Heart Disease Prediction Using XGBOOST Classifier

Sharada.A., Dr. Priyanka H
CSE Department,M tech Student,PES University
CSE Department,Associate Professor,PES University
sharadaa620@gmail.com
priyankah@pes.edu

ABSTRACT : The goal of this paper is to predict heart disease using xgboost classifier. A thorough examination of this topic is presented, with the aim of devising a reliable heart disease prediction analysis for clinical decision support systems. Key processes explored in this inquiry include data pre-processing, feature extraction, and the application of classifiers such as logistic regression and XGBoost algorithms. The Statlog (Heart) dataset and the Cleveland dataset, both from the UCI Machine, are the primary data sources driving the study.The prediction models are trained and tested using real-world data from the Learning Repository datasets, which contain valuable information. To measure the effectiveness of the models, various metrics are utilized, such as the confusion matrix. The experimental outcome shows that the xgboost classifier is in comparison with logistic regression is highly effective, achieving high prediction accuracy on both the Cleveland and Statlog datasets. The potential of machine learning techniques in heart disease prediction is highlighted in the paper, emphasizing their significance in clinical decision-making.

Index Terms: Heart prediction, logistic regression, XGBoost classifier, machine learning.

1.INTRODUCTION

Heart disease prediction is an important undertaking in healthcare as it facilitates in early detection, analysis, and treatment planning. Machine learning (ML) models have proven top notch capability in improving the accuracy and performance of heart disorder prediction. In this paper, the goal to expand a heart disease prediction version using ML techniques and examine its overall performance the usage of the Statlog and Cleveland datasets. [4]. ML techniques have shown promise in appropriately predicting coronary heart disease primarily based on affected person facts.Logistic regression presents interpretability, permitting insights into characteristic importance, at the same time as XGBoost classifier harnesses the collective strength of multiple vulnerable for stepped forward predictive performance [6].DBSCAN set of guidelines to choose out ability clusters or patterns in the datasets. DBSCAN, which stands for density-based spatial clustering of applications with noise, is a clustering set of rules commonly applied in records assessment and anomaly detection. It agencies together data elements which can be close to every different within the characteristic vicinity and identifies outliers or noise elements that don't belong to any cluster. In this study, it incorporate the SMOTEENN set of rules to deal with magnificence imbalance and the DBSCAN set of rules to pick out potential clusters or patterns within the datasets. These techniques helps decorate the prediction accuracy and advantage insights into the underlying characteristics and elements related to heart ailment. By leveraging the Statlog and Cleveland datasets along with the SMOTEENN and DBSCAN algorithms, goal is to develop a study heart disorder prediction version that can provide accurate and reliable predictions for improved scientific decision-making[2]-[3].The SMOTEENN set of policies, brief for synthetic minority over-sampling technique edited nearest neighbors, is a combination of famous strategies used to address elegance imbalance in datasets. Class imbalance occurs while the variety of instances belonging to 1 elegance is appreciably smaller than the alternative. In coronary heart disease prediction, the presence of coronary coronary heart sickness instances is regularly fairly lower than non-sickness instances.SMOTEENN combines the SMOTE set of guidelines, which generates synthetic samples for the minority beauty (heart sickness times), and the Edited Nearest Neighbors (ENN) set of regulations, which removes noisy samples from the bulk elegance (non-disease instances). By oversampling the minority magnificence and doing away with noisy samples, SMOTEENN allows to balance the elegance distribution, leading to stepped forward general performance and higher prediction accuracy.The examine makes use of a heart disorder dataset and evaluates the models the use of numerous overall performance metrics to evaluate their effectiveness in coronary heart ailment prediction [2]. Heart disorder is a large health state of affairs globally, and its accurate prediction performs a critical characteristic in improving affected man or woman outcomes have confirmed promising outcomes in several healthcare programs, together with coronary heart disorder prediction.By reading massive datasets and identifying complicated styles, ML models can help in early detection, evaluation, and treatment planning.

This paper targets to broaden an effective coronary heart illness prediction model using ML strategies, specially logistic regression and XGBoost classifier, to decorate clinical selection-making.Heart sickness prediction has received big interest because of its ability to improve. ML models leverage the strength of records analysis and pattern popularity to perceive hidden relationships among affected person attributes and the probability of coronary heart illness.By integrating multiple predictors, at the side of age, gender, levels of cholesterol, and blood stress, those parametres can provide insights to clinicians and resource in chance stratification.The primary targets of this study a strong coronary heart disease prediction version and study its overall performance using actual-global datasets.

The Statlog dataset is a famous dataset typically utilized in heart disorder prediction research. It contains a set of scientific features and patient attributes, including age, cholesterol levels, and resting blood strain. The dataset is labeled, with each example indicating the presence or absence of heart ailment. By analyzing this dataset, are able to reach ML models to study patterns and relationships among those capabilities and the presence of coronary heart disease[23]-[24]. These datasets comprise a whole set of patient abilties, consisting of scientific and demographic information, taking into account comprehensive analysis and accurate prediction.

This paper specializes in the utility of logistic regression and XGBoost classifier, aiming to evaluate the overall performance and suitability for coronary heart disorder prediction [5].

