Open Peer Review on Qeios

Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis

Moshood A. Hambali[1], Ishaku Andrew[1]

1 Federal University Wukari

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

In today's market, customers have a plethora of options available to them when deciding where to invest their money. Consequently, customer churn and engagement have emerged as prominent concerns. With an increasing number of service providers targeting the same customer base, it is imperative for providers to understand evolving customer behavior and heightened expectations to retain their clientele. Numerous studies have addressed the issue of customer churn, with data mining frequently employed to predict bank customer attrition. While many researchers have proposed various approaches for predicting customer churn, some machine learning (ML) algorithms have struggled to deliver the required performance in identifying customer churn accurately most especially when the dataset is imbalance data. Therefore, this paper presents an application of Synthetic Minority Over Sampling Technique (SMOTE) on bank churn dataset. The SMOTE algorithm was employed to address the problem of data imbalance and Genetic Algorithm (GA) was applied to select most informative features from the original dataset. The selective features were evaluate using four (4) different classification algorithms: Random Forest (RF), K-Nearnest Neighbor (KNN), Artificial Neural Network (ANN) and Adaboost algorithms. The KNN model demonstrated superior performance compared to other models in terms of accuracy (96%), precision (96%), and F-measure (96%) respectively. Furthermore, we compared our results with existing models that utilized the same dataset, and our proposed strategy outperformed them.

Moshood A. Hambali*, and **Ishaku Andrew**

Federal University Wukari, PMB 1020, Katsina Ala Road, Wukari, Nigeria

*Corresponding Author: hamberlite@gmail.com

1. Introduction

In the current landscape, the market is characterised by its dynamic nature and intense competition, driven by the proliferation of service providers. These providers face the challenge of adapting to shifting customer behaviours and escalating expectations. In the realm of business, customers are considered the most valuable asset. The aspirations of today's consumers, along with their diverse demands for connectivity and innovative, personalised approaches, mark a significant departure from previous generations. These consumers are well-educated and well-informed about emerging trends, which has led to a trend of over-analysis in their purchasing decisions, ultimately aiding in more informed choices (Saheed & Hambali, 2021; Tan, 2023).

Consequently, this presents a considerable challenge for newer service providers to devise innovative strategies to meet and exceed customer expectations. Customer churn, a phenomenon rooted in dissatisfaction, is observed across various industries including insurance, banking, and telecommunications. To address this, companies employ Customer Relationship Management (CRM) models to foster long-term relationships and prevent customers from switching to competitors(Saheed & Hambali, 2021).

With the escalating competition and diverse product offerings in the industrial market, many companies turn to data mining techniques to gauge customer churn rates. Data mining, aimed at uncovering significant patterns within large datasets, offers various techniques such as estimation, classification, association, and clustering. Classification, particularly crucial in data mining, serves to predict data classes. However, the challenge of class imbalance arises, especially in industries like banking where churn rates are low. The Synthetic Minority Over Sampling Technique (SMOTE) is often employed to address this issue without compromising data integrity (Hambali & Gbolagade, 2016).

Moreover, classification accuracy on high-dimensional data often suffers due to the abundance of attributes. Attribute selection methods, such as genetic algorithms, are employed to enhance accuracy by reducing attribute complexity while retaining pertinent information. Genetic algorithms, known for their ability to streamline attributes in high-dimensional data, operate based on an evolutionary process to find optimal solutions.

2. Review of Related Works

Leung & Chung, (2020) introduced a novel technique to refine model specifications by incorporating time-series predictors, utilizing data spanning diverse timeframes, and identifying rare occurrences, all aimed at enhancing the accuracy of churn prediction. The study employed a unique dataset comprising three years' worth of records, totalling 32,000 transactions from a retail bank based in Florida, USA. The methodology involved trend modelling to capture the evolution of customer behaviour over time. The results indicated that integrating data from multiple timeframes resulted in improvements in both model precision and recall. Specifically, for six months, the accuracy was 94.78%, precision was 24.8%, and recall stood at 31.14%. Meanwhile, for four months, the accuracy was 90.81%, precision was 13.89%, and recall was 30.32%. Moreover, this dynamic approach to churn prediction can be extended to other sectors requiring long-term customer data analysis. However, one limitation of this approach is the potential issue of non-independence stemming from multiple observations of the same individuals. As the training data is gathered from various periods, data from the same customers is repeatedly included. Despite potential variations in customer behaviour across different timeframes, the static predictors remain constant, which may challenge the assumption of independence.

Rahman & Kumar, (2020) utilized machine learning (ML) techniques to estimate turnover among bank clients. By analyzing client behaviour, researchers aimed to forecast churn. Employed classifiers included K-nearest neighbour (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. Several feature selection strategies were implemented to identify crucial characteristics and evaluate system performance. The Kaggle churn modelling dataset served as the basis for analysis. Comparative analysis of the outcomes facilitated the identification of a precise and reliable model. Random Forest exhibited superior accuracy compared to other models post-oversampling. Results demonstrated that combining the Random Forest classifier with oversampling yielded superior outcomes, achieving an accuracy of 95.74%. It is noteworthy that feature selection techniques did not significantly impact tree classifiers like Decision Trees and Random Forests. As evidenced by the results, reducing features (via feature selection) tended to diminish the predictive performance of tree classifiers.

Domingos et al., (2021) investigated the impact of various hyperparameters in the application of Deep Neural Networks (DNNs) for churn prediction within the banking sector. The findings from three experiments revealed that the DNN model outperformed the Multilayer Perceptron (MLP) particularly when the rectifier activation function was used in the hidden layers, and the sigmoid function was employed in the output layer. Furthermore, the DNN demonstrated enhanced performance with a smaller batch size relative to the test dataset size, while the root mean square propagation (RMSprop) training algorithm exhibited superior accuracy compared to alternatives such as stochastic gradient descent (SGD), adaptive gradient algorithm (AdaGrad), Adaptive Movement Estimation (Adam), Adadelta, and AdaMax. Results indicated that the MLP achieved its highest performance accuracy (84.5%) when trained using the RMSProp algorithm, whereas its lowest performance (79.65%) was observed with the use of AdaGrad. Similarly, for the DNN, the highest performance (86.45%) was achieved with RMSProp as the training algorithm, while the lowest performance (83.1%) was recorded with SGD. However, a limitation of the study was the exclusive use of a fabricated dataset obtained from a public repository, possibly originating from a single bank over a limited timeframe. Therefore, caution should be exercised when generalizing the findings to other banking institutions. Furthermore, the dataset exhibited an imbalance in distribution, with 2000 churners and 8000 non-churners, potentially affecting the accuracy of machine learning classifiers in predicting outcomes despite the application of stratified cross-validation techniques to maintain a proportional representation of each category.

