THE AMERICAN
UNIVERSITY IN CAIRO
الجـامـعة الأمـريكـيـة بالقـاهـرة

# Exploring Machine Learning and Data Mining Techniques for Heart Disease Prediction: A Comprehensive Analysis

A Thesis Submitted by

Maya Hany Elshweikhy

Supervised by

Dr. Noha Youssef

May 2024

## Table of Contents

## Abstract

This study explores the advancements, methodologies, and challenges in heart disease prediction using machine learning (ML) techniques. With heart disease being a leading global health concern, the integration of ML tools holds immense potential for enhancing prediction accuracy and facilitating early intervention. The review synthesizes key studies, discusses various machine learning models and outlines current challenges in this critical field.

*Keywords:* machine learning, cardiovascular diseases, supervised learning algorithms.

# Exploring Machine Learning and Data Mining Techniques for Heart Disease Prediction

## 1. Introduction

The exponential growth of data in the medical industry has become challenging in managing and extracting relevant information for accurate decision-making. Addressing this challenge, machine learning techniques emerge as a promising solution as Machine Learning (ML) algorithms can help with the early detection and prevention of developing CVDs. Machine learning techniques play a crucial role in solving real-world problems by uncovering hidden patterns and extracting pertinent information from extensive datasets. Particularly in the context of heart failure symptoms, which can manifest at any age, the need to detect and predict such conditions becomes paramount. Older individuals, in particular, face a comparatively higher risk of experiencing heart failure symptoms. Leveraging data mining techniques can identify previously unknown patterns and strongly linked characteristics, facilitating the accurate prediction of heart disease presence from extensive datasets. This study focuses on implementing various data mining techniques and utilising different ML models to determine a person's risk of developing CVDs by using their personal lifestyle factors and other features. These machine learning (ML) models Logistic Reggression, Descison Trees and Random Forest, to predict early-stage cardiac disease.

The study starts by providing an overview of the heart disease prediction in Section 1. A literature review is presented in Section 3. Section 4, lists all ML procedures to be applied on the data set. In addition, Section 4 reviews oversampling techniques that can be applied on the imbalanced data sets.

## 2. Overview of Heart Disease Prediction

Cardiovascular diseases (CVDs), including heart disease, continue to be a leading cause of morbidity and mortality worldwide. Cardiovascular diseases (CVDs) encompass a range of heart and blood vessel disorders. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) claimed an estimated 17.9 million lives in 2019, constituting 32% of total global deaths. Heart attacks and strokes account for 85% of these CVD-related deaths (2021). Of the 17 million premature deaths under the age of 70 attributed to noncommunicable diseases in 2019, 38% were linked to CVDs (2021). World Heart Federation reported in 2023 that over half a billion individuals worldwide are affected by cardiovascular diseases, contributing to 20.5 million deaths in 2021, accounting for almost a third of global deaths—an increase from an estimated 12.1 million CVD deaths (2023). Addressing behavioral risk factors such as tobacco use, unhealthy diet, obesity, physical inactivity, and harmful alcohol consumption could prevent a majority of cardiovascular diseases.

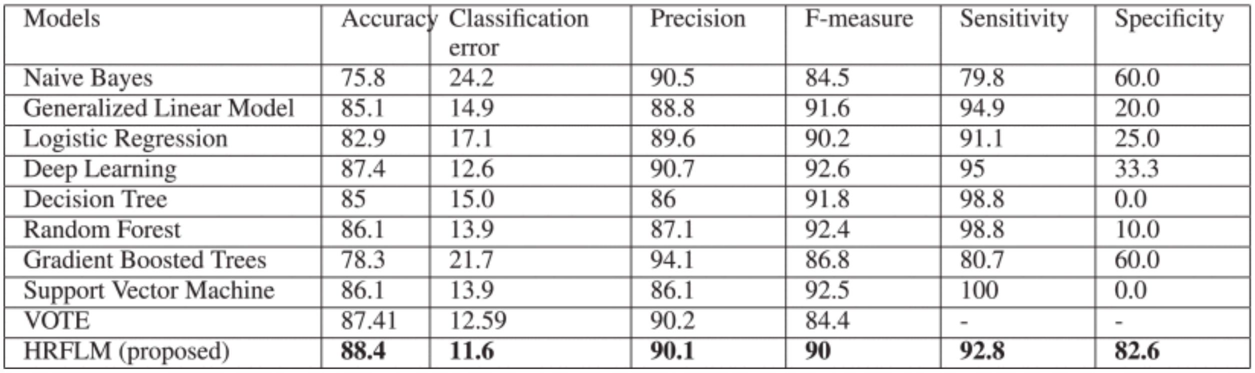
## 3. Literature Review

An early concentrate on the utilisation of ML techniques for heart disease prediction was done by (Shah et al. 2020). Shah et al. (2020)implemented KNN and used machine learning techniques to predict heart disease. They obtained the maximum accuracy using the KNN model, 90.789%. In their study, Singh and Kumar (2020) further delve into the use of ML methods in the conjecture of cardiovascular disease. According to their study, they attained an accuracy of 87%, which was the best the KNN model could deliver.

Additionally, Ramalingam et al. (2018) assessed several ML models for CVDs prediction. Their research offered a careful assessment of the models that are currently being involved and proposed possible areas for development. In this study, Ramalingam et al. discussed that with

SVM gives an accuracy of 98.9%, KNN gives an accuracy of 83.16%, and DT has the worst performance with an accuracy of 77.55%.

Furthermore, Mohan et al. (2019) presented hybrid ML procedures for more compelling heart disease predictions. Distributed in IEEE Access, their review investigated various algorithms to upgrade forecast exactness. The algorithms that showed high-performance evaluation were Hybrid Random Forest with Linear Model (HRFLM), VOTE, Deep Learnnnig, Support Vector Machine, Random Forest, and Generalized Linear Model, respectively; as shown below, the Models' Performance Evaluation, including Accuracy, Classification error, Precision, recall, F-Measure, Sensitivity, and Specificity were calculated for each model.



| Models | Accuracy | Classification error | Precision | F-measure | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Naive Bayes | 75.8 | 24.2 | 90.5 | 84.5 | 79.8 | 60.0 |
| Generalized Linear Model | 85.1 | 14.9 | 88.8 | 91.6 | 94.9 | 20.0 |
| Logistic Regression | 82.9 | 17.1 | 89.6 | 90.2 | 91.1 | 25.0 |
| Deep Learning | 87.4 | 12.6 | 90.7 | 92.6 | 95 | 33.3 |
| Decision Tree | 85 | 15.0 | 86 | 91.8 | 98.8 | 0.0 |
| Random Forest | 86.1 | 13.9 | 87.1 | 92.4 | 98.8 | 10.0 |
| Gradient Boosted Trees | 78.3 | 21.7 | 94.1 | 86.8 | 80.7 | 60.0 |
| Support Vector Machine | 86.1 | 13.9 | 86.1 | 92.5 | 100 | 0.0 |
| VOTE | 87.41 | 12.59 | 90.2 | 84.4 | - | - |
| HRFLM (proposed) | **88.4** | **11.6** | **90.1** | **90** | **92.8** | **82.6** |

Likewise, Jindal et al. (2021) utilized ML algorithms, as discussed in the IOP Conference Series, to foster a high-level prescient model for heart diseases; the best model was KNN with 88.52% accuracy. Furthermore, Sharma et al. (2020) examined methods influencing ML accuracy for precise predictions. Their experimental results demonstrate that SVM and Random Forest outperformed Gaussian Naive Bayes and Decision Tree. SVM models achieve 98% accuracy, which is 8% higher than Nave Bayes and about 13% higher than Decision Tree. Similarly, the Random Forest model produces the best prediction results with 99% accuracy, which is more accurate than our second-best SVM model for heart disease prediction.

