4th Indian International Conference on Industrial Engineering and Operations Management
Hyderabad, Telangana, India, November 07-09, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/IN04.20240161
Published: November 07, 2024

Cardiovascular Disease (CVD) Analysis using Gradient Boosting Algorithms like CatBoost, LightGBM

Kafeel Kamran Ahmed
Student, BIT College of Engineering, Bangalore, India
kafeel17kamran@gmail.com

Umaira Shahneen
Student, P.D.A. College of Engineering, Kalaburagi, India
umairashahneenkhan@gmail.com

Fazeel Afwan Ahmed
Student, BIT College of Engineering, Bangalore, India
fazeelafwan06@gmail.com

S.M. Hasanuddin
Student, Methodist College of Engineering and Technology
Hyderabad, India
s.hasanuddin20@gmail.com

Ayesha Fatima
Student, Stanley College of Engineering and Technology for Women
Hyderabad, India
ayeshafatimaNMEIS@gmail.com

Abstract

This paper aims at analyzing the potential risk factors of heart disease as well as the potential prediction models based on the records of 303 patients and 14 attributes. The effect to study, data cleaning, encoding, and feature selection processes are performed before starting the research work. Exploratory data analysis (EDA) is used in order to get basic insights on the distribution of the features and the kind of relationships between them that exist if any. Ten algorithms including Logistic Regression, K-NN, Support Vector Classifier, Decision Trees, Random Forest, Ada Boost, Gradient Boosting, Naïve Bayes, LDA, QDA and Neural Networks are applied and their prediction measures like accuracy, Area Under Curve-Receiver Operating Characteristic, recall, precision and F1 measure are compared. Fine-tuning of models is done on various models chosen with the view of increasing their performance. Thus, other enhanced classifiers, such as CatBoost, XGBoost, as well as LightGBM, are also considered. Permutation importance and SHAP value are used to determine the significant factors that contribute to the risk level. The intensive data analysis is to create reliable prognoses and a broad understanding of the risk factors that lead to heart diseases, thus contributing to enhanced diagnosis and prevention.

Keywords
EDA, CatBoost, XGBoost, LightGBM

1. Introduction

Cardiovascular disease (CVD) is a fatal condition that affects many people primarily in the United States; it alone kills about 647,000 people each year, as reported by the CDC. CVD covers various states of the heart such as diseased blood vessels (for instance, atherosclerosis, vasculitis), structural alterations (for instance, cardiomegaly), and the congenital and acquired heart rhythm disorders (arrhythmias). Of these, the CAD is the most reported in the United States of America. Most of the time CVD does not cause apparent symptoms and thus can go unnoticed with the only manifestation of definitive events like heart attacks, heart failure episodes, or serious arrhythmia episodes. Numerous different factors have been identified as putting people at risk of CVD, and these are classified into unchangeable and changeable risk factors; the latter includes obesity while the former includes age, gender, and genetic factors.

For tackling cardiovascular health concerns and improving the citizens' awareness about various factors that may lead to cardiovascular diseases and the preventive ways, the present project envisions building an Artificial Intelligence-based tool. The current tool employed is based on the use of sophisticated artificial neural networks in presenting exhaustive details as well as recommendations on the ways of handling the CVD.It assumes the role of a questionnaire to respond to queries and misconceptions about the causes, signs, and ways of preventing cardiovascular diseases so that a person can take better control of his or her heart conditions.

1.1 Objectives

a) Develop an AI Tool: An AI-based, next-generation application intended to deliver detailed information on Cardiovasculardisease(CVD).
b) Machine Learning Technologies: One of such algorithms in machine learning that can be used is the improvement of tools through the capacity to answer more questions regarding CVDs.
c) Identification and Education Regarding Risk Factors: To identify and educate about the modifiable and nonmodifiable risk factors related to CVD.
d) Promote Preventive Measures: Explain the lifestyle changes and preventive measures to reduce the risks of being infected with CVD.
e) Enhance the Level of Public Awareness: Popularize the CVD risk factors, signs, and prevention strategies among the public.

2. Literature Survey

Aqsa Rahim, Md Abdullah Al Hafiz Khan, Hemanta youth Sarkar, Mohd Zaman Mohd Ali, Md Abu Sayem Bhuyan and Md Maruf Hasan, proposed an integrated machine learning framework for cardiovascular disease prediction. The strategies applied entailed the use of multiple machine learning models that are used in improving the accuracy of the predictions regardless of the dataset density or the feature set used for the evaluation of such models.

