**CS1138**

**Machine Learning**

**Early-Stage Heart Disease Prediction using Machine Learning Algorithms**

**REPORT**

Submitted By:

Aishani Goyal 2022BTech007

Akshita Bhandari 2022BTech009

Bhoomi Gupta 2022BTech025

Varun Pareek 2022BTech108

Submitted To: Dr. Arpan Gupta

Assistant Professor | Computer Vision

IET JKLU

**INDEX**

|  |  |  |
| --- | --- | --- |
| Serial No. | Title | Page |
| 1 | Abstract | 2 |
| 2 | Introduction | 2-3 |
| 3 | Problem Statement | 3 |
| 4 | Literature Review | 3-4 |
| 5 | Proposed Methodology | 5-6 |
| 6 | Experiments | 6-9 |
| 7 | Results and Discussion | 10-21 |
| 8 | Conclusion and Future Scope | 22-23 |
| 9 | References | 23 |
| 10 | Appendix | 24 |

**Abstract**

Heart disease, a prevalent cardiovascular disease globally, prioritizes efficient prediction and diagnosis. With the help of machine learning, our study proposes an "Early-Stage Heart Disease Prediction Model." Using SMOTE technique for data balance and Random Forest for most relevant feature identification and further predicting heart disease, our model enhances accuracy. Evaluation against diverse algorithms emphasizes the model's efficacy, promising improved clinical decision-making and timely interventions, thereby reducing mortality rates due to heart diseases. The study outlines task description, novel approaches, literature review, model design, performance evaluation, and future research directions.

**Introduction**

Heart disease is a type of cardiovascular disease (CVDs) that is a major health issue globally, with heart attacks and strokes being the main causes of CVD-related deaths, particularly affecting individuals under 70 years old. According to the WHO, by 2030, CVDs will cause about 22 million deaths every year globally, which is more than the current 12 million. The risk of heart disease is greatly influenced by lifestyle choices like eating poorly, not exercising, smoking, and drinking too much alcohol.

The WHO recommends a range of measures such as preventive strategies, standardized care protocols, improved healthcare infrastructure, and effective disease monitoring. Healthcare benefits greatly from machine learning, which provides essential abilities in identifying, diagnosing, and predicting diseases. By utilizing data mining and machine learning methods, clinicians can gain valuable information to enhance patient results. It has been observed that machine learning has significant uses in cardiology, especially in automating the identification of irregularities in electrocardiogram readings, thanks to the increasing availability of wearable devices.

Machine learning is not only used for detecting anomalies, but also for categorizing risks in order to predict early heart disease. Machine learning algorithms can accurately categorize individuals into various risk groups by examining a wide range of patient information, including medical history, lifestyle factors, and diagnostic test results. This allows healthcare providers to identify patients who are at a higher risk and implement specific interventions and ongoing monitoring.

In this study, we propose an "Early-Stage Heart Disease Prediction Model" comprising several crucial components. Our approach includes the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalanced training data distributions. Additionally, we employ a Random Forest Classifier to identify the key factors contributing to heart disease. Using Random Forests for predicting heart disease, with their ensemble learning approach to understand complex relationships between features, improves the accuracy of the model.

Developing this model posed several challenges, particularly in feature selection and achieving a balanced distribution of the training dataset while averting overfitting. Identifying the most relevant features among a potentially large set of variables while avoiding noise or irrelevant information is crucial as determining the optimal number of features and their interaction can impact model performance.

To evaluate our proposed model, we worked on publicly available dataset from Kaggle, comparing its performance against various other machine learning algorithms, including Logistic Regression, K-Nearest Neighbours, Decision Trees, Naïve Bayes, Support Vector Machines, Stochastic Gradient Descent, and Gradient Boosting.

We believe that the developed model will enable clinicians to diagnose patients effectively and efficiently, thus improving clinical decision-making. Timely detection facilitated by the model can facilitate prompt interventions, potentially preventing fatalities stemming from delayed diagnoses.

The study is structured as follows: Section II mentions the description of the task that we aimed to solve and the novel approaches in our work. Section III summarized the literature review. Section IV presents the proposed model including datasets description, overall design, and modules of the proposed model as well as analysis of the novel approaches in the work. Section V discusses the details of the models trained, datasets worked on and different experiments that were executed. Section VI presents the performance evaluation of the proposed model and the visualization of results that were obtained. Finally, the concluding remarks and future research directions are presented in Section VII.

**Problem Statement:**

To develop a predictive model using various health indicators to assess an individual’s risk of getting heart disease at an early stage in order to reduce the mortality rate in the world making an efficient model with less errors by introducing novelty in feature selection, computational efficiency and handling class imbalance using SMOTE (Synthetic Minority Oversampling Technique).

**Literature Review:**

A lot of research and studies are undergoing till now to reduce the mortality rate worldwide by using machine learning algorithms to detect and predict heart diseases at an early stage. Below are the literature reviews that are taken from the informative research and studies that were done in the past years which helped us in adding another efficient model in the aim to reduce the mortality rates.

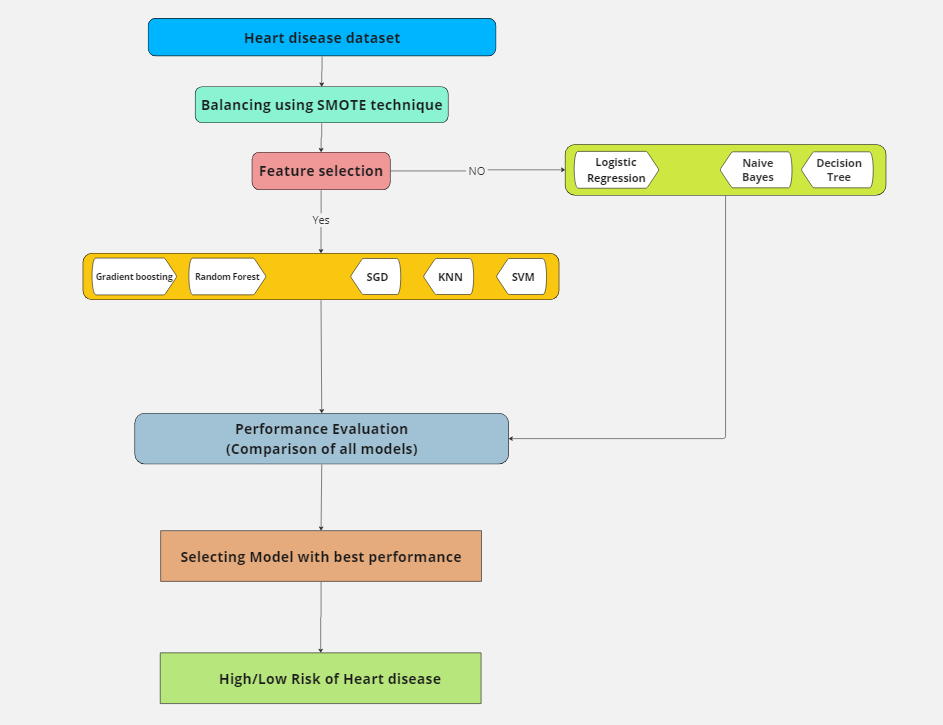
Fitriyani et al. (2020) developed a HDPM, an integrated predictive model within a clinical decision support system. They utilized two publicly available heart disease datasets that have been widely used to compare the performance of prediction models among researchers.: Statlog, comprising 270 instances, and Cleveland, containing 303 instances, each characterized by 13 attributes. Their novel approach included DBSCAN for outlier detection and elimination, SMOTE-ENN to balance the distribution of training data, particularly to address class imbalance and XGBoost for heart disease prediction. Their proposed model achieved high accuracy (Statlog: 95.90%, Cleveland: 98.40%) outperforming other models and previous study results. Furthermore, their model achieved AUC score of up to 1.00 (both datasets), higher than other state-of-the-art models.

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022) developed two machine learning models KNN and MLP (Multi-layer perceptron) where they checked and compared the performances of both the models and chose the one which performed efficiently in predicting and detecting the Cardio-vascular disease at an early stage The goal was to find out whether a person would have heart disease or not. The confusion matrix of each model was created. Where 80% (243 samples) of the dataset was used to train the model and 20% (60 samples) used to test the model. The dataset used was taken from UCI (University of California Irvine) repository. It had 303 samples and 76 attributes, where only 13 attributes were selected using feature selection technique. The dataset was imbalanced and contained both categorical and numerical features. After data pre-processing, class balance and careful and most important features’ selection, the performance of the two models were obtained where the evaluation matrixes showed that MLP (82.47 %) performed better and predicted more TPs and fewer FPs than KNN (73.77 %). The novel approach used here was, focusing more on data pre-processing and improving model’s performance through basic feature selection techniques, like correlation matrix, plotting various graphs between various independent features and plotting graphs between each independent feature and target. To increase the MLP’s performance, they focused on increasing and finding the optimal number of hidden layers to make an efficient MLP model.