Another study by Kavitha et al. (2021) followed different approaches and implemented various ML models for heart disease predictions. Their review was introduced at the sixth International Conference on Inventive Computation Technologies. By implementing the Hybrid model (random forest and decision tree combined), the experimental findings suggest that the heart disease prediction model using the hybrid model has an accuracy level of 88.7%.

## 4. Data Description and Analysis

### 4.1 Dataset Overview

The Cardiovascular Diseases Risk Prediction Dataset was collected in 2023 and obtained from Kaggle.com. This data was sourced from the Behavioral Risk Factor Surveillance System (BRFSS) by the World Health Organization (WHO). BRFSS is the first national telephone survey system to collect state-level data on health risk behaviours, chronic diseases and preventive service utilization. The dataset contains records related to personal lifestyle factors, encompassing a wide range of attributes that could influence cardiovascular health. Among the features included are demographic information such as age, gender (Sex), medical history (Diabetes), and self-reported health status (General Health). Additionally, the dataset comprises various other attributes related to lifestyle choices, habits, and health indicators. It serves as a collection of diverse data points aiming to capture various aspects of an individual's life and health that could contribute to the risk of developing cardiovascular diseases.

### 4.1.2 Dataset Details

- **Number of Variables: 19**
  - Categorical Variables: 7
  - Numerical Variables: 12
- **Number of Observations: 308,854**

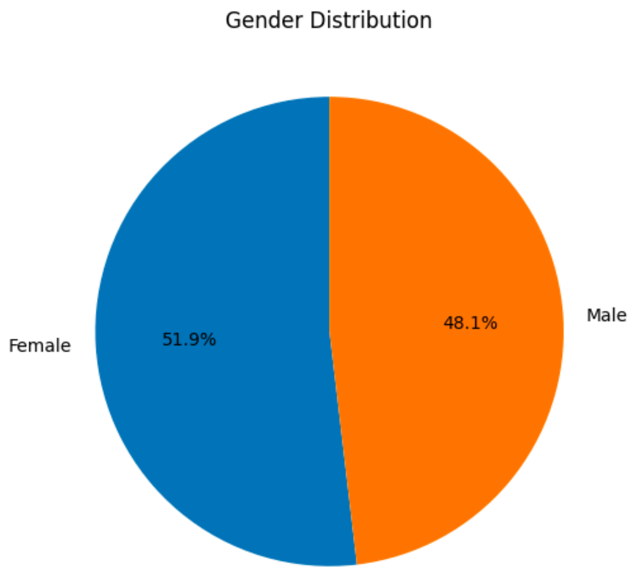
The variables and their descriptions are presented in the table below:

| Variable | Type | Description | Units of Measurement |
|---|---|---|---|
| General_Health | Qualitative (Categorical) | General health condition of an individual | Categories ('Poor' 'Very Good' 'Good' 'Fair' 'Excellent') |
| Checkup | Qualitative (Categorical) | Time since last medical checkup | Intervals ('Within the past 2 years' 'Within the past year' '5 or more years ago' 'Within the past 5 years' 'Never') |
| Exercise | Qualitative (Categorical) | For past month, other than regular job, participation in any physical activities or exercises | Binary ('Yes' or 'No') |
| Heart_Disease (target variable) | Qualitative (Categorical) | Indicates presence of heart disease | Binary ('Yes' or 'No') |
| Skin_Cancer, Other_Cancer, Depression, Arthritis | Qualitative (Categorical) | Presence or absence of corresponding condition | Binary ('Yes' or 'No') |
| Diabetes | Qualitative (Categorical) | Respondents that reported having a diabetes. If yes, what type of diabetes it is/was. | Categories ('Yes'  'No' 'Yes, but female told only during pregnancy' or 'No, pre-diabetes or borderline diabetes') |
| Sex | Qualitative (Categorical) | Gender of the individual | Categories ('Male' or 'Female') |
| Age_Category | Qualitative (Categorical) | Categorization of individuals into age groups | Categories ('70-74' '60-64' '75-79' '80+' '65-69' '50-54' '45-49' '18-24' '30-34' '55-59' '35-39' '40-44' '25-29') |
| Height_(cm) | Quantitative (Numeric) | Height of the individual in centimetres | Centimetres |

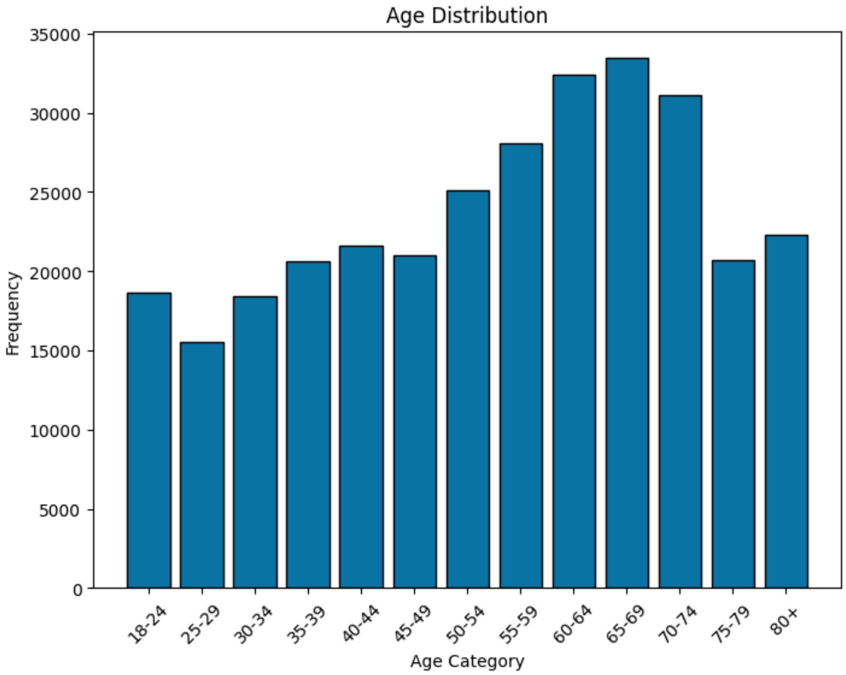
| | | | |
|---|---|---|---|
| Weight_(kg) | Quantitative (Numeric) | Weight of the individual in kilograms | Kilograms |
| BMI | Quantitative (Numeric) | Body Mass Index gives an indication of whether an individual's weight is healthy for their height | Derived value from weight and height |
| Smoking_History | Qualitative (Categorical) | Description of smoking habits of the individual | Binary ('Yes' or 'No') |
| Alcohol_Consumption | Quantitative (Numeric) | Average units of alcohol consumed per week | Units per week |
| Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption | Quantitative (Numeric) | Number of consumption in a month | Units per month Example: a 128 response would be a representative of a person who consumes vegetables at least 4-5 times a day |

