CARDIOVASCULAR DISEASES RISK PREDICTION

by

ABHIJITH DAMERUPPALA

Abstract— **The project, "Cardiovascular Diseases Risk Prediction", is an effort in the realm of healthcare analytics. It seeks to leverage a comprehensive dataset from Kaggle, which has a broad spectrum of health metrics and lifestyle habits, to predict the risk of heart disease in individuals. This project aligns with the global challenge posed by cardiovascular diseases (CVD) - a leading cause of mortality worldwide. Through predictive modeling, the project aims to enable healthcare providers to identify at-risk individuals early, paving the way for preemptive strategies and personalized treatment plans. This not only holds the potential to transform individual patient care but also to broadly impact public health initiatives.**

I. Introduction

According to the World Heart Federation, cardiovascular diseases are a principal cause of death globally, spanning developed and developing nations. The ability to detect early signs and risks associated with CVD can dramatically improve patient outcomes and enhance life quality. Utilizing a Kaggle dataset with health metrics and habits, this project aims to predict individual heart disease risk. The objective is to create a predictive model using patient data, including general health, lifestyle habits, and other health conditions, to foresee heart disease risk. This model will aid healthcare providers in early identification of at-risk individuals, facilitating preventive measures and treatment.

Aim

I aim to create a predictive model that uses patient data, such as their general health, lifestyle habits, and other health conditions, that are available in the dataset, to predict the risk of developing heart disease. By working on this project, I aim to help health care providers in early identification of at-risk individuals, and this can lead to preventative measures and treatment. Healthcare providers can offer personalized guidance based on individual risk profiles.

To ensure the robustness and generalizability of the predictive model, a cross-dataset validation approach will be implemented.

II. Discussion of related work

1. "Risk prediction of cardiovascular disease using machine learning classifiers" is a study focusing on developing a predictive model for cardiovascular diseases (CVD) using machine learning techniques. The researchers used publicly available data from the University of California Irvine repository. Their study employs two machine learning classifiers: Multi-Layer Perceptron (MLP) and K-Nearest Neighbor (K-NN), and these were chosen for their reliability in disease detection. And the models' performances were evaluated and compared using a confusion matrix and accuracy, sensitivity, F1-score, and AUC-ROC.

   The MLP model demonstrated higher accuracy and AUC values compared to the K-NN model, indicating its superior performance in CVD prediction. The research concludes that the MLP model is more effective for automatic CVD detection compared to the K-NN model.
2. "Prediction of cardiovascular disease risk based on major contributing features" is another research study that focuses on developing a machine learning model for predicting risk of CVD. The model, named XGBH, is evaluated for its effectiveness in risk prediction.

   The authors compared the XGBH model with four other machine learning models: logistic regression, support vector machine, random forest, and XGBoost. They evaluated models using ROC curves, calibration curves, and clinical impact curves and used Net Benefit and High-Risk Threshold metrics for clinical evaluation.

Overall, the XGBH model seemed to show high accuracy (AUC = 0.81) and outperformed the baseline models. The researchers also created a simpler model requiring only age, systolic blood pressure, and cholesterol status, with a slight reduction in accuracy (AUC = 0.79) as a simplified diagnostic model. The XGBH model effectively predicts CVD risk, offering improvements over existing models in performance and computing time.

3. The research paper "Machine Learning Prediction in Cardiovascular Diseases," which was published in Scientific Reports, is a thorough meta-analysis that assesses how well different machine learning algorithms perform in terms of cardiovascular illness prediction. A thorough evaluation and analysis of several research articles published between 1966 and March 2019 were part of the study. Following a thorough screening procedure, 55 studies with 103 cohorts and more than 3 million participants were included in the analysis. The study focuses on several cardiovascular disease categories, such as heart failure, coronary artery disease (CAD), cardiac arrhythmias, and stroke. It evaluates the predictive power of several machine learning models, such as boosting methods, support vector machines (SVMs), convolutional neural networks (CNNs), and custom-built algorithms.
4. The paper "Cardiovascular Disease Risk Prediction using Automated Machine Learning" presents a novel approach to predicting cardiovascular disease risk using a large-scale dataset from the UK Biobank. With 423,604 participants, this prospective study highlights the application of automated machine learning methods to improve prediction accuracy. The study presents AutoPrognosis, a group of three machine learning pipelines that use several methods for feature processing, data imputation, calibration, and classification. Principal component analysis (PCA) for data reduction, the MissForest method for missing data imputation, and a variety of classification techniques, including random forest, are important components of this study. The study also looks at how well certain characteristics predict cardiovascular events for both men and women using the UK Biobank dataset, indicating the relative significance of these variables.

