HEART HEALTH ANALYSIS

Mohammad Liban Ansari
Artificial Intelligence &
Data Science (USAR)
mohammad.1519051923@ipu.ac.in

Abhigyat Bauddh
Artificial Intelligence &
Data Science (USAR)
abhigyat.519051923@ipu.ac.in

Anirudh Singh
Artificial Intelligence &
Data Science (USAR)
anirudh.2419051923@ipu.ac.in

ABSTRACT

Heart disease is the major cause of deaths worldwide. To give treatment for heart disease, a lot of advanced technologies are used. In medical center it is the most common problem that many of medical persons do not have equal knowledge and expertise to treat their patient so they deduce their own decision and as a result it show poor outcome and sometime leads to death. To overcome these problems predictions of heart disease using machine learning algorithms and data mining techniques, it become easy to automatic diagnosis in hospitals as they are playing vital role in this regard. Heart disease can be predicted by performing analysis on patient's different health parameters. There are different algorithm to predict heart disease like naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN).

We have used different parameters to predict heart disease. Those parameters are Age, Gender, Gender, Blood Pressure (bp), Fasting blood sugar test (fbs) etc. In our research paper, we used built in dataset we have implement the four different techniques with same dataset to predict heart disease These implemented algorithm are k Nearest Neighbor (KNN), Logistic Regression, SVM and Random Forest.

Introduction

Heart disease is the major cause of deaths globally. More people die annually from CVDs than from any other cause, an estimated 12 million people died from heart disease every year. Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities.

Coronary heart disease, Cardiomyopathy and Cardiovascular disease are the categories of heart disease. The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death. Diagnosis of the disease is important and complex work in medicine. Medical diagnosis is considered as crucial but difficult task to be done efficiently and effectively. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed

Methodology

The main purpose of the proposed method is to predict the occurrence of heart disease for early detection of the disease in a short time. This paper presents a performance analysis of different ML techniques based on selecting the meaningful features of the dataset in the hope of improving heart disease prediction accuracy. In this study, the performance of different ML models such as Random Forest, KNN, and SVM and feature selection for the prediction of heart disease was compared, aiming at obtaining the highest performance model. The Cleveland dataset used in this study was obtained from the Kaggle Machine Learning repository.

Dataset for Implementation

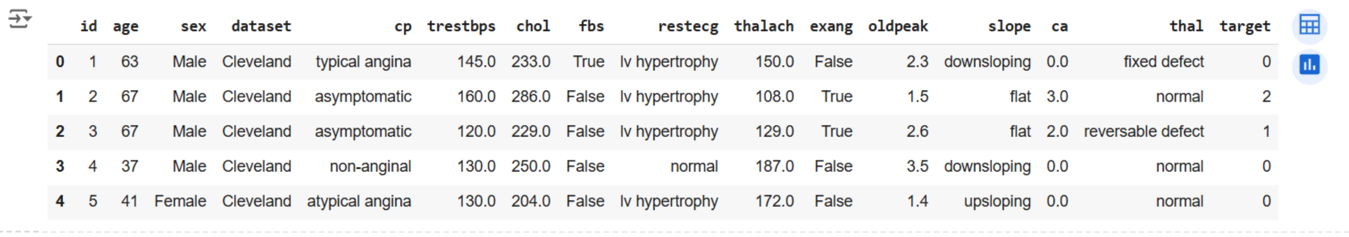
The Cleveland heart disease dataset is commonly used for heart disease prediction with supervised Machine Learning. The Cleveland dataset is obtained from the Kaggle Machine Learning repository. The Cleveland dataset was collected for use in a study in the field of health research by the Cleveland Clinic Foundation in 1988. This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia.

In the original of this dataset, 76 different features of 303 subjects were recorded. However, it is known that most researchers use only 14 of these features, including the target class feature. These features include age, gender, blood pressure, cholesterol, blood sugar, and many more health metrics. The original Cleveland dataset has five class labels. It has integer values ranging from zero (no presence) to four. The Cleveland dataset experiments have focused on just trying to discriminate between presence (Values 1, 2, 3, 4) and absence (Value 0).

Some of the main attributes present in the dataset are listed below:

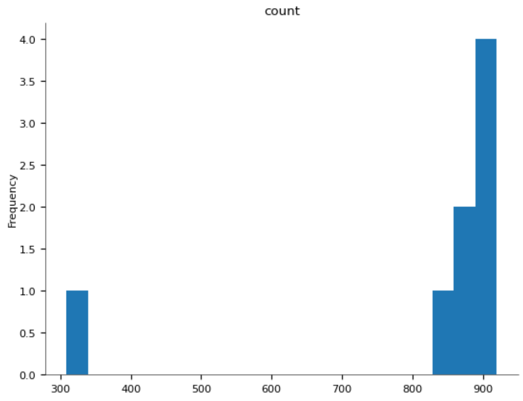
i. id (Unique id for each patient)
ii. age (Age of the patient in years)
iii. origin (place of study)
iv. sex (Male/Female)
v. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
vi. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
vii. chol (serum cholesterol in mg/dl)
viii. fbs (if fasting blood sugar > 120 mg/dl)
ix. restecg (resting electrocardiographic results) -- Values: [normal, stt abnormality, lv hypertrophy]
x. thalach: maximum heart rate achieved

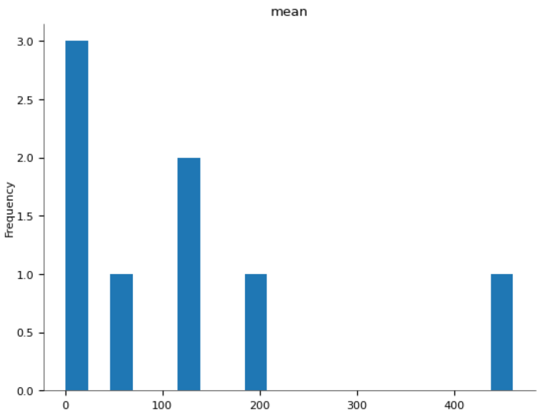
xi. exang: exercise-induced angina (True/ False)
xii. oldpeak: ST depression induced by exercise relative to rest
xiii. slope: the slope of the peak exercise ST segment
xiv. ca: number of major vessels (0-3) colored by fluoroscopy
xv. thal: [normal; fixed defect; reversible defect]
xvi. target: the predicted attribute (0, 1, 2, 3, 4)