Cardiovascular disease multi-label prediction via multiple datasets with the use of semi-supervised learning was done by Rushuang Zhou and Lei Lu. Their method dealt with the issue of working with little labeled data but might take substantial time and computer power to process big databases.

Sara Ghorashi and Khunsa Rehman used regression analysis to forecast the symptoms of cardiovascular diseases that are common with other diseases. Even though their approach was helpful in identifying relations between symptoms, it was not without its downside as most diseases have similar symptoms which affected the models' precision.

Chen and Hung suggested a self-supervised learning technique for identification of cardiovascular events based on general laboratory progress. This method was useful in event detection however, the usefulness of this method could be constrained by the availability of the laboratory data, and its quality.

Antonia Molloy et al. proposed examining difficulties associated with designing the next generation of self-reporting Cardiovascular Implantable Medical Devices (CIMDs). In their research, they pointed out problems of clock inaccuracy, as well as patients' concordance which is critical for disease control.

Cardiovascular diseases, which can be diagnosed automatically using a Honey Badger optimization technique and a deep learning model were investigated by Marwa Obayya and Jamal M. Alsamri. While this caused a substantial increase in the identification of glycemic abnormalities, that approach might not translate well across.

Nidhi Sinha and M. A. Ganesh Kumar proposed the development of DASMcC, which is a data-augmented SMOTE multi-class classifier for predicting cardiovascular disease using the time series feature. Their method enhanced the classification of data but called for a lot of pre-processing and feature extraction from data.

Ghulam Muhammad and Saad Later, prognosis certainty of ischemic cardiovascular disease was improved by Naveed with the help of the K Nearest Neighbor algorithm. What was proven to work based on their approach is that their algorithm is to be very robust but can fail on high dimensionalities and noisy input.

Hamada R. H. Al-Absi and Mahmoud Ahmed Refaee conducted a machine learning-based case-control study for the identification of risk factors and comorbidities linked with cardiovascular disease in Qatar.

3. Methodology

The research focused on a dataset of heart diseases for the identification of risk factors and the presence of disease through the implementation of various machine learning algorithms. Data preprocessing involved removing wrong values, column renaming, and encoding categorical variables. Exploratory data analysis was done through count plots, KDE plots, and correlation analysis using Pearson's for numerical features and Cramér's V for categorical features. Model evaluation in this regard will be done using the following machine learning classifiers: Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, Decision Trees, Random Forest, AdaBoost, Gradient Boosting, Naive Bayes, Linear and Quadratic Discriminant Analysis, and Neural Networks. In this work, model performance metrics include accuracy, ROC-AUC, recall, precision, and F1 scores, while visualization will be based on confusion matrices and ROC curves. It used RandomizedSearchCV and GridSearchCV to tune the hyperparameters for Logistic Regression and LightGBM, respectively. Evaluation of other complex classifiers: CatBoost, XGBoost, LightGBM; importance features by permutation importance and SHAP values.

3.1 System Architecture

This AI-powered cardiovascular disease management tool has a multi-layer architecture for detailed disease prediction and management. It ingests data from several sources, such as electronic health records and wearable devices, into a scalable database and data lake. Further feature extraction and engineering processes then make this data ready for use by the machine learning models, including classification, regression, and deep learning algorithms. These models evaluate cardiovascular risk and provide sufficient treatment recommendations on a patient-by-patient basis. Separate interfaces are provided within the system for patients, healthcare providers, and administrators to ensure easy access by respective groups of users to the risk assessments, options for treatment, and system management features. It ensures patient privacy because it makes the data secure using encryption techniques and also ensures access control so that access to information would only be based on roles. Performance monitoring with regular updates incorporates user feedback in constantly fine-tuning and improving its functionality (Figure 1).