III. PROJECT DESCRIPTION

The dataset in use comprises 308,854 records with 19 attributes related to health, exercise habits, and existing health conditions. The primary attribute of focus is "Heart_Disease," indicating the presence of heart disease. The project's timeline includes an eight-week plan, starting from the proposal to the final presentation. Key milestones involve data familiarization, preprocessing, exploratory analysis, feature and model selection, model evaluation and refinement, and final training and validation. The project also references significant research studies in the field, including those employing machine learning classifiers like Multi-Layer Perceptron (MLP) and K-Nearest Neighbor (K-NN) for CVD prediction. Proposed models for this project include Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, and Gaussian Naïve Bayes, with a comprehensive performance evaluation plan using metrics like accuracy, precision, recall, confusion matrix, sensitivity, F1-score, and AUC-ROC. The project is anchored in relevant literature, ensuring a methodical and scientifically grounded approach to tackling one of the most pressing health challenges of our time.

IV. DATA DESCRIPTION

A. Data Source

Data Set we have obtained from Kaggle data repository but the actual source of data in Kaggle was gathered from the official e-government website of the Republic of Korea linked at https://www.data.go.kr/data/15007122/fileData.do. This open-source data was collected from around 1 million people each year. The data collection method has not been disclosed. And there are no restrictions on the scope of use of the data.

B. Data Description

The dataset contains 308,854 records with 19 attributes related to general health, exercise habits, existing health conditions, and more. The primary attribute of interest in my project is "Heart_Disease", which indicates if a person has heart disease.

V. METHODOLOGY

1. Data Preprocessing:

Preprocessing the data typically includes handling missing values if any and ensuring that the dataset is balanced to prevent model bias towards a particular class.

2. Exploratory Data Analysis (EDA):

- Analyzing features and understanding the dataset better.
- Getting good visualizations to get conclusions out of.

3. Choosing the Model and Training:

Several machine learning models were trained and evaluated, including Naïve Bayes, Support Vector Classifier, Random Forest, Logistic Regression, Decision Tree, and XGBoost Classifiers. Each model's performance was assessed using metrics like accuracy, precision, recall, and F1 score.

4. Evaluating and comparing models and selecting the best model:

The models were evaluated based on their ability to accurately classify individuals based on their ‘Heart_Disease’ parameter, considering various performance metrics.

VI. IMPLEMENTATION

A. BASIC EXPLORATORY DATA ANALYSIS

After data cleaning and preparation, I continued with some Exploratory Data Analysis to know about the feature distributions in dataset by using graphs.

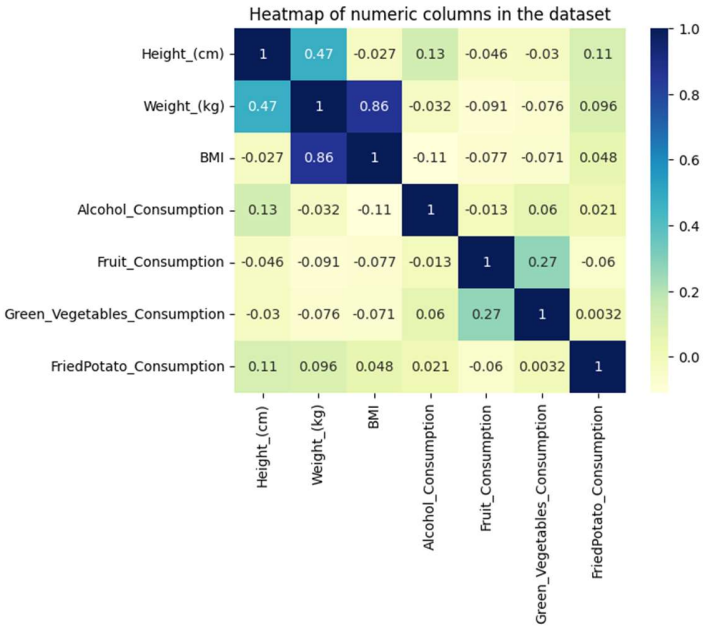

Fig-1: Heat map of features along with their correlation scores

Figure-1 shows the heatmap of features along with their correlation scores. Consumption of fruits and green vegetables has a positive correlation, and there's a very strong positive correlation between weight and BMI, which is expected as BMI is calculated based on weight and height. I then moved on to dive deep into features such as Age, BMI and more.

