generate local data points: LIME generates a hard and fast of perturbed data points round the selected patient record to create a neighborhood community. This community represents the range of statistics points which can be similar to the selected file and on which the local explanation might be based totally.
4) Train a local interpretable model: A local interpretable model is trained on the nearby of perturbed data points. This model is designed to be more obvious and interpretable than the original black-box model and can be used to offer an explanation for the prediction made on the selected affected person report.
5) Generate feature importance weights: LIME generates characteristic importance weights that suggest how a good deal each function contributes to the prediction made on the chosen patient record. these weights are primarily based on the coefficients of the neighborhood interpretable model and are used to perceive the most important risk features riding the prediction.
6) Present the explanation: The features importance weights are supplied to the clinician or patient in a human-readable layout, such as a listing of the most important features or a visualization that indicates how different functions contribute to the general prediction.

©IJRASET: All Rights are Reserved

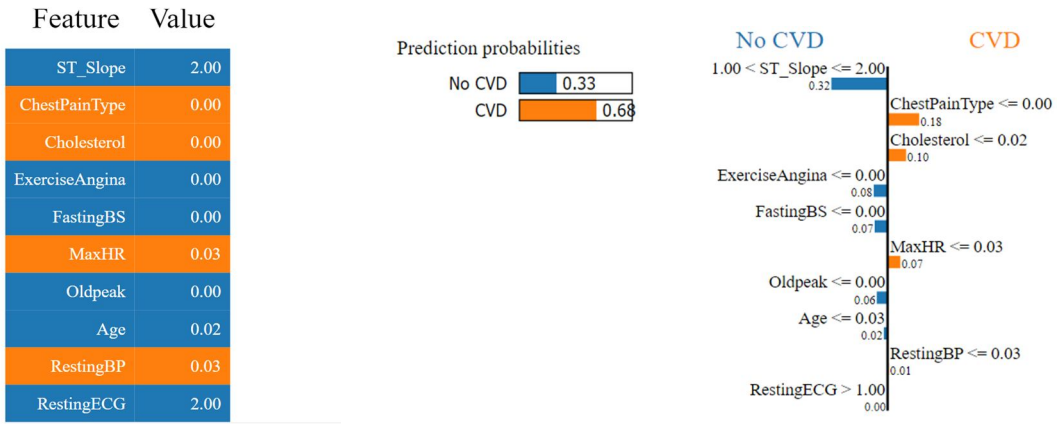


Fig 2.Lime Explanations genrated on the dataset

B. SHAP

Shapley Additive exPlanations (SHAP) values determine how each attribute contributes to the model's prediction. The technique assigns an importance value to each feature or risk factor based on how much it contributes to the prediction made by the model. This importance value is then used to create an explanation for the model's prediction. In the context of cardiovascular disease analysis, SHAP can help identify the most important risk factors associated with the disease. These risk factors may include age, gender, blood pressure, cholesterol levels, smoking status, and family history of heart disease. By understanding the importance of these risk factors, doctors and healthcare professionals can make more informed decisions about treatment and prevention strategies for patients.

SHAP sets a mean prediction (base value) of the model and identifies the relative contribution of every feature to the deviation of the target from the base. It can give both local as well as global explanations[3].

Mathematical Formula:

$$SHAP_{feature}(x) = \sum_{set:feature \in set} [|set| \times \binom{F}{|set|}]^{-1} [Predict_{set}(x) - Predict_{set \setminus feature}(x)]$$