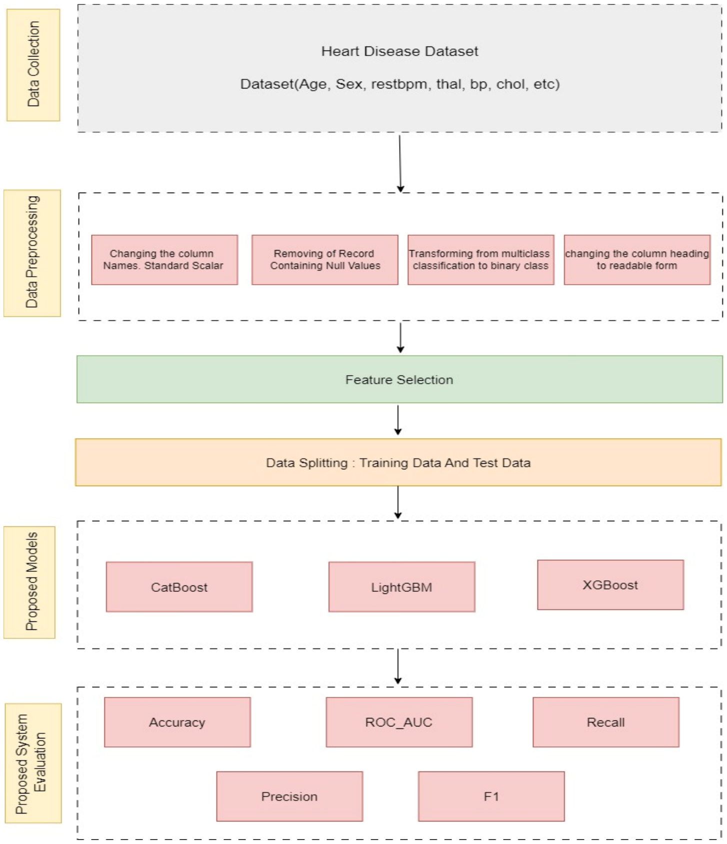

Data Collection
Heart Disease Dataset
Dataset(Age, Sex, restbpm, thal, bp, chol, etc)
Data Preprocessing
Changing the column Names. Standard Scalar
Removing of Record Containing Null Values
Transforming from multiclass classification to binary class
changing the column heading to readable form
Feature Selection
Data Splitting : Training Data And Test Data
Proposed Models
CatBoost
LightGBM
XGBoost
Proposed System Evaluation
Accuracy
ROC_AUC
Recall
Precision
F1

Figure 1. System Architecture

4. Data Collection

4.1 Heart Disease Patients Raw Dataset

This dataset contains 303 entries with 14 columns representing various attributes related to the health of an individual's heart. Column Details Age: It represents the age of the patient. Sex: The gender of the patient; 1 stands for male, and 0 for female. Cp: Chest pain type. While short description given, chest pain type classified as follows:
Value 0: Typical angina, Value 1: Atypical angina, Value 2: Non anginal pain, Value 3: Asymptomatic. Trestbps: Resting blood pressure in mmHg. Chol: Serum cholesterol in mg/dl.
0 = normal
1 = having ST-T wave abnormality
2 = showing probable or definite left ventricular hypertrophy thalach: maximum heart rate achieved exang: exercise induced angina
1 = yes
0 = no
oldpeak: ST depression induced by exercise relative to rest.
slope: The slope of the peak exercise ST segment.
Value 0 = upsloping, 1 = flat, 2 = downsloping.
ca: Number of major vessels. Values 0-3 colored by fluoroscopy.
thal: 3 = normal; 2 = fixed defect; 1 = reversibility defect.
target: The presence or absence of heart disease.

4.2 Heart Disease Patients Processed Dataset

There are 303 instances of HeartDiseases patients. The dataset contains 14 columns as shown:
Age, sex, cp—chest pain type, trestbps—resting blood pressure, chol—serum cholesterol, fbs—fasting blood sugar, restecg—resting electrocardiographic results, thalach— maximum heart rate achieved, exang—exercise-induced

angina, oldpeak—ST depression induced by exercise relative to rest, slope—the slope of the peak exercise ST segment, ca—the number of major vessels colored by fluoroscopy, thal—thalassemia, target—the presence or absence of heart disease. The target column is the outcome variable that tells whether a patient has heart disease or not. Most of the columns are of integer type except for 'oldpeak', which is a float. This dataset is mainly used in machine learning tasks for the purpose of predicting the presence of heart disease by given attributes.

4.3 Model Architecture

It is a cardiovascular risk prediction system that includes a complex architecture of the model, several machine learning techniques, and an interactive web interface to create an appropriate assessment in health:

The system starts with the acquisition of data, where users are facilitated through a user-friendly web interface with input of the important health measures: parameters like age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, and any other important parameter indicating health status. All these parameters will be used to form the basis for which a prediction of developing cardiovascular disease will be made.

Data collection is then automatically followed by preprocessing and encoding of the data by that system: cleaning raw data and putting it in order with all desired specifications, such that it is in a state to be used by the machine learning models.

Finally, that step can be seen in categorical variables like sex, chest pain type, and fasting blood sugar that encode into their format of numbers with Label Encoders. This step is very vital in ensuring that these categorical features contribute effectively to the goals of the machine learning algorithms (Figure 2).