2.LITERATURE SURVEY

Heart disease prediction has been drastically studied inside the subject of machine mastering and healthcare. Several researchers have explored various techniques and algorithms to enhance the accuracy and effectiveness of heart disorder prediction models. In this literature survey, its an overview of some relevant studies and their key findings as follows:

This HDPM i.e, Heart Disease Prediction Model is explained in N. L. Fitriyani et.al,[1]. This study centered on addressing elegance imbalance, a common challenge in coronary coronary heart sickness prediction, the use of the synthetic minority over-sampling technique (SMOTE). The authors performed the SMOTE set of rules to oversample the minority elegance (heart disorder times) and balance the dataset.They in comparison the general overall performance of various classifiers, together with logistic regression, choice trees, and adequate-nearest associates, earlier than and after applying SMOTE. Results confirmed that SMOTE successfully improved the prediction accuracy, particularly for classifiers that have been touchy to elegance imbalance.This observe investigated the usage of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) set of rules for anomaly detection in coronary heart disorder prediction. The authors applied DBSCAN to select out functionality clusters or patterns inside the dataset, specializing in bizarre instances or outliers. By studying the diagnosed anomalies, the acquired insights into the characteristics and elements associated with coronary heart ailment. The outcomes showed that DBSCAN successfully detected anomalies and supplied treasured facts for understanding the underlying patterns in coronary coronary heart ailment prediction.

.In [3], D. Bertsimas,et.Al, explains that how it could implement novel generation to extract ECG facts. This observe explored the use of ensemble studying strategies, consisting of AdaBoost, bagging, and stacking, for heart disorder prediction. Experimental effects showed that ensemble mastering strategies outperformed man or woman classifiers, attaining better accuracy and robustness in heart disease prediction.

In [6], Mythili, T. et al. This observe focused on function choice strategies and assist vector machines (SVM) for coronary heart sickness prediction. The authors carried out a genetic algorithm to pick out the maximum applicable functions from a massive set of clinical attributes. SVM models have been then educated using the chosen functions. The effects tested that the genetic algorithm-primarily based function choice stepped forward the prediction accuracy of the SVM models as compared to the usage of the complete feature set.This look at in comparison the overall performance of a couple of system gaining knowledge of algorithms, along with logistic regression, decision trees, random forests, support vector machines, and artificial neural networks, for heart disease prediction. The authors utilized a dataset comprising uses.

3. PROPOSED METHODOLOGY

The proposed method uses XGBOOST that is to predict heart disease,DBSCAN is to detect outlier of parameters,SMOTTEENN is to balance the whole dataset.Here statlog and Cleveland dataset is used.

The ML models are used like logistic regression and XGBOOST classifier as comparison study.Steps involved in this methodology is as follow:

Data Collection: The first step in the proposed methodology is to gather the required dataset for heart disease prediction. In this study, two datasets will be utilized: the Cleveland dataset and the Statlog dataset[23][24].The Cleveland dataset contains various clinical and non-clinical attributes related to heart disease, while the Statlog dataset provides a comprehensive set of features for heart disease prediction.

Data Preprocessing: Once the datasets are collected, preprocessing steps will be applied to ensure the data is suitable for training and testing machine learning models[14][15]. This includes handling missing values, normalizing or standardizing features, and encoding categorical variables if necessary.

Feature Selection:Common feature selection methods include correlation analysis, information gain, and recursive feature elimination.

Model Selection: Several machine learning algorithms will be considered for heart disease prediction. The selection of models will be based on their suitability of the problem.

Results Analysis: The results obtained from the evaluation and comparison, and testing phases will be analyzed and interpreted.The accuracy, precision, recall, and other relevant metrics will be reported, along with any insights gained from the study [28].

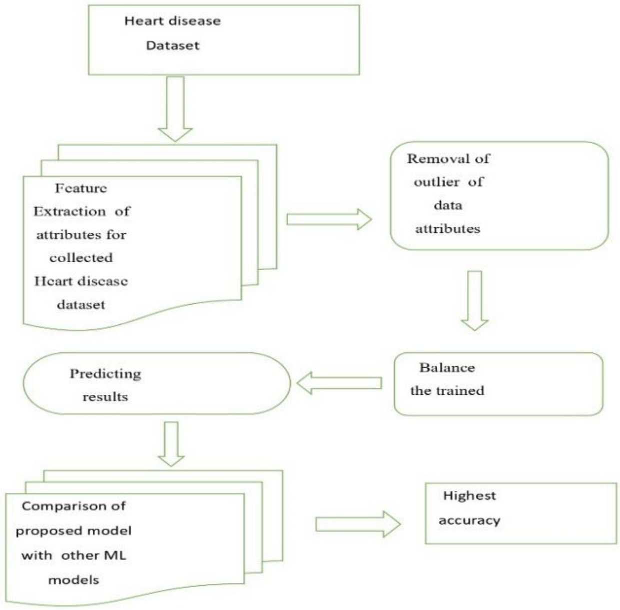

Heart disease Dataset
Feature Extraction of attributes for collected Heart disease dataset
Removal of outlier of data attributes
Balance the trained
Predicting results
Comparison of proposed model with other ML models
Highest accuracy

Fig 1: Architecture and proposed implemented

The fig 1 shows the complete method of proposed model,first by collecting dataset (statlog and Cleveland) and by implementing some methods like DBSCAN for detecting outlier,SMOTTENN for balancing dataset,XGBOOST for predicting disease logistic regression and XGBOOST classifier is ML model which is used in this work.