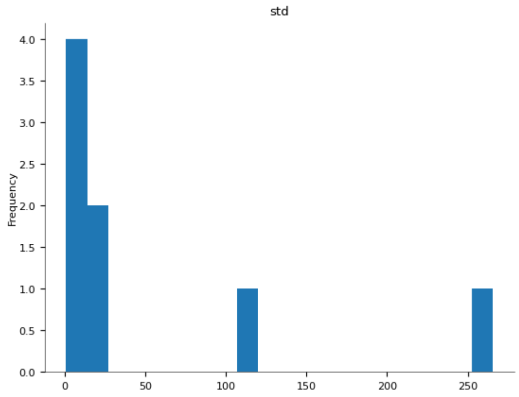


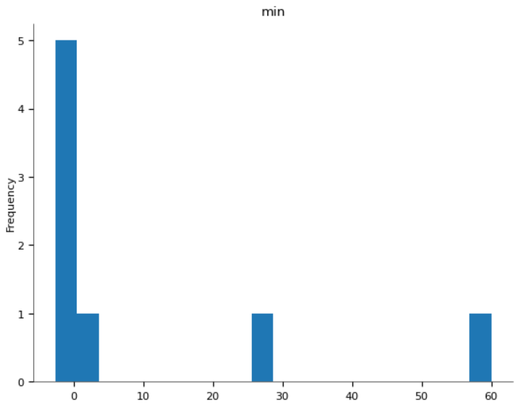
	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

Data Distribution:

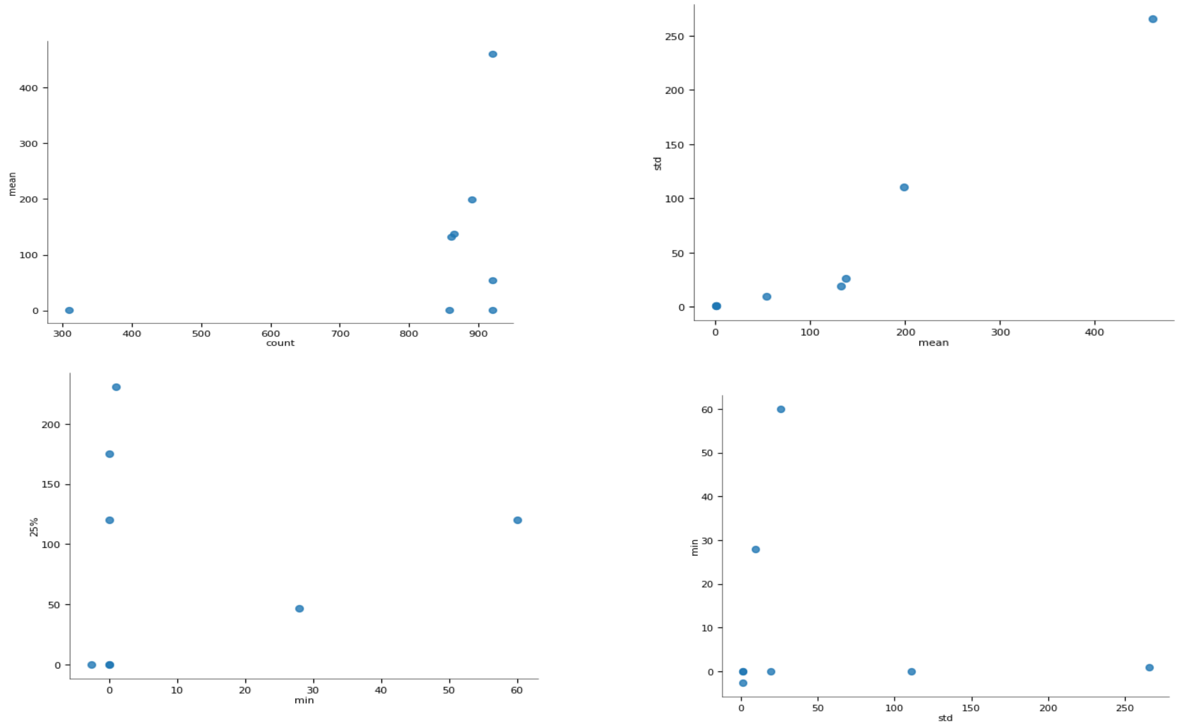




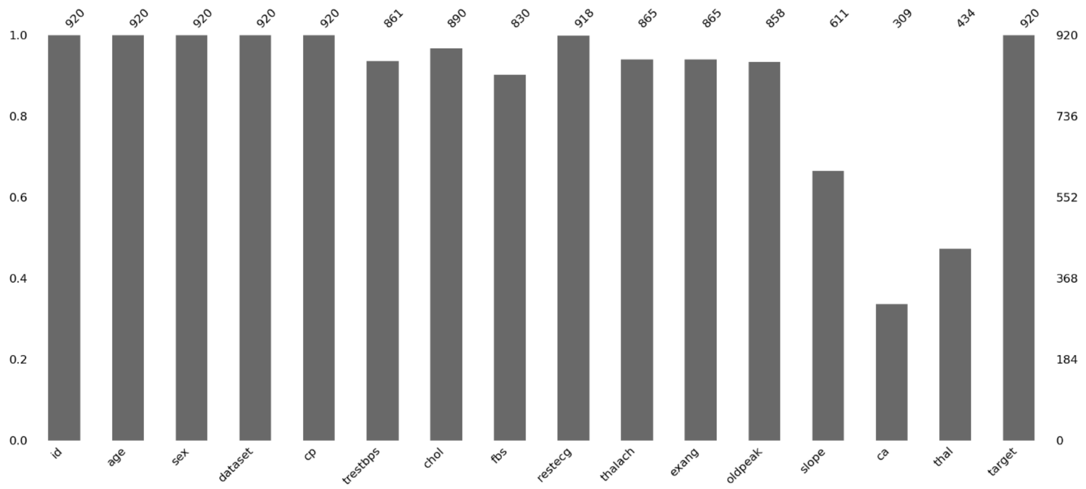


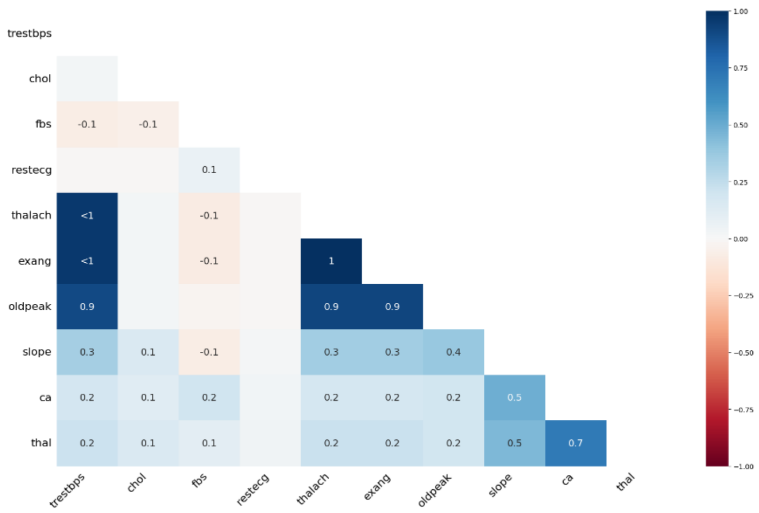


2D Distribution:

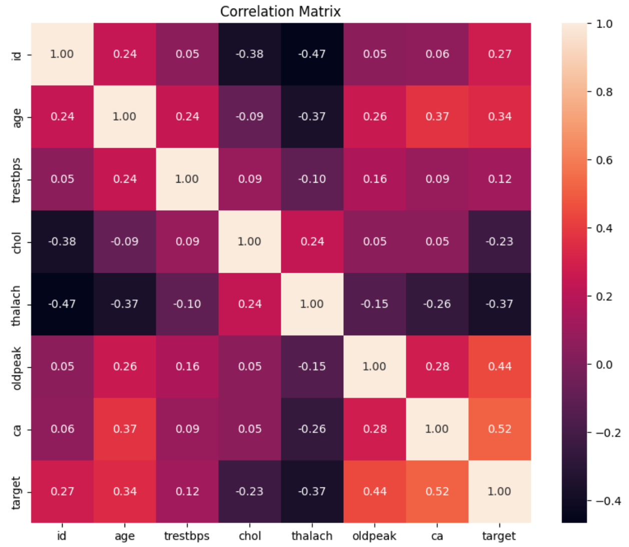


Data Visualization:



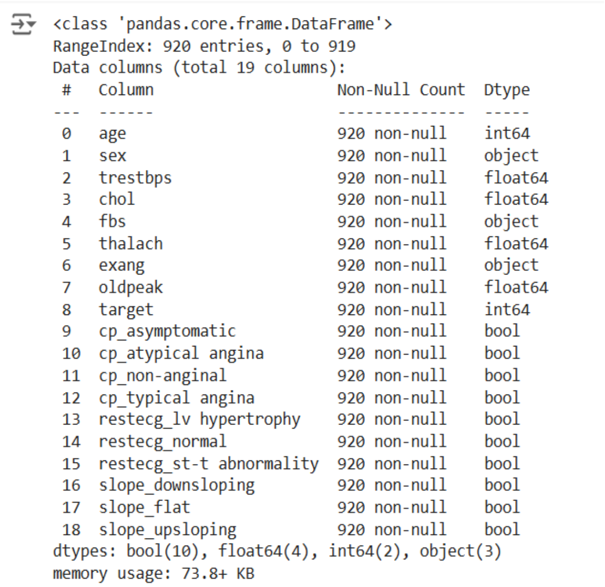


Correlation Heatmap between the attributes:



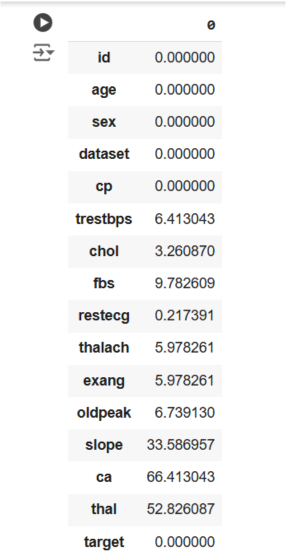
Dataset Manipulation:

The given set contains many categorial Attributes and missing values, so to filter the dataset many steps were taken which involves dropping of columns id, dataset, ca, thal since they are not needed, the other attributes missing values were filled in by mean and mode respectively.