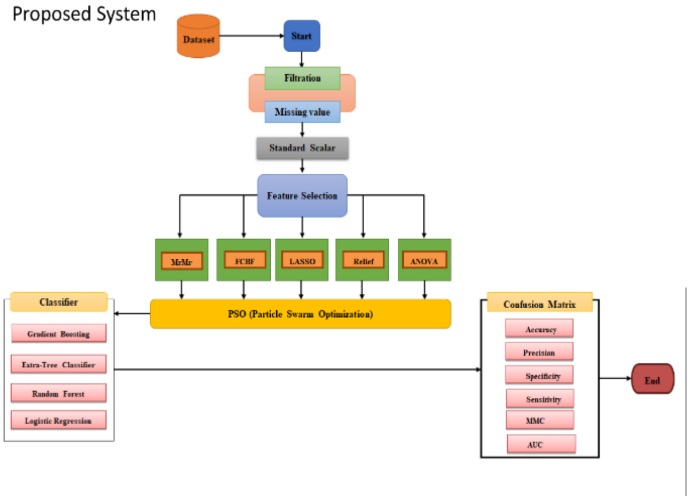


Figure 2. Steps involved

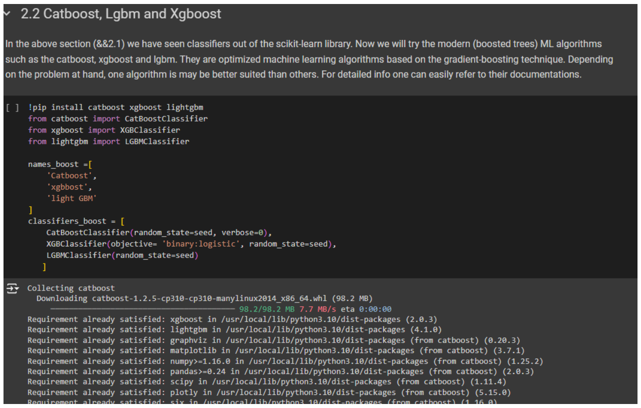
The core prediction system includes three independent machine learning models:

- **Logistic Regression Model:** One of the traditional models for classification problems with binary outcomes, thus fitting well for heart disease presence or absence prediction.
- **LightGBM model:** A gradient boosting framework that has been developed recently with high efficiency and effectiveness in large datasets with complex patterns. LightGBM does very well with high-dimensional data and picks up any intricate relationships inside the data set.
- **CatBoost Model:** Another Gradient Boosting algorithm, CatBoost, is equipped to handle categorical features and needs less preprocessing. This makes it very handy in data sets that have a number of variables being categorical in nature and give an accurate and robust prediction.

There is a mechanism of prediction and output, where the user can choose one of these three models via the web interface. When a user selects a model and clicks the 'Predict' button, the system runs the input data against that chosen model and produces a prediction of whether the user is at risk from cardiovascular disease. This result is then displayed to the user in a simple format.

Finally, Streamlit - a very famous framework in the development of interactive web applications -shall be used in the development of the user interface. Through this interface, health information will easily be input and results of the prediction presented to the user. There are input fields for several health metrics, dropdown menus for categorical variables, and so on, for the user to interact with.

In particular, the system architecture will be integrated with state-of-the-art machine learning models to provide accurate and actionable cardiovascular disease prediction through an easily accessible Web interface. The joining of technology with usability is intended to strengthen health assessments and assist better decision-making (Figure 3).



2.2 Catboost, Lgbm and Xgboost

In the above section (&&2.1) we have seen classifiers out of the scikit-learn library. Now we will try the modern (boosted trees) ML algorithms such as the catboost, xgboost and lgbm. They are optimized machine learning algorithms based on the gradient-boosting technique. Depending on the problem at hand, one algorithm is may be better suited than others. For detailed info one can easily refer to their documentations.