4. Implementation

The implementation combines these algorithms handle class imbalance, perform classification with logistic regression and XGBoost, and apply clustering using DBSCAN, quantitative measure models.[1].

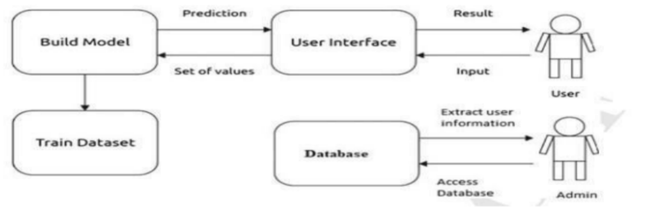


Fig 2 : Architecture of the system

The above fig 2 describes the whole methodology ,this model is trained by using experimental data and that is trained to use identify the data and its values. In this section whole implementation of work will be explained step by step in ML model languages. Also, the exact design methodology which is held in the whole implementation part.First and foremost thing is to do preprocessing and loading the data as shown in fig 3.below.

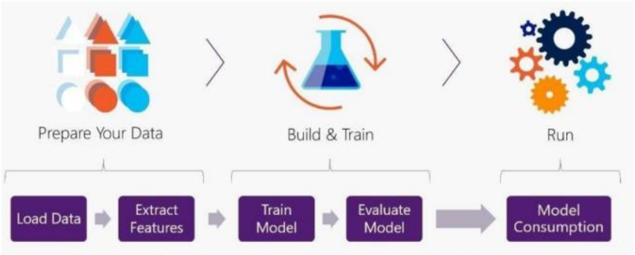


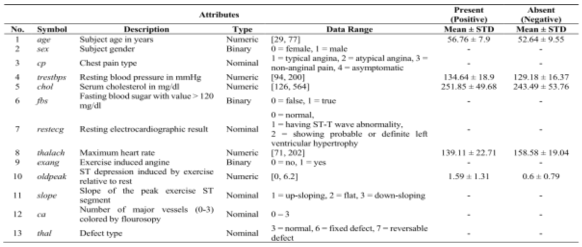
Fig 3: procedure of implementing ML model

The steps will be followed while building and training ML model with the flow.

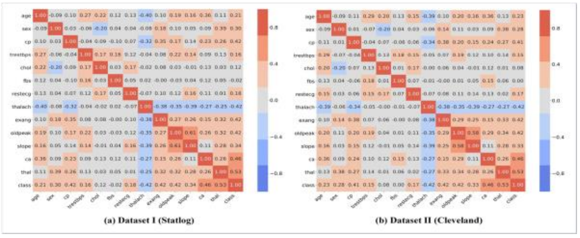
A. HEART DISEASE DATASET

Dataset II is given in Table 1 along with description below.It is having two datasets namely Statlog and Cleveland.This dataset is having12attributeslike:age,gender,thalach,exang,ca,chol,restecg,fbs ,thal,restecg,trestbps,target.

table 1: description of both datasets (Cleveland and statlog)



Attributes					Present (Positive)	Absent (Negative)
No.	Symbol	Description	Type	Data Range	Mean ± STD	Mean ± STD
1	*age*	Subject age in years	Numeric	[29, 77]	56.76 ± 7.9	52.64 ± 9.55
2	*sex*	Subject gender	Binary	0 = female, 1 = male	-	-
3	*cp*	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	-	-
4	*trestbps*	Resting blood pressure in mmHg	Numeric	[94, 200]	134.64 ± 18.9	129.18 ± 16.37
5	*chol*	Serum cholesterol in mg/dl	Numeric	[126, 564]	251.85 ± 49.68	243.49 ± 53.76
6	*fbs*	Fasting blood sugar with value > 120 mg/dl	Binary	0 = false, 1 = true	-	-
7	*restecg*	Resting electrocardiographic result	Nominal	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy	-	-
8	*thalach*	Maximum heart rate	Numeric	[71, 202]	139.11 ± 22.71	158.58 ± 19.04
9	*exang*	Exercise induced angine	Binary	0 = no, 1 = yes	-	-
10	*oldpeak*	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]	1.59 ± 1.31	0.6 ± 0.79
11	*slope*	Slope of the peak exercise ST segment	Nominal	1 = up-sloping, 2 = flat, 3 = down-sloping	-	-
12	*ca*	Number of major vessels (0-3) colored by flourosopy	Nominal	0 – 3	-	-
13	*thal*	Defect type	Nominal	3 = normal, 6 = fixed defect, 7 = reversable defect	-	-



(a) Dataset I (Statlog)

(b) Dataset II (Cleveland)

Fig 4:Using heatmap how graph will be plotted using both datasets attributes(statlog and Cleveland)

The fig 4(a) and (b) shows how graph is plotted using both datasets 1 and 2 by having attributes of it.

The DBSCAN is very useful for removal of outlier unwanted attributes of dense regions of datasets[1].

B. DBSCAN TECHNIQUE RULES AND REGULATIONS

In the implementation, DBSCAN is applied to the scaled training data to perform clustering.The unique clusters are identified, and for each cluster, the majority label is determined by counting the labels of the data points within that cluster.This information can be used to gain insights into the clusters and their majority labels.