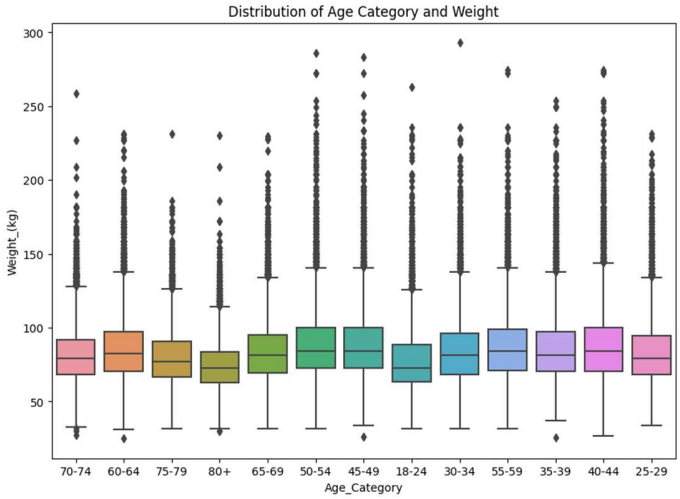

Fig-2: Box plot of Distribution of Age Category and Weight

Even though we do not observe any specific trends, from Figure 2, the box plot does show that the median weights are quite consistent across the age categories.

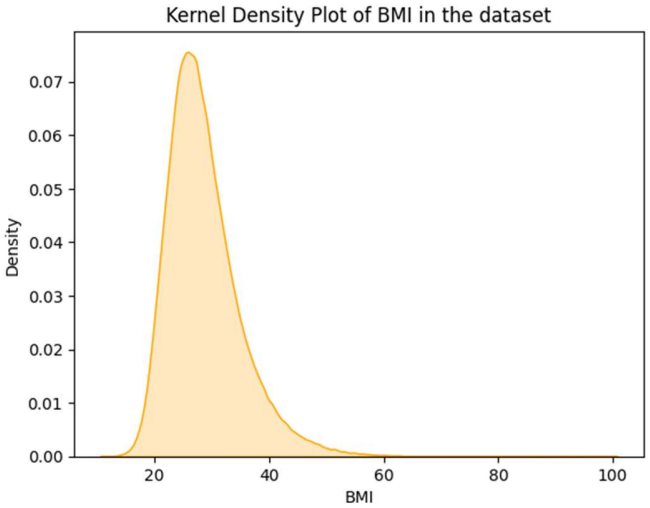

Fig-3: KDE Plot of BMI in the dataset

Figure-3 represents a density plot of BMI from the dataset. We observe that there is a peak at the lower end of the BMI scale, meaning that there is a large proportion of data that is between 20 ~ 45. And the peak of the curve signifies the mode of the distribution.

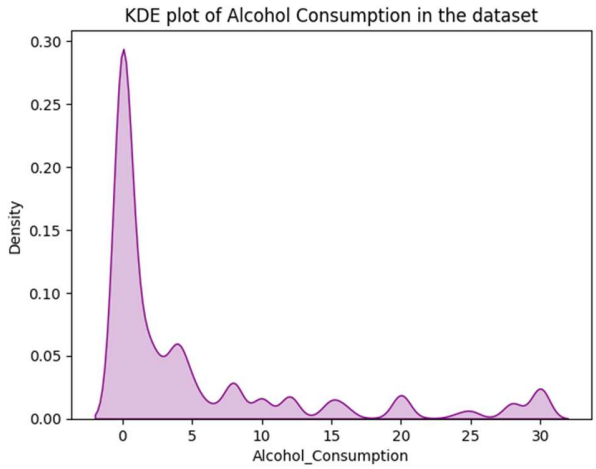

Fig-4: KDE plot of Alcohol consumption

Figure-4 is a KDE plot for Alcohol Consumption, has a sharp peak close to 0, meaning that there is a large number of individuals with low alcohol consumption values. And this plot is right skewed, meaning that most of the observations are concentrated towards 0, there are samples with higher consumption values as well.