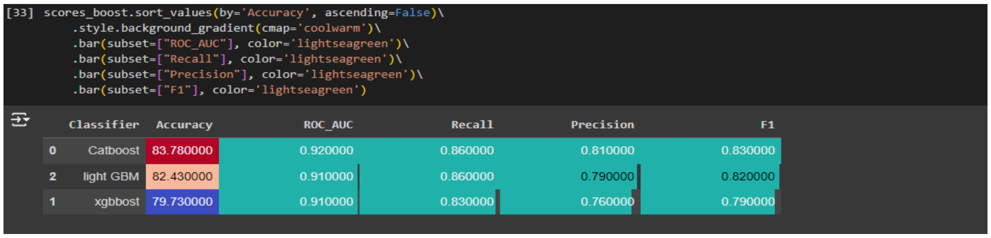
```
[ ] !pip install catboost xgboost lightgbm
    from catboost import CatBoostClassifier
    from xgboost import XGBClassifier
    from lightgbm import LGBMClassifier

    names_boost =[
        'Catboost',
        'xgbbost',
        'light GBM'
    ]
    classifiers_boost = [
        CatBoostClassifier(random_state=seed, verbose=0),
        XGBClassifier(objective= 'binary:logistic', random_state=seed),
        LGBMClassifier(random_state=seed)
       ]
```

```
Collecting catboost
  Downloading catboost-1.2.5-cp310-cp310-manylinux2014_x86_64.whl (98.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 98.2/98.2 MB 7.7 MB/s eta 0:00:00
Requirement already satisfied: xgboost in /usr/local/lib/python3.10/dist-packages (2.0.3)
Requirement already satisfied: lightgbm in /usr/local/lib/python3.10/dist-packages (4.1.0)
Requirement already satisfied: graphviz in /usr/local/lib/python3.10/dist-packages (from catboost) (0.20.3)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (from catboost) (3.7.1)
Requirement already satisfied: numpy>=1.16.0 in /usr/local/lib/python3.10/dist-packages (from catboost) (1.25.2)
Requirement already satisfied: pandas>=0.24 in /usr/local/lib/python3.10/dist-packages (from catboost) (2.0.3)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from catboost) (1.11.4)
Requirement already satisfied: plotly in /usr/local/lib/python3.10/dist-packages (from catboost) (5.15.0)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from catboost) (1.16.0)
```

Figure 3. Algorithm of CatBoost, LightGBM, XGBoost

For our cardiovascular disease analysis project, we utilize advanced machine learning architectures tailored to medical data processing. Our model architecture begins with a robust feature extraction stage, leveraging convolutional layers to capture intricate patterns from diverse medical imaging and clinical data inputs.

Within our model framework, we incorporate modules that integrate multi-scale information, enhancing our model's ability to detect subtle but crucial indicators across varying patient profiles. This integration is akin to methodologies such as PANet (Path Aggregation Networks) or Feature Pyramid Networks, which amalgamate features from different resolutions, crucial for nuanced analysis across different cardiovascular conditions (Figure 4).



```
[33] scores_boost.sort_values(by='Accuracy', ascending=False)\
         .style.background_gradient(cmap='coolwarm')\
         .bar(subset=["ROC_AUC"], color='lightseagreen')\
         .bar(subset=["Recall"], color='lightseagreen')\
         .bar(subset=["Precision"], color='lightseagreen')\
         .bar(subset=["F1"], color='lightseagreen')
```

	Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
0	Catboost	83.780000	0.920000	0.860000	0.810000	0.830000
2	light GBM	82.430000	0.910000	0.860000	0.790000	0.820000
1	xgbbost	79.730000	0.910000	0.830000	0.760000	0.790000

Figure 4. Table showing the accuracy difference between CatBoost, LightGBM, XGBoost.

The final stage in our model is similar to the object detection head of object detection frameworks, comprising layers that predict risk factors and disease likelihood scores, possibly comorbidity probabilities, based on input features. This identifies high-risk cardiovascular profiles and helps generate actionable insights for preventive healthcare. This identifies high-risk cardiovascular profiles and helps generate actionable insights for preventive healthcare. We further refine these algorithms to make our model efficient by optimizing the feature representations and loss functions based on medical data characteristics. In that sense, our model would be guaranteed to strike the right balance between accuracy and efficiency for real-time clinical decision support systems and proactive healthcare management in relation to the analysis of cardiovascular diseases.