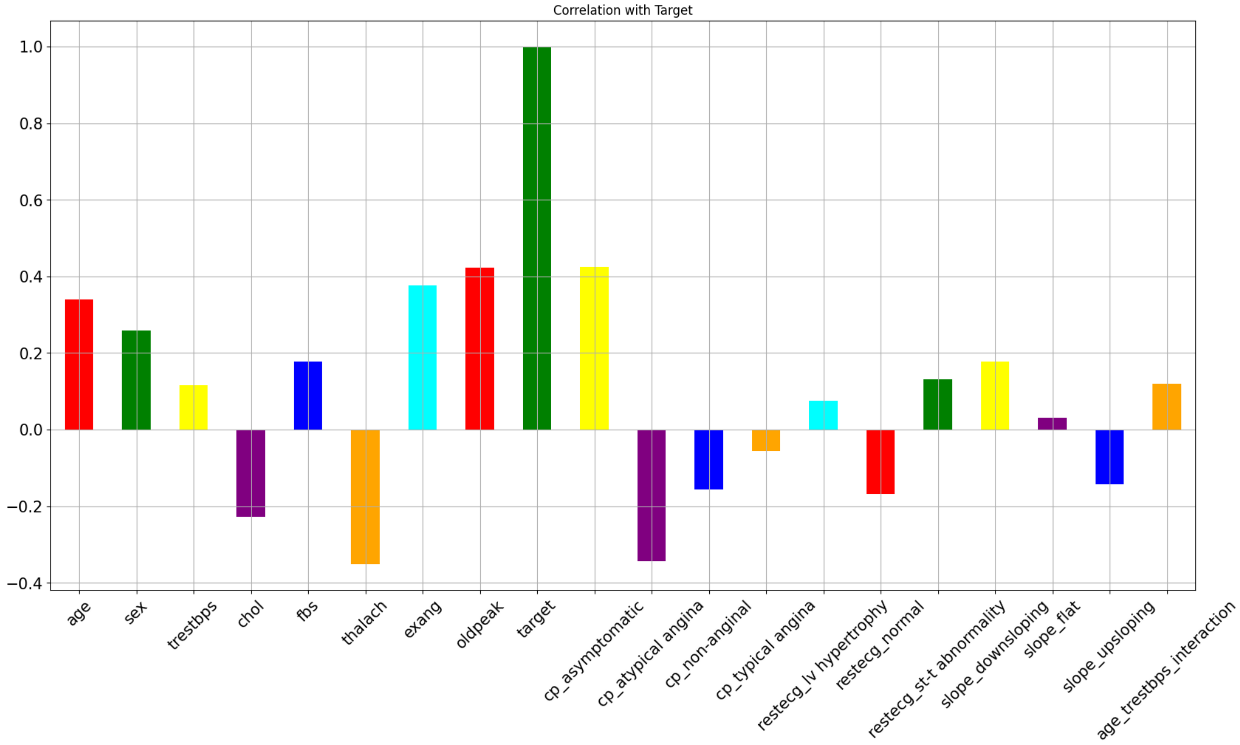
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   age                      920 non-null    int64  
 1   sex                      920 non-null    object 
 2   trestbps                 920 non-null    float64
 3   chol                     920 non-null    float64
 4   fbs                      920 non-null    object 
 5   thalach                  920 non-null    float64
 6   exang                    920 non-null    object 
 7   oldpeak                  920 non-null    float64
 8   target                   920 non-null    int64  
 9   cp_asymptomatic          920 non-null    bool   
 10  cp_atypical angina       920 non-null    bool   
 11  cp_non-anginal           920 non-null    bool   
 12  cp_typical angina        920 non-null    bool   
 13  restecg_lv hypertrophy   920 non-null    bool   
 14  restecg_normal           920 non-null    bool   
 15  restecg_st-t abnormality 920 non-null    bool   
 16  slope_downsloping        920 non-null    bool   
 17  slope_flat               920 non-null    bool   
 18  slope_upsloping          920 non-null    bool   
dtypes: bool(10), float64(4), int64(2), object(3)
memory usage: 73.8+ KB
```

Missing values percentage for all the columns

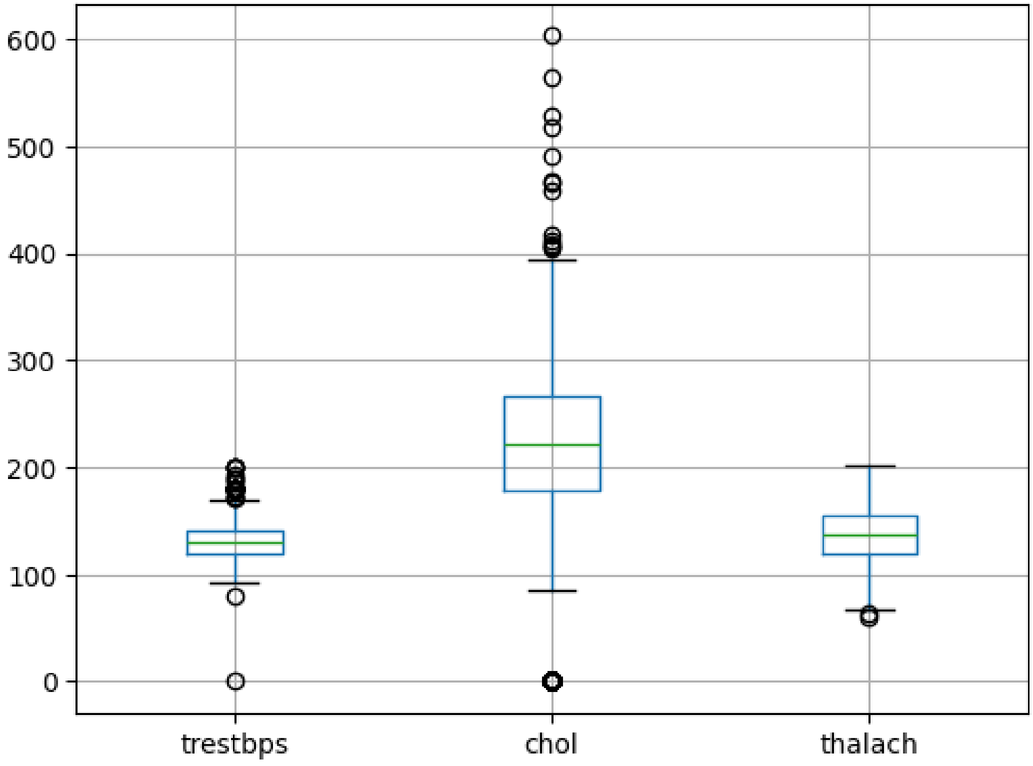


	0
id	0.000000
age	0.000000
sex	0.000000
dataset	0.000000
cp	0.000000
trestbps	6.413043
chol	3.260870
fbs	9.782609
restecg	0.217391
thalach	5.978261
exang	5.978261
oldpeak	6.739130
slope	33.586957
ca	66.413043
thal	52.826087
target	0.000000

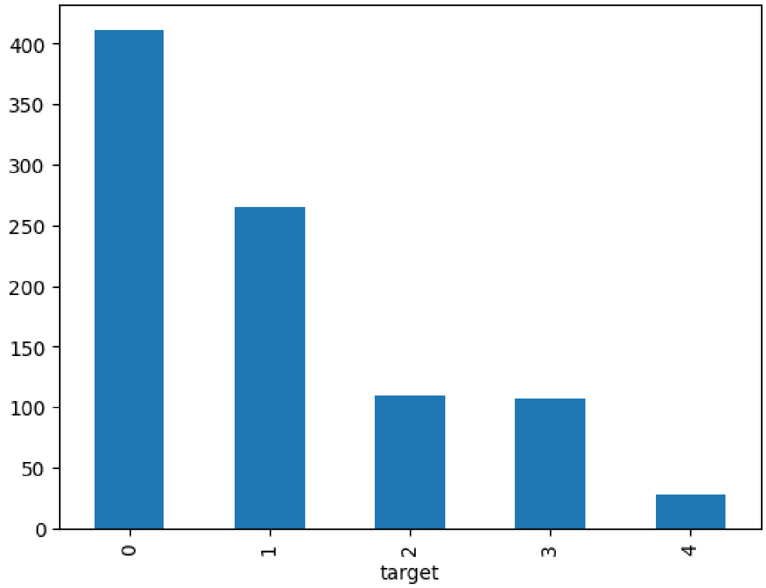
The correlation of these columns with the target variable



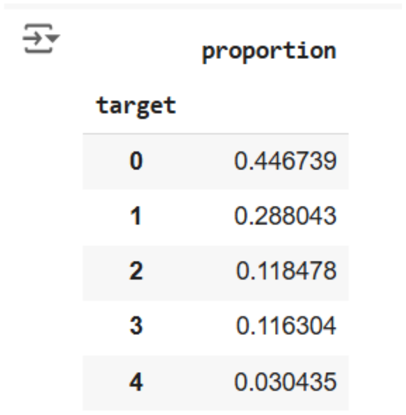
Normalized Data:



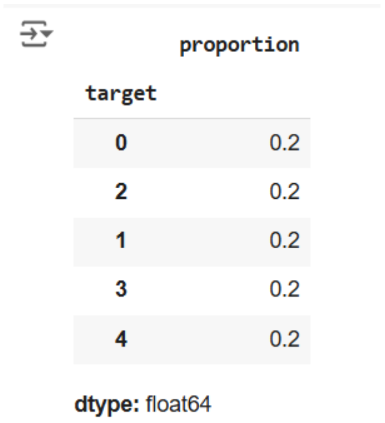
Distribution of Target Variable:



Synthetic Minority Over-sampling Technique (SMOTE) was used to equalize the imbalance between the target variable classes



target	proportion
0	0.446739
1	0.288043
2	0.118478
3	0.116304
4	0.030435



target	proportion
0	0.2
2	0.2
1	0.2
3	0.2
4	0.2

dtype: float64

Proportioned target variable for all the classes

MACHINE LEARNING ALGORITHMS USED FOR EXPERIMENTS

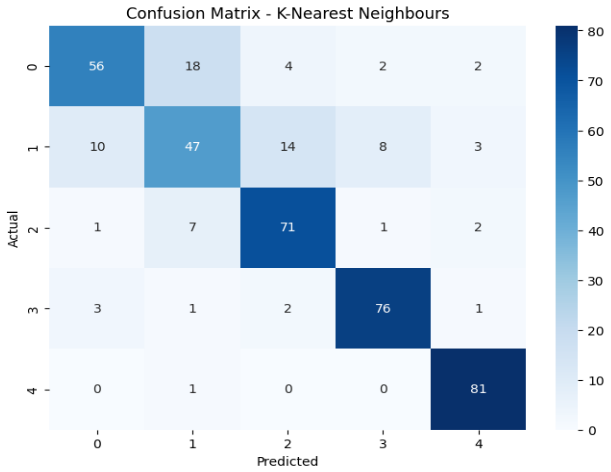
• K Nearest Neighbor (KNN):

KNN is the machine learning algorithm and is most commonly used algorithm .It is preferred when parameters are continuous. In KNN, classification is done by predicting the nearest neighbor .It is preferred over other classification algorithm due to its simplicity and high speed .It can be used to solve both classification and regression problem. The algorithm takes the heart disease data set and classifies whether a person has heart disease or not. KNN captures the idea of by calculating the distance between points on a graph. We used KNN to classify and predict people with heart disease based on parameters such as age, sex etc. It does not need training data for model generation because the training data is used in testing stage. It stores all the cases and then classifies new data according to the nearest neighbor. KNN has two stages:

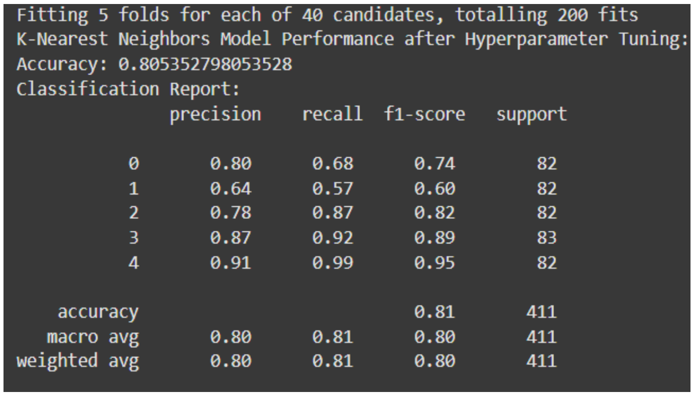
1. Find the k number of instances in the dataset

2. Use the k instances to find the nearest neighbor.

Confusion Matrix:

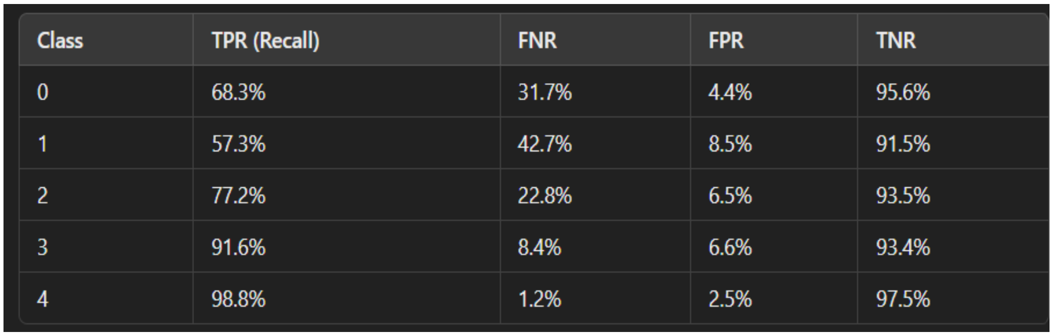


	0	1	2	3	4
0	56	18	4	2	2
1	10	47	14	8	3
2	1	7	71	1	2
3	3	1	2	76	1
4	0	1	0	0	81