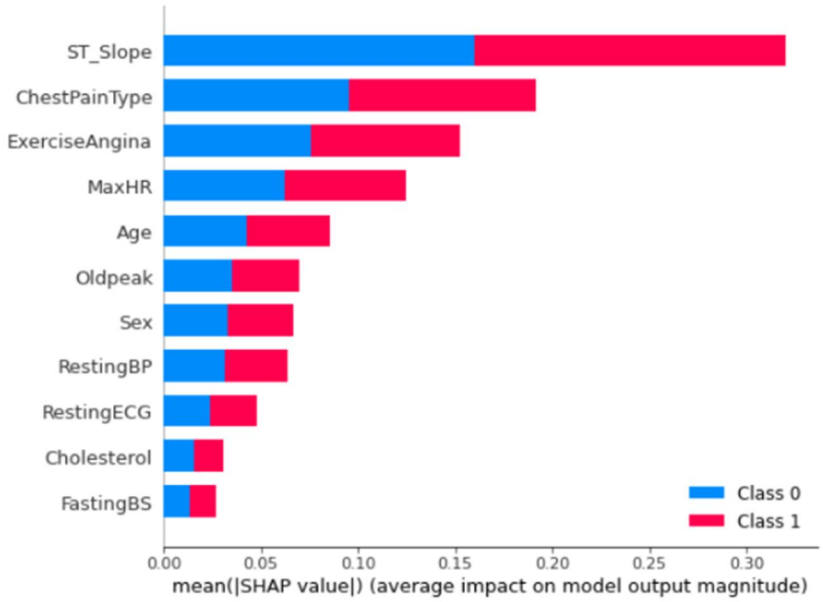


Fig 3. Contribution of each feature in the output using SHAP

1) Local Explanations

We will execute on several instances in order to show how the SHAP values behave as a local explanation method. Local explanations refer to the explanation of a single prediction made by a machine learning model. Specifically, it involves calculating the contribution of each feature (or variable) to the prediction made by the model for a particular instance (or observation) in the dataset.

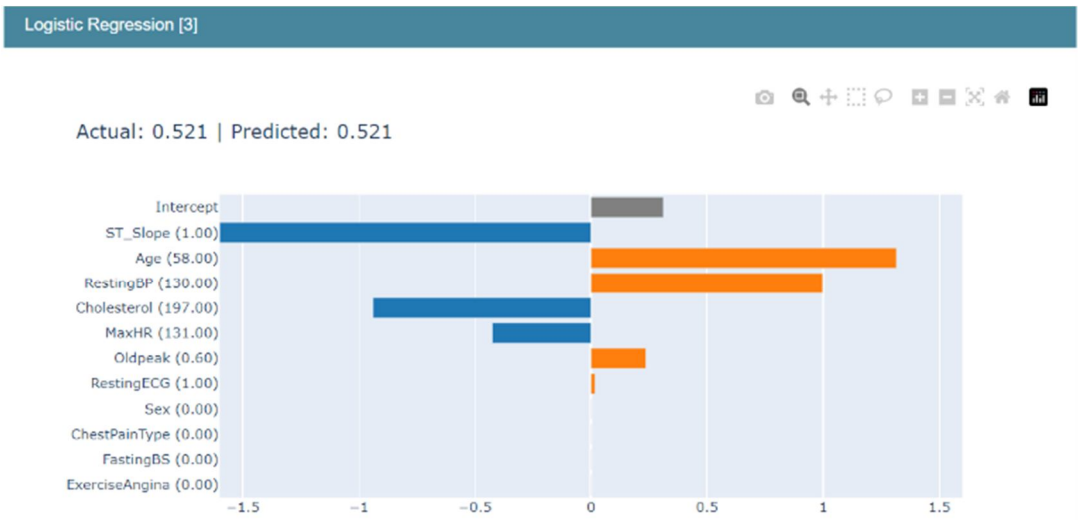


Fig 4: contribution of each feature on a particular sample

2) Global Explanations

The collective SHAP values got shows how much each feature contributes, how it contributes i.e. positively or negatively to the final prediction. Global SHAP explanations refer to an explanation of how the model behaves on average across the entire dataset. This means that for each feature, the global SHAP value represents its average contribution to the model output over all the instances in the dataset. Global SHAP values can be visualized using a SHAP summary plot, which shows the features sorted by importance and the direction of their impact on the model's output.