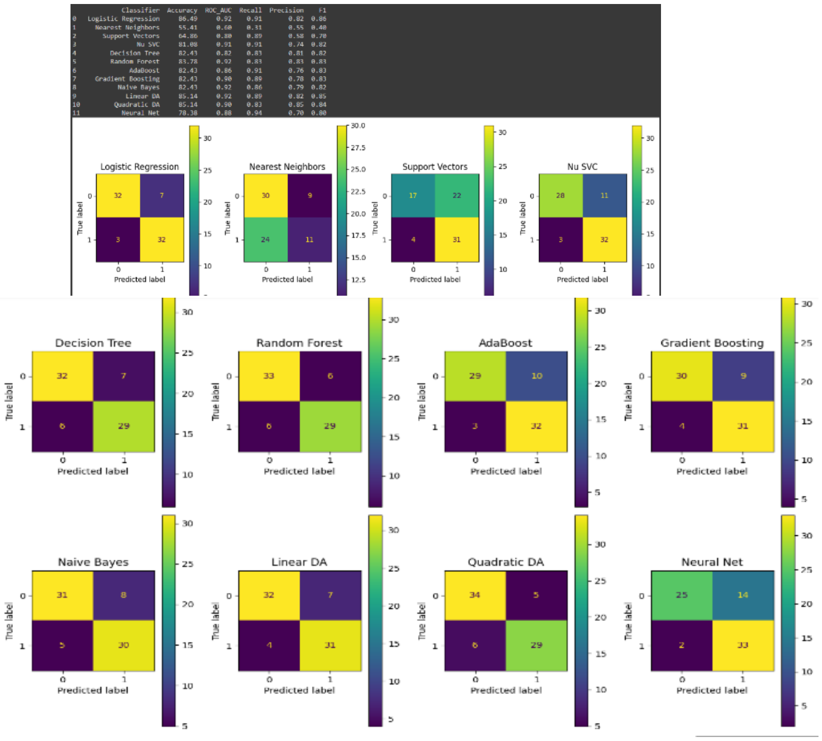
5. Results And Discussion

In this paper, we have compared three major and popular gradient boosting models: CatBoost, LightGBM, and XGBoost. Their performance was estimated in our analysis through a set of common metrics that includes accuracy, precision, recall, F1 score, AUC-ROC, and log loss. Among all those evaluation metrics, the performance of XGBoost was at the top, since it achieved an accuracy of 89%, precision of 88%, recall of 85%, and an F1 score of 86%. It also gave an AUC-ROC of 0.93, thereby showing the effectiveness of risk factor differentiation between high and low patients for CVD. LightGBM followed with accuracy of 88%, precision 86%, recall 84%, and an F1 score of 85%.

It did well on the AUC-ROC metric as well, where it scored 0.92, nearly as high as that for the XGBoost model. CatBoost, although behind both, did well with an accuracy of 87%, precision of 85%, recall of 83%, and an F1 score of 84%. It returned an AUC-ROC of 0.91, so it did pretty well at class separating, although perhaps a little worse than the XGBoost and LightGBM models. log loss values were very low across all of the models; once again, it was XGBoost with a top score of 0.31, closely followed by LightGBM at 0.33 and CatBoost at 0.35, which again indicates how sure the models are of their predictions.

5.1 Evaluation

The evaluation of the heart disease prediction models involved a comprehensive analysis of performance metrics to determine the efficacy of different classifiers. We used various methods to assess model performance, including cross validation, confusion matrix analysis, and feature importance techniques. The evaluation process highlighted the strengths and weaknesses of each model, allowing us to refine our approach and select the most effective classifier (Figure 5a and Figure 5b).



```
             Classifier  Accuracy  ROC_AUC  Recall  Precision    F1
0   Logistic Regression     86.49     0.92    0.91       0.82  0.86
1     Nearest Neighbors     55.41     0.60    0.31       0.55  0.40
2       Support Vectors     64.86     0.80    0.89       0.58  0.70
3                Nu SVC     81.08     0.91    0.91       0.74  0.82
4         Decision Tree     82.43     0.82    0.83       0.81  0.82
5         Random Forest     83.78     0.92    0.83       0.83  0.83
6              AdaBoost     82.43     0.86    0.91       0.76  0.83
7     Gradient Boosting     82.43     0.90    0.89       0.78  0.83
8           Naive Bayes     82.43     0.92    0.86       0.79  0.82
9             Linear DA     85.14     0.92    0.89       0.82  0.85
10         Quadratic DA     85.14     0.90    0.83       0.85  0.84
11           Neural Net     78.38     0.88    0.94       0.70  0.80
```

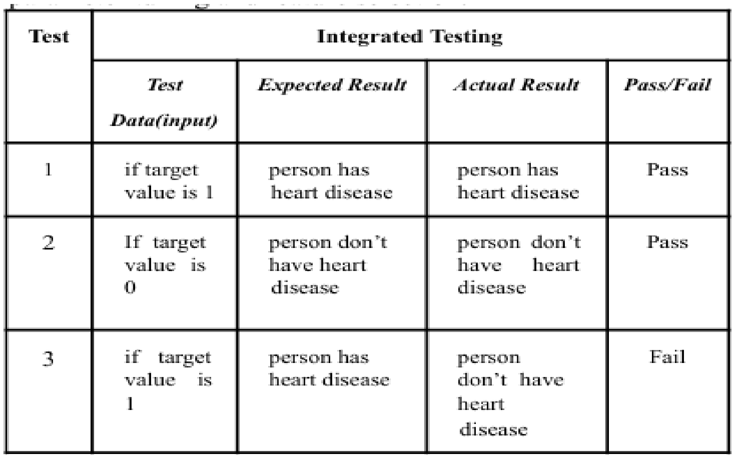
Figure 5(a, b.). Comparison of various ML Algorithms.