## 4.2 Data Analysis and Visualization

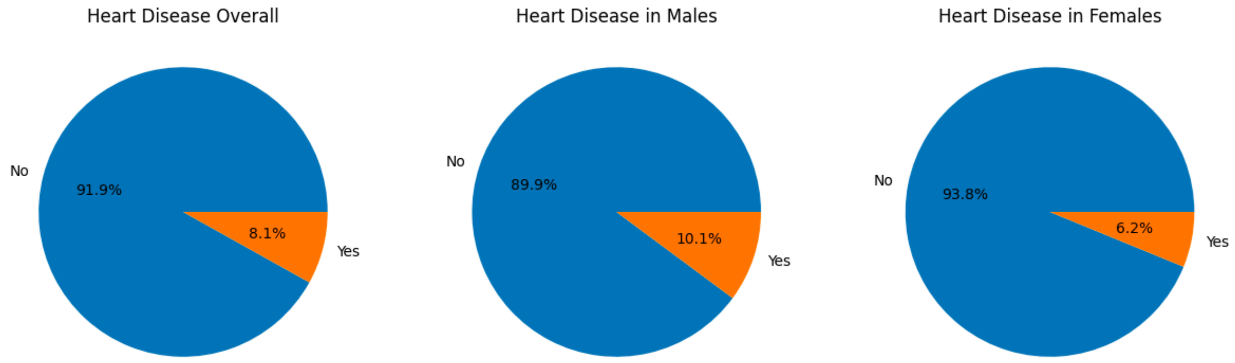
### Analysis of Different Features



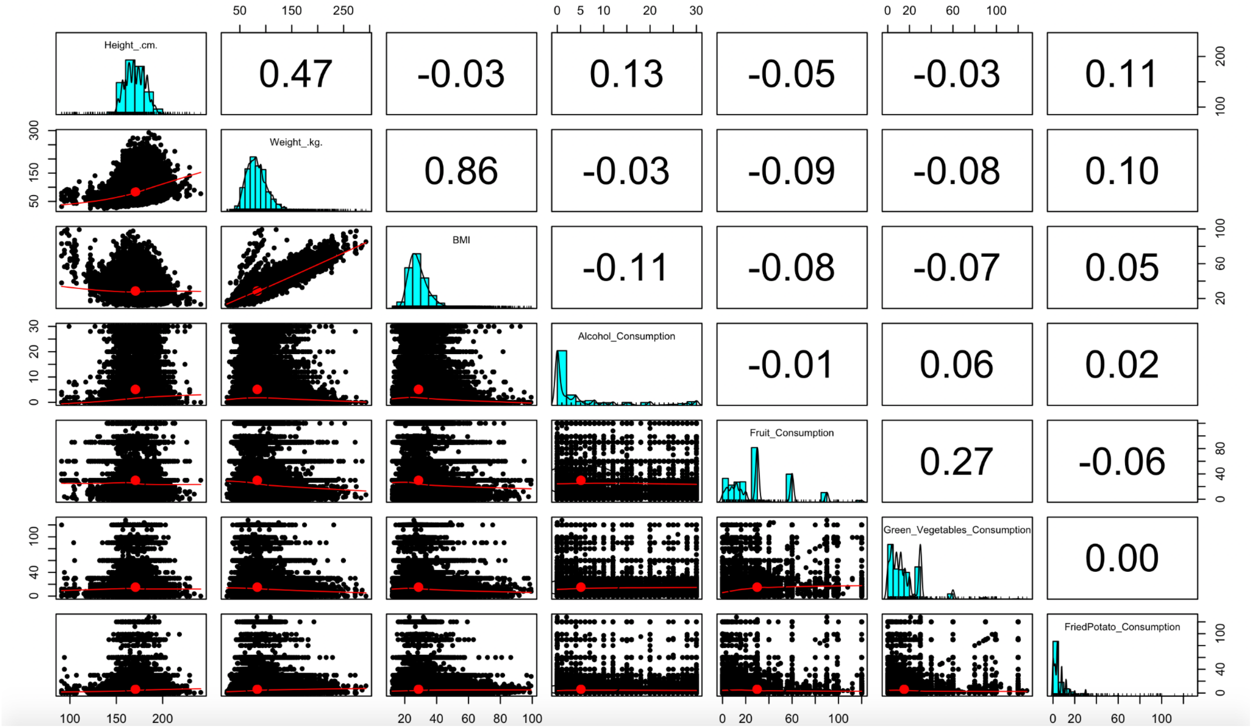
For gender,  there is a balance between male and female, in which 51.9% are female and 48.1 are male.



As shown in the age distribution histogram, a significant proportion of the records are from ages 50 to 74. To illustrate this, between the age categories of 18 to  34, there are around 55,000 registrations. Between the age categories of 60 and 74, there are approximately 95,000 records.



As shown in the pie charts above, the first pie chart represents heart disease, which is the target variable; 25,017  patients have heart disease, which represents 8.1% of the data, and 283,837 patients who do not have heart disease, which represents 91.9% of the data. This shows that there is a class imbalance in the target variable. As shown in the second and third pie charts, the heart disease percentage of patients by gender, we can also see that men, which represents 10.1%, are more affected than women, which represents 6.2%.



As shown in the figure above, there are some variales that are highly correlated. For example weight and BMI with correlation coefficient 0.86 which shows that there is a strong correlation. Another correlation is between weight and height with correlation coefficient 0.47 which shows that there is a positive moderate correlation. While other variables have weak correraltions. It is also shown fromthe density plots that some variales are normally distributed.