Surendra Reddy Vinta (2023) developed a leveraging machine learning technique for improving heart disease prediction using feature selection. The dataset used in the model was of the Cleveland heart disease dataset which has 303 instances and is available on UCI machine learning repository. The novel approach of the research includes a novel hybrid feature selection technique in which maximum relevance and minimum redundancy (mRMR), Relief, a genetic algorithm, and least absolute shrinkage and selection operator (LASSO) were used for feature selection. Then many classifiers were used in the creation of the prediction system, including support vector machines, random forests, logistic regression, and naive bayes. The research outcome showcased that the best accuracy and sensitivity was achieved with the Random Forest algorithm (89% and 78.2% respectively). The study's findings show that by using feature selection algorithms we can improve the prediction system's accuracy, sensitivity, specificity, and processing speed.

The paper "Improving Heart Disease Detection and Patients’ Survival" by Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C. O., Chuah, J. H., & Gheni, H. M. was published in IEEE Access on June 21, 2022, with DOI 10.1109/ACCESS.2022.3185129.They Introduced an approach that combined supervised infinite feature selection and improved weighted random forest model to enhance heart disease detection and patient survival prediction. The methodology addresses class imbalance and high dimensionality, improving accuracy by through feature selection and hyperparameter optimization. The innovative use of machine learning techniques demonstrates potential for improved healthcare decision-making and patient care. This research enhanced the healthcare system and served as a valuable tool for healthcare practitioners.

**Proposed Methodology:**



The following proposed model was developed to provide high performance prediction in the presence or absence of heart disease given the current condition of the subjects:

1. Heart disease dataset: The process starts with a collection of data on heart disease. This data likely consists of patient information including medical history, demographics, and biological measurements.

2. Balancing using SMOTE technique: This step addresses a potential imbalance in the data. Data imbalance can occur when there are significantly more instances of one class (e.g., healthy patients) than the other (e.g., patients with heart disease). The SMOTE (Synthetic Minority Oversampling Technique) technique is used to address this by creating synthetic data points for the under-represented class.

3. Feature selection: The data contains 21 features (data points). This step involves choosing a subset of features that are most relevant to predicting heart disease. Models that have been trained with feature selection are KNN, SVM, SGD, Random Forest and Gradient boosting. Other models (Logistic Regression, Naive Bayes and Decision Tree) have been trained without feature selection.

4. Splitting the data: The data is divided into two sets: a training set and a testing set. The model is trained on the training data, and its performance is evaluated on the testing data.

5. Model selection: Here, different machine learning algorithms are used to create models to predict heart disease. The algorithms that we worked on include Logistic Regression, KNN, SVM, SGD, Naive Bayes, Decision Tree, Random Forest and Gradient Boosting.

6. Performance Evaluation: The performance of each model is assessed on the testing data. This likely involves metrics that measure how well each model can distinguish between patients with and without heart disease.

7. Selecting Model with best performance: The model with the best performance on the testing data is chosen as the final model.

8. High/Low Risk of Heart disease: This final step depicts the model being used in predicting the likelihood of heart disease for new patients. The input would be a patient’s information, and the output would be a prediction of their risk of having heart disease.

In our work, we introduce a novel approach that focuses mainly on three key aspects of building a robust heart disease prediction model. Firstly, we address class imbalance in the data using the SMOTE technique. This ensures the model is trained on a balanced dataset, leading to more accurate predictions. Secondly, we apply a strategy to avoid overfitting. By using the ensemble learning approach of random forest, we mitigate this risk. Lastly, we incorporated feature importance analysis. This indicates which factors in the data have the strongest influence on the model's predictions. By understanding these key factors, we gain valuable knowledge about the key contributors to early-stage heart disease. This additional insight can be crucial for future research and preventative measures.

**Analysis of the Approaches:**

Each proposed algorithm is evaluated with and without balancing the dataset using SMOTE. Performance metrics such as accuracy, precision, recall, and F1 score are compared between balanced and unbalanced datasets for each classifier. Then the accuracy of all the algorithms after balancing are compared.  
At last the algorithm with the highest accuracy comes out to be Random Forest (91.67%), which means that the model best works with this classifier.

**Experiments:**

**1. Decision Tree:**

a) Imbalanced Data Training: The first step was to train the original dataset with a decision tree classifier.

b) Balanced Data Training using SMOTE: Afterwards, we use SMOTE technique to balance the original dataset and then use decision tree classifiers on the balanced dataset.

c) Balanced Dataset with maximum depth: We also set the maximum depth of the decision tree to be 3 as it is easier for visualization and computationally effective.

At last, we compare the performance metrics of both the balanced and imbalanced dataset and see which one is the best.

**2. Naive Bayes:**

a) Imbalanced Data Training: The first step was to train the original dataset with naïve bayes classifier.

b) Balanced Data Training using SMOTE: Afterwards, we use SMOTE technique to balance the original dataset and then use naïve bayes classifier on the balanced dataset.

At last, we compare the performance metrices of both the balanced and imbalanced dataset and see which one is the best.

**3. Stochastic Gradient Descent (SGD):**

a) Imbalanced Data Training: The first step was to train the original dataset with SGD classifier without class balancing the training data.

b) Balanced Data Training using SMOTE: Afterwards, we use SMOTE technique to balance the original dataset. Due to overfitting, model performed poorly with SMOTE technique.

c) Balanced Data Training using SMOTE-ENN: SMOTE-ENN was used to balance the training data in which samples for minority class are generated and some samples from majority class are removed, achieving a balance between the two. Significant increase in model’s performance observed with an increase in TPR and a decrease in FNR.

d) Feature Selection using ANOVA: Then SGD classifier was trained with feature selection technique (in this case used ANOVA) on the balanced dataset. But not much difference was observed in the model’s performance. Instead, the FNR was slightly increased which indicated some features having an impact on target variable were removed by ANOVA which was its limitation.

**4． Support Vector Machine (SVM):**

a) Imbalanced Data Training: Initially after data preprocessing, the SVM model is trained on the unbalanced data.

b) Balanced Data Training using SMOTE: Afterwards, we use SMOTE technique to balance the original dataset. A decrease in model’s accuracy was observed but alongside a balance between other evaluation metrics was found. Also, an increase in AUC score from 0.55 to 0.74 which indicated that the performance was improved in correctly identifying the minority class.

c) Although the true predictions were increased after balancing the training data but there was a significant decrease in the accuracy of the SVM model as this classifier is sensitive to unbalanced data due to which after balancing it prioritizes maximizing the margin between the two classes resulting in poor prediction for minority class. Hence, model’s performance is reduced.

**5. Logistic Regression:**

a) Imbalanced Data Training: The first step was to train the original dataset with Logistic Regression.

b) Balanced Data Training using SMOTE and Weight Balance: Afterwards, we used SMOTE technique and Weight balance technique to balance the original dataset and then used Logistic Regression on the balanced dataset.

At last, we compared the performance metrics of both the balanced and imbalanced dataset and compared the results between the two (After balancing the dataset using weight balance and SMOTE). And used the SMOTE as the model’s performance was better in this.

**6. K-nearest neighbours (KNN):**

a) Imbalanced Data Training: The first step was to train the original dataset with KNN classifier.

b) Balanced Data Training using SMOTE: Afterwards, we used SMOTE technique to balance the original dataset and then chose the optimal value of K, where the recall was high for class 1 and FNs were low.

Then, we compared the model's performance with and without feature selection and chose the one which gave the better results.

**7. Random Forest**

Since the objective was to classify and predict a person has early signs of heart risk, we required an efficient machine learning model that not only classifies our input collected data but works well on test data.

It is an ensemble learning model which uses multiple decision trees to classify different classes, it uses the concept of “*bagging*” and “*boosting*” implying it first divides the complete dataset into smaller sub parts and work on them individually and then further learn from the previous mistakes and combines the results of all trees. This process helps it come up with a solid prediction, discarding any wrong ideas as it goes.

The Random Forest classifier is good at handling complicated data. It can deal with different relationships between features without getting stuck on the details. And it doesn't get too caught up in the training data, so it can make good predictions with new data too.

Further we observed there was a significant class imbalance between our majority class and minority class thus Random Forest was initially biased towards the majority class and was not able to predict the accurate model. Therefore, we used the concept of class balancing using SMOTE analysis.

This imbalance could potentially skew the model's performance, as it may prioritize the majority class at the expense of the minority class.