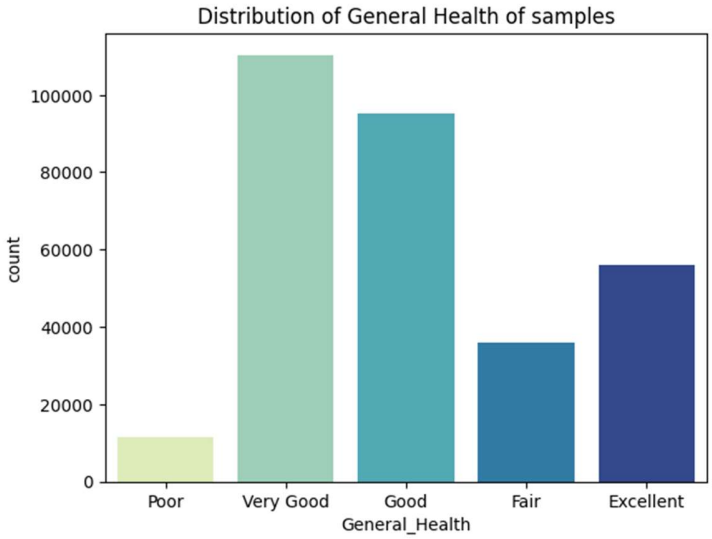

Fig-5: Distribution of General Health

Also, by looking at the categorized distribution of General Health from the dataset of 308854 samples, we observe that a large number of individuals rated their health as 'Very Good', followed by 'Good'. While the 'Poor' category has the smallest count. This suggests that most of the population from where the survey was conducted considers their health to be positively maintained.

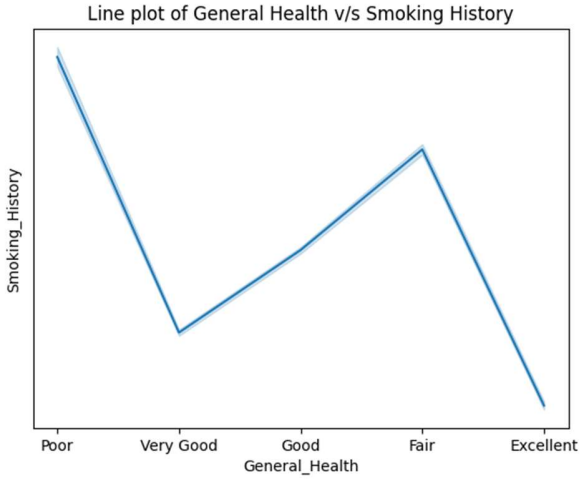

Fig-6: Line Plot of general v/s smoking history

Now that we know most people consider their health to be positive, I compared people with smoking history and their thoughts on their general health condition. And as we might expect, people with a higher history of smoking do consider their health to be in Poor condition, while the people with lowest history of smoking consider their health to be in Excellent condition.

B. ENCODING

I started to preprocess the dataset and observed that there are just 7 numeric features of the 19 features. So, I encoded all the categorical features using one-hot encoding and increased the numeric columns to 42. This could potentially help capture more information for the model.

C. INITIAL MODEL TRAINING

While training a Logistic Regression Model, I encountered an error in convergence. So, I implemented a pipeline that has a standard scalar to normalize the data before feeding the data back to the Logistic Regression model.

The following are the results of the model:

	precision	recall	f1-score	support
0.0	0.92	0.99	0.96	56774
1.0	0.51	0.06	0.12	4997
accuracy			0.92	61771
macro avg	0.72	0.53	0.54	61771
weighted avg	0.89	0.92	0.89	61771

Fig-7: Classification Report of Logistic Regression

Looking at the reported scores, we observe that there must be an error. I confirmed it by plotting the AUC-ROC curve:

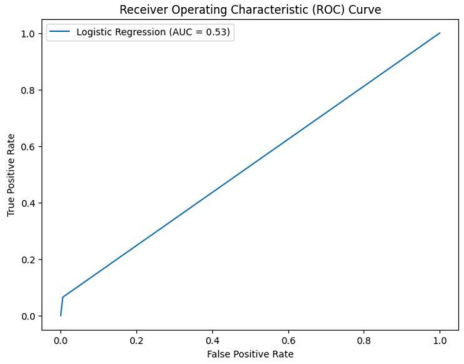

Fig-8: Line Plot of general v/s smoking history

We observe that the ROC curve shows an AUC of 0.53, which is just slightly better than random guessing which would result in an AUC of 0.5. This shows that the model is not reliable for prediction, and it cannot differentiate between positive and negative classes.