## 5. Machine Learning and Data Mining Techniques

Early detection of CVD is crucial, allowing for timely intervention through counselling and medication. Machine learning, with its ability to analyze big datasets and discern complex patterns, has emerged as a promising tool for enhancing the accuracy and efficiency of heart disease prediction models. However, a research gap persists in the comparative analysis of machine learning models, hyperparameter tuning influence, and identifying crucial personal attributes in CVD risk prediction. To address these gaps, this study evaluates multiple machine learning models, including **Logistic Regression (LR), Decision Tree Classifier (DT), and Random Forest (RF).**

### 5.1 Supervised Learning Approaches

Supervised learning is a type of machine learning where the algorithm learns from a labeled dataset, which means it is provided with input-output pairs during the training process. In supervised learning, the algorithm aims to establish a mapping between the input data and corresponding output labels, allowing it to make predictions or classifications when presented with new, unseen data. The term "supervised" refers to the process of guiding the algorithm's learning by providing it with labeled examples, essentially showing it the correct answers.

#### 5.1.1 Data Splitting in Supervised Learning

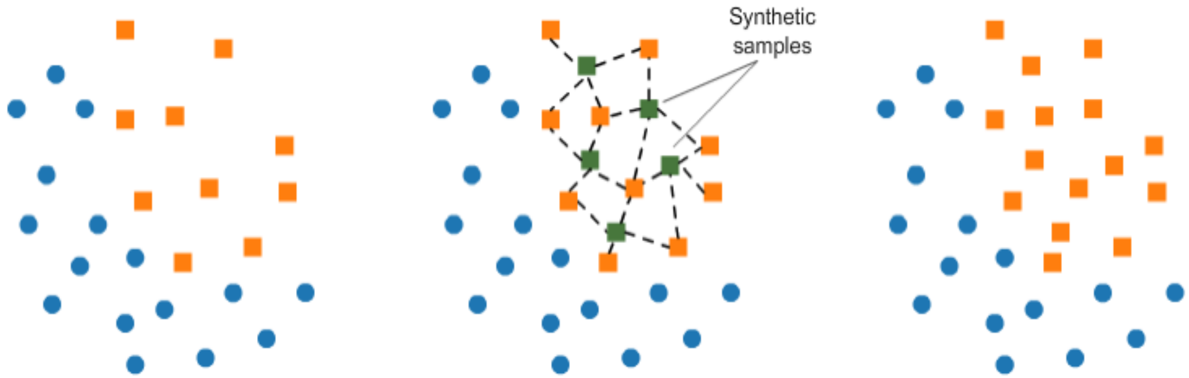
In supervised learning, evaluating the performance of the trained model on data it has never seen before is crucial. To achieve this, the labeled dataset is typically divided into two subsets: the training and test sets. The training set is used to train the supervised learning algorithm. The model learns the underlying patterns and relationships between input features and output labels by adjusting its parameters based on the labeled examples in the training set. On the other hand, the test set is kept separate from the training set and is not used during the training phase. After the model has been trained, it is evaluated on the test set to assess its ability to generalize to new, unseen data. This evaluation provides insights into the model's performance and helps estimate how well it will perform on the unseen data. The data splitting usually follows:  80% of the observations are for training, and the other 20% to test the model performance. The model performance is tested on 20% of the observations for each model. The metrics of accuracy, precision, recall, and F1 score are used to evaluate the performance of each model.

### 5.2 Model Training and Validation

#### 5.2.1 Handling Class Imbalance Techniques:

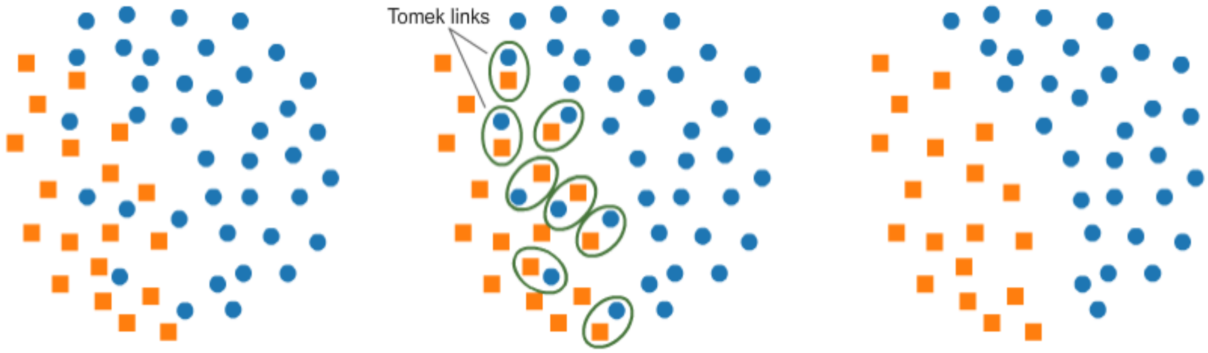
In this section, class imbalance issues in the dataset were addressed through the implementation of various resampling techniques. These techniques included:

1. **SMOTE (Synthetic Minority Over-sampling Technique)**



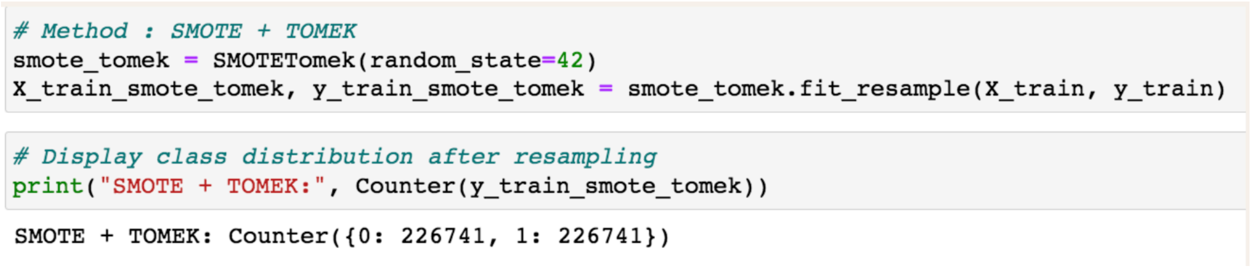
**SMOTE (Synthetic Minority Over-sampling Technique)** method as shown in the figure above involved generating synthetic samples for the minority class to rebalance the class distribution. SMOTE is a widely-used technique for addressing class imbalance. It works by generating synthetic samples for the minority class by interpolating between existing minority class instances. This approach helps to rebalance the class distribution by increasing the number of minority class samples. The newly generated synthetic samples are created along the line segments joining k minority class nearest neighbors in the feature space.

2. **TOMEK Links**



As shown in the figure above, TOMEK Links is a technique used for undersampling the majority class by removing samples that form TOMEK links with minority class samples. A TOMEK link exists between two samples of different classes if they are nearest neighbors of each other and cannot be classified correctly. By removing these majority class samples that are close to minority class samples, TOMEK Links aim to improve the separation between classes and alleviate the effects of class imbalance.