Thus, we could go with either random sampling or SMOTE sampling.

**Random Sampling**

i) Begin with a defined population, which could be anything from the dataset.

ii)Use a random selection process to choose individuals from the population to include in the sample. Every member of the population has an equal likelihood of being chosen for the sample.

iii) We may choose to either replace them back into the population after selection or leave them out. We can analyse it to draw conclusions or make conclusions about the larger population from which it was drawn.

**8. Gradient Boosting Classifier**

Alone random forest was not able to prove the above observations, therefore we tried Gradient Boosting Classifier. It is used for classification tasks. It works in the following manner:

i) It starts by identifying the minority class in the dataset.

ii) It iteratively learns from the data and focuses on improving the prediction accuracy of the minority class during each iteration.

iii) GBC operates by training weak learners in a sequential manner. Each weak learner aims to rectify the mistakes of its predecessors, giving more attention to misclassified instances from the minority class

iv) During each iteration, it adjusts the weights of the training instances based on the gradient of the loss function. This effectively gives more importance to misclassified instances, particularly those from the minority class, in subsequent iterations so as to converge faster and reach the optimum.

v) We combine the predictions of all weak learners and predict the final outcome. It assigns higher weights to weak learners that perform well on misclassified instances from the minority class, resulting in a model that is better at classifying the minority class.

**Results and Observations:**

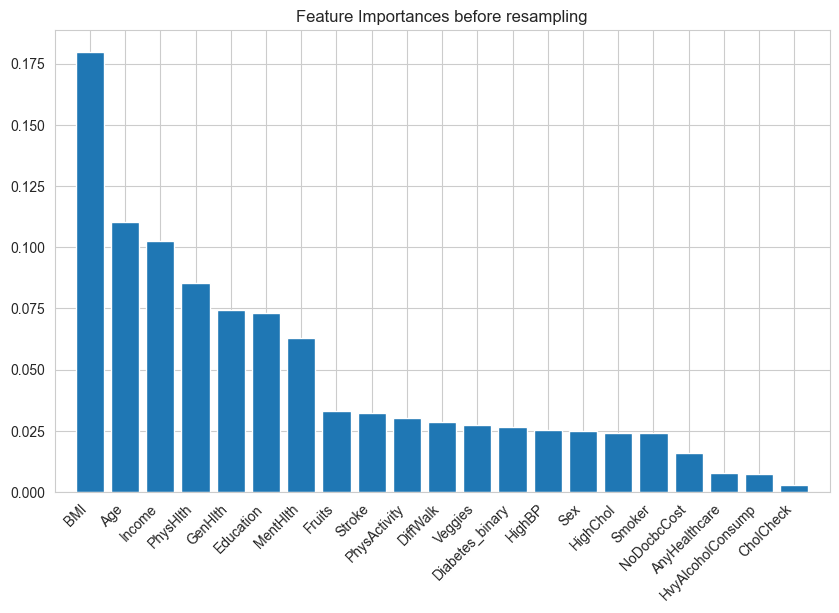
**There were significant results observed:**

**1.**

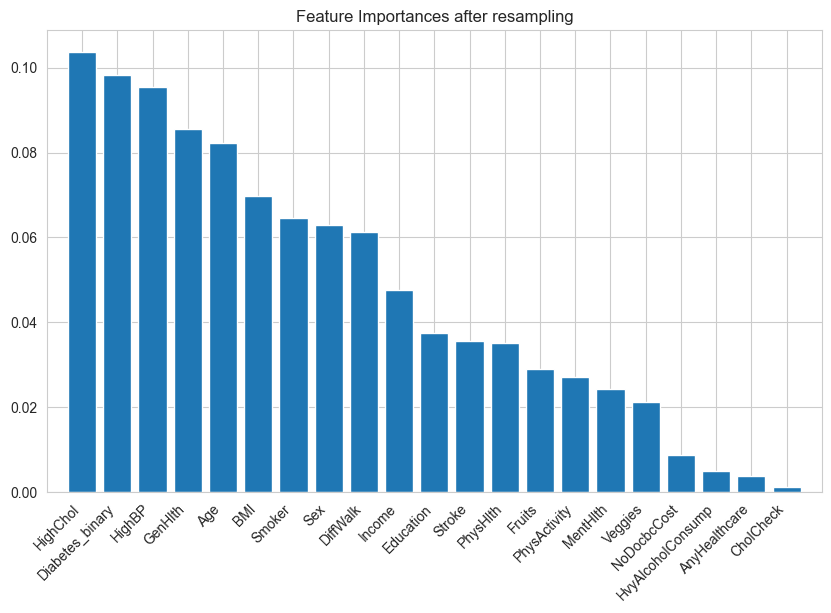
\* There was significant class imbalance observed