And the confusion matrix also confirms that this is the case since there are a significant number of false negatives which are 4673 out of 4997.

Analyzing the issue:

After investigating for the low AUC-ROC score, and classification metrics along with the confusion matrix, I learnt that the dataset has class imbalance, and this adversely affected in the model's performance.

Application of SMOTE:

To address this problem of imbalance of classes in the dataset, I wanted to perform the SMOTE technique to the dataset, which leads to a balanced dataset.

Revised Model training:

With the newly obtained balanced dataset, I trained a Logistic Regression model again and this time, I observed the following results:

	precision	recall	f1-score	support
0.0	0.80	0.74	0.77	56701
1.0	0.76	0.81	0.79	56853
accuracy			0.78	113554
macro avg	0.78	0.78	0.78	113554
weighted avg	0.78	0.78	0.78	113554

Fig-8: Classification Report of Revised Logistic Regression model.

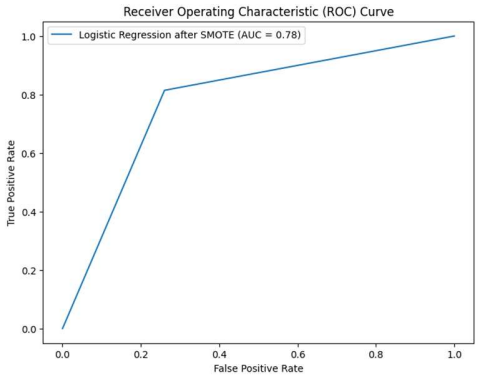

Fig-9: AUC-ROC plot for Logistic Regression after SMOTE analysis.

This is a significant improvement to 0.78 AUC-ROC score. Which indicates that the model could now much better discriminate between the positive and the negative classes.

D. MODEL TRAINING:

1. Gaussian Naïve Bayes

Here are the results:

	precision	recall	f1-score	support
0.0	0.98	0.60	0.75	56774
1.0	0.16	0.85	0.27	4997
accuracy			0.62	61771
macro avg	0.57	0.73	0.51	61771
weighted avg	0.91	0.62	0.71	61771

Fig-10: Classification Report for Gaussian Naïve Bayes

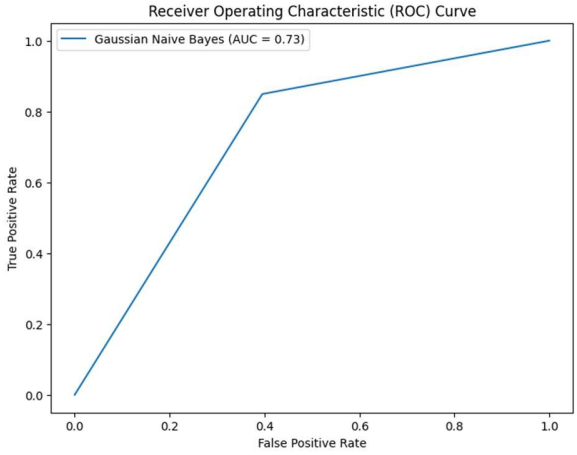

Fig-11: AUC-ROC plot for Gaussian Naïve Bayes before SMOTE analysis.

And new that we already know that the dataset is imbalanced, after applying SMOTE techniques, here are the classification report, confusion matrix and the AUC-ROC plot:

	precision	recall	f1-score	support
0.0	0.84	0.58	0.69	56701
1.0	0.68	0.89	0.77	56853
accuracy			0.74	113554
macro avg	0.76	0.74	0.73	113554
weighted avg	0.76	0.74	0.73	113554

```
[[33136 23565]
 [ 6268 50585]]
```

Fig-12: Classification Report of Gaussian NB after SMOTE analysis.

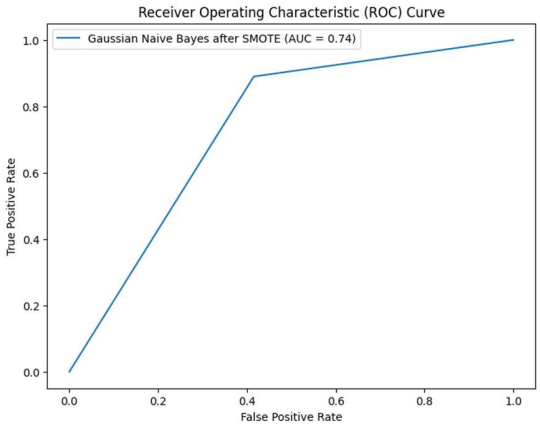

Fig-13: AUC-ROC plot for Gaussian NB after SMOTE analysis.