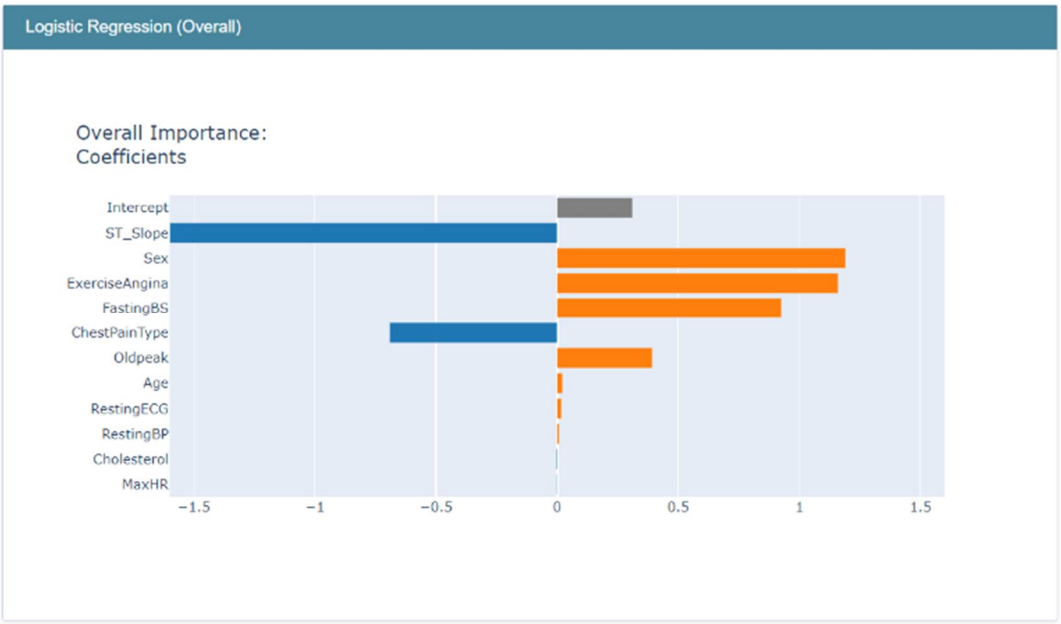


Fig 5: Impact of feature with scale positive and negative in the prediction

3) TensorFlow Decision Forest

TensorFlow Decision Forest (TF-DF) is an open-source framework for training and deploying large-scale decision forests in TensorFlow. TF-DF is designed to be scalable and efficient, allowing you to train decision forests on large datasets with millions or even billions of examples. TF-DF also provides tools for visualizing and interpreting the resulting models, as well as for deploying them in production environments.

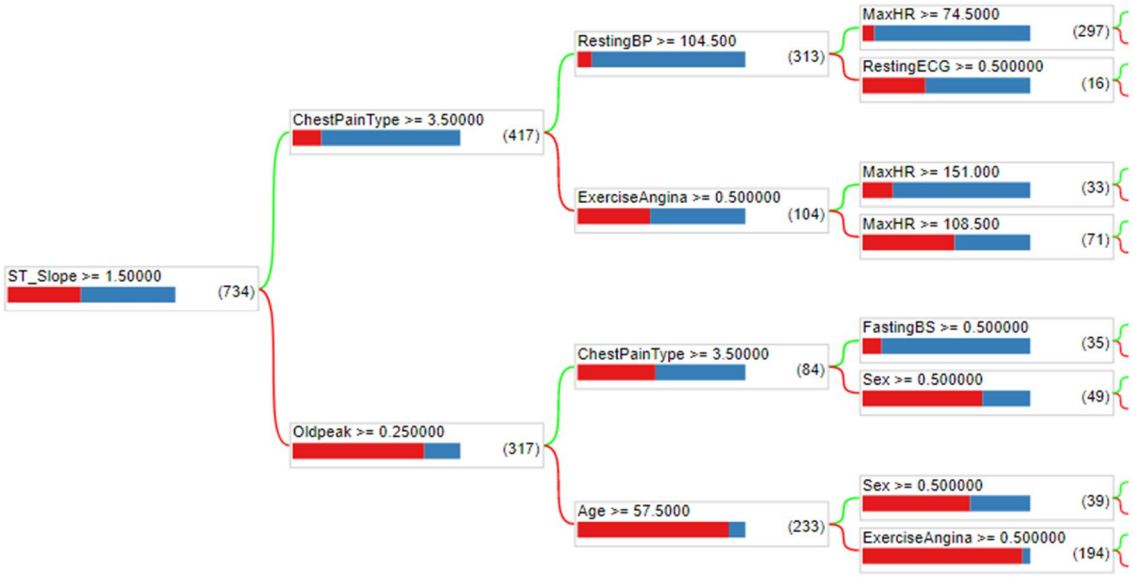


Fig 6. TensorFlow Decision Forest

C. What-if Tool

The What-If Tool (WIT) provides an easy-to-use interface for expanding understanding of a black-box classification or regression ML model. What-If Tool is an interactive and user-friendly visual interface used for exploring and understanding machine learning models. It allows users to perform "what-if" analysis on a trained model by modifying input features and observing the effect on the model's predictions in real-time.