First the dataset to identify the points of attributes dataset[28]. The algorithm 1 is showing the pseudocode of DBSCAN and fig 4 is showing DBSCAN will remove outlier of dense regions.

Table 2 is showing the results after removing the dense regions of Outliers and parameters.

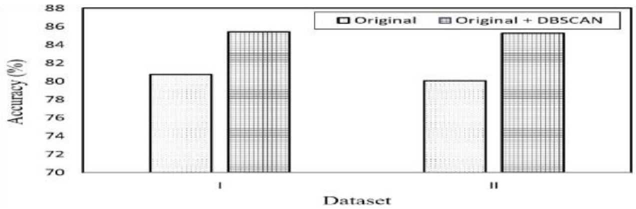


Fig 5: The graph of after removing outliers of parameters using DBSCAN

```
Algorithm 1 DBSCAN Pseudocode
  Input: dataset, D; minimum point, minPts; radius, eps
  Output: clustered C and un-clustered data UC
  for each sample point SP in dataset D do
      if SP is not visited then
          mark SP as visited
          neigbrPts ← samples points in ε-neighborhood of SP
          if sizeof(neigbrPts) < minPts then
              mark SP as UC
          end
          else
              add SP to new cluster C
              for each sample point SP' in neigbrPts do
                  if SP' is not visited then
                      mark SP' as visited
                      neigbrPts' ← samples points in
                      ε-neighborhood of SP'
                      if sizeof(neigbrPts') ≥ minPts then
                          neigbrPts ← neigbrPts + neigbrPts'
                      end
                  end
                  if SP' is not a member of any cluster then
                      add SP' to cluster C
                  end
              end
          end
      end
  end
```

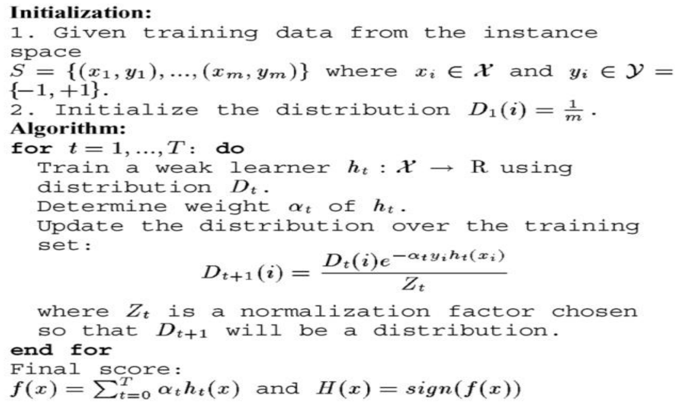
table 2: The results of parameters using DBSCAN outliers

Dataset	*MinPts*	*eps*	# Outlier Data
Dataset I (Statlog)	5	9	3
Dataset II (Cleveland)	5	8	6

XGBOOST IMPLEMENTATION:

XGBoost (Extreme Gradient Boosting) is a powerful and widely used machine learning algorithm that belongs to the ensemble learning family. It is particularly popular in data science competitions and has gained significant attention due to its high predictive performance and scalability. XGBoost is an extension of the gradient boosting method that incorporates several advanced features to enhance model accuracy and efficiency.The algorithm works by building an ensemble of weak prediction models, typically decision trees, in a sequential manner. Each subsequent model is trained to correct the mistakes made by the previous models. The final prediction is obtained by aggregating the predictions of all individual models.Algorithm 3 is pseudocode of XGBOOST implementation. XGBoost is an implementation of gradient boost choice bushes designed for velocity and performance.

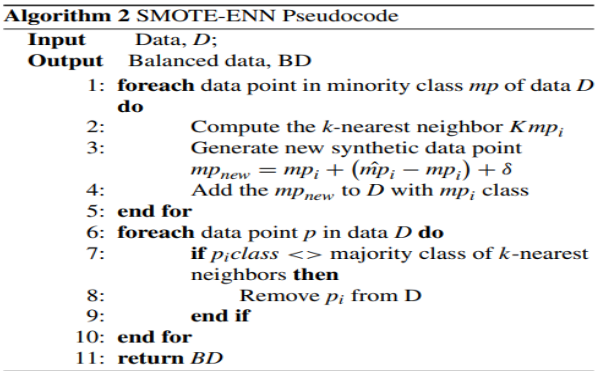
Algorithm 3: Pseudocode of XGBOOST Classifier



```
Initialization:
1. Given training data from the instance space
$S = \{(x_1, y_1), ..., (x_m, y_m)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1, +1\}$.
2. Initialize the distribution $D_1(i) = \frac{1}{m}$.
Algorithm:
for $t = 1, ..., T$: do
  Train a weak learner $h_t : \mathcal{X} \to \mathrm{R}$ using distribution $D_t$.
  Determine weight $\alpha_t$ of $h_t$.
  Update the distribution over the training set:
      $D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$
  where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution.
end for
Final score:
$f(x) = \sum_{t=0}^{T} \alpha_t h_t(x)$ and $H(x) = sign(f(x))$
```

C.SMOTTEENN BALANCED METHODS

In the implementation, SMOTE-ENN is applied to the scaled training data to handle class imbalance.The fit_resample() function is called to resample the data, resulting in a balanced training set with synthetic examples and potentially removed noisy examples.