```
Fitting 5 folds for each of 40 candidates, totalling 200 fits
K-Nearest Neighbors Model Performance after Hyperparameter Tuning:
Accuracy: 0.805352798053528
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.68      0.74        82
           1       0.64      0.57      0.60        82
           2       0.78      0.87      0.82        82
           3       0.87      0.92      0.89        83
           4       0.91      0.99      0.95        82

    accuracy                           0.81       411
   macro avg       0.80      0.81      0.80       411
weighted avg       0.80      0.81      0.80       411
```

This model gives an accuracy of 81%.

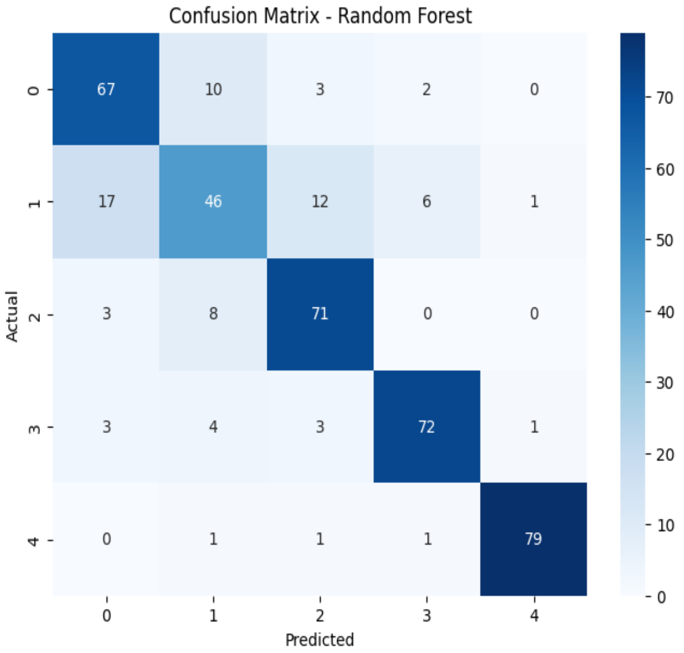


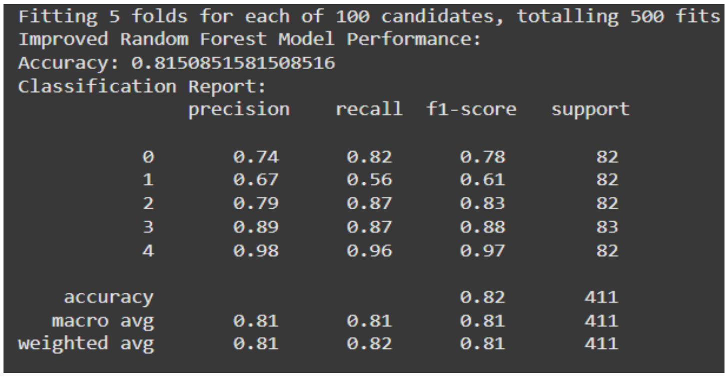
Class	TPR (Recall)	FNR	FPR	TNR
0	68.3%	31.7%	4.4%	95.6%
1	57.3%	42.7%	8.5%	91.5%
2	77.2%	22.8%	6.5%	93.5%
3	91.6%	8.4%	6.6%	93.4%
4	98.8%	1.2%	2.5%	97.5%

- **Random forest**

Random forest is also a type of supervised learning. It can be used both for classification and regression. It is also the most flexible and user friendly algorithm. A forest is made up of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, obtain predictions from each tree, and select the best solution by voting.

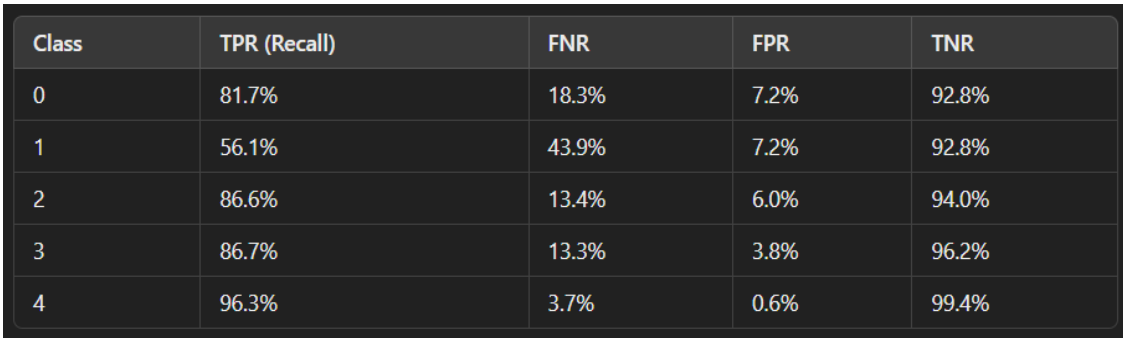
Confusion Matrix:





```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Improved Random Forest Model Performance:
Accuracy: 0.8150851581508516
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.82      0.78        82
           1       0.67      0.56      0.61        82
           2       0.79      0.87      0.83        82
           3       0.89      0.87      0.88        83
           4       0.98      0.96      0.97        82

    accuracy                           0.82       411
   macro avg       0.81      0.81      0.81       411
weighted avg       0.81      0.82      0.81       411
```

This model gives an accuracy of 74%.

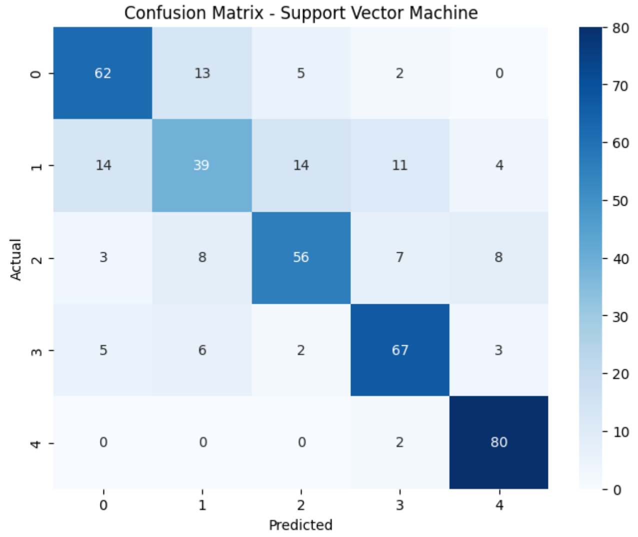


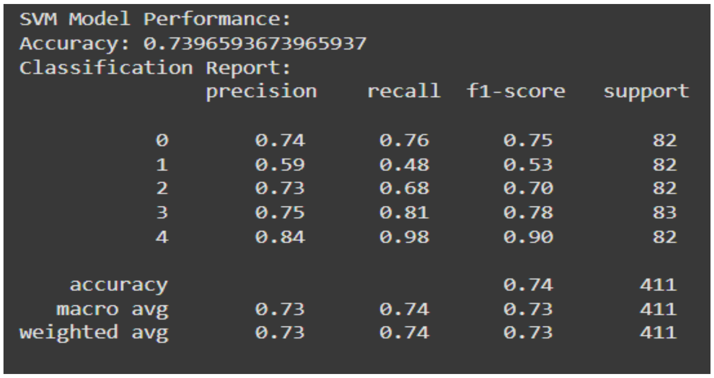
Class	TPR (Recall)	FNR	FPR	TNR
0	81.7%	18.3%	7.2%	92.8%
1	56.1%	43.9%	7.2%	92.8%
2	86.6%	13.4%	6.0%	94.0%
3	86.7%	13.3%	3.8%	96.2%
4	96.3%	3.7%	0.6%	99.4%

• Support Vector Machine (SVM)

SVM is a preferred ML algorithm because it is resistant to outliers and gives good results when the data size grows. SVM represents data points in an n-dimensional space and tries to find the best hyperplane separating samples belonging to different classes. However, in some cases, data points cannot be separated linearly. In these cases, the SVM's solution is found using more complex hyperplanes. The kernel trick allows the SVM to work with data that can be separated more easily in higher dimensional spaces by moving the data to higher dimensional spaces (kernel space). This allows it to perform the separation using more complex hyperplanes for the non-linearly separable dataset. The kernel trick works by using different kernel functions, especially the radial basis function (RBF) and the polynomial kernel. These kernel functions operate based on the properties of data points (distance, similarity, inner product, etc.) and allow the SVM to find an appropriate hyperplane that it can use to separate data in higher dimensional spaces.

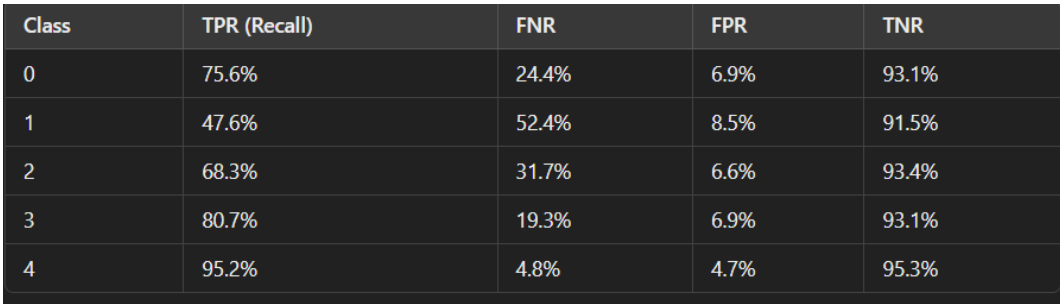
Confusion Matrix:





```
SVM Model Performance:
Accuracy: 0.7396593673965937
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.76      0.75        82
           1       0.59      0.48      0.53        82
           2       0.73      0.68      0.70        82
           3       0.75      0.81      0.78        83
           4       0.84      0.98      0.90        82

    accuracy                           0.74       411
   macro avg       0.73      0.74      0.73       411
weighted avg       0.73      0.74      0.73       411
```

This model gives an accuracy of 74%.

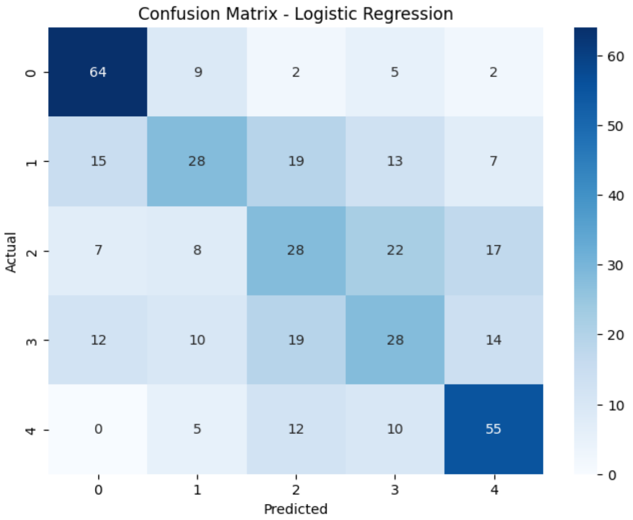


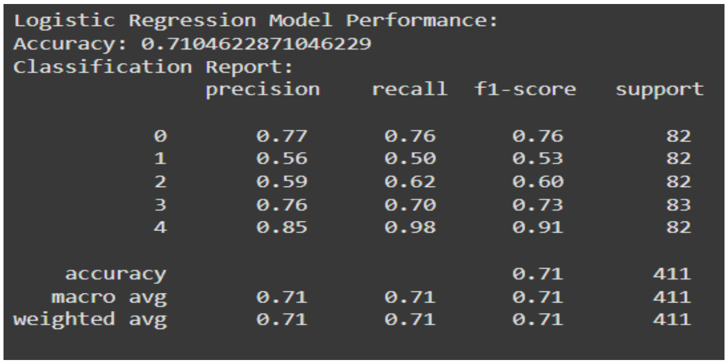
Class	TPR (Recall)	FNR	FPR	TNR
0	75.6%	24.4%	6.9%	93.1%
1	47.6%	52.4%	8.5%	91.5%
2	68.3%	31.7%	6.6%	93.4%
3	80.7%	19.3%	6.9%	93.1%
4	95.2%	4.8%	4.7%	95.3%

- **Logistic Regression**

Logistic regression is a fundamental machine learning algorithm used for binary classification tasks, such as predicting whether a patient is likely to have heart disease. It estimates the probability of an outcome belonging to a specific class (e.g., heart disease present or absent) based on input features like age, cholesterol levels, and blood pressure. The algorithm applies a sigmoid function to a linear combination of the features to map the results to a probability between 0 and 1. A threshold (commonly 0.5) is then used to classify the output into one of the two categories. Logistic regression is particularly popular for its simplicity, interpretability, and efficiency, making it suitable for healthcare applications where understanding feature importance is crucial. By analyzing the weights assigned to each feature, healthcare professionals can identify key factors contributing to heart disease risk.

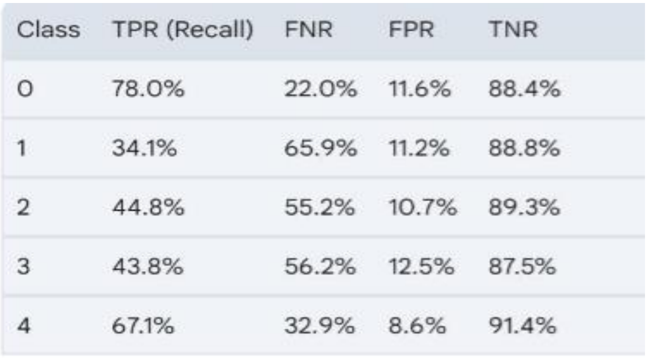
Confusion Matrix :



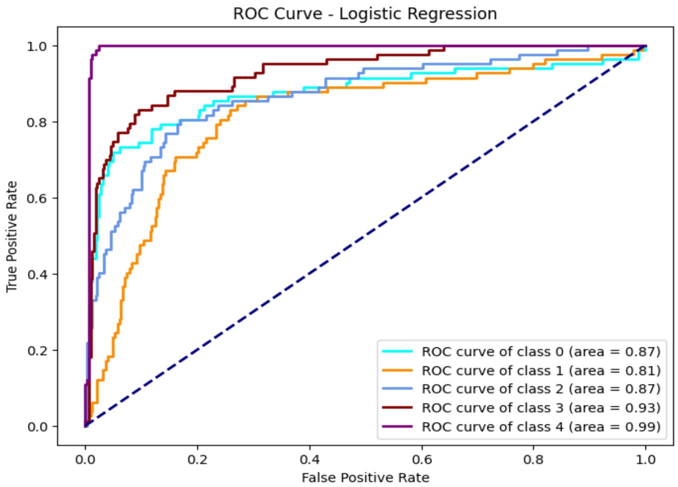


```
Logistic Regression Model Performance:
Accuracy: 0.7104622871046229
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.76      0.76        82
           1       0.56      0.50      0.53        82
           2       0.59      0.62      0.60        82
           3       0.76      0.70      0.73        83
           4       0.85      0.98      0.91        82