The What-If Tool provides several features that help users to understand how the model works and how to improve its performance. Some of the key features of the What-If Tool include:

1) Data exploration: Users can visualize and explore the dataset used to train the model, and filter data based on different criteria.
2) Model analysis: Users can examine the model's performance on different subsets of the data, and evaluate its accuracy and bias.
3) Input modification: Users can modify the input features to the model and observe the effect on the prediction. This allows users to test different scenarios and understand how the model responds.
4) Counterfactual analysis: Users can generate counterfactual examples by modifying input features to achieve a desired outcome. This can help users understand how to improve the model's performance and make it more fair and transparent.

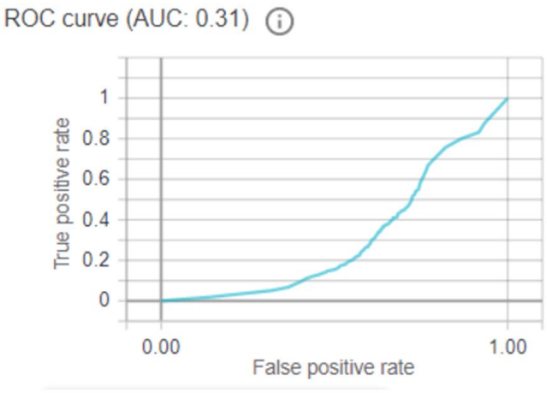
The What-If Tool is an open-source software developed by Google, and can be used with a wide range of machine learning models and input data types. It can be used in various applications such as healthcare, finance, and marketing, to provide insights and recommendations for decision-making.

In proposed system What-if analysis is done using TensorFlow decision-forest algorithm.

4) ROC Curve

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classifier. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different threshold values of the classifier.

The true positive rate (TPR) is the proportion of actual positive samples that are correctly classified as positive by the classifier. The false positive rate (FPR) is the proportion of actual negative samples that are incorrectly classified as positive by the classifier.



The area under the ROC curve (AUC) is a widely used metric to evaluate the overall performance of a binary classifier. The AUC is a number between 0 and 1, with a higher value indicating better classifier performance. An AUC of 1 corresponds to a perfect classifier, while an AUC of 0.5 corresponds to a random classifier.

5) Confusion Matrix

Confusion Matrix

	Predicted Yes		Predicted No		Total	
Actual Yes	7.4%	(68)	13.7%	(126)	21.1%	(194)
Actual No	49.9%	(458)	29.0%	(266)	78.9%	(724)
Total	57.3%	(526)	42.7%	(392)		

1) True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
2) True negatives (TN): We predicted no, and they don't have the disease.
3) False positives (FP): We predicted yes, but they don't actually have the disease.
4) False negatives (FN): We predicted no, but they actually do have the disease.

Edit - Datapoints 880 and 820

Search features

Feature	Value(s)	Counterfactual value(s)
Age	52	52
ChestPainType	NAP	TA
Cholesterol	199	298
ExerciseAngina	N	N
FastingBS	1	1
HeartDisease	0	0
MaxHR	162	178
Oldpeak	0.5	1.2
RestingBP	172	152
RestingECG	Normal	Normal
Sex	M	M
ST_Slope	Up	Flat

Fig 7. Datapoint editor in WIT showing actual and counterfactual values for every feature