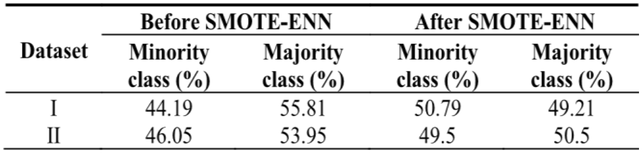


Algorithm 2 SMOTE-ENN Pseudocode

```
Input   Data, $D$;
Output  Balanced data, BD
foreach data point in minority class $mp$ of data $D$ do
    Compute the $k$-nearest neighbor $Kmp_i$
    Generate new synthetic data point
    $mp_{new} = mp_i + (\hat{mp}_i - mp_i) + \delta$
    Add the $mp_{new}$ to $D$ with $mp_i$ class
end for
foreach data point $p$ in data $D$ do
    if $p_i class <>$ majority class of $k$-nearest neighbors then
        Remove $p_i$ from D
    end if
end for
return $BD$
```

This oversampling techniques is used to data balancing and it is divided into three categories namely over-sampling,under-sampling and hybrid sampling.

table 3: SMOTEENN having two phases



Dataset	Before SMOTE-ENN		After SMOTE-ENN	
	Minority class (%)	Majority class (%)	Minority class (%)	Majority class (%)
I	44.19	55.81	50.79	49.21
II	46.05	53.95	49.5	50.5

The above table 3 shows how SMOTEENN is used in two phases having both minority and majority classes.

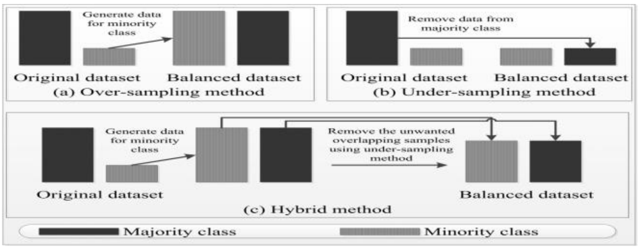

Generate data for minority class
Original dataset
Balanced dataset
(a) Over-sampling method
Remove data from majority class
Original dataset
Balanced dataset
(b) Under-sampling method
Generate data for minority class
Remove the unwanted overlapping samples using under-sampling method
Original dataset
Balanced dataset
(c) Hybrid method
Majority class
Minority class

Fig 6: DBSCAN-eliminating the outlier of sampling method

fig 6 shows how the graph will be plotted after removing outlers using DBSCAN by using both datasets [28]-[29].This also shows both original dataset and balanced dataset.

D.XGBOOST IMPLEMENTATION METRICS

Basically xgboost algorithm will be on some rules with regulations by calculating some formulae and equations to be solved for example shown below equation (1)

$$L(\phi) = \sum_i l(\hat{y}_{i,}, y_i) + \sum_k \Omega(f_k);$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \qquad (1)$$

The term l here is the differentiable convex loss function that calculates the difference between the prediction y^i and the target yi. While the regularized term "Ω" penalizes the complexity of the model and the number of leaves in the tree are represented using T . Furthermore, each fk corresponds to an independent tree structure q and leaf weight w. Finally, the term γ corresponds to the threshold and pre-pruning is performed while optimizing to limit the growth of the tree and λ is used to smooth the final learned weights to prevent overfitting.

It is implemented XGBoost using the XGBoostV0.81 python library.HDPM will be implemented to the datasets will represents with positive results by raising with prediction accuracy by comparing with other models[1].Here it have done comparison study on logistic regression with XGBOOST classifier and implementing with confusion matrix. In this section all implemented code with resultant graph and plots will be shown and will be explained in detail. First by collecting heart disease dataset, here by using UCI Statlog and Cleveland dataset[23][24] is involved. By doing feature selection of attributes in dataset and DBSCAN is implemented to remove the outlier of clusters and noise of attributes.

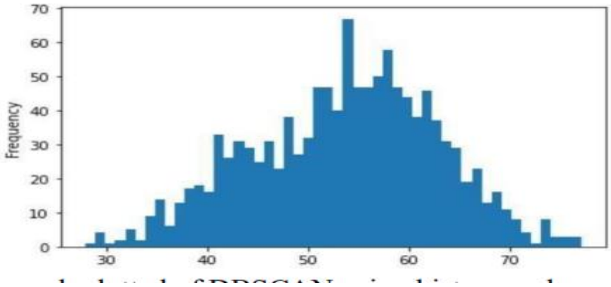
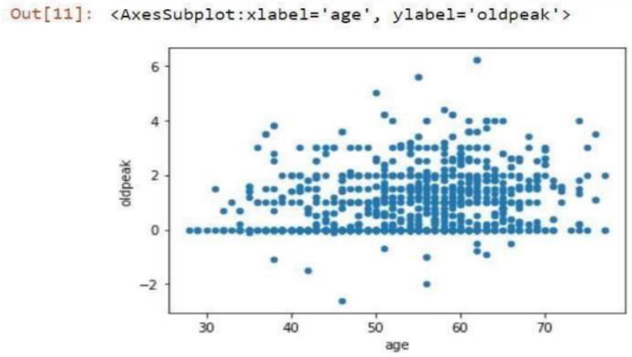


Fig 7: DBSCAN using histogram by extracting age in dataset.