3. **SMOTE + TOMEK**



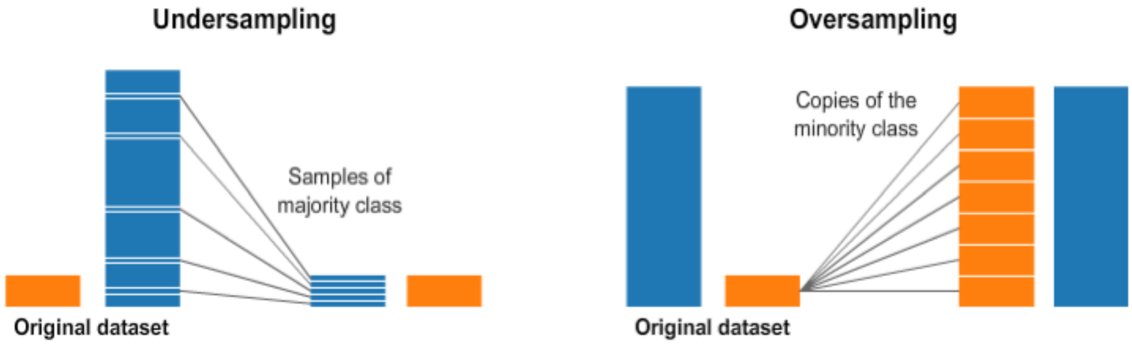
```
# Method : SMOTE + TOMEK
smote_tomek = SMOTETomek(random_state=42)
X_train_smote_tomek, y_train_smote_tomek = smote_tomek.fit_resample(X_train, y_train)
```

```
# Display class distribution after resampling
print("SMOTE + TOMEK:", Counter(y_train_smote_tomek))
```

```
SMOTE + TOMEK: Counter({0: 226741, 1: 226741})
```

**SMOTE + TOMEK** is a combination of SMOTE over-sampling and TOMEK under-sampling techniques was employed to achieve a balanced class distribution as shown in the figure above. First, SMOTE is used to generate synthetic samples for the minority class, thus increasing its representation in the dataset. Then, TOMEK Links are applied to remove samples that form links between the minority and majority classes. This combined approach aims to achieve a balanced class distribution while also improving the separability between classes.

4. **Random Over-sampling and Random Under-sampling**



**Random Over-sampling:** This technique randomly replicated minority class samples to balance the class distribution. This method increases the number of minority class instances by duplicating existing samples. While simple to implement, random over-sampling may lead to overfitting if not used cautiously, especially with small datasets.

**Random Under-sampling:** This technique addresses class imbalance by randomly removing samples from the majority class to balance the class distribution. This method reduces the number of majority class instances in the dataset. While it can effectively balance class proportions, random under-sampling may lead to information loss if important samples are removed.

**The Class Distribution:**

**Class distribution before resampling**: Counter({0: 227087, 1: 19929})

**Class distribution after resampling:**

**Random Over-sampling**: Counter({0: 283800, 1: 283800})

**Random Under-sampling**: Counter({0: 24971, 1: 24971})

**SMOTE:** Counter({0: 283800, 1: 283800})

**TOMEK:** Counter({0: 274278, 1: 24971})

**SMOTE + TOMEK:** Counter({0: 282212, 1: 282212})

### 5.2.2 Implemented ML Models

1. **Logistic Regression (LR)**

Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The parameters are L2 regularization and C. A higher C means less regularization strength, which can easily lead to overfitting. While a lower C means higher regularization strength, which means that the fitted line is robust to outliers and will not change much by the highly influential observations.

2. **Decision Tree (DT)**

The Decision Tree algorithm is represented as a flowchart, where inner nodes depict dataset attributes, and outer branches signify potential outcomes. Decision Trees are preferred for their speed, reliability, ease of interpretation, and minimal data preparation requirements. In the decision tree structure, the prediction of a class label begins at the root, comparing the root attribute's value to the record's attribute. Depending on the comparison result, the algorithm follows the corresponding branch, advancing to the next node.

### 3. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. Random Forest combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. The algorithm has the ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning. In addition, Random Forest can handle the data sets containing continuous variables, in the case of regression, and categorical variables, in the case of classification.

### 5.2.3 Cross-Validation

Stratified k-fold cross-validation was utilized to assess the models. This approach ensured that each fold of the dataset retained the same class distribution as the original dataset, thus mitigating the impact of class imbalance.

### 5.2.4 Grid Search CV for Hyperparameter Tuning

To optimize the performance of each model, grid search cross-validation was employed. This method systematically explored a specified parameter grid to identify the best combination of hyperparameters. Hyperparameters explored for each model included regularization strength (C) for Logistic Regression, maximum depth for Decision Tree, number of estimators and, number of neighbors for K-Nearest Neighbor, and C and gamma for SVM.

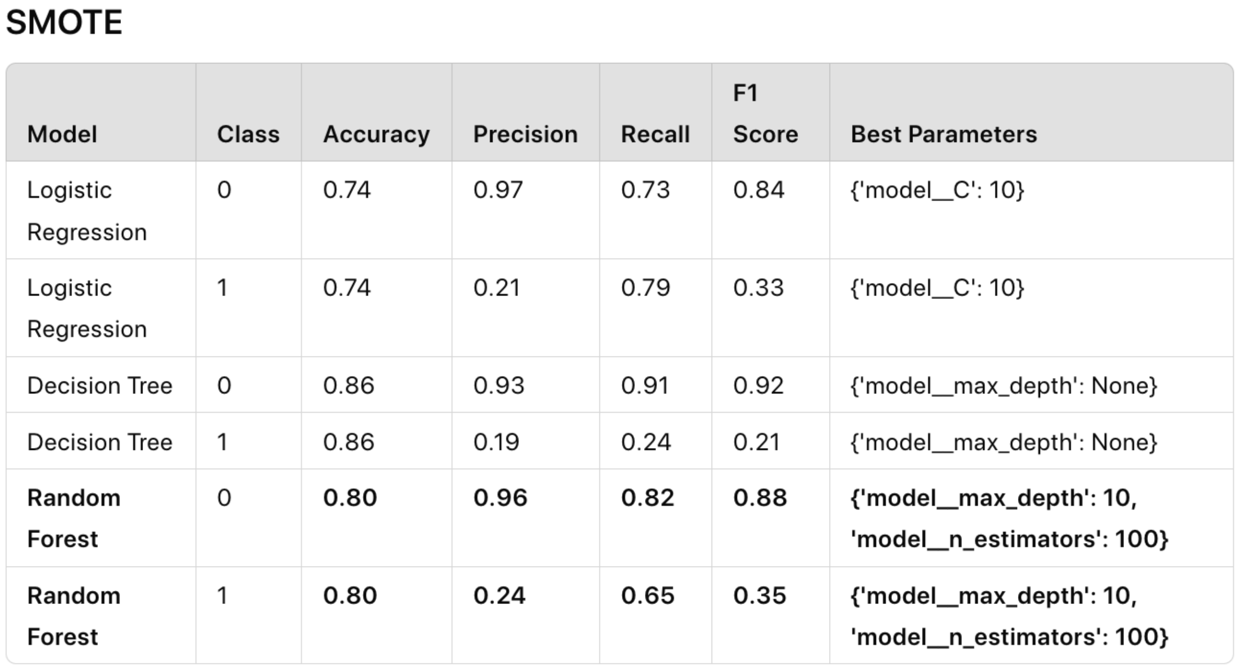
#### 5.2.5 Model Evaluation

The performance of each model with 5 different imbalance handling techniques was evaluated using the following metrics:

- Classification Report: This report provided precision, recall, F1-score, and support for each class, as well as average scores across all classes.
- Confusion Matrix: The confusion matrix was visualized for each model, depicting true positive, false positive, true negative, and false negative values. This facilitated an assessment of the models' classification performance for each class.