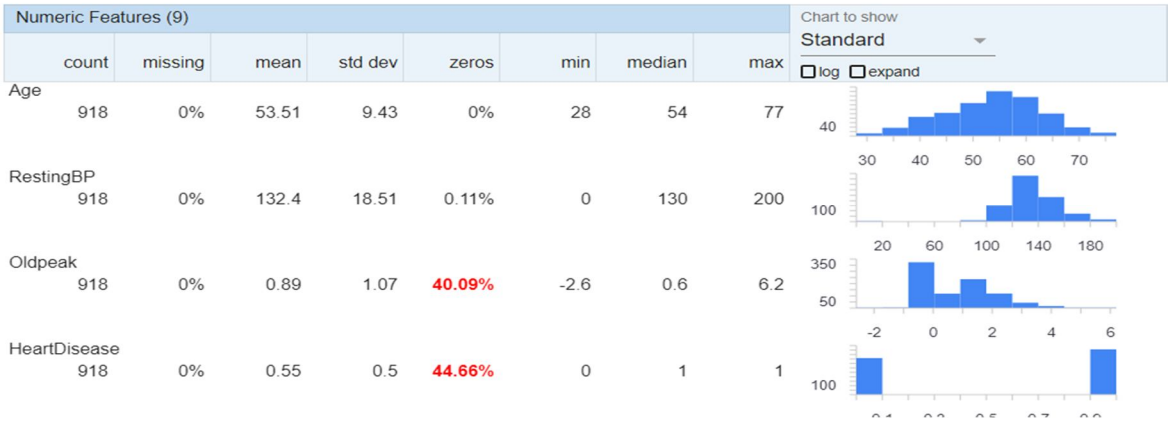
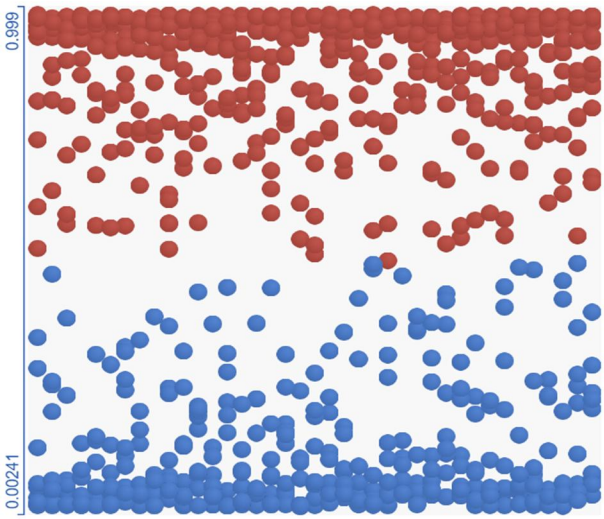