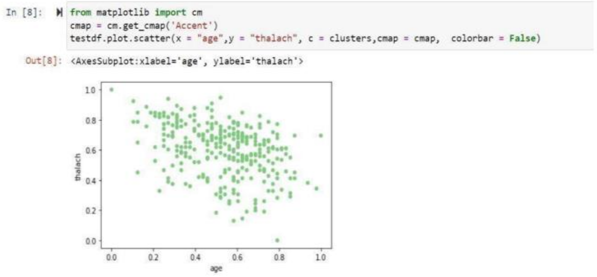
The fig 7 shows the DBSCAN is used hist feature of extract age attribute in dataset.



```
Out[11]:  <AxesSubplot:xlabel='age', ylabel='oldpeak'>
```

Fig 8: removal of outliers in scattered shape.

The fig 8 shows how DBSCAN can be used outliers using scattered usually it starts in a oscillating beginning assumption values and neighbourhood in the pt grabbed using for all values with distance between two points. Henceforth, the value will be noted as planes and in two phases it will be pointed as “visited”.

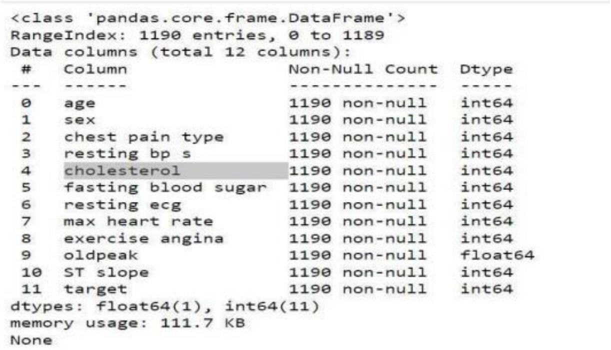


```
In [8]:  from matplotlib import cm
         cmap = cm.get_cmap('Accent')
         testdf.plot.scatter(x = "age",y = "thalach", c = clusters,cmap = cmap,  colorbar = False)

Out[8]:  <AxesSubplot:xlabel='age', ylabel='thalach'>
```

Fig 9: implementation of scatter points using different attributes.

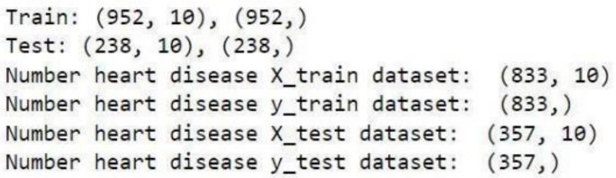
The fig 9 shows about DBSCAN is involving scatter points for different y points using attribute “thalach” and checks distance.This fig is showing the scattered points.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  1190 non-null   int64
 1   sex                  1190 non-null   int64
 2   chest pain type      1190 non-null   int64
 3   resting bp s         1190 non-null   int64
 4   cholesterol          1190 non-null   int64
 5   fasting blood sugar  1190 non-null   int64
 6   resting ecg          1190 non-null   int64
 7   max heart rate       1190 non-null   int64
 8   exercise angina      1190 non-null   int64
 9   oldpeak              1190 non-null   float64
 10  ST slope             1190 non-null   int64
 11  target               1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
None
```

Fig 10: balancing data from overlapping on one another

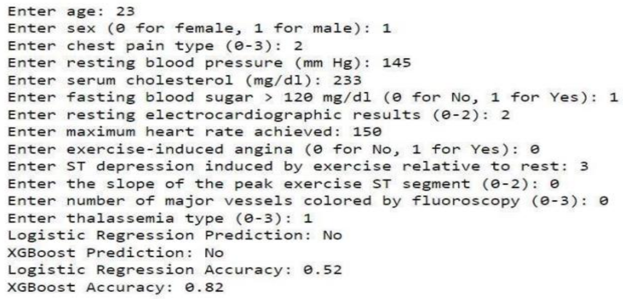
By implementing SMOTEENN the below fig 10 shows how this SMOTEENN will balance the attributes using dataframe (df) size along will nulltypes and counting the number and it will also prevent data from overlapping on one another.



```
Train: (952, 10), (952,)
Test: (238, 10), (238,)
Number heart disease X_train dataset:  (833, 10)
Number heart disease y_train dataset:  (833,)
Number heart disease X_test dataset:  (357, 10)
Number heart disease y_test dataset:  (357,)
```

Fig 11:the split of 70:30 ratio and also describes info about train and test set

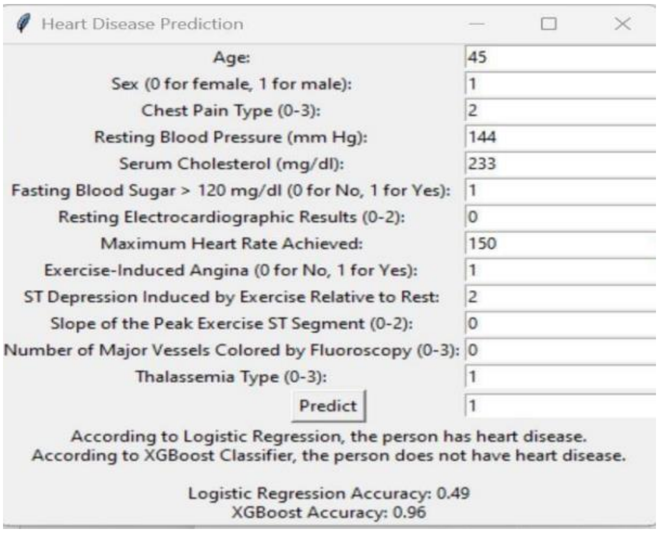
The fig 11 shows the SMOTEENN will train and balance the data by splitting in ratio 70:30 both testing and training the data and also describes the same.