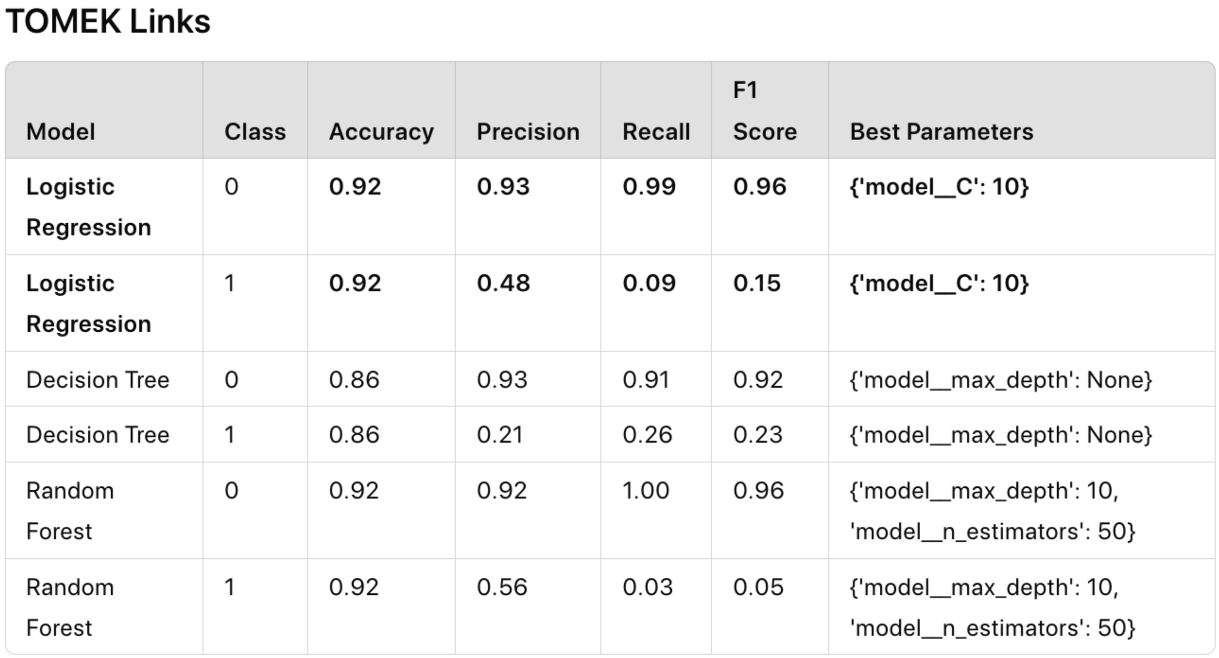
By employing these techniques and conducting comprehensive model evaluations, the aim was to effectively address class imbalance challenges and identify the most suitable model for the dataset under consideration.

## 6. Results



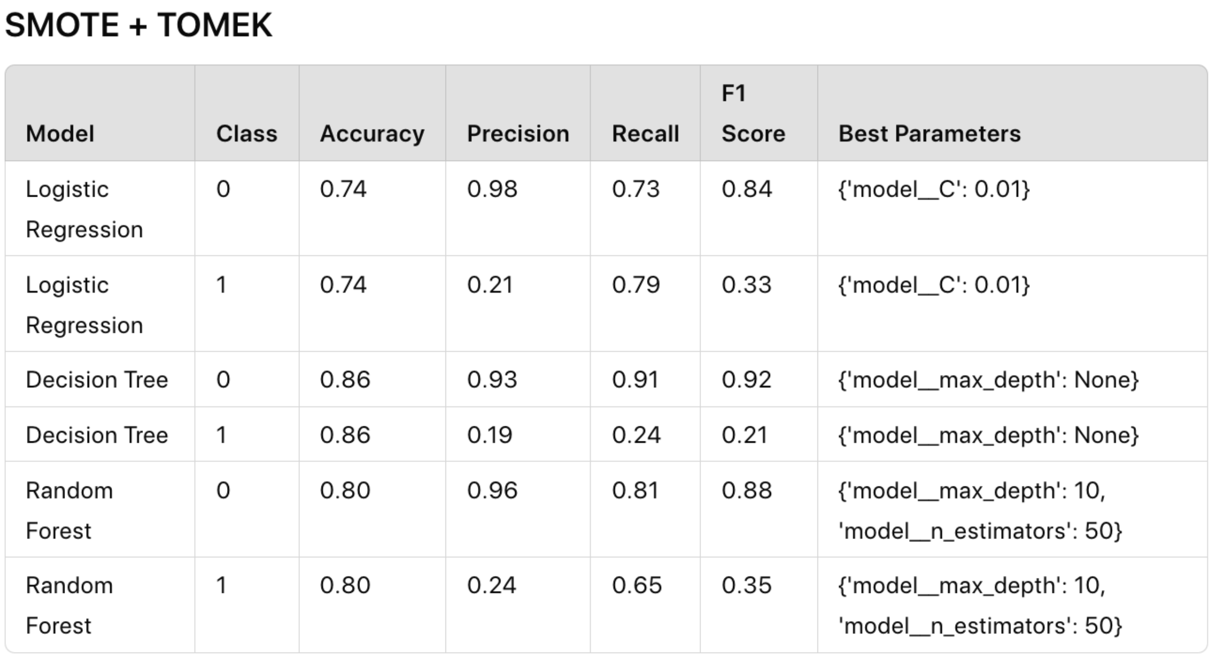
### SMOTE

| Model | Class | Accuracy | Precision | Recall | F1 Score | Best Parameters |
|---|---|---|---|---|---|---|
| Logistic Regression | 0 | 0.74 | 0.97 | 0.73 | 0.84 | {'model__C': 10} |
| Logistic Regression | 1 | 0.74 | 0.21 | 0.79 | 0.33 | {'model__C': 10} |
| Decision Tree | 0 | 0.86 | 0.93 | 0.91 | 0.92 | {'model__max_depth': None} |
| Decision Tree | 1 | 0.86 | 0.19 | 0.24 | 0.21 | {'model__max_depth': None} |
| **Random Forest** | 0 | **0.80** | **0.96** | **0.82** | **0.88** | **{'model__max_depth': 10, 'model__n_estimators': 100}** |
| **Random Forest** | 1 | **0.80** | **0.24** | **0.65** | **0.35** | **{'model__max_depth': 10, 'model__n_estimators': 100}** |