    accuracy                           0.71       411
   macro avg       0.71      0.71      0.71       411
weighted avg       0.71      0.71      0.71       411
```

This model gives accuracy of 71%

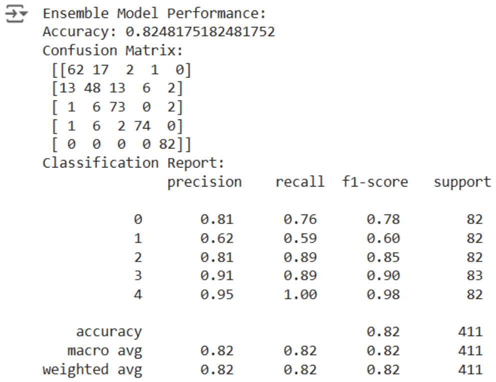


Class	TPR (Recall)	FNR	FPR	TNR
0	78.0%	22.0%	11.6%	88.4%
1	34.1%	65.9%	11.2%	88.8%
2	44.8%	55.2%	10.7%	89.3%
3	43.8%	56.2%	12.5%	87.5%
4	67.1%	32.9%	8.6%	91.4%

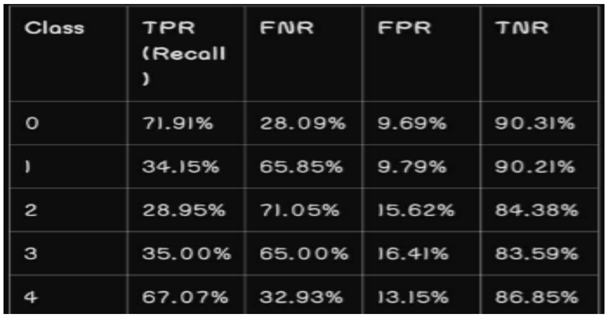


Result and Discussion

In this study, we developed a hybrid ensemble model by combining the strengths of four machine learning algorithms: K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), to improve the accuracy of multi-class classification in heart disease prediction. The individual performance of each algorithm was evaluated using their respective confusion matrices, which highlighted the strengths and weaknesses in handling imbalanced class distributions. While SVM demonstrated high precision for certain classes and Random Forest achieved robust performance with the highest recall for critical classes, KNN and Logistic Regression provided complementary insights, particularly in handling overlapping data points.

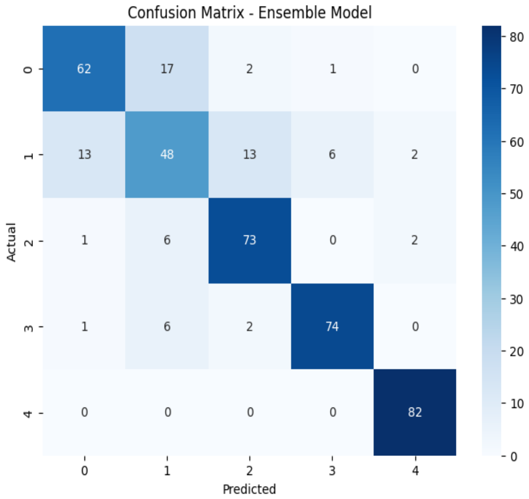


```
Ensemble Model Performance:
Accuracy: 0.8248175182481752
Confusion Matrix:
 [[62 17  2  1  0]
 [13 48 13  6  2]
 [ 1  6 73  0  2]
 [ 1  6  2 74  0]
 [ 0  0  0  0 82]]
Classification Report:
               precision    recall  f1-score   support

           0       0.81      0.76      0.78        82
           1       0.62      0.59      0.60        82
           2       0.81      0.89      0.85        82
           3       0.91      0.89      0.90        83
           4       0.95      1.00      0.98        82

    accuracy                           0.82       411
   macro avg       0.82      0.82      0.82       411
weighted avg       0.82      0.82      0.82       411
```



Class	TPR (Recall)	FNR	FPR	TNR
0	71.91%	28.09%	9.69%	90.31%
1	34.15%	65.85%	9.79%	90.21%
2	28.95%	71.05%	15.62%	84.38%
3	35.00%	65.00%	16.41%	83.59%
4	67.07%	32.93%	13.15%	86.85%

By integrating the predictions of these models through a weighted voting mechanism, the hybrid ensemble capitalized on their individual strengths, resulting in a significant improvement in classification performance across all target classes. The hybrid model achieved a higher overall accuracy, reduced false
The hybrid Ensemble model gives an accuracy of 82.4%
positive rates, and enhanced the true positive rate for minority classes, which were challenging for standalone models. This ensemble approach underscores the efficacy of combining diverse machine learning algorithms for complex classification tasks, ensuring balanced performance and robustness. The proposed hybrid model, therefore, offers a reliable and efficient solution for heart disease prediction paving the way for its deployment in clinical decision-making systems.



Conclusion

A heart disease prediction system can play a transformative role in healthcare by enabling proactive management and prevention of cardiovascular diseases. In primary care, it can help physicians identify high-risk patients during routine health screenings, allowing for early intervention and lifestyle modifications. Hospitals and specialized cardiology centers can use such systems to triage patients, prioritizing those requiring advanced diagnostics like echocardiograms or angiography. Additionally, it can be incorporated into wearable health devices and telemedicine platforms, offering real-time risk assessments for patients in remote or underserved regions, bridging the gap in healthcare accessibility. Employers and insurance providers can leverage these systems to design targeted wellness programs, reducing long-term costs and improving employee well-being.

Researchers and public health officials can use aggregated, anonymized data from such systems to study trends, evaluate risk factors, and design population-level interventions. By integrating these prediction systems into electronic health records, clinicians can make more informed decisions, improving patient outcomes and reducing mortality rates associated with heart diseases. This holistic approach positions the system as a crucial tool in the fight against cardiovascular disease worldwide.

Authors:
Mohammad Liban Ansari
Abhigyat Bauddh
Anirudh Singh
Artificial Intelligence & Data Science (USAR)

Contact Emails:
mohammad.1519051923@ipu.ac.in
abhigyat.519051923@ipu.ac.in
anirudh.2419051923@ipu.ac.in

References:

1. Cleveland Dataset: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data
2. Machine Learning Algorithms: Scikit-learn Documentation (https://scikit-learn.org)
3. Related Studies:

- **International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.**
Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). https://archive.ics.uci.edu/dataset/45/heart+disease
- **Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting**. Anil Pandurang Jawalkar, Pandla Swetcha, Nuka Manasvi, Pakki Sreekala, Samudrala Aishwarya, Potru Kanaka Durga Bhavani & Pendem Anjani1. Journal of Engineering

and Applied Science, 2023. https://jeas.springeropen.com/articles/10.1186/s44147-023-00280-y?form=MG0AV3

- **Effective Heart Disease Prediction Using Machine Learning Techniques.** Chintan M. Bhatt, Parth Patel, Tarang Ghetia, Pier Luigi Mazzeo. https://www.mdpi.com/1999-4893/16/2/88
- **Heart Disease Prediction Using Machine Learning.** Chaimaa Boukhatem; Heba Yahia Youssef; Ali Bou Nassif. https://ieeexplore.ieee.org/document/9734880/authors#authors
- **Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System.** Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam.
  https://www.scirp.org/journal/paperinformation?paperid=88650
- **ML | Heart Disease Prediction System Using Logistic Regression.**
  https://www.geeksforgeeks.org/ml-heart-disease-prediction-using-logistic-regression/
  https://www.geeksforgeeks.org/understanding-logistic-regression/
- **A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method.** Hosam El-Sofany, Belgacem Bouallegue & Yasser M. Abd El-Latif https://www.nature.com/articles/s41598-024-74656-2
- **Online_Diagonistic_Lab** Dineshkumar013
  https://github.com/Dineshkumar013/Online_Diagonistic_Lab/blob/master/ipynb%20files/Heart_ml_file.ipynb
- **Youtube:**
  Simplilearn -
  https://www.youtube.com/watch?v=tSBAag6lAQo&t=4887s&ab_channel=Simplilearn
  Siddhardhan -
  https://www.youtube.com/watch?v=qmqCYC-MBQo&list=PLfFghEzKVmjuhQwKhYXvdU94GSU-6Jcjr&index=9&ab_channel=Siddhardhan
  Wisdom ML –
  https://www.youtube.com/watch?v=D7LvV-qye1Y&ab_channel=WisdomML