```
Enter age: 23
Enter sex (0 for female, 1 for male): 1
Enter chest pain type (0-3): 2
Enter resting blood pressure (mm Hg): 145
Enter serum cholesterol (mg/dl): 233
Enter fasting blood sugar > 120 mg/dl (0 for No, 1 for Yes): 1
Enter resting electrocardiographic results (0-2): 2
Enter maximum heart rate achieved: 150
Enter exercise-induced angina (0 for No, 1 for Yes): 0
Enter ST depression induced by exercise relative to rest: 3
Enter the slope of the peak exercise ST segment (0-2): 0
Enter number of major vessels colored by fluoroscopy (0-3): 0
Enter thalassemia type (0-3): 1
Logistic Regression Prediction: No
XGBoost Prediction: No
Logistic Regression Accuracy: 0.52
XGBoost Accuracy: 0.82
```

Fig 12: Manually giving values in data attributes and getting accuracy of models.

The fig 12 shows how the accuracy will be calculated manually by having comparison study of both models having accuracy it showing xgboost having highest accuracy compared to logistic regression. (0.82 or 82%).



Heart Disease Prediction

Age: 45
Sex (0 for female, 1 for male): 1
Chest Pain Type (0-3): 2
Resting Blood Pressure (mm Hg): 144
Serum Cholesterol (mg/dl): 233
Fasting Blood Sugar > 120 mg/dl (0 for No, 1 for Yes): 1
Resting Electrocardiographic Results (0-2): 0
Maximum Heart Rate Achieved: 150
Exercise-Induced Angina (0 for No, 1 for Yes): 1
ST Depression Induced by Exercise Relative to Rest: 2
Slope of the Peak Exercise ST Segment (0-2): 0
Number of Major Vessels Colored by Fluoroscopy (0-3): 0
Thalassemia Type (0-3): 1

1

According to Logistic Regression, the person has heart disease.
According to XGBoost Classifier, the person does not have heart disease.

Logistic Regression Accuracy: 0.49
XGBoost Accuracy: 0.96

Fig 13: GUI to predict the disease

In fig 13 by using GUI to predict the heart disease and to have accuracies of comparison of both models (XGBOOST and Logistic regression).In this also XGBOOST is posing with highest percentage of accuracy and predicting disease of patient.Some attributes are like:age,thalach,resting blood pressure,exang,ca,chol,cp,thal and target.

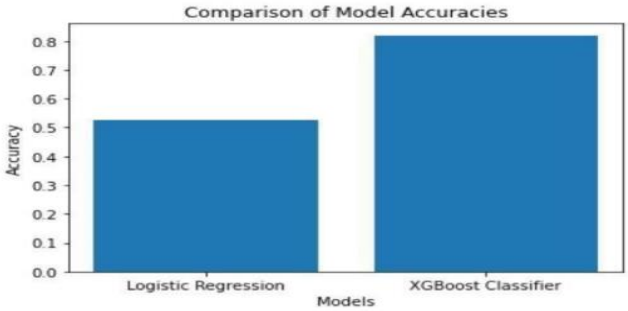


Fig 14: comparison of model accuracies.

The fig 14 shows the graphical plot of accuracies of comparison percentage of both proposed model and another model (i.e., XGBOOST vs Logistic regression).Here also its showing XGBOOST having highest accuracy frequency.

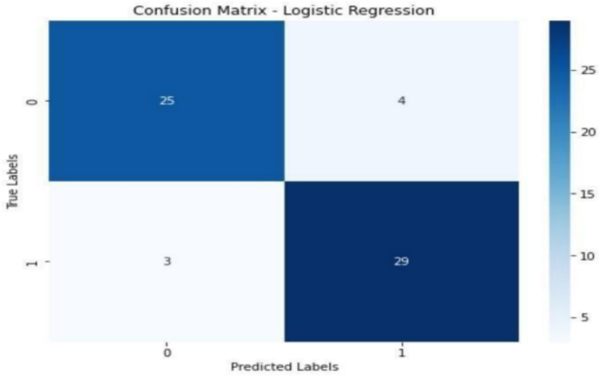


Fig 15: logistic regression accuracy using confusion matrix.

The fig 15 shows the accuracy percentage in predicting disease by implementing in confusion matrix using heat map correlation.

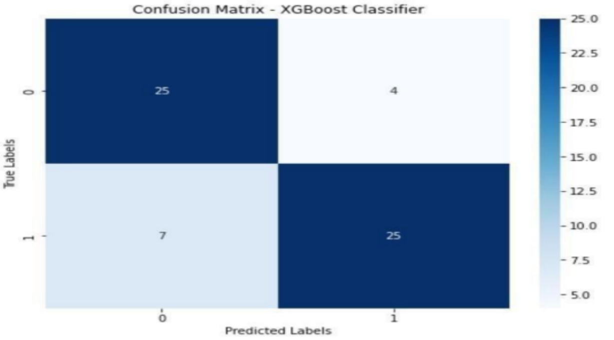


Fig 16: XGBOOST classifier using confusion matrix.

The fig 16 shows the accuracy percentage in predicting heart disease and displaying the accuracy percentage using heat map correlation.

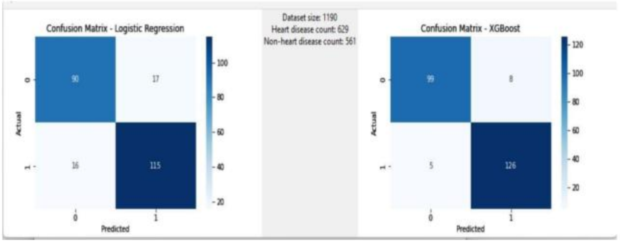


Fig 17: displaying the size of dataset and exact count of patients having disease and also not having disease.

In fig 17 is predicting the heart disease through dataset by calculating the size of whole dataset along with the exact count of number of patients who is having disease and who are not having disease by implementing in GUI along with comparison study and deploying in confusion matrix plotting in heatmap graph.

CONCLUSION:

The goal of the study was to develop an effective prediction model to assist in clinical decision-making for the early detection and management of heart disease.The research utilized the XGBoost classifier, a powerful machine learning algorithm known for its ability to handle complex data patterns. The classifier was trained and evaluated using the Statlog and Cleveland datasets, which provided a diverse range of patient features and heart disease labels.The experimental results demonstrated that the XGBoost classifier achieved high accuracy and robust performance in predicting heart disease. The model effectively captured the underlying patterns and relationships between the input features and the target variable, enabling accurate predictions of heart disease presence or absence.