Fig 8. figure tab in WIT, summarizing dataset features



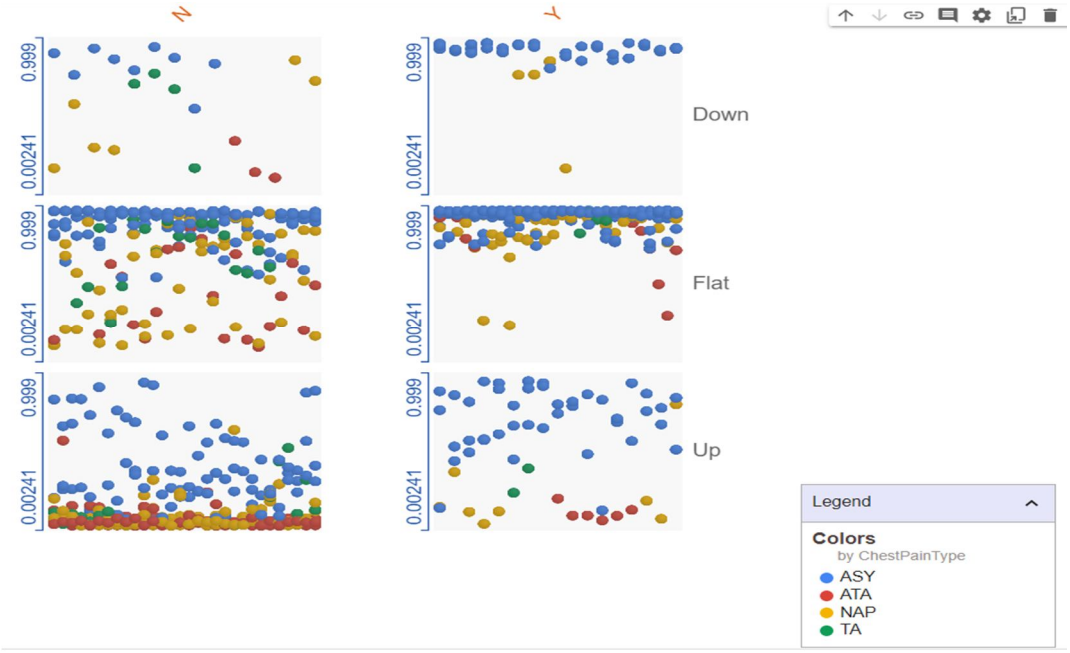


Fig 9. Scatter plot of ST-slope and exercise angina colored by Chest pain type

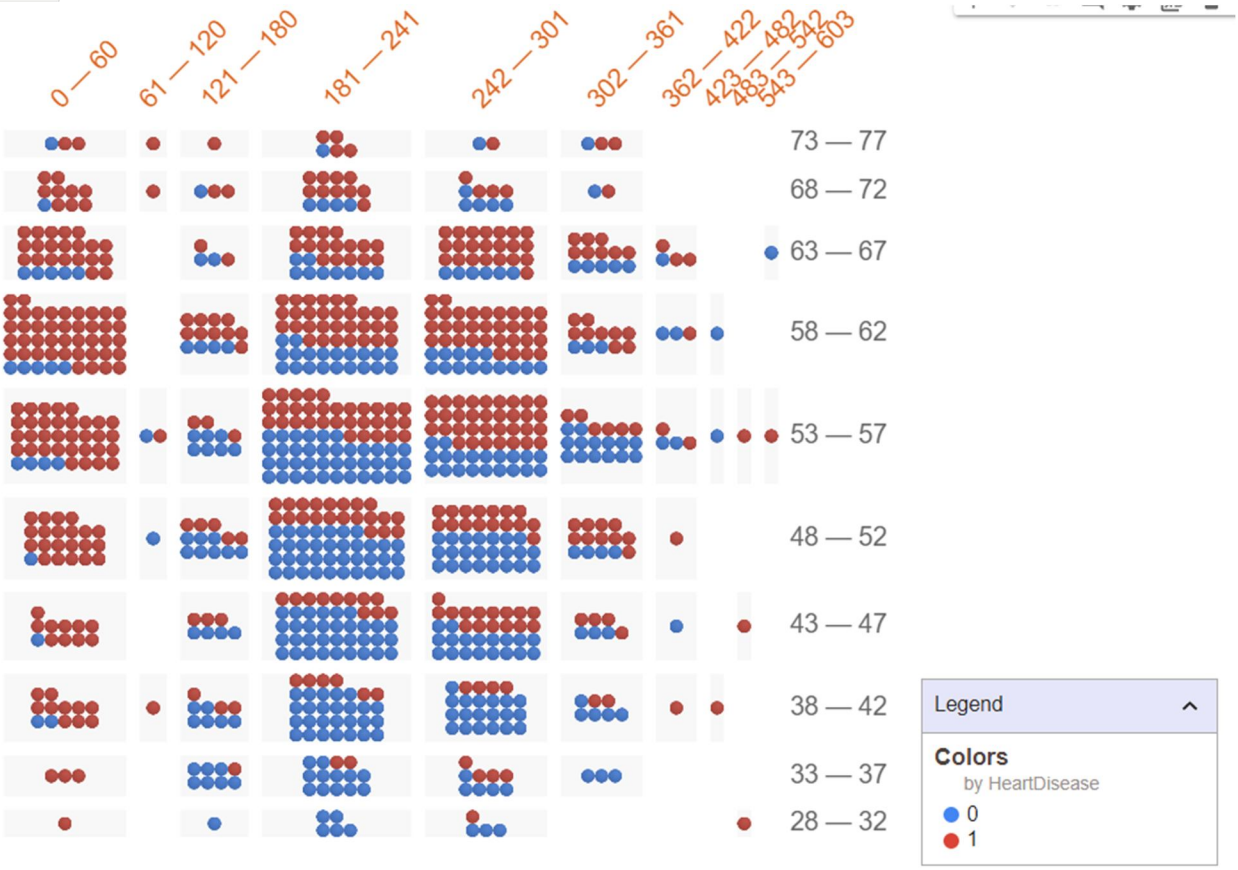


Fig 10. two-dimensional histogram of age and cholesterol

III. RESULT

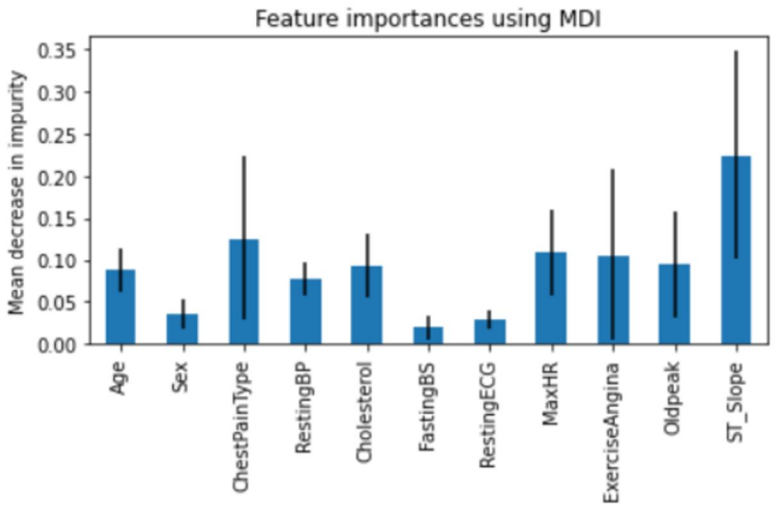
Algorithm	Accuracy%	True Positive Rate %	False Positive Rate %	Error%
Random Forest	87.5	92.92	9.21	12.5
Decision Tree	81.52	85.85	17.72	18.47
Logistic Regression	86.95	90.90	11.39	13.04
KNeighbors	88.04	89.89	12014	11.95
SVM	86.95	92.92	9.33	13.04.

Table 1. Performance Analysis of algorithms by pre-processing dataset

By analysis of the above algorithms performed on the dataset, we observed that the accuracy is majorly increased due to Random Forest after preprocessing data using normalization, label encoding and k-fold validation.

Random Forest is an ensemble learning method that creates multiple decision trees and aggregates their predictions to make a final prediction. The feature importance score is calculated based on the reduction in impurity that is achieved when a feature is used to split the data. The feature importance score can be used to identify the most important features in a dataset and to understand which features have the greatest impact on the model's predictions.

In scikit-learn, the feature importance score is normalized to add up to 1, and can be accessed through the feature_importances_ attribute of the trained RandomForestRegressor or RandomForestClassifier object. This attribute returns an array of importance scores, with the index of each score corresponding to the index of the input feature.