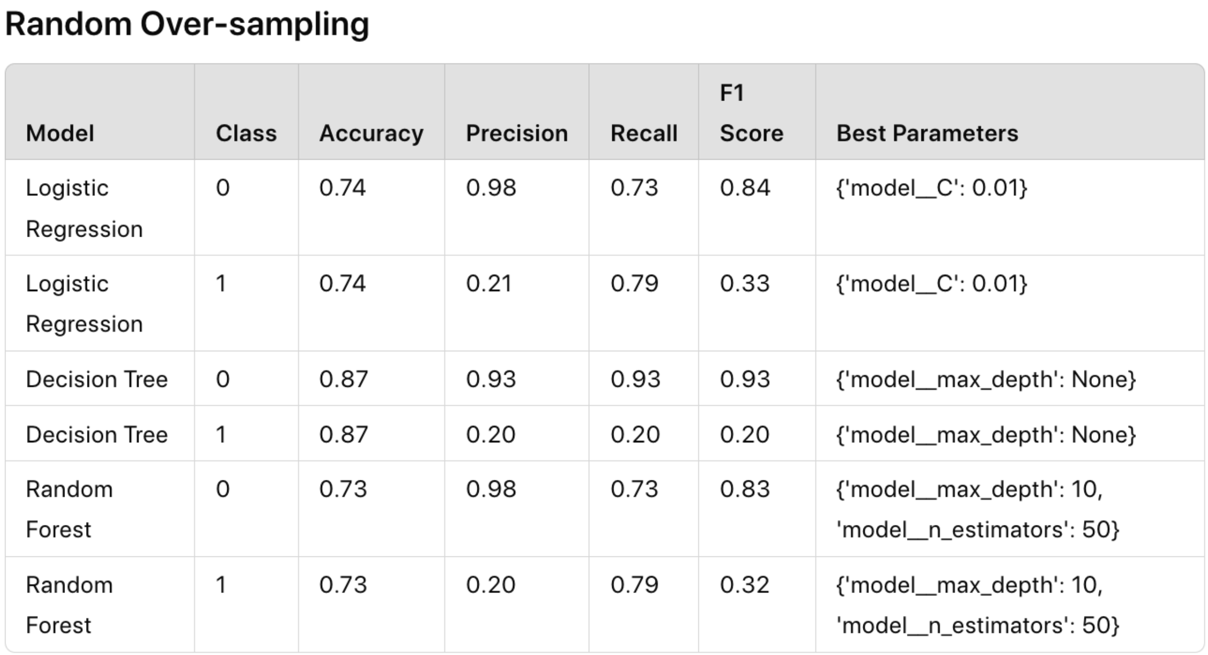
### TOMEK Links

| Model | Class | Accuracy | Precision | Recall | F1 Score | Best Parameters |
|---|---|---|---|---|---|---|
| **Logistic Regression** | 0 | **0.92** | **0.93** | **0.99** | **0.96** | **{'model__C': 10}** |
| **Logistic Regression** | 1 | **0.92** | **0.48** | **0.09** | **0.15** | **{'model__C': 10}** |
| Decision Tree | 0 | 0.86 | 0.93 | 0.91 | 0.92 | {'model__max_depth': None} |
| Decision Tree | 1 | 0.86 | 0.21 | 0.26 | 0.23 | {'model__max_depth': None} |
| Random Forest | 0 | 0.92 | 0.92 | 1.00 | 0.96 | {'model__max_depth': 10, 'model__n_estimators': 50} |
| Random Forest | 1 | 0.92 | 0.56 | 0.03 | 0.05 | {'model__max_depth': 10, 'model__n_estimators': 50} |



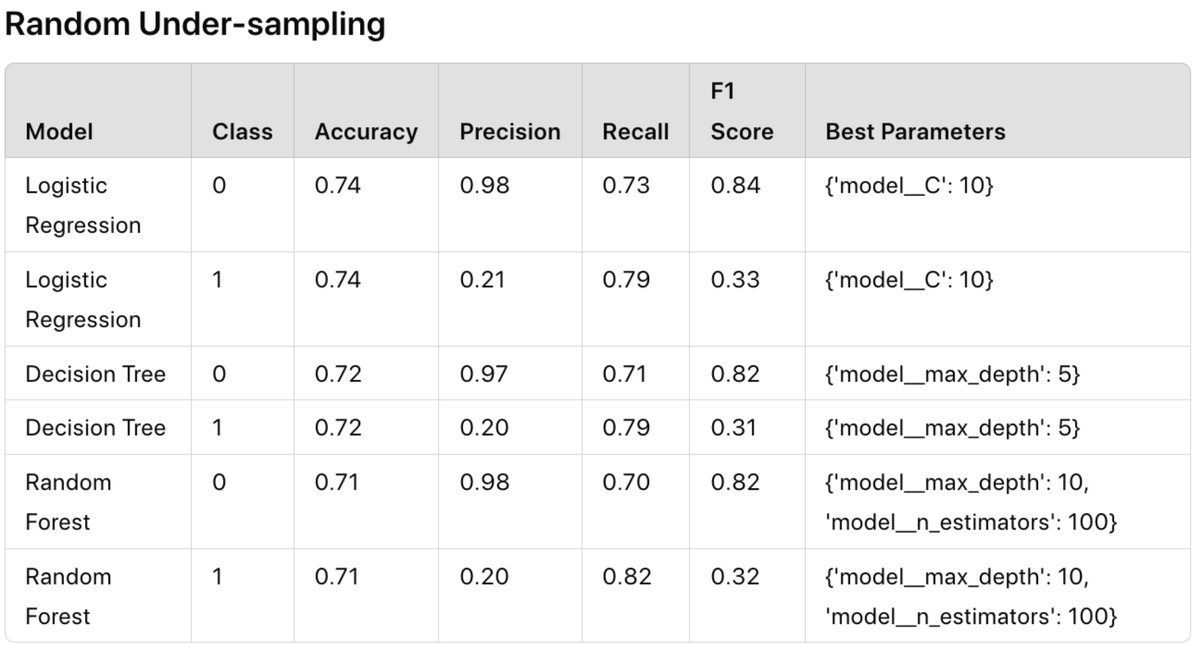
## SMOTE + TOMEK

| Model | Class | Accuracy | Precision | Recall | F1 Score | Best Parameters |
|---|---|---|---|---|---|---|
| Logistic Regression | 0 | 0.74 | 0.98 | 0.73 | 0.84 | {'model__C': 0.01} |
| Logistic Regression | 1 | 0.74 | 0.21 | 0.79 | 0.33 | {'model__C': 0.01} |
| Decision Tree | 0 | 0.86 | 0.93 | 0.91 | 0.92 | {'model__max_depth': None} |
| Decision Tree | 1 | 0.86 | 0.19 | 0.24 | 0.21 | {'model__max_depth': None} |
| Random Forest | 0 | 0.80 | 0.96 | 0.81 | 0.88 | {'model__max_depth': 10, 'model__n_estimators': 50} |
| Random Forest | 1 | 0.80 | 0.24 | 0.65 | 0.35 | {'model__max_depth': 10, 'model__n_estimators': 50} |