2. After using SMOTE Analysis:

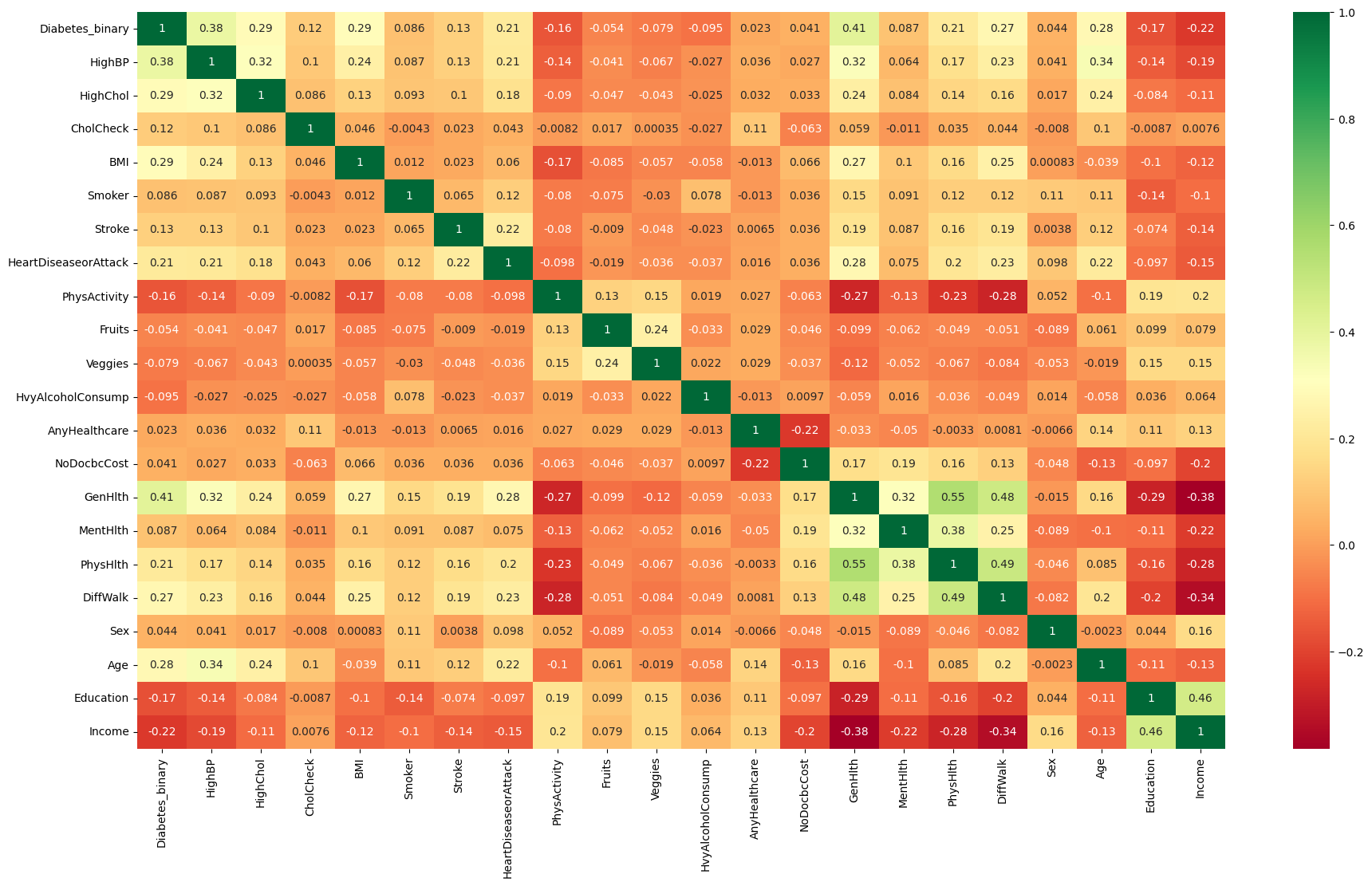


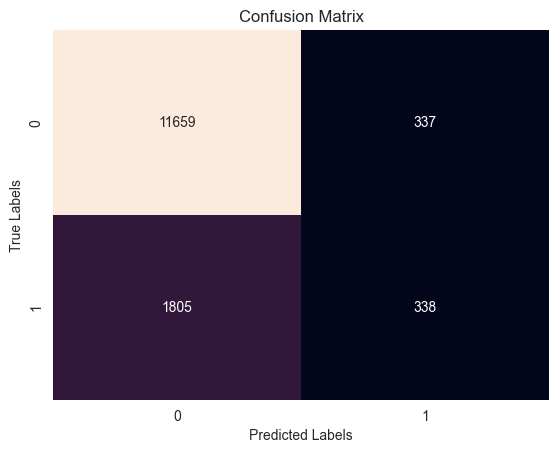
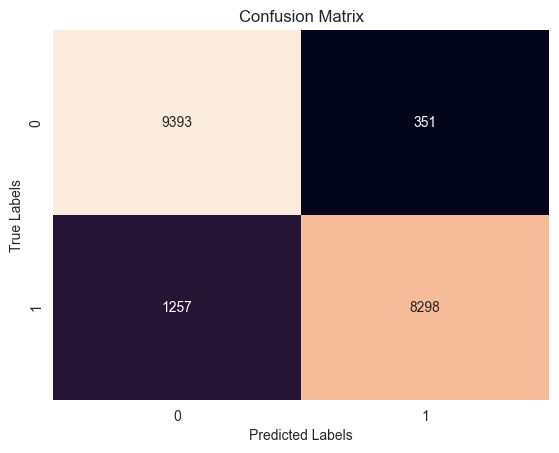
3. Feature Importance before balancing

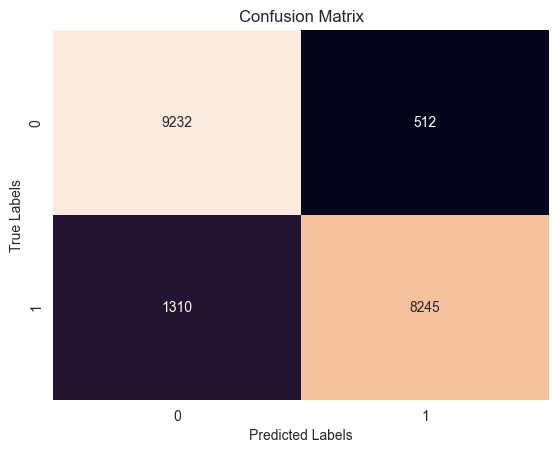
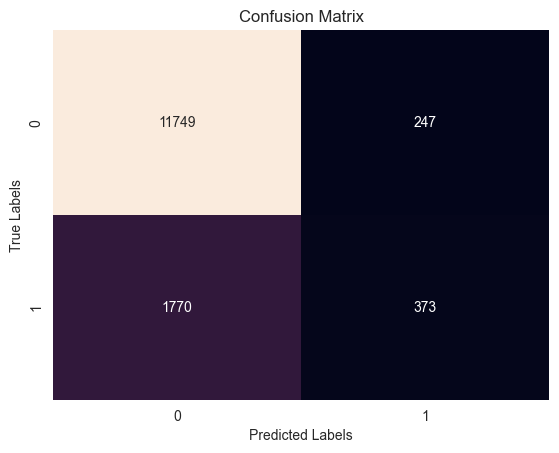
4. Feature importance after balancing

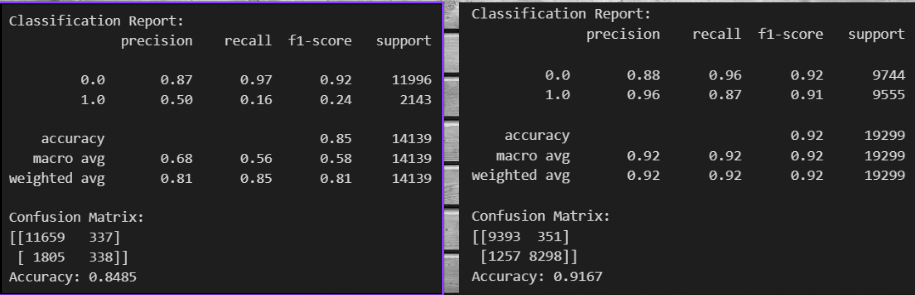


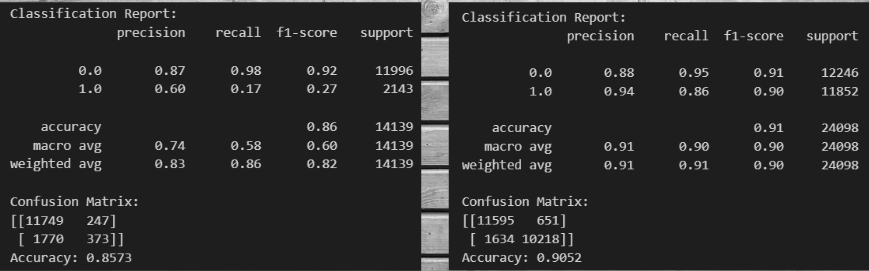
5.The above results can be verified from the correlation matrix



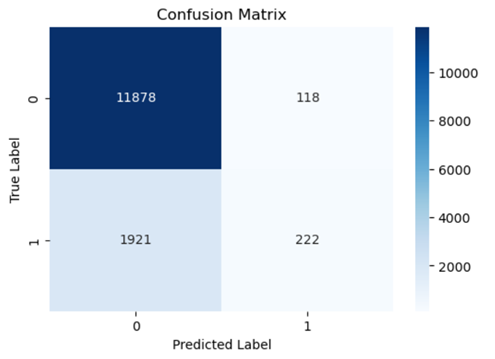
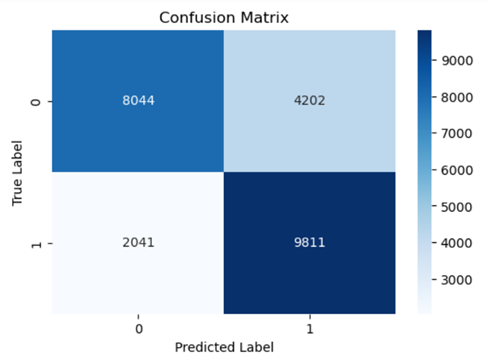
6. Confusion Matrix for RF before and after class balancing:

7. Confusion Matrix for Gradient Boosting Classifier before and after class balancing:

8. The classification report for Random Forest classifier:

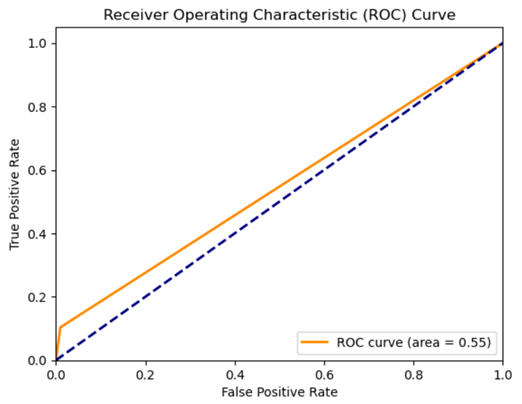
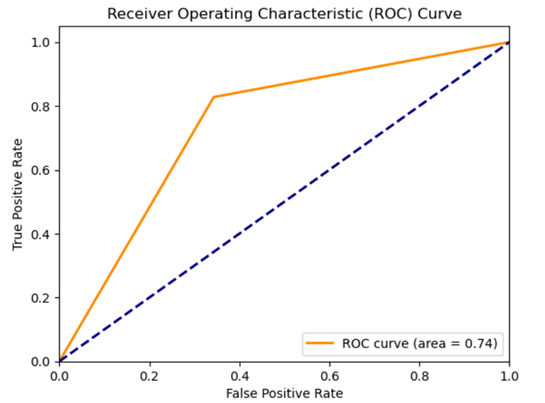
9. The classification report for Gradient Boosting Classifier:

10 (a).. SVM Model Comparison on Unbalanced and Balanced data:

**Fig1 (a). Confusion Matrix for unbalanced dataset Fig1 (b). Confusion matrix for balanced training data**

10 (b). ROC curve:

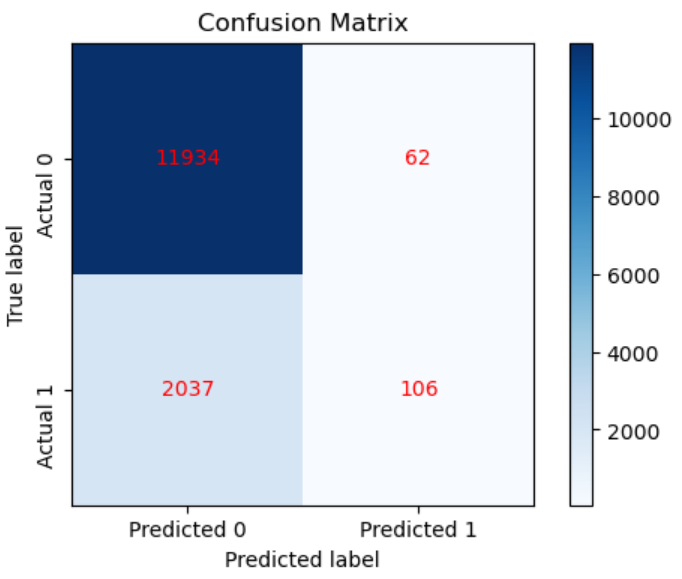
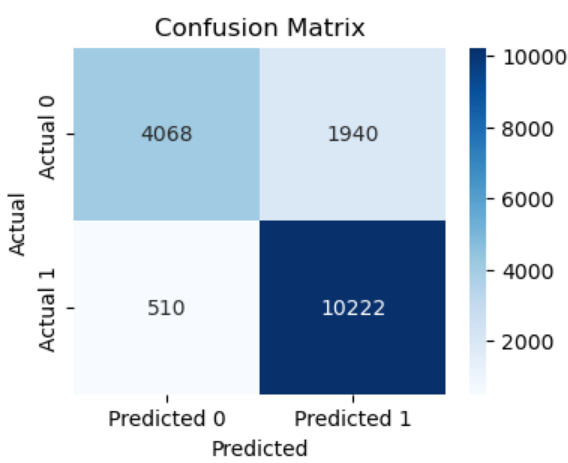
 

**Fig2 (a). ROC curve for unbalanced dataset Fig2 (b). ROC curve for balanced training data**

**Observations for SVM model:**

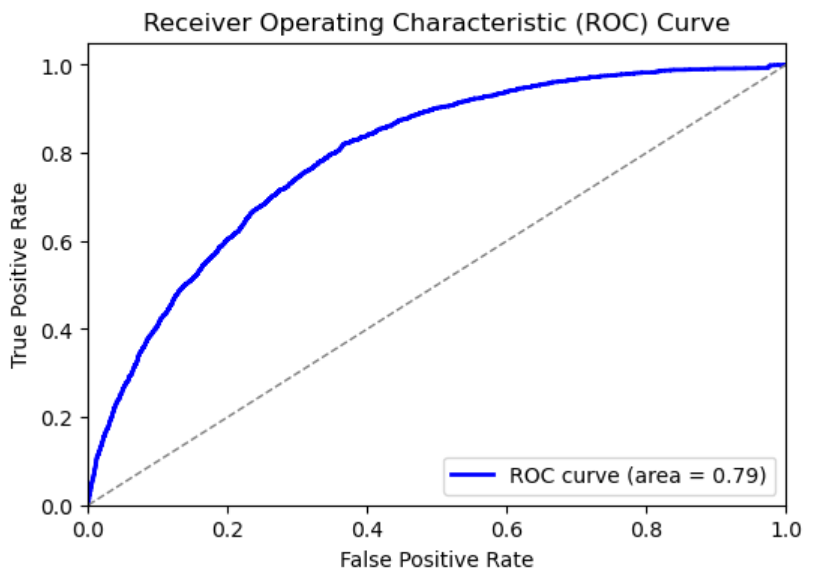
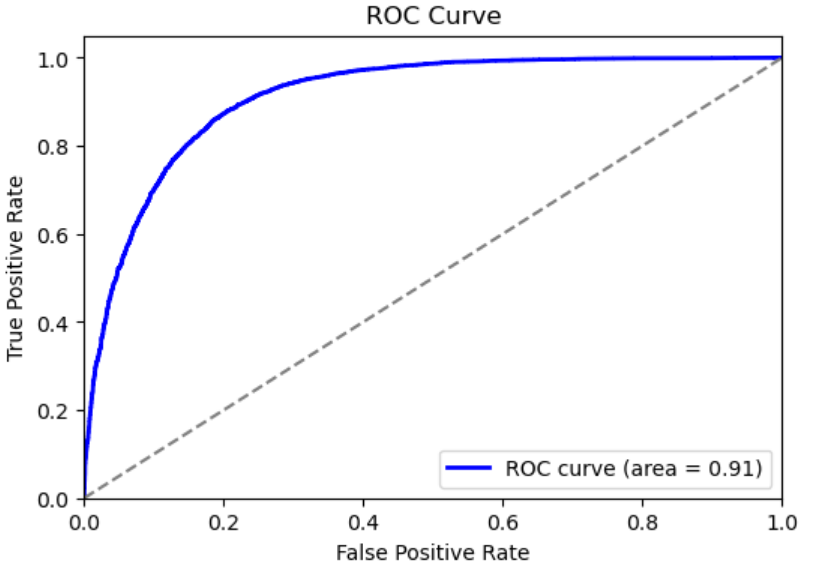
* As seen from the Fig1 and Fig2 , when SVM model is trained on unbalanced data poor performance is obtained in correctly identifying the minority class (class 1)
* When SMOTE based balancing is applied to training data, performance is improved in correctly identifying the minority class.
* AUC score is improved from 0.55 to 0.74, indicating better discrimination between classes.

11 (a) SGD (Stochastic Gradient Descent Classifier) on unbalanced and balanced data:

**Fig1 (a). Confusion Matrix for unbalanced dataset Fig1 (b). Confusion matrix for balanced training data**

12 (b) ROC curve:

**Fig2 (a). ROC curve for unbalanced dataset Fig2 (b). ROC curve for balanced training data**

12 (c) Feature Selection using ANOVA in SGD:

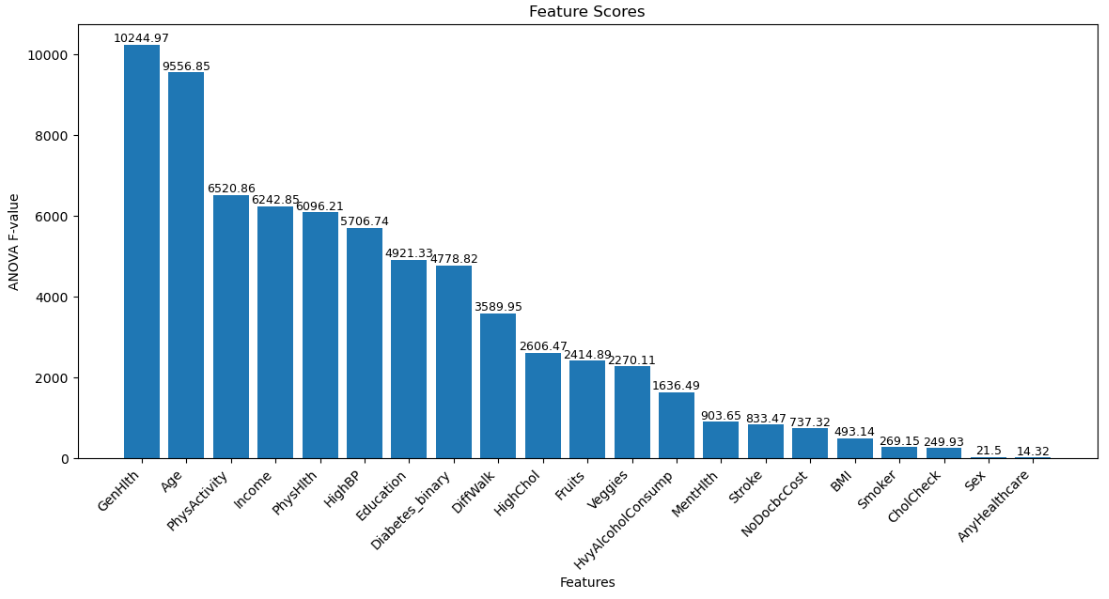


Fig1 (a). Before applying feature selection

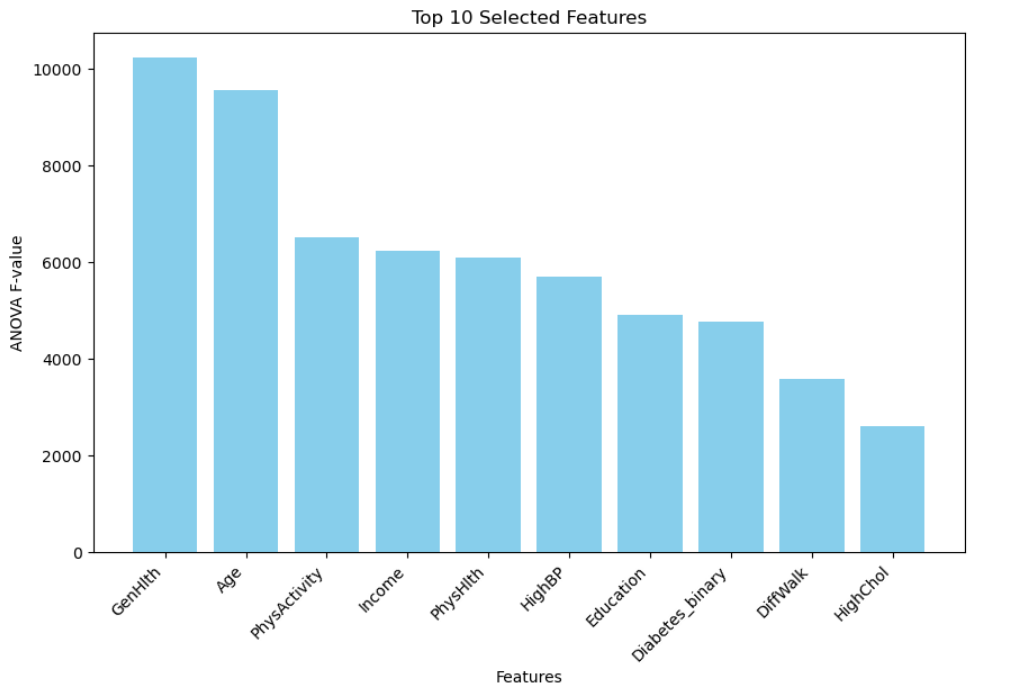
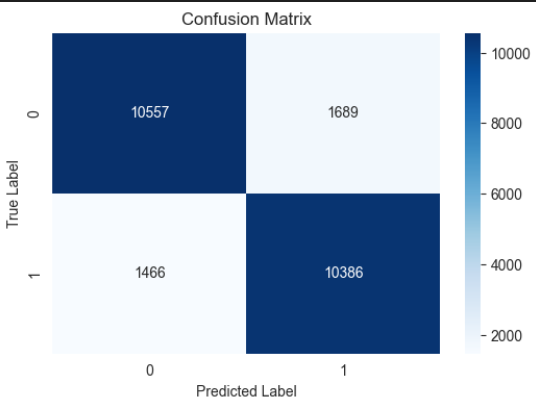
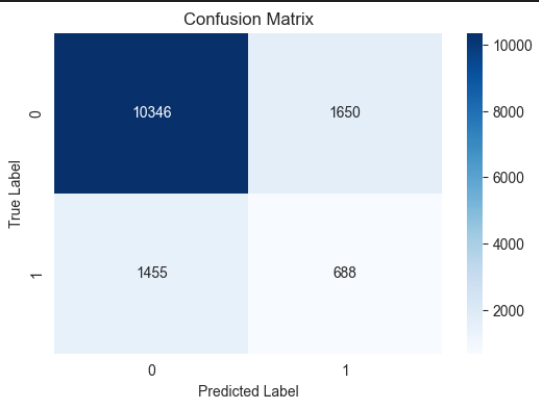


Fig1 (b). After applying feature selection using ANOVA

**Observations for SDG model:**

* SMOTE-ENN achieves a higher balance between precision and recall for both classes. TPR is significantly increased and FNR was reduced that was the major objective.
* After feature selection, similar level of accuracy was maintained, instead TPR was slightly reduced and FNR was increased than before which could be due to high dependency of SDG classifier on feature importance. With removal of some features that may have an impact in predicting the target variable in combination with some other features and not individually, the model’s performance is reduced.
* Apart from this , the SDG classifier has shown an increase in the correct predictions after balancing the training data using SMOTE-ENN technique with maximum computational efficiency for large dataset.

13 (a) Decision Tree Model Comparison on Unbalanced and Balanced data:

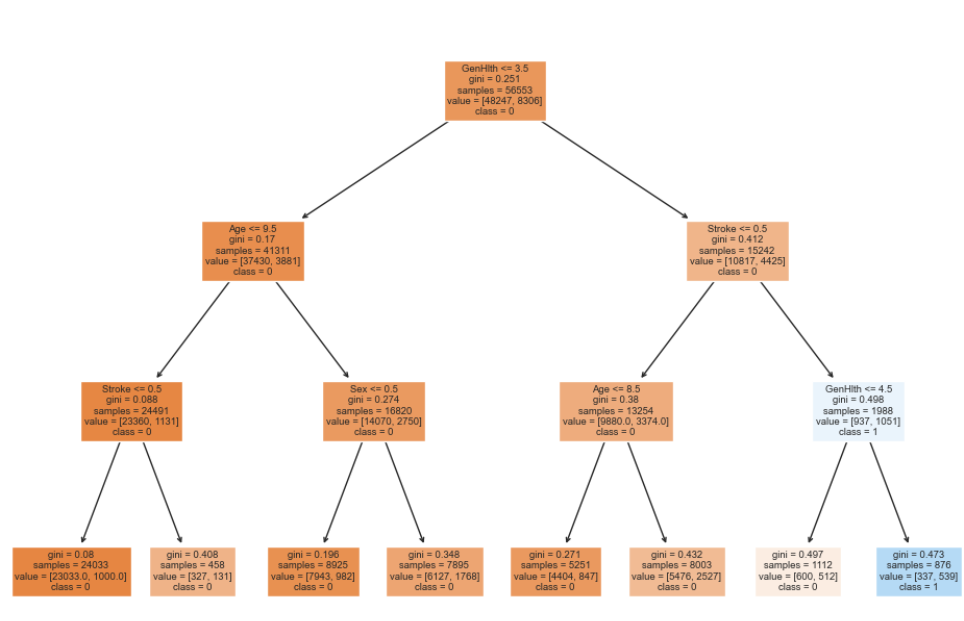


**Fig 1(a). Confusion matrix without balancing Fig 1(b). Confusion matrix with balancing**

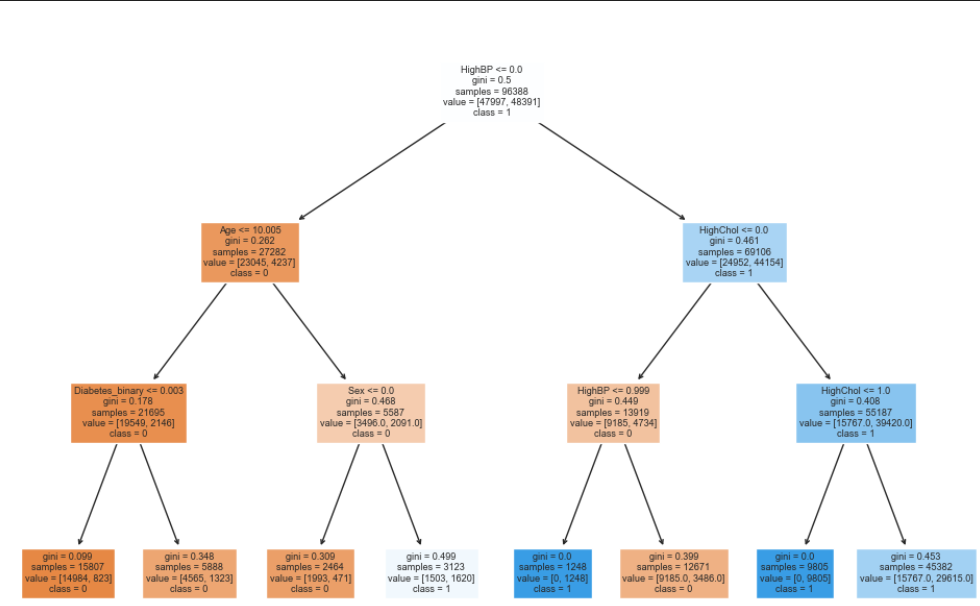
**Observations for Decision Tree model:**

* As seen from the Fig1 and Fig2 , when decision tree model is trained on unbalanced data poor performance is obtained in correctly identifying the minority class (class 1)
* When SMOTE based balancing is applied to training data, performance is improved in correctly identifying the minority class.
* The accuracy was improved from 78.03% to 86.90%.
* For visualization the maximum depth of the decision tree was set to 3 for better visualization and it is also computationally effective.