Tékouabou et al., (2022) outlined a comprehensive procedure for constructing a Machine Learning (ML) model, incorporating cross-validation of the Synthetic Minority Oversampling Technique (SMOTE) to rectify data imbalances and ensemble modelling. With balanced data, the random forest (RF) model achieved an accuracy of 0.86 and an f1-score of 0.86. Notably, "Age" emerged as the most crucial attribute in both the developed and optimized models, whereas "HasCrCard" held the least significance. The authors suggested utilizing the developed model to inform bank customer loyalty decisions. They acknowledged that classification algorithms faced challenges arising from factors such as data volume limitations, data diversity, non-numeric data, and class imbalances.

Mirabdolbaghi & Amiri, (2022) employed Principal Component Analysis (PCA), Autoencoders, Latent Dirichlet Allocation (LDA), T-Stochastic Neighbor Embedding (T-SNE), and Xgboost for Bank churn prediction. They introduced a structured model aimed at forecasting churn in Gradient Boosting Machine (GBM) applications, encompassing five distinct phases.

Initially, data preprocessing was conducted to address missing and corrupted values, ensuring clean and appropriately scaled input data. Subsequently, feature selection techniques were employed to reduce data dimensionality and retain relevant features. The third step involved fine-tuning the hyperparameters of the Light GBM model using Bayesian and genetic optimization methods to optimize performance. Post-training, an interpretability phase analyzed the model's behaviour and predictions, leveraging Shapley additive explanation (SHAP) to understand feature influence on model outputs. Finally, the model was applied to rank potential churners based on estimated customer lifetime value, aiding businesses in prioritizing retention efforts. This comprehensive approach aimed to bolster the efficacy of GBM churn prediction and offer actionable insights for retention strategy decisions. Evaluation using four established datasets showcased its superiority over Adaptive Boosting (AdaBoost), SVM, and decision tree algorithms across seven evaluation metrics, including accuracy, Area under the ROC Curve (AUC), Kappa, Matthews Correlation Coefficient (MCC), Brier score, F1 score, and Economic Model Predictive Control (EMPC). Their algorithm demonstrated superior handling of imbalanced datasets in churn prediction based on assessment metrics.

3. Research Methodology

For the evaluation of bank customer churns, it is essential to devise a methodology. Hence, the proposed research methodology consists of three main phases: data preprocessing, classification model applications, and performance evaluation. In detail, the data preprocessing phase is targeted at preparing and cleaning the customer churn dataset collected from the Kaggle Machine Learning repositories for effective analysis. Essentially, at the preprocessing stage tasks such as handling missing values, normalizing data, and removing noise or outliers were carried out. It is important to note that a data balancing technique using the SMOTE was applied to address the imbalanced data distribution of the Bank churn dataset. Also, GA was applied to select informative features from available features in the original dataset. Upon the cleansing of the dataset, the cleaned dataset is split into train and test proportions resulting in the next phase called the classification models application. The classification phase deals with the application of four distinct machine-learning algorithms including ANN, AdaBoost, KNN, and Random Forest to train and learn from the training dataset. The result of this phase is a classification model. In the last phase identified as the performance evaluation phase, the models built from the previous phase are evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. This step ensures the model's reliability and effectiveness in identifying churn. The methodology approach proposed can be visualised in Figure 1.

3.1. Dataset Descriptions

This study draws upon the Bank Customer Churn Dataset obtained from the Kaggle machine learning repositories, accessible at the following link: https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset. This dataset encompasses comprehensive information related to customer churn occurrences within the ABC Multistate bank, specifically tailored for the prediction of customer churn. Comprising a total of 12 features, including customer_id, credit_score, country, gender, age, tenure, balance, products_number, credit_card, active_member, estimated_salary, and churn (serving as the target variable), the dataset is structured such that a value of 1 indicates that a client has departed the bank within a specified period, while a value of 0 signifies that the client has not churned. With a voluminous dataset incorporating 10,000 churn records.

3.2. Data Balancing

To rectify the class imbalance inherent in the customer churn dataset, this study advocates for the utilization of the Synthetic Minority Over-sampling Technique (SMOTE) as a data balancing method. SMOTE, an acronym for Synthetic Minority Over-sampling Technique, is devised to ameliorate the challenges associated with imbalanced datasets (Hambali & Gbolagade, 2016; Zhou et al., 2022). Unlike conventional approaches that involve replicating existing instances of the minority class, SMOTE introduces synthetic samples using interpolation between extant instances (Hambali & Gbolagade, 2016). This process entails the selection of a minority class instance and its proximate neighbours, followed by the creation of synthetic instances along the line segments connecting the chosen instance with its neighbouring counterparts. The adoption of SMOTE is motivated by its capacity to effectively address class imbalance, thereby enhancing the performance of machine learning models, particularly in scenarios where the minority class is of particular significance.