We observe that there is not much of an improvement for this model as we had for the Logistic Regression model. After researching the reasons for this situation, here are some of the top reasons that I learnt:

1. Gaussian Naïve Bayes assumes that all the features are independent. While SMOTE generates synthetic data points based on feature space similarity. And both these points do not align with each other.
2. Gaussian NB relies on the probability distribution of the features to make predictions. And adding synthetic points using SMOTE do not alter the overall distribution.
3. Naïve Bayes is less sensitive to class imbalance since it considers the probability of each class and the conditional probability of each feature separately.

Now that I understood why SMOTE technique does not improve Gaussian Naïve Bayes, I also researched as to why it affects Logistic Regression. And here is what I learnt:

1. As we might have guessed, Logistic Regression is more sensitive to class imbalance. Because it relies on the optimization of a loss function that is directly related to the number of instances of each class.
2. Logistic Regression creates a decision boundary between classes. And if the dataset is imbalanced, this boundary may be biased towards the majority class.

2. Bernoulli Naïve Bayes

I also proceeded to train another Naïve Bayes model, which is the Bernoulli's Naïve Bayes model to confirm the conclusions. And here are the classification report, confusion matrix and the AUC-ROC curve before applying SMOTE:

	precision	recall	f1-score	support
0.0	0.95	0.91	0.93	56774
1.0	0.28	0.42	0.34	4997
accuracy			0.87	61771
macro avg	0.61	0.66	0.63	61771
weighted avg	0.89	0.87	0.88	61771

```
[[51460  5314]
 [ 2914  2083]]
```

Fig-14: Classification Report of Bernoulli Naïve Bayes

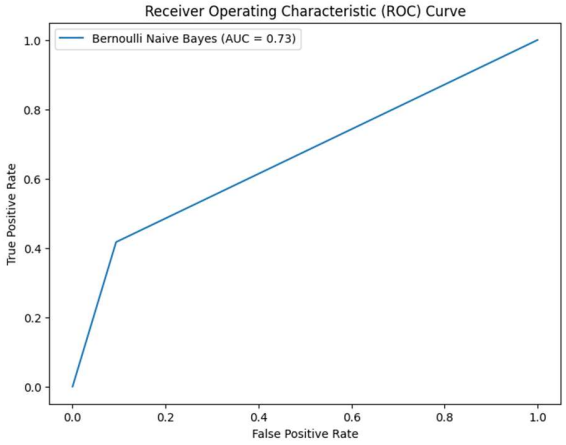

Fig-15: AUC-ROC plot of Bernoulli Naïve Bayes before SMOTE

And here are the results after applying the SMOTE technique:

	precision	recall	f1-score	support
0.0	0.88	0.82	0.85	56701
1.0	0.83	0.89	0.86	56853
accuracy			0.85	113554
macro avg	0.86	0.85	0.85	113554
weighted avg	0.86	0.85	0.85	113554

```
[[46432 10269]
 [ 6383 50470]]
```

Fig-16: AUC-ROC plot of Bernoulli Naïve Bayes after SMOTE.

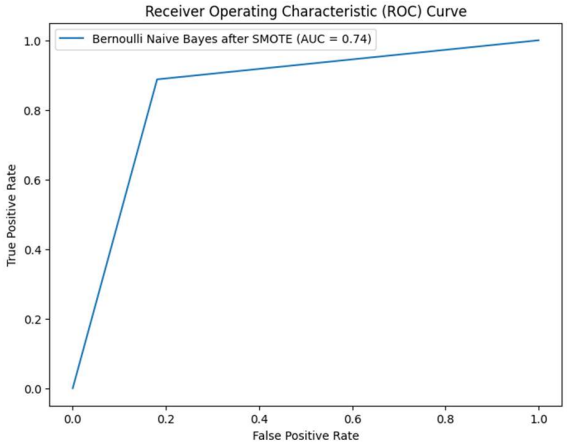

Fig-17: AUC-ROC plot of Bernoulli Naïve Bayes after SMOTE.

So, we can conclude that the Naïve Bayes models' performance remains consistent before and after SMOTE, as shown by the above ROC curves. This also suggests that the class imbalance did not affect the model's ability to discriminate between classes, due to the above-mentioned reasons.