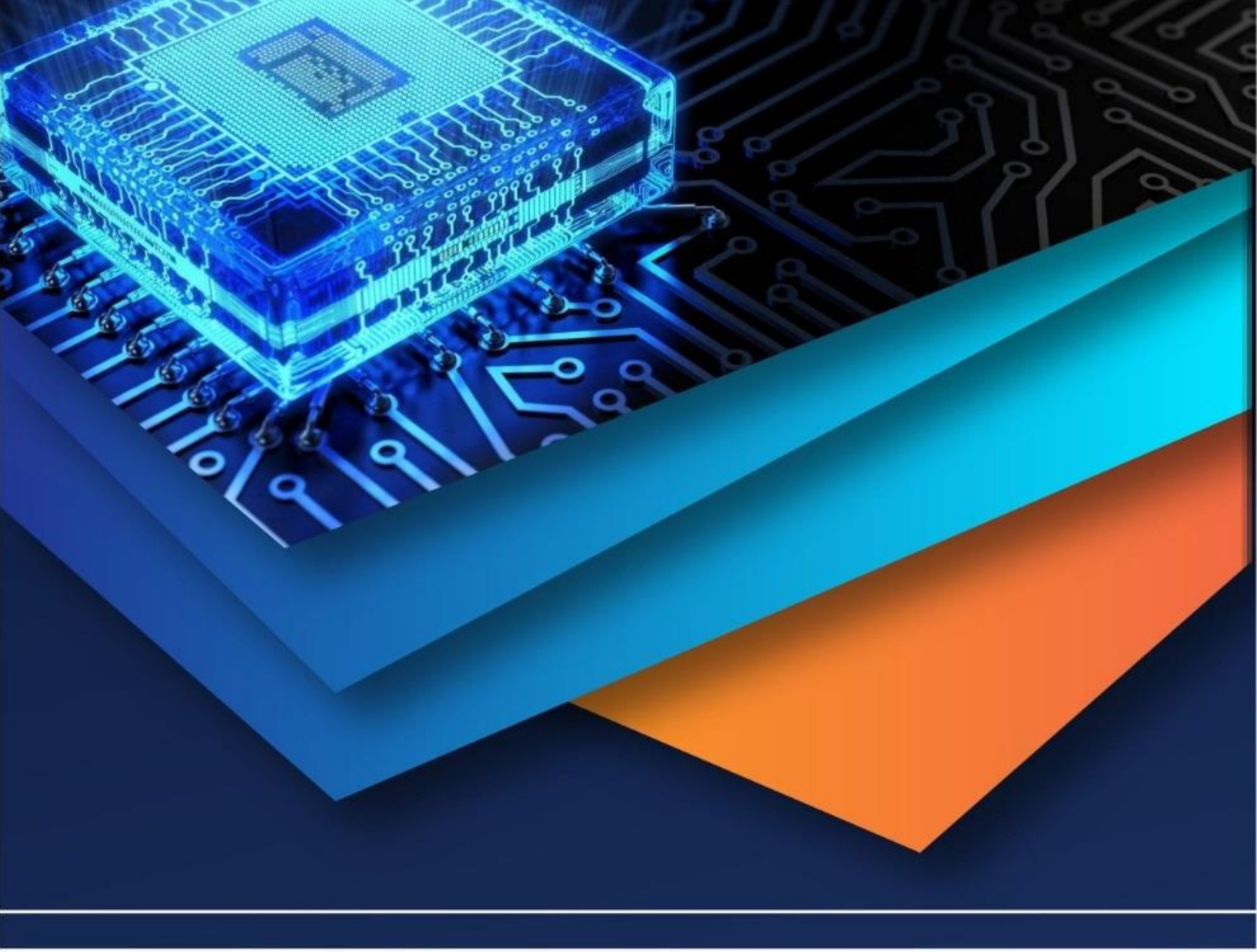


IV. CONCULSION

An application of this concept is in healthcare sector to provide addition information about a disease on the basis of previous data. The explanations generated can correct or incorrect in both cases useful insights are gained. If explanations are correct the trust in model increase hence, encouraging use of XAI and its methodologies in healthcare. If the explanations are incorrect it implies our model is not trained enough. This gives scope for improvement and development of better models and experimentation. The clinical can get better insights about a health condition by looking at the data.

REFERENCES

[1] Urja Pawar, Donna O'Shea, Susan Rea, Ruairi O'Reilly "Explainable AI in Healthcare".
[2] Marco Tulio Riberio, Sameer Singh, Carlos "Why Should I Trust You?" Explaining the Predictions of Any Classifier.
[3] Scott M. Lundberg, Su-In Lee "A Unified Approach to Interpreting Model Predictions".
[4] Sandra Wachter, Brent Mittelstadt & Chris Russell "CounterFractual Explanations Without opening the black box: Automated Decision and the GDPR.
[5] Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N. Bokoro & Ravi Sharma ""Explainable AI for healthcare 5.0: opportunities and challenge".
[6] Wagle, V., Kaur, K., Kamat, P., Patil, S., & Kotecha, K. (2021). Explainable ai for multimodal credibility analysis: Case study of online beauty health (mis)-information. IEEE Access, 9, 127985-128022.
[7] J. Wexler, et al.,"The What-If Tool: Interactive Probing of Machine Learning Models".



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)