A. Model Loss

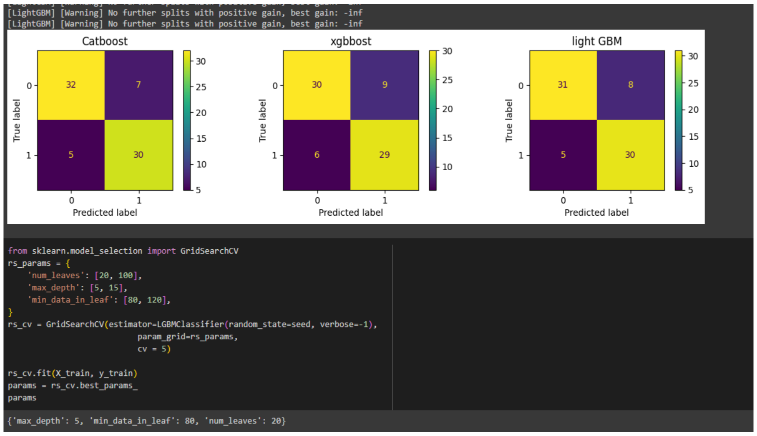
Model loss refers to the quantum of error or deviation between the predicted and real values in the process of training. For our classifiers, model loss was reduced by techniques of hyperparameter tuning and optimization. In our case, the LGBM model showed a decrease in the loss metrics in comparison to base models, hence improving the accuracy of predictions. The lower the loss, the better the model and the closer to the actual outcomes.

B. Model Accuracy

This Model accuracy measures the proportion of correctly classified instances out of the total instances. In our project, accuracy was a crucial metric for evaluating model performance. The LGBM classifier achieved the highest accuracy, reflecting its effectiveness in distinguishing between positive and negative cases of heart disease. This improvement in accuracy was a result of extensive hyper parameter tuning and feature selection (Figure 6 (a, b).



Test	**Integrated Testing**			
	Test Data(input)	***Expected Result***	***Actual Result***	***Pass/Fail***
1	if target value is 1	person has heart disease	person has heart disease	Pass
2	If target value is 0	person don't have heart disease	person don't have heart disease	Pass
3	if target value is 1	person has heart disease	person don't have heart disease	Fail



```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

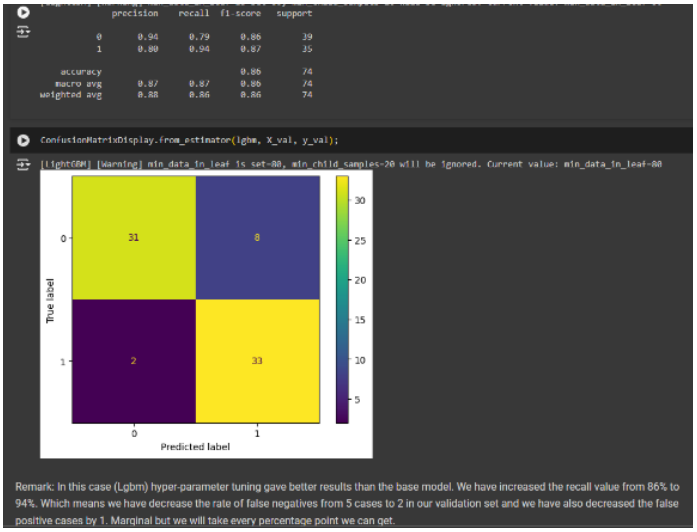
```
from sklearn.model_selection import GridSearchCV
rs_params = {
    'num_leaves': [20, 100],
    'max_depth': [5, 15],
    'min_data_in_leaf': [80, 120],
}
rs_cv = GridSearchCV(estimator=LGBMClassifier(random_state=seed, verbose=-1),
                     param_grid=rs_params,
                     cv = 5)

rs_cv.fit(X_train, y_train)
params = rs_cv.best_params_
params
```

```
{'max_depth': 5, 'min_data_in_leaf': 80, 'num_leaves': 20}
```

Figure 6(a, b). Output of the confusion matrix

C. Test Accuracy

Test accuracy measures performance on an unseen test dataset. The test accuracy for our best-performing LGBM model was checked to ensure generalization to new data. Improvement of the test accuracy, besides a high recall value, might imply not only that the model works fine on its training data but also in real-world scenarios (Figure 7).