REFERENCES:

[1] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511

[2] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.30011

[3] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[4] D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.

[5] Mythili, T. et al. “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL).” International Journal of Computer Applications 68 (2013): 11-15.

[6] A. Bhowmick, K. D. Mahato, C. Azad and U. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 60-65, doi: 10.1109/AIC55036.2022.9848885.

[7] A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan and S. M. Marium, "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033683.

[8] C. Bemando, E. Miranda and M. Aryuni, "Machine-LearningBased Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 232-237, doi: 10.1109/ICSECS52883.2021.00049.

[9] H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-5, doi: 10.1109/ATSIP49331.2020.9231760.

[10] N. N. Itoo and V. K. Garg, "Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 599-604, doi: 10.1109/MECON53876.2022.9751883.

[11] D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752947.

[12] Reddy, Kummita Sravan Kumar and K. V. Kanimozhi. “Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM.” 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (2022): 1-5.

[13] S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319, doi: 10.1109/AEMCSE55572.2022.00071.

[14] G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh, "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022, pp. 01-06, doi: 10.1109/MLSP55214.2022.9943373.

[15] K. G, K. G and D. M. Raja S, "Modelling an Efficient Heart Disease Prediction System using Norm- and Regularization based Learning Approach," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 1923-1927, doi: 10.1109/ICACCS54159.2022.9785202.

[16] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, doi: 10.1109/icETITE47903.2020.242.

[17] D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.

[18] D. Swain, S. K. Pani and D. Swain, "A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-6, doi: 10.1109/ICACAT.2018.8933603.

[19] 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.

[20] M. Mamun, M. M. Uddin, V. Kumar Tiwari, A. M. Islam and A. U. Ferdous, "MLHeartDis:Can Machine Learning Techniques Enable to Predict Heart Diseases?," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0561-0565, doi: 10.1109/UEMCON54665.2022.9965714.

[21] C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[22] K. Joshi, G. A. Reddy, S. Kumar, H. Anandaram, A. Gupta and H. Gupta, "Analysis of Heart Disease Prediction using Various Machine Learning Techniques: A Review Study," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 105-109, doi: 10.1109/DICCT56244.2023.10110139.

[23] Statlog (Heart) Data Set. Accessed: Oct. 2, 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

[24] Heart Disease Data Set. Accessed: Oct. 2, 2019. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Heart+Disease..

[25] K.-A. Toh, J. Kim, and S. Lee, "Maximizing area under ROC curve for biometric scores fusion," *Pattern Recognit.*, vol. 41, no. 11, pp. 3373–3392, Nov. 2008, doi: 10.1016/j.patcog.2008.04.002.

[26] S. H. Jee *et al.*, "A coronary heart disease prediction model: The korean heart study," *BMJ Open*, vol. 4, no. 5, May 2014, Art. no. e005025, doi: 10.1136/bmjopen-2014-005025.

[27] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185–196, Nov. 2019, doi: 10.1007/s12065-019-00327-1.

[28] B. R. Kirkwood, J. A. C. Sterne, and B. R. Kirkwood, *Essential* Medical Statistics, 2nd ed. Malden, MA, USA: Blackwell Science, 2003.

[29] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The differences and similarities between two-sample T-test and paired T-test," *Shanghai Arch, Psychiatry*, vol. 29, no. 3, pp. 184–188, Jun. 2017, doi: 10.11919/j.issn.1002-0829.217070.

[30] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018, doi: 10.3390/s18072183.