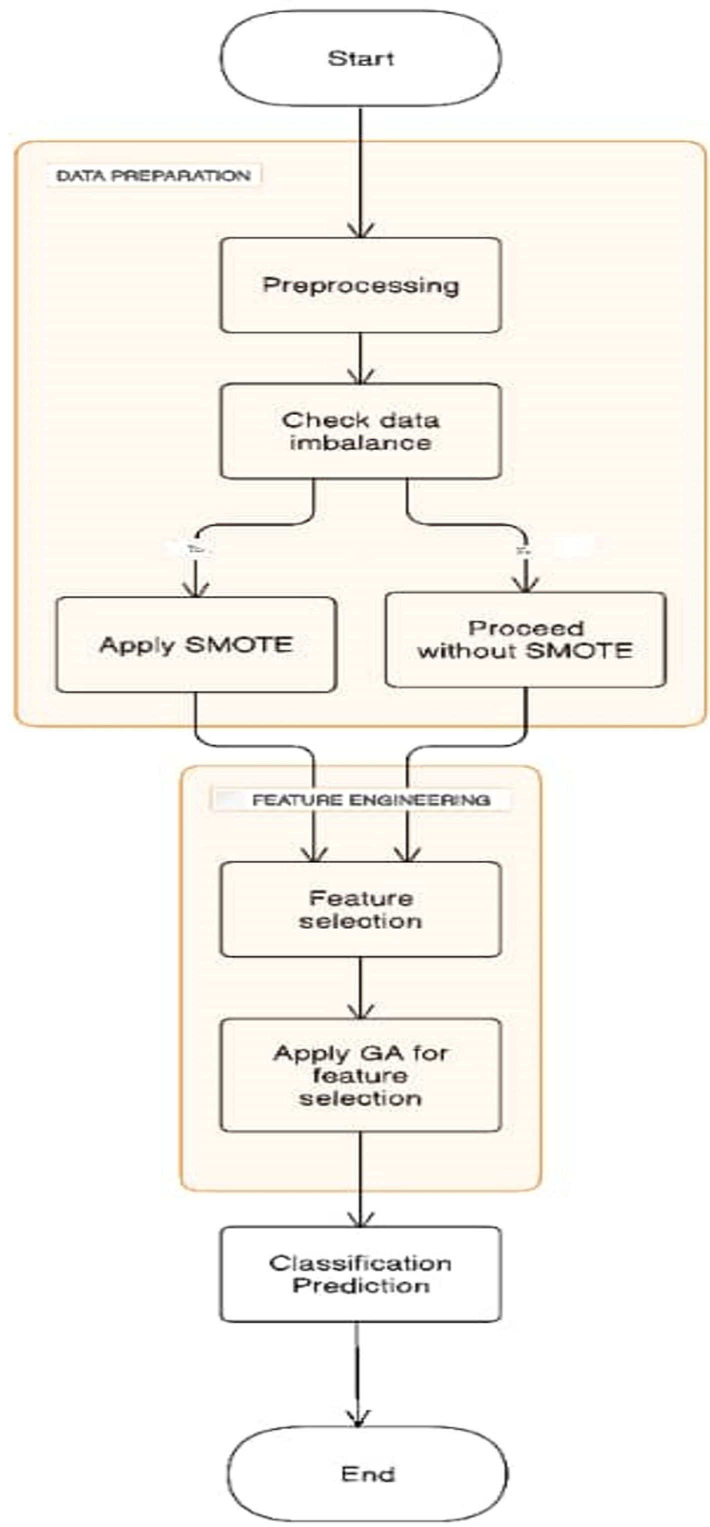


Figure 1. Research Methodology Framework

3.3. Generic Algorithm

The Genetic Algorithm (GA) technique is widely recognized in evolutionary computation research, mirroring the principles of natural selection. Its application spans various fields including business, engineering, and numerous other domains, aiming to derive optimal solutions to problems. GA comprises three fundamental operators: selection, crossover, and mutation. Selection identifies the fittest individuals within the population set based on a fitness function. Crossover involves merging the genetic material of two-parent records, while mutation randomly alters specific bits to introduce diversity.

The steps involved in GA are as follows:

- Initialization: Create a population of random chromosomes, each representing a unique solution to the problem.
- Fitness Evaluation: Assess the fitness of each chromosome.
- Reproduction: Generate a new population through selection, crossover, and mutation operations.
- Selection: Choose individuals with higher fitness as parents.
- Crossover: Produce new offspring by combining genetic material from parent chromosomes.
- Mutation: Randomly alter certain bits to create novel offspring.
- Replacement: Replace the current population with the new one to continue the evolutionary process.
- Termination Criteria: Stop the algorithm when certain conditions are met, such as reaching a maximum number of generations or achieving a desired level of solution quality.
- Iteration: Repeat the process from step 2 until the termination criteria are satisfied.

The parameters of the GA are configured as shown in Table 1:

Table 1. Parameter Setting

Parameter	Value Selected
Population Size	20
Number of Generations	20
Crossover Probability	0.6
Mutation Probability	0.033
Repetition Frequency	20
Seed for Random Number Generation	1

At the end of the feature selection process, GA selects the most informative attributes and discards the other features.

3.4. Classification Algorithm

Considering that the problem of Bank customer churn prediction is identified to be a classification problem, four classification algorithms were proposed. These algorithms include the Random Forest (RF), K-Nearest Neighbor (KNN), AdaBoost, and Artificial Neural Network (ANN).

3.4.1. Random Forest

The RF classifier employs a divide-and-conquer approach, relying on the random subspace method. Multiple trees are created, and each decision tree is trained by randomly selecting attributes from the predictor attribute set (Abdulsalam et al., 2015; Hambali et al., 2022; Sekulić et al., 2020). Each tree matures based on the available attributes or parameters, and the final decision is made through weighted averages. Notably, it can effectively manage a large number of input parameters without requiring deletion and is capable of handling missing values within the training dataset for predictive model training.

3.4.2. K-Nearest Neighbour

The k-nearest Neighbor (KNN) method stands out as a straightforward and highly effective non-parametric approach to classification within supervised learning (Rahman & Kumar, 2020). KNN operates by identifying the *k* closest samples from an established dataset. When a new and unfamiliar sample emerges, the method classifies this new sample based on its similarity to the existing dataset. In essence, the classification algorithm categorizes the test sample into a group determined by the *k*-training samples that serve as its nearest neighbours, ultimately assigning it to the class with the highest likelihood.

3.4.3. AdaBoost

The Ada-boost, similar to the Random Forest Classifier, is an ensemble classifier composed of multiple algorithms whose outputs are combined. While a single algorithm may exhibit poor classification performance, the Ada-Boost enhances accuracy by iteratively retraining the algorithm (Lalwani et al., 2022). This involves selecting the training set based on the accuracy of previous training iterations and assigning appropriate weights in the final voting process.

3.4.4. Artificial Neural Network (ANN)

The proposed ANN algorithm simulates the working of the biological neural system and its architecture primarily consists of key components, including inputs, weights, the sum function, activation function, and outputs. Within the working mechanism of an ANN, the input signals undergo modification by the weights and are subsequently summed with the bias term, as indicated in equation 3.1 (Olagoke et al., 2016; Upreti et al., 2022). A crucial aspect of ANN design lies in its sum function, incorporating both inputs and weights. The operational process of an ANN involves the transformation of input signals through weight adjustments, and the resultant modified signals, when combined with the bias term, are aggregated together, as depicted in equation 1.

$$y = f\left(\sum_{i=0}^{n} x_i w_i - b\right)$$

where "f" represents the activation function and the "w" weight of the*ith* input neuron, *n* represents the number of neurons, and *b* is the bias term, with the result being denoted as y; the output of the neural network (equation 1) includes an input for the activation function "f" which is referred to as "sum" in the equation.

3.5. Performance Evaluation Metric

To evaluate the performance of the model the accuracy, precision, recall, and f1-score metrics were utilized. The metrics are based on some parameters such as:

i. **True Positive (TP):** defines customers who are churning (positive) and classified as churn (positive).
ii. **True Negative (TN):** defines customers who are not churning (negative) and classified as not churning (negative).
iii. **False Positive (FP):** identifies customers who are not churning (negative) and classified as churn (positive).
iv. **False Negative (FN):** identifies customers who are churning (positive) and classified as not churning (negative).