## Random Over-sampling

| Model | Class | Accuracy | Precision | Recall | F1 Score | Best Parameters |
|---|---|---|---|---|---|---|
| Logistic Regression | 0 | 0.74 | 0.98 | 0.73 | 0.84 | {'model__C': 0.01} |
| Logistic Regression | 1 | 0.74 | 0.21 | 0.79 | 0.33 | {'model__C': 0.01} |
| Decision Tree | 0 | 0.87 | 0.93 | 0.93 | 0.93 | {'model__max_depth': None} |
| Decision Tree | 1 | 0.87 | 0.20 | 0.20 | 0.20 | {'model__max_depth': None} |
| Random Forest | 0 | 0.73 | 0.98 | 0.73 | 0.83 | {'model__max_depth': 10, 'model__n_estimators': 50} |
| Random Forest | 1 | 0.73 | 0.20 | 0.79 | 0.32 | {'model__max_depth': 10, 'model__n_estimators': 50} |



### Random Under-sampling

| Model | Class | Accuracy | Precision | Recall | F1 Score | Best Parameters |
|---|---|---|---|---|---|---|
| Logistic Regression | 0 | 0.74 | 0.98 | 0.73 | 0.84 | {'model__C': 10} |
| Logistic Regression | 1 | 0.74 | 0.21 | 0.79 | 0.33 | {'model__C': 10} |
| Decision Tree | 0 | 0.72 | 0.97 | 0.71 | 0.82 | {'model__max_depth': 5} |
| Decision Tree | 1 | 0.72 | 0.20 | 0.79 | 0.31 | {'model__max_depth': 5} |
| Random Forest | 0 | 0.71 | 0.98 | 0.70 | 0.82 | {'model__max_depth': 10, 'model__n_estimators': 100} |
| Random Forest | 1 | 0.71 | 0.20 | 0.82 | 0.32 | {'model__max_depth': 10, 'model__n_estimators': 100} |

From the results shown above and after apply 5 imbalance handling techniques, it can be concluded for Class 0 (Majority Class), the highest Precision and F1 Score was TOMEK Links with Logistic Regression shows high precision (0.93) and F1 score (0.96) and Random Over-sampling with Decision Tree and Random Forest also show high precision (0.93 and 0.98, respectively) and good F1 scores (0.93 and 0.83, respectively). On the other hand, Class 1 (Minority Class) the highest Precision and F1 Score was SMOTE and SMOTE + TOMEK with Random Forest show better balance in precision (0.24) and higher F1 scores (0.35) and Random Over-sampling with Logistic Regression and Random Forest shows reasonable recall (0.79) and F1 scores (0.33 and 0.32, respectively).

Therefore best model for class 1 is Random Forest with SMOTE and SMOTE + TOMEK techniques, given their higher F1 scores. While best Imbalance Technique is SMOTE and SMOTE + TOMEK techniques generally perform better in balancing precision and recall, leading to higher F1 scores, particularly for the minority class. The metrics used was F1 Score which is the most informative metric as it balances precision and recall, especially important for the minority class in imbalanced datasets. Recall is crucial to ensure that the minority class is detected correctly, hence also important for evaluating models in this context.

## 7. Conclusion and Recommendation

As machine learning models evolve and gain prominence, there is a growing opportunity to address cardiovascular health on a broader scale, potentially alleviating the global burden of heart diseases. This study encompassed various methods, machine learning models, and key findings from several real studies, showcasing the significant approaches in heart disease prediction using machine learning. While the study discussed the diverse methodologies employed to enhance predictive accuracy, it becomes evident that a notable gap exists in the current research landscape. This gap signals the need for further exploration and investigation to advance our understanding and improve the effectiveness of machine learning models in predicting heart diseases. The aggregate discoveries highlight the capability of ML to change medical services, especially in the domain of cardiovascular diseases. As technology advances and datasets expand, the synergistic integration of artificial intelligence (AI) and machine learning holds promising opportunities for additional research to refine the accuracy and reliability of machine learning models in the prediction of heart diseases. For future work, the analysis would be applied without the outliers and compare the two analysis. Another recommendation is to implement more ML models and test their performance.

## References

Gandhi, Rohith. "Support Vector Machine - Introduction to Machine Learning Algorithms." Medium, Towards Data Science, 5 July 2018, https://towardsdatascience.com/support-vectormachine-introduction-to-machine-learning-algorithms-934a444fca47.

Jindal, H., Agrawal, S., Khera, R., Jain, R. & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. & Suraj, R.S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333).

Mohan, S., Thirumalai, C. & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, pp.81542-81554.

Ramalingam, V.V., Dandapath, A. & Raja, M.K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), pp.684-687.

Shah, D., Patel, S. & Bharti, S.K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, pp.1-6.

Sharma, V., Yadav, S. & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)* (pp. 177-181).

Singh, A. & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457).

Webtunix Solutions - Business Solution Provider. “K-Nearest Neighbors Classifier.” KNearest Neighbors Classifier, https://www.ris-ai.com/k-nearest-neighbors-classification.

World Health Organization. (2021).Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

World Heart Federation. (2023). Retrieved from https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.

Yadav, A.L., Soni, K. & Khare, S. (2023, July). Heart Diseases Prediction using Machine Learning. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7).

Alphiree. (2023). *Cardiovascular diseases risk prediction dataset*. Retrieved from https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data

Mohan, S., Thirumalai, C. & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, pp.81542-81554.

Ahmed, Intisar, "A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING AND DATA MINING" (2022). Electronic Theses, Projects, and Dissertations. 1591. https://scholarworks.lib.csusb.edu/etd/1591

Khetan, N., Wells, Q. S., & Kraus, W. E. (2019). Pathobiology of obesity. In K. R. Feingold, B. Anawalt, A. Boyce, G. Chrousos, K. Dungan, A. Grossman, ... & D. P. Wilson (Eds.), Endotext. MDText.com, Inc. https://www.ncbi.nlm.nih.gov/books/NBK541070/

Banerjee, A. (2020, July 9). 10 techniques to deal with class imbalance in machine learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/