3. Decision Tree Classifier

I was looking for alternate models that could benefit the class balancing without costing Accuracy, by also improving other classification metric parameters. And I decided to use Non-Linear model for this. And Decision Tree Classifier seems to be the perfect model for this dataset.
So, I first proceeded to train the model using the initial dataset which was not sampled. And here are the observed results:

	precision	recall	f1-score	support
0.0	0.93	0.92	0.92	56774
1.0	0.20	0.23	0.22	4997
accuracy			0.86	61771
macro avg	0.57	0.58	0.57	61771
weighted avg	0.87	0.86	0.87	61771

```
[[52082  4692]
 [ 3829  1168]]
```

Fig-18: Classification Report of Decision Tree Classifier before applying SMOTE.

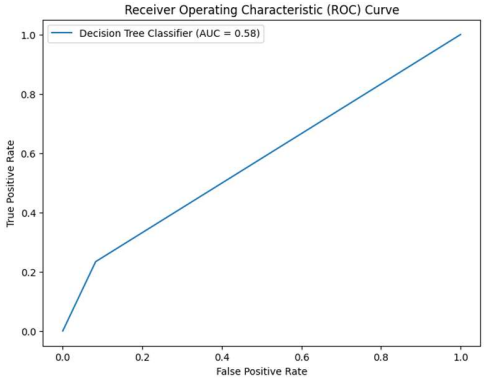

Fig-19: AUC-ROC of Decision Tree Classifier before applying SMOTE.

And after applying SMOTE technique, I observed the following classification report:

	precision	recall	f1-score	support
0.0	0.93	0.92	0.92	56701
1.0	0.92	0.93	0.92	56853
accuracy			0.92	113554
macro avg	0.92	0.92	0.92	113554
weighted avg	0.92	0.92	0.92	113554

```
[[51906  4795]
 [ 3941 52912]]
```

Fig-20: Classification Report of Decision Tree Classifier after SMOTE technique

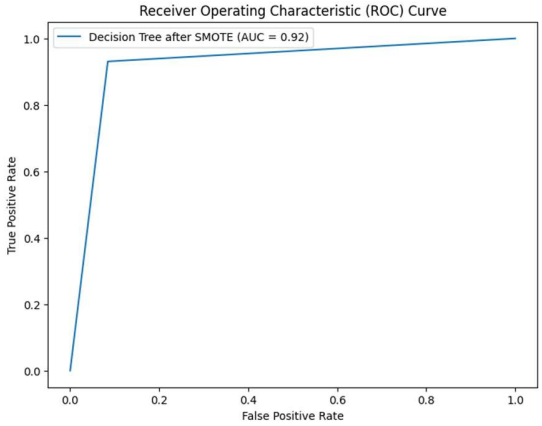

Fig-21: AUC-ROC plot for Decision Tree Classifier after SMOTE technique

VII. CONCLUSION

As we expected, there is significant improvement in Accuracy, precision scores for the minority class, and the overall model performs best for this dataset when it is balanced. So, we can conclude that Decision trees (non-linear models) can handle more complex patterns. And by using SMOTE, we provide the model with more set of examples to learn from, which improve its performance.

We can conclude that this model is well-calibrated for both the classes, from the high precision, recall and f1-scores. The almost equal number of instances in both the classes show that the model is not biased towards a particular class. And an accuracy score of 92% suggests that the model performs strongly for the dataset.

REFERENCES

[1]. Krittanawong, C., Virk, H.U.H., Bangalore, S. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 10, 16057 (2020). https://doi.org/10.1038/s41598-020-72685-1.

[2]. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019 May 15;14(5):e0213653. doi: 10.1371/journal.pone.0213653. PMID: 31091238; PMCID: PMC6519796

[3]. Peng, M., Hou, F., Cheng, Z. et al. Prediction of cardiovascular disease risk based on major contributing features. Sci Rep 13, 4778 (2023). https://doi.org/10.1038/s41598-023-31870-8

[4]. Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022 Jun 17;17(1):1100-1113. doi: 10.1515/med-2022-0508. PMID: 35799599; PMCID: PMC9206502.

[5]. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944.

[6]. Baharvand-Ahmadi B, Bahmani M, Zargaran A. A brief report of Rhazes manuscripts in the field of cardiology and cardiovascular diseases. Int J Cardiol. 2016;207:190–1.