13(a) Decision Tree visualization: -

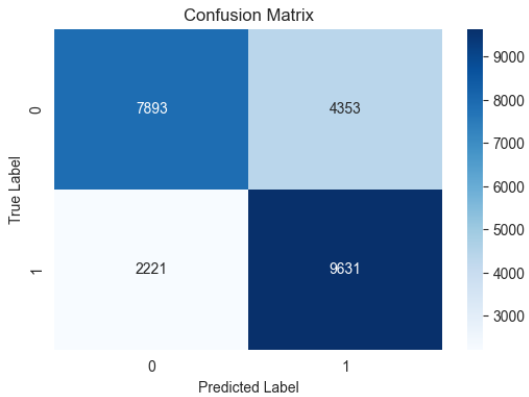
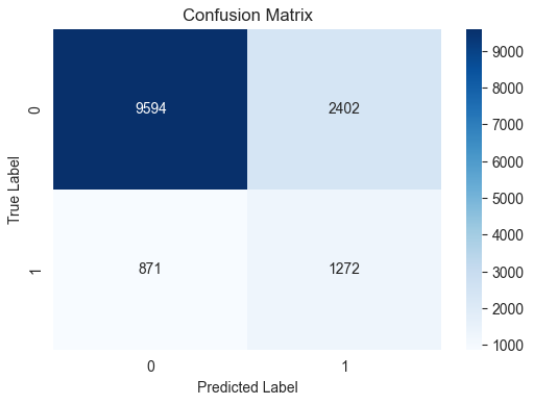


**Fig 2(a). Decision Tree visualization without balancing**

****

**Fig 2(b). Decision Tree visualization with balancing**

13 (a) Naive Bayes Model Comparison on Unbalanced and Balanced data:



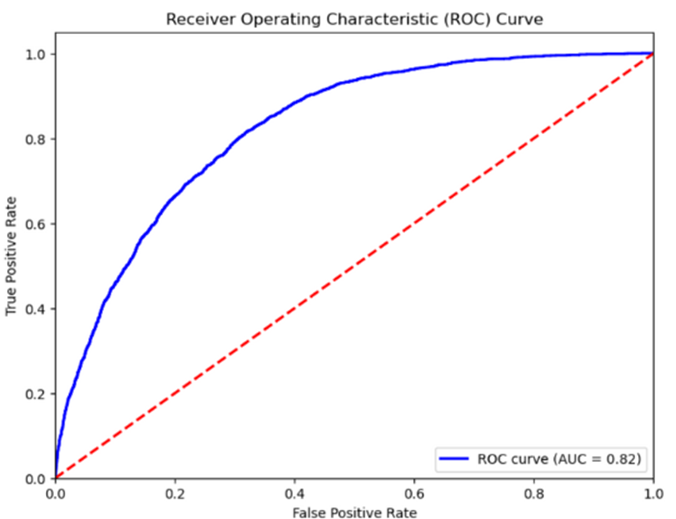
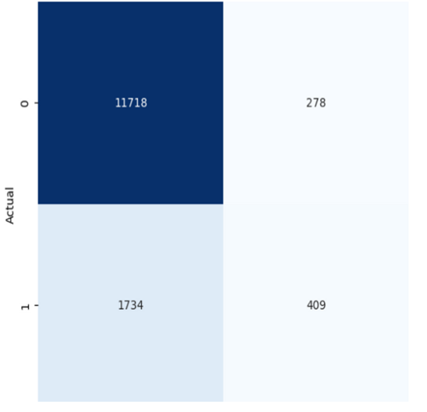
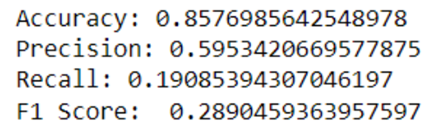
**Fig 2(b). Confusion matrix without balancing Fig 2(b). Confusion matrix with balancing**

**Observations for Naive Bayes model:**

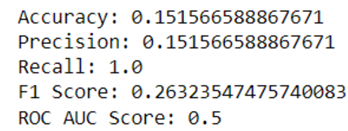
* As seen from the Fig1 and Fig2 , when Naive model is trained on unbalanced data, poor performance is obtained in correctly identifying the minority class (class 1)
* When SMOTE based balancing is applied to training data, performance is improved in correctly identifying the minority class.
* Though the accuracy was decreased from 76.85% to 72.71%, the overall precision and recall score improved.

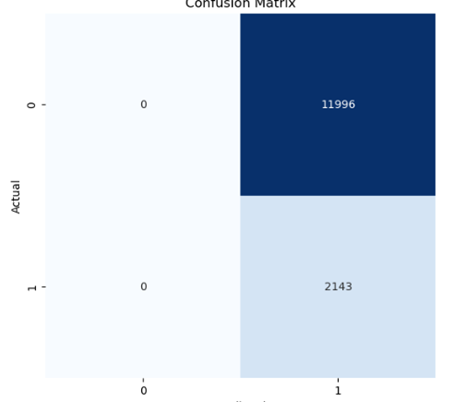
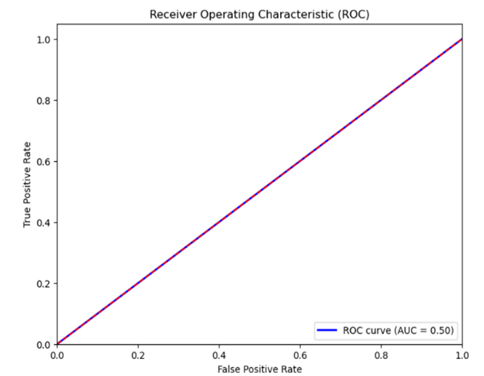
**14. Logistic Regression Model :**

**(a) Results of the model before data pre-processing:**

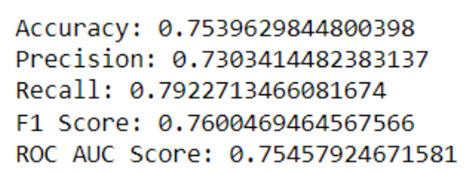
**** ****

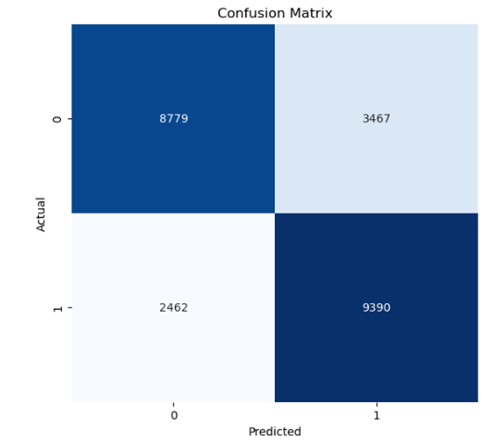
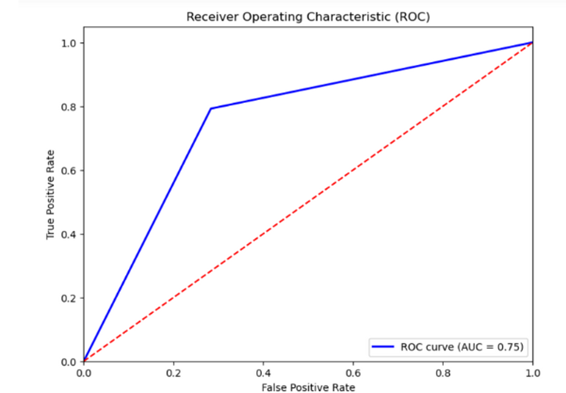
**(b) Results of the model after balancing the data using Weight Balance technique:**

****

** **

**(c) Results of the model after applying SMOTE as balancing technique:**

****

** **

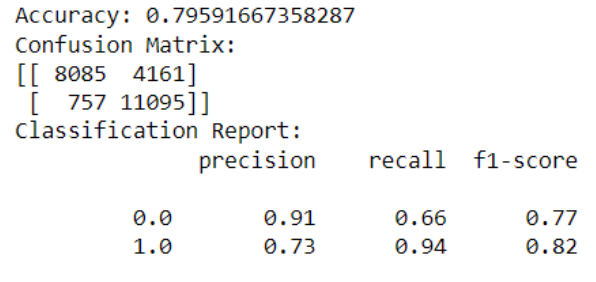
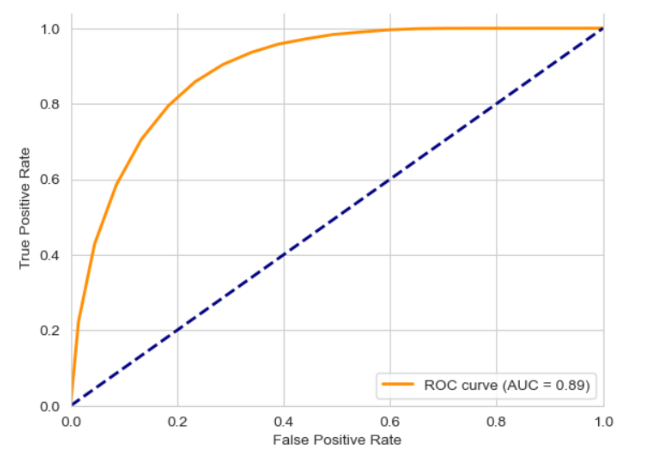
**Observations for Logistic Regression Model:**

As seen from the results of (a), even though the accuracy is high, the model is detecting the high number of instances wrongly. This is because the dataset is imbalanced. So. by applying the two types of balancing techniques, the Weight balancing (b) is performing very poorly, even though it reduced the number of FNs greatly. This is because the weight balancing technique could be biased towards the minority class and gives poor results. The SMOTE (c) gave better performance without being biased.