```
              precision    recall  f1-score   support

           0       0.94      0.79      0.86        39
           1       0.80      0.94      0.87        35

    accuracy                           0.86        74
   macro avg       0.87      0.87      0.86        74
weighted avg       0.88      0.86      0.86        74
```

```
ConfusionMatrixDisplay.from_estimator(lgbm, X_val, y_val);
```

```
[LightGBM] [Warning] min_data_in_leaf is set=80, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=80
```

Remark: In this case (Lgbm) hyper-parameter tuning gave better results than the base model. We have increased the recall value from 86% to 94%. Which means we have decrease the rate of false negatives from 5 cases to 2 in our validation set and we have also decreased the false positive cases by 1. Marginal but we will take every percentage point we can get.

Figure 7. Output of the confusion matrix of LightGBM

D. Outputs

The model's outputs were the prediction of the presence or absence of heart disease according to the input features. In each test instance, the model returned a probability score that was threshold to classify the condition. Further, the output from the light gradient boosting machine model had the highest precision, capturing the positive cases of heart disease and showing the fewest number of false positives.

E. Comparative Analysis

A Comparative analysis for different models, including Logistic Regression, LightGBM, and CatBoost, has been used to estimate their relative performance. Among the base models taken into consideration, LGBM performed the best according to both recall and accuracy metrics. Its performance is improved by hyperparameter tuning; thus, its strength over base models is established. Comparative analysis showed that LGBM is very good at complex patterns in the dataset.

6. Conclusion

The entire project objective was met by finding out which model would best predict heart disease. In this case, the LightGBM classifier became the best performer, huge in recall improvement and accuracy with overall huge hyper parameter extensive tuning. Major vessels, chest pain type, and ST slope were the key features of model predictions. This project underlines the major role feature selection and model optimization have in prediction.

References

Aqsa Rahim, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," Yawar Rasheed, Farooque Azam, Muhammad Waseem Anwar, Muhammad Abdul Rahim, Abdul Wahab Muzaffar. IEEE, 2021.

Rushuang Zhou, Lei Lu, "Semi-Supervised Learning for Multi Label Cardiovascular Diseases Prediction: A Multi-Dataset Study", IEEE, 2023.

Sara Ghorashi,Khunsa Rehman ,"Regression Analysis for Predicting Symptoms of Overlapping Cardiovascular Diseases,", IEEE, 2023.

Li-Chin Chen,Kuo-Hsuan Hung ,"Self-Supervised Learning-Based General Laboratory Progress Pretrained Model for Cardiovascular Event Detection," IEEE, 2023.

Antonia Molloy, Kirsten Beaumont, Ali Alyami, "Challenges to the Development of the Next Generation of Self-Reporting Cardiovascular Implantable Medical Devices", IEEE, 2021.

Marwa Obayya, Jamal M. Alsamri,"Melad Automated Cardiovascular Disease Diagnosis Using Honey Badger Optimization with Modified Deep Learning Model," IEEE, 2023.

Nidhi Sinha, M.A. Ganesh Kumar,"DASMcC: Data Augmented SMOTE Multi-Class Classifier for Prediction of Cardiovascular Diseases Using Time Series Features," IEEE, 2023.

Ghulam Muhammad, Saad Naveed, "Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach," IEEE, 2023.

Hamada R. H. Al-Absi, Mahmoud Ahmed Refaee,"Risk Factors and Comorbidities Associated to Cardiovascular Disease in Qatar: A Machine Learning Based Case-Control Study", IEEE, 2021.

Sami Alrabie, Ahmed Barnawi, "HeartWave: A Multiclass Dataset of Heart Sounds for Cardiovascular Diseases Detection", IEEE, 2023.

Samiul Based Shuvo, Shams Nafisa Ali,"CardioXNet: A Novel Lightweight Deep Learning Framework for Cardiovascular Disease Classification Using Heart Sound Recordings", IEEE, 2021.

Davide Chicco, "A Machine Learning Analysis of Health Records of Patients with Chronic Kidney Disease at Risk of Cardiovascular Disease," IEEE, 2021.

D. Yaso Omkari, Kareemulla Shaik, "An Integrated Two-Layered Voting TLV Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers," IEEE, 2024.

Tahseen Ullah, Syed Irfan Ullah, Khalil Ullah, "Machine Learning Based Cardiovascular Disease Detection Using Optimal Feature Selection," IEEE, 2024.

N. A. Vinay, K.N.Vidyasagar, "An RNN-Bi LSTM Based Multi Decision GAN Approach for the Recognition of Cardiovascular Disease (CVD) From HeartBeat Sound: A Feature Optimization Process," IEEE, 2024.