Accuracy: Accuracy computes the percentage of customers who are correctly churned or not churned. Accuracy can be mathematically expressed as:

$$\text{Accuracy} = \frac{TP + FN}{TP + FP + TN + FN}$$

Precision: defines the ratio of correctly classified positive samples (True Positive) to the total number of classified positive samples (either correctly or incorrectly). Essentially precision is the percentage of TP to the sum of TP. Precision can be mathematically expressed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: is defined mathematically as the ratio between the numbers of positive samples correctly classified as positive to the total number of positive samples. The recall measures the model's ability to detect positive samples.

$$TPR = \frac{TP}{TP + FN}$$

F1 Score: is the harmonic mean of precision and recall equation. The maximum possible F1 score is 1, which indicates perfect recall and precision.

$$F_{\text{Score}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Result and Discussion

This section elucidates the findings resulting from a practical exploration and application aimed at predicting customer churn using the Bank Customer Churn Dataset sourced from the Kaggle machine learning repositories. The research methodology entails the application of machine learning algorithms, specifically the RF, KNN, AdaBoost, and ANN algorithms. Following model construction, a thorough validation process is conducted, evaluating performance metrics including precision, recall, and F1-score.

4.1. Environmental Setup

The bank churn prediction model was developed utilizing the computational capabilities of the Anaconda programming environment. The model's complexities were smoothly incorporated into Python, running on a Windows operating system equipped with a dual-core Intel Core i5 processor and 4GB of RAM. Implementation was facilitated by the Sklearn API, designed for sophisticated machine-learning tasks. Additionally, essential Python libraries like NumPy were utilized for intricate numerical computations, while pandas managed dataset parsing and integration seamlessly. To visually represent the behaviour of the meticulously crafted machine learning models, Matplotlib, a versatile visualization tool, was employed.

4.2. Exploratory Data Analysis

Upon conducting exploratory data analysis (EDA) on the dataset, from Figure 2, it was observed that 20.4% of 10,000 customers tended to churn, while the majority, constituting 79.6%, appeared likely to be retained. Further examination of churn rates by gender and country revealed notable insights as visualized in Figure 3. Specifically, among females, there was a churn tendency of 20.43%, while males exhibited a slightly lower churn rate of 12.71%. When considering the geographical aspect, within Germany, females displayed a significantly higher churn tendency of 37.55%, contrasting with males who exhibited a churn rate of 27.81%. Similarly, among Spanish customers, females demonstrated a churn tendency of 21.21%, while males exhibited a relatively lower churn rate of 13.11%. Essentially, the rate of churn based on gender signifies that females have a higher tendency to churn than males as shown in Figure 4.

4.3. Classification Results Without SMOTE

Table 2 presents the classification performance of various models for predicting bank customer churn, with metrics including training accuracy and testing accuracy. The following is the analysis of each model's performance:

Random Forest achieved perfect accuracy during training, indicating that the model was able to perfectly fit the training data. However, the testing accuracy of 86% suggests that the model may have slightly overfitted the training data, as it did not generalize as well to unseen data. Nonetheless, with an accuracy of 86%, Random Forest still performed relatively well on the testing dataset.

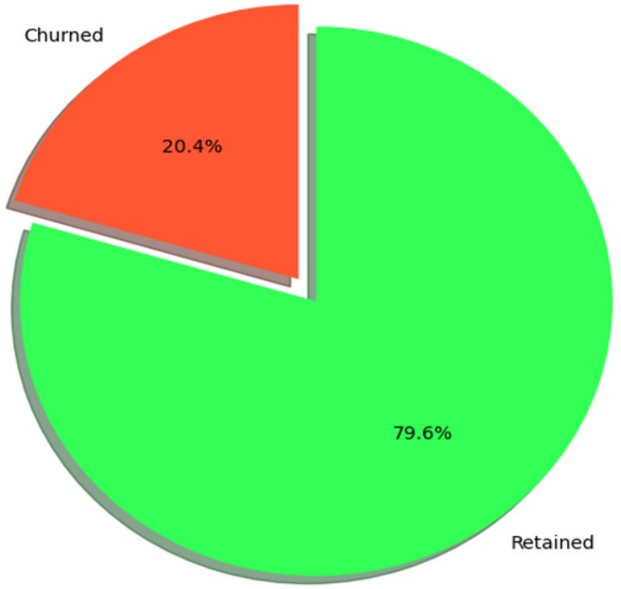


Figure 2. Churn Ratio

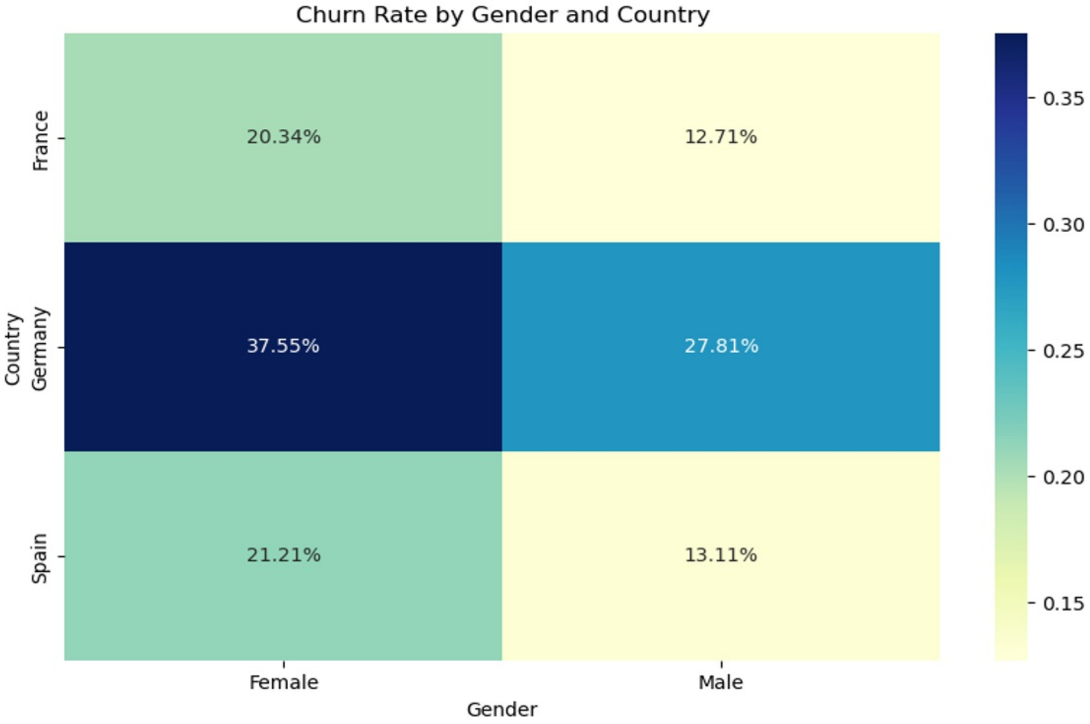


Figure 3. Churn by Country Genders

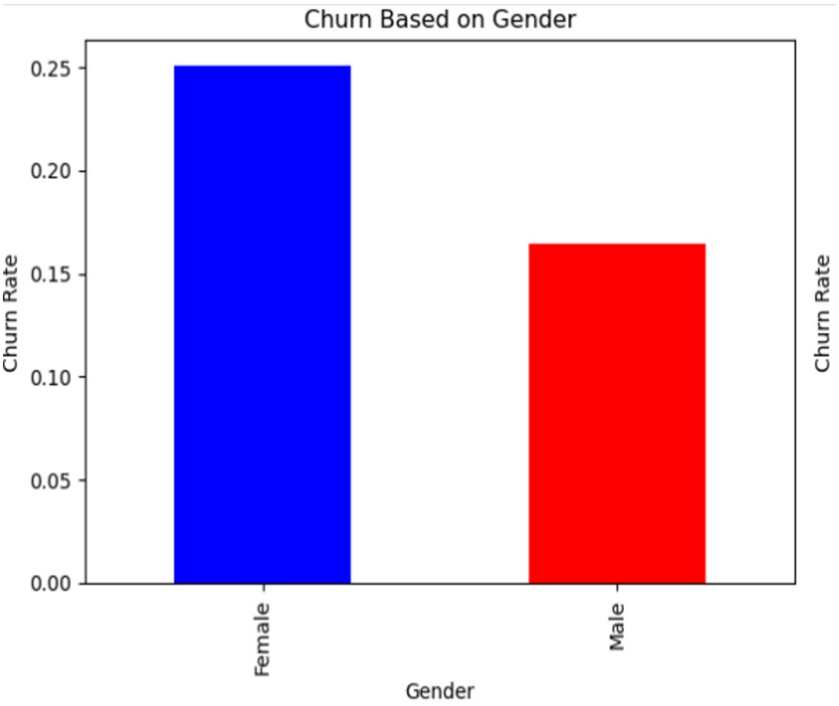


Figure 4. Churn by Genders

K-Nearest Neighbors achieved an accuracy of 87% during training, which suggests a reasonable fit to the training data. However, the testing accuracy of 85% indicates that the model performed slightly worse on unseen data compared to its performance on the training data. KNN's performance suggests that it may not generalize as well to new instances compared to other models.

Table 2. Bank Churn Models Results without SMOTE

Models	Training Accuracy (%)	Testing Accuracy (%)
RF	1.00	0.86
KNN	0.87	0.85
Ada	0.86	0.85
ANN	0.87	0.87

AdaBoost achieved an accuracy of 86% during training, indicating a decent fit to the training data. However, similar to KNN, the testing accuracy of 85% suggests that AdaBoost may not generalize as effectively to new instances. Nonetheless, it still performed reasonably well on the testing dataset.

ANN achieved an accuracy of 87% during training, suggesting a good fit for the training data. Impressively, the testing accuracy of 87% indicates that the model performed equally well on unseen data, demonstrating strong generalization capabilities. ANN appears to be the most robust model in terms of classification performance, achieving the highest testing accuracy among the models listed.

Overall, while Random Forest demonstrated perfect training accuracy, it slightly lagged in generalization to unseen data compared to ANN, which exhibited consistent performance across both training and testing datasets. KNN and AdaBoost also performed reasonably well but showed slightly lower testing accuracies, indicating potential limitations in their ability to generalize to new instances. Figure 5 depicts the accuracy comparison of algorithms used.

Table 3. Bank Churn Models Performance Metric without SMOTE

Models	Precision	Recall	F1-Score
RF	0.85	0.86	0.85
KNN	0.83	0.85	0.83
Ada	0.84	0.85	0.85
ANN	0.86	0.87	0.85

Table 3 presents the performance metrics of different classification models in predicting bank customer churn, focusing on precision, recall, and F1-score, alongside the corresponding training and testing accuracy. RF yields a Precision of 0.85, Recall of 0.86 and F1-Score of 0.85. RF exhibits high accuracy in training but slightly lower accuracy in testing, suggesting a potential issue of overfitting. However, its precision, recall, and F1-score are relatively consistent and balanced, indicating its effectiveness in correctly identifying churners while minimizing false positives.

KNN yields Precision of 0.83, Recall of 0.85 and F1-Score of 0.83. KNN demonstrates similar performance to RF in terms of testing accuracy. However, it has slightly lower precision, recall, and F1-score, indicating a marginally higher rate of misclassification and lower ability to accurately identify churners.

AdaBoost yields Precision of 0.84, Recall of 0.85 and F1-Score of 0.85. AdaBoost also exhibits comparable performance to RF and KNN in terms of testing accuracy. Its precision, recall, and F1-score are also similar to KNN, indicating a moderate level of predictive capability.

ANN result showed Precision of 0.86, Recall of 0.87 and F1-Score of 0.85. ANN demonstrates the highest accuracy in both training and testing among the models listed. Its precision, recall, and F1-score are also slightly higher than those of RF, indicating a robust ability to correctly classify churners and non-churners.

In summary, while all models achieve relatively high accuracy, ANN appears to outperform the others slightly in terms of both accuracy and the precision-recall trade-off. However, it's essential to consider factors such as computational complexity and interpretability when selecting the most suitable model for real-world applications.

4.4. Classification Results Without SMOTE

To address the class imbalance within the dataset, the SMOTE data augmentation technique was employed. The reason for the application of the SMOTE data balancing was that the class labels for the churn and retained labels as seen in

Figure 3 are not balanced. Hence, the SMOTE approach aimed to create a more equitable distribution between class labels. As a result of applying the SMOTE algorithm, both the churn and not churn labels were balanced, with each totalling 7963 instances. The augmentation process led to the generation of additional records, as depicted in Figure 6.

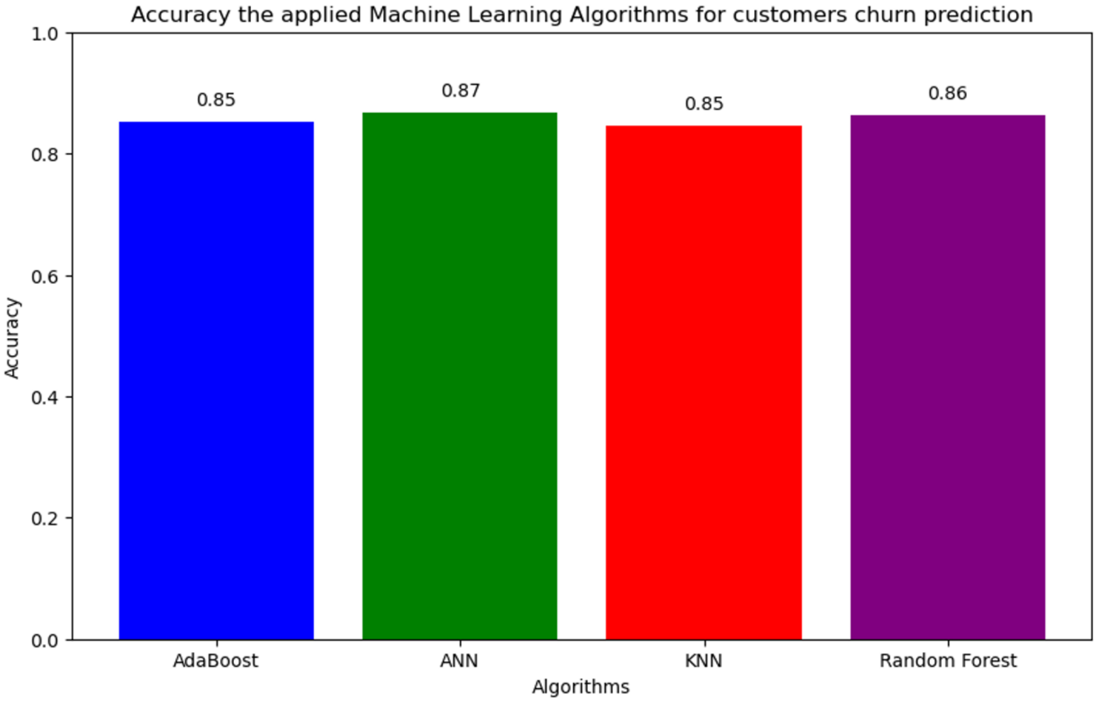
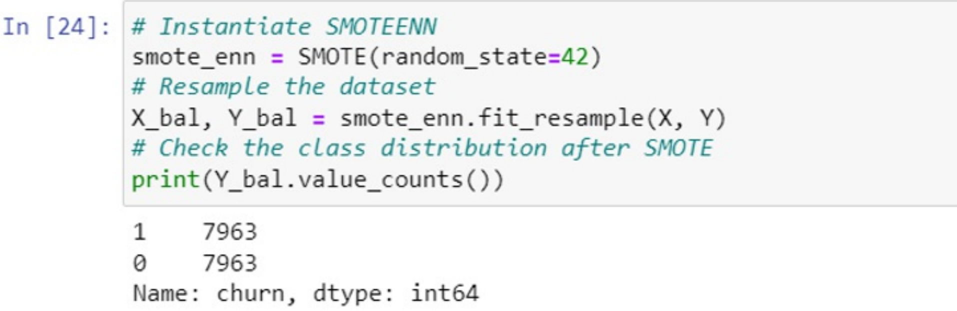


Figure 5. Comparison of Accuracy results of Classification Algorithms without SMOTE



```
In [24]: # Instantiate SMOTEENN
         smote_enn = SMOTE(random_state=42)
         # Resample the dataset
         X_bal, Y_bal = smote_enn.fit_resample(X, Y)
         # Check the class distribution after SMOTE
         print(Y_bal.value_counts())

         1    7963
         0    7963
         Name: churn, dtype: int64
```

Figure 6. SMOTE Data Balancing

Table 4 presents the results of different churn prediction models, detailing their training and testing accuracies. Random Forest (RF) achieved a perfect training accuracy of 99%, indicating that it perfectly fits the training data. However, its testing accuracy slightly dropped to 95%, suggesting a minor degree of overfitting but still maintaining strong predictive performance on unseen data. K-Nearest Neighbors (KNN) demonstrated a high training accuracy of 98%, indicating a good fit to the training data. Its testing accuracy remained high at 96%, indicating its ability to generalize well to new data while maintaining a lower risk of overfitting compared to RF. AdaBoost (Ada) exhibited a training accuracy of 89%, which indicates a relatively good fit to the training data but lower than RF and KNN. Its testing accuracy was 88%, suggesting a slight drop in performance on unseen data, potentially indicating some overfitting or limitations in generalization. Artificial

Neural Network (ANN) showed a training accuracy of 91%, indicating a reasonably good fit to the training data. Its testing accuracy was also 91%, reflecting consistent performance on both training and testing datasets.

Table 4. Churn Models Result with SMOTE

Models	Training Accuracy (%)	Testing Accuracy (%)
RF	1.00	0.95
KNN	0.98	0.96
Ada	0.89	0.88
ANN	0.91	0.91

Overall, the results suggest that KNN and ANN models performed relatively well in terms of both training and testing accuracies, with RF also showing strong predictive performance despite a slight decrease in testing accuracy. Hence, taking into cognizance their testing performances, Figure 7 shows a graph of each model performance with their respective scores appended to bars.

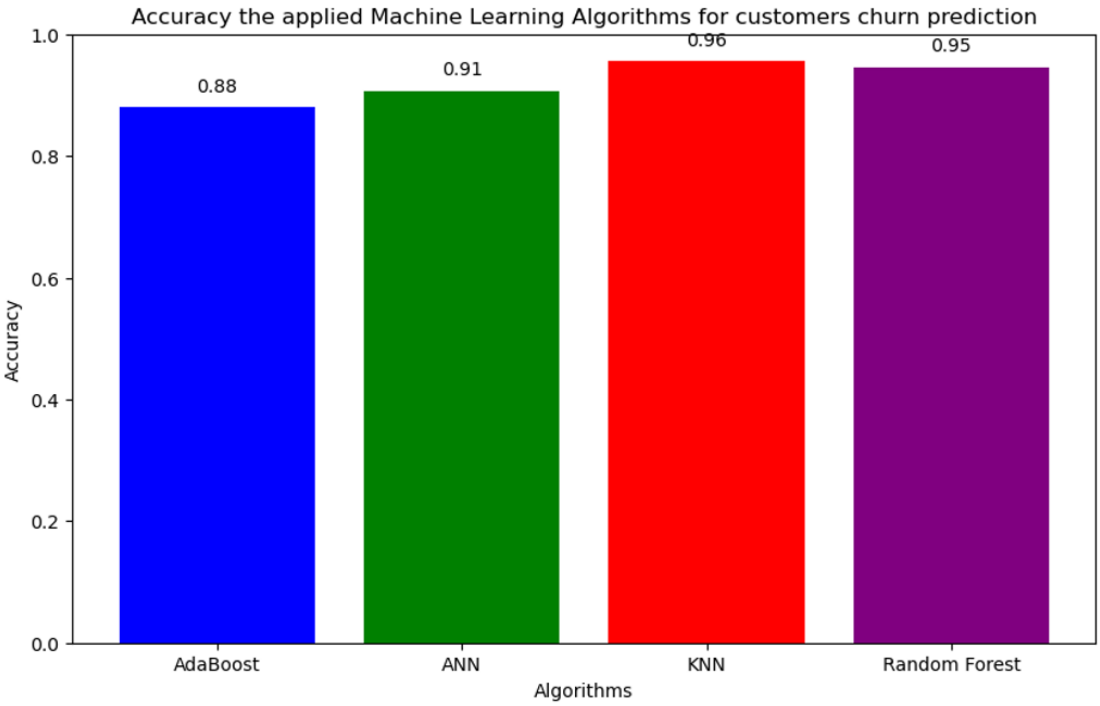


Figure 7. Comparison of Accuracy results of Classification Algorithms with SMOTE

Figure 8 compares the results of classification algorithms with and without SMOTE. The experimental results shown that, there significant improvement in accuracy when using SMOTE to address the data imbalance in the dataset.

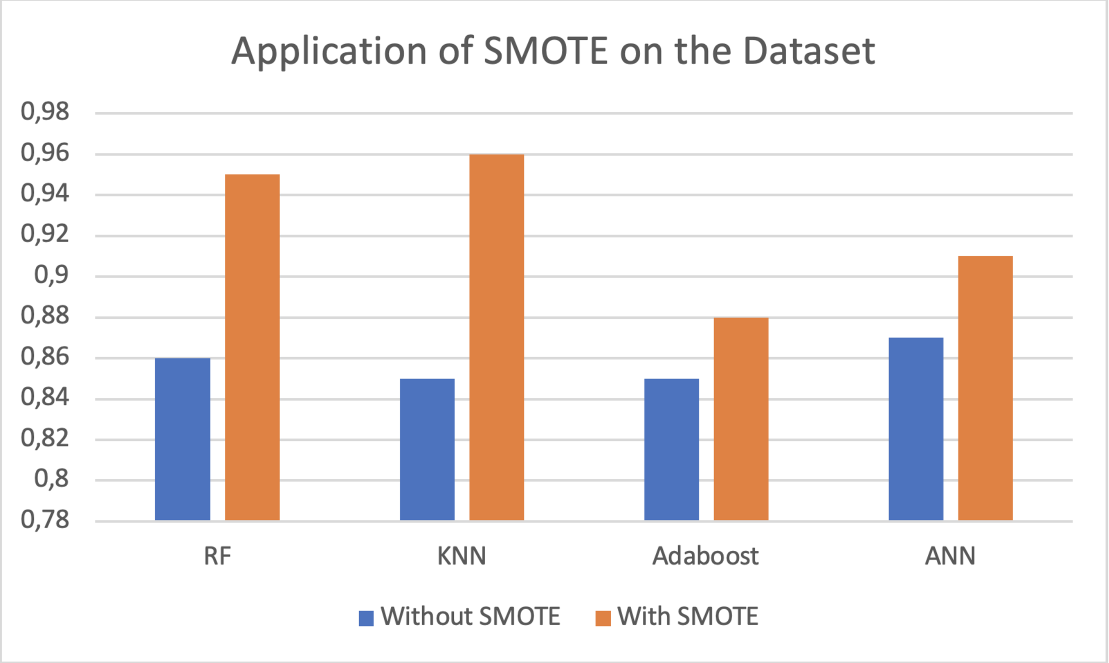


Figure 8. Comparison of Accuracy results of Classification with and without SMOTE

To further analyze the performance of the model, other performance metrics were employed. The metrics that provide the result of each model are precision, recall, and f1-scores for the respective label of churn and not churn.

Table 5. Other Performance Metrics

Models	Precision	Recall	F1-Score
RF	0.95	0.95	0.95
KNN	0.96	0.96	0.96
Ada	0.88	0.88	0.88
ANN	0.91	0.91	0.91

From Table 5, the Random Forest (RF) classification model yielded promising results in predicting customer churn. With an overall accuracy of 95%, the model demonstrates a high level of performance in distinguishing between customers likely to churn (class 1) and those likely to be retained (class 0). The precision, which measures the proportion of true positive predictions among all positive predictions, is 94% for class 0 and 95% for class 1(see Appendix). This indicates that the RF model is highly precise in identifying both retained and churned customers. Moreover, the recall, or true positive rate, is 93% for class 0 and 96% for class 1, signifying that the model effectively captures the majority of actual churn instances while minimizing false negatives. The F1-score, which considers both precision and recall, is 94% for class 0 and 95% for class 1, reflecting a balanced performance between precision and recall. Overall, the RF model demonstrates robustness in predicting customer churn, as evidenced by its high accuracy and balanced performance across various evaluation metrics.

The K-Nearest Neighbors (KNN) classification model for customer churn prediction yielded promising results, as evidenced by the classification report presented in Table 5. For customers labelled as not churning (class 0), the model achieved a precision of 0.97 and a recall of 0.92, indicating that among the instances predicted as not churning, 97% were correctly classified, and 92% of actual not churning instances were identified by the model (see Appendix). The f1-score, which considers both precision and recall, was 0.95 for this class, reflecting a balanced performance. Similarly, for customers labelled as churning (class 1), the model exhibited high precision (0.94) and recall (0.98), indicating accurate identification of churn instances. The f1-score for class 1 was 0.96, suggesting robust performance in capturing churn instances. Overall, the model demonstrated an accuracy of 0.96, indicating that it correctly classified approximately 96% of the instances in the dataset. The macro average and weighted average of precision, recall, and f1-score were also high, further underscoring the model's effectiveness in customer churn prediction. These results suggest that the KNN classifier shows promise in accurately identifying customers at risk of churn, providing valuable insights for customer retention strategies.

The AdaBoost classifier's performance for customer churn prediction was also evaluated through precision, recall, and F1-score metrics from Table 5. With regards to class 0 (customers likely to be retained), the classifier achieved a precision of 0.87 and recall of 0.86, indicating that among the instances predicted as not churning, 87% were correctly classified, and it successfully captured 86% of the actual instances of customers likely to be retained. The F1-score, which considers both precision and recall, stands at 0.86 for this class. For class 1 (customers with a tendency to churn), the precision, recall, and F1-score are slightly higher, with values of 0.89, 0.90, and 0.89, respectively. This indicates that the classifier performed slightly better in identifying customers with a churn tendency, correctly classifying 89% of the instances predicted as churning and capturing 90% of the actual churn instances. Overall, the classifier achieved an accuracy of 0.88, suggesting that it correctly classified 88% of all instances in the dataset. The macro average and weighted average of precision, recall, and F1-score are all 0.88, indicating balanced performance across both classes. These results suggest that the AdaBoost classifier shows promising performance in predicting customer churn, with a balanced trade-off between precision and recall for both classes.

Also, the classification report for the ANN on the customer churn reveals the precision for class 0 (customers not churning) is 0.91, indicating that among the instances predicted as not churning, 91% were correctly classified. Similarly, for class 1 (customers churning), the precision is 0.90, signifying that 90% of the instances predicted as churning were accurately classified (see Appendix). The recall for class 0 is 0.88, suggesting that 88% of the actual instances of not churning were correctly identified by the model. For class 1, the recall is 0.93, indicating that the model successfully identified 93% of the actual instances of churning. Lastly, the F1-score, which is the harmonic mean of precision and recall, is 0.89 for class 0 and 0.92 for class 1. These scores reflect the balance between precision and recall, with higher values indicating better model performance.

4.4.1. Confusion Matrix

The confusion matrix further provides valuable insight into the performance of the RF, KNN, AdaBoost, and ANN models in predicting churn and non-churn instances. It consists of four quadrants representing different combinations of predicted and actual class labels: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In the context of churn prediction. TP are instances where the model correctly predicts churn (positive class) among customers who churned. In other words, these are customers correctly identified as likely to churn. FP occurs when the model incorrectly predicts churn among customers who did not churn (negative class). These are instances where the model incorrectly flags customers as likely to churn when they remain with the service. TN are instances where the model correctly predicts non-churn among customers who did not churn. In essence, these are customers correctly identified as likely to be retained. FN occur when the model incorrectly predicts non-churn among customers who churned. These instances represent missed opportunities where the model fails to identify customers who are likely to churn. Hence, the count of each quadrant from Figure 9 to Figure 12 defines the confusion matrices for the RF, KNN, AdaBoost, and ANN.

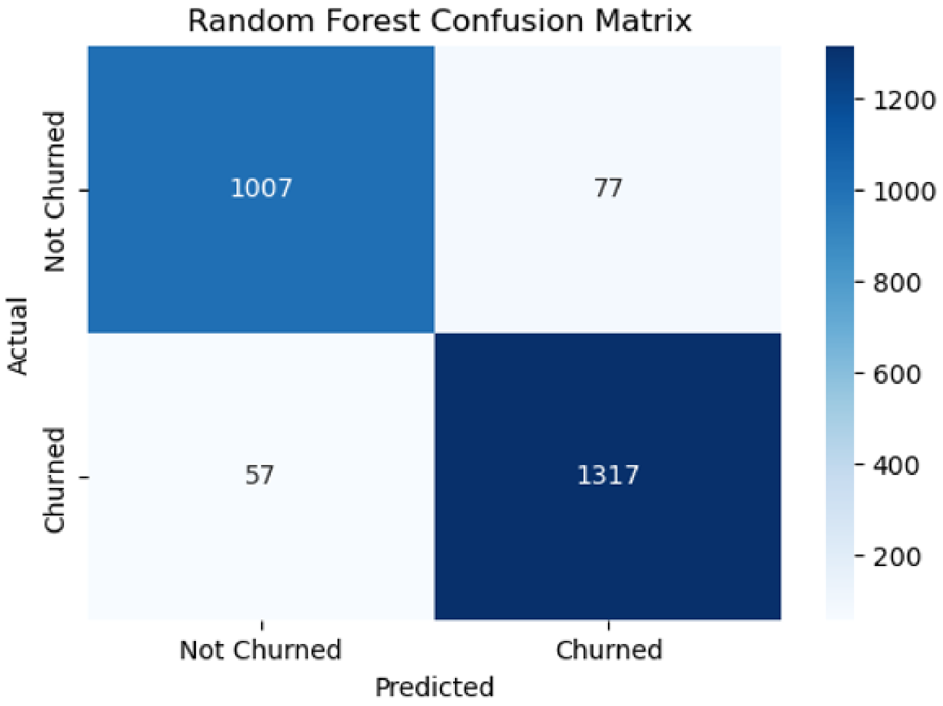


Figure 9. RF Confusion Matrix

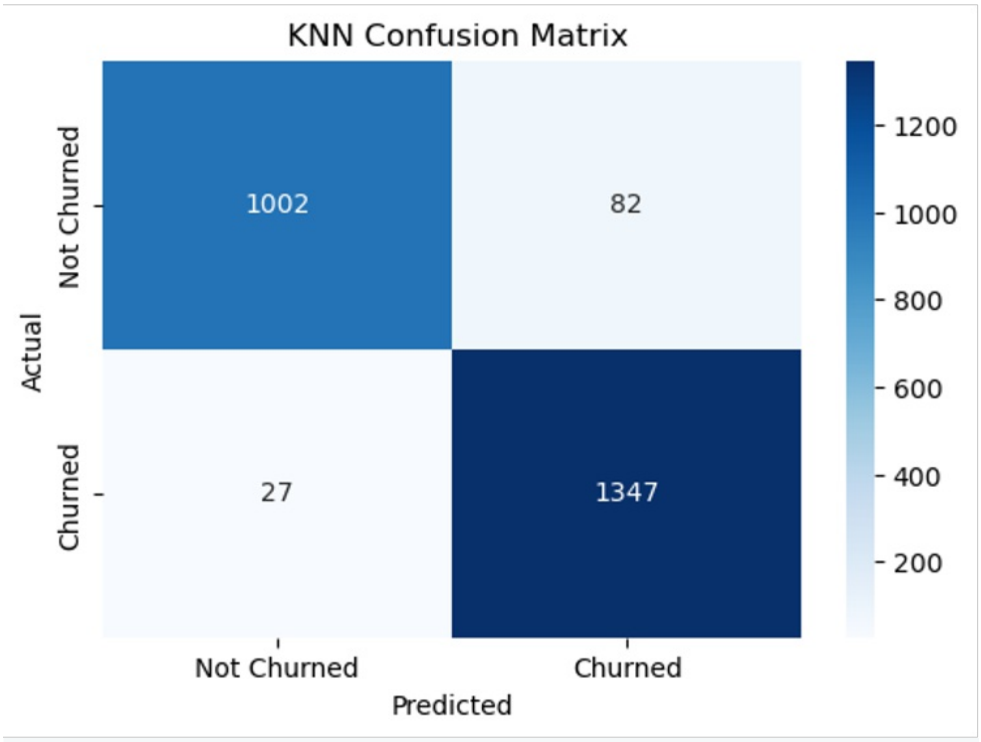


Figure 10. KNN Confusion Matrix