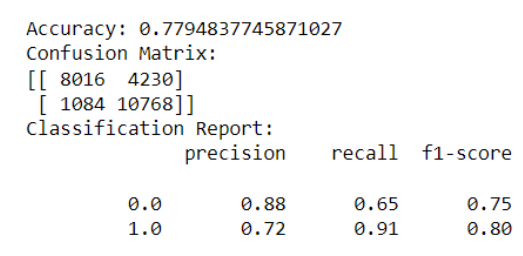
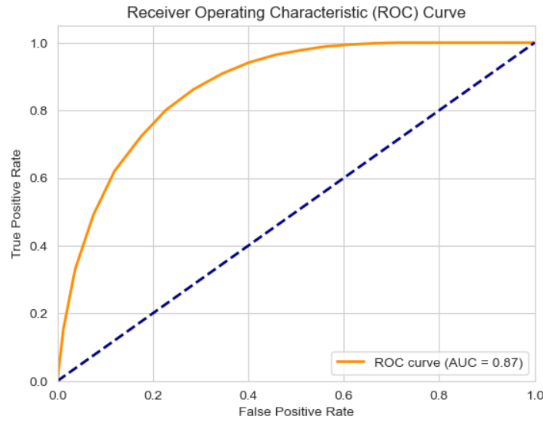
For the logistic model, the feature selection didn’t play much importance and gave approx similar results. Therefore, all features were selected.

**KNN MODEL -**

1. **Results of the model without feature selection technique:**

** **

1. **Results of the model after feature selection technique:**

** **

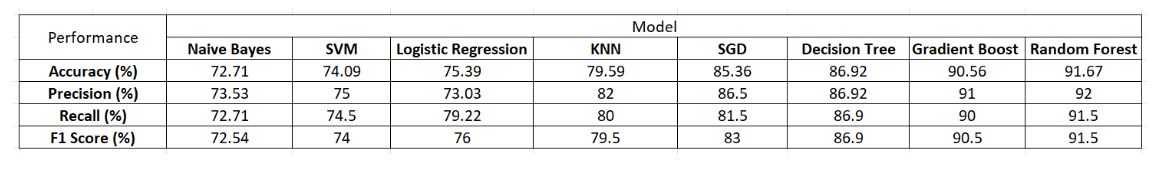
**fig 1**

fig 1: Random Forest classifier used to extract most top 10 important features.

**Observations for KNN Model:**

Without feature selection the model’s overall performance came out better compared to after feature selection. This means that all the independent features provided the necessary information to generate the output.

**Performance Evaluation:**

****

In conclusion, it is evident that the Random Forest model produces the most favorable outcomes for our problem. Our focus revolves around refining the model to near-perfection, with the aim of effecting positive societal change with awaring people before a heart disease of attack.

**Conclusions:**

Provided the above performance evaluation table, we observed various things

Random Forest Classifier works best for the prediction of heart disease. This leads to several results.

a. The Random Forest Classifier shows the highest accuracy in predicting heart disease compared to other models we used. This suggests that its ensemble learning approach, which uses many decision trees, worked for the complex relationships between features and the target variable that is ‘HeartAttackorRisk’ with overall 91.67% accuracy.

b. The Random Forest Classifier identifies key features such as age, cholesterol levels, blood pressure, and general health as significant factors.

c. We noticed that when we balanced the class imbalance, the importance of features changed. Before balancing, our model showed bias towards **BMI** which affected its accuracy. But after balancing, all features had a similar importance. This helped us optimize the model to consider all features equally, reducing bias and improving its performance.

d. Another key observation we noticed is that where other models were prone to overfitting and to some extent after balancing, they got overfitted, Random Forest showed its robustness and mitigated overfitting as the ensemble learning played a crucial role in achieving this accuracy.

e. Another keen observation we observed is that this model is prone to overfitting as well after balancing as the MSE for degree 2 and MSE for degree 3 are almost equal suggesting that 21 features overfits data as after hyperparameter tuning the model accuracy went down.

f. The Random Forest Classifier's predictions match up with what doctors already know about heart disease risk factors. By accurately spotting people at high risk, it helps doctors’ step in early with treatments and preventive actions. This means better outcomes for patients and less money spent on treating heart disease later on.

**Future Directions:**

* We observed that Gradient Boosting Classifier predicted very similar to Random Forest but was less than it because of the complexity of features and after class balance it degraded so we would explore ensemble techniques such as XGBoost to enhance the performance of it.
* Logistic Regression assumes a linear relationship between the features and the target variable but there is a possibility for non-linearity. So, we would focus on feature engineering techniques to create new
* features that capture nonlinear relationships with the target variable.
* For KNN our future scope would be to find the best optimal value of **K** to produce the optimal solution.
* Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This strong assumption may not hold true so we would look for advanced Naïve Bayes.
* SVM maximizes the margin between classes when optimizing the decision boundary. For imbalanced data, where minority class has fewer samples than the majority class, SVM may focus more on correctly classifying the majority class, leading to poorer predictions for the minority class. So, we would be looking to assign different weights to classes so as to balance the difference out.
* For SGD it is dependent on features, and after the change in feature importance the model may not perform well therefore the future scope will be to continuously monitor the changes adapt alternative approaches

**References:**

1. Vinta, S. R., Anbalagan, E., Basavaraddi, C. C. S., & Mazumdar, N. (2023). Leveraging Machine Learning Techniques for Improving Heart Disease Prediction Systems Using Feature Selection. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(8s), 560-567.

2. Alalawi, H. H., & Alsuwat, M. S. (2021). Detection of cardiovascular disease using machine learning classification models. *International Journal of Engineering Research & Technology*, *10*(7), 151-7.

3. Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100-1113.

4. Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C. O., Chuah, J. H., & Gheni, H. M. (2022). Improving the heart disease detection and patients’ survival using supervised infinite feature selection and improved weighted random forest. *IEEE Access*, *10*, 67363-67372.

5. 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.

6. ProjectPro. (2024). Heart Disease Prediction using Machine Learning Project. ProjectPro.

7. Analytics Vidhya. (2020). Feature Selection Techniques in Machine Learning. Analytics Vidhya.

8. [Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

**Appendix**

* Dataset: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>
* Presentation:<https://www.canva.com/design/DAGD30YqLVc/_zufArzF72KSCKp1A-Kmsg/edit?utm_content=DAGD30YqLVc&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>
* Code Snippet for Random Forest Model:

# pip install imbalanced-learn

# pip install xgboost

# pip install pdpbox --user

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score,classification\_report,confusion\_matrix

from imblearn.over\_sampling import SMOTE

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.preprocessing import PolynomialFeatures

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

X\_train\_bal, X\_test\_bal, y\_train\_bal, y\_test\_bal = train\_test\_split(X\_resampled, y\_resampled, test\_size=0.2, random\_state=42)

rf\_classifier\_bal = RandomForestClassifier(n\_estimators=100, random\_state=42)

rf\_classifier\_bal.fit(X\_train\_bal, y\_train\_bal)

y\_pred\_bal = rf\_classifier\_bal.predict(X\_test\_bal)

print("Classification Report:")

report\_bal = classification\_report(y\_test\_bal, y\_pred\_bal)

print(report\_bal)

print("Confusion Matrix:")

cm\_balance = confusion\_matrix(y\_test\_bal, y\_pred\_bal)

print(cm\_balance)

accuracy\_bal = accuracy\_score(y\_test\_bal, y\_pred\_bal)

print(f"Accuracy: {accuracy\_bal:0.4f}")

sns.heatmap(cm\_balance, annot=True, fmt="d", cbar = False)

plt.title("Confusion Matrix")

plt.xlabel("Predicted Labels")

plt.ylabel("True Labels")

plt.savefig('graph3.png', dpi=300, bbox\_inches='tight')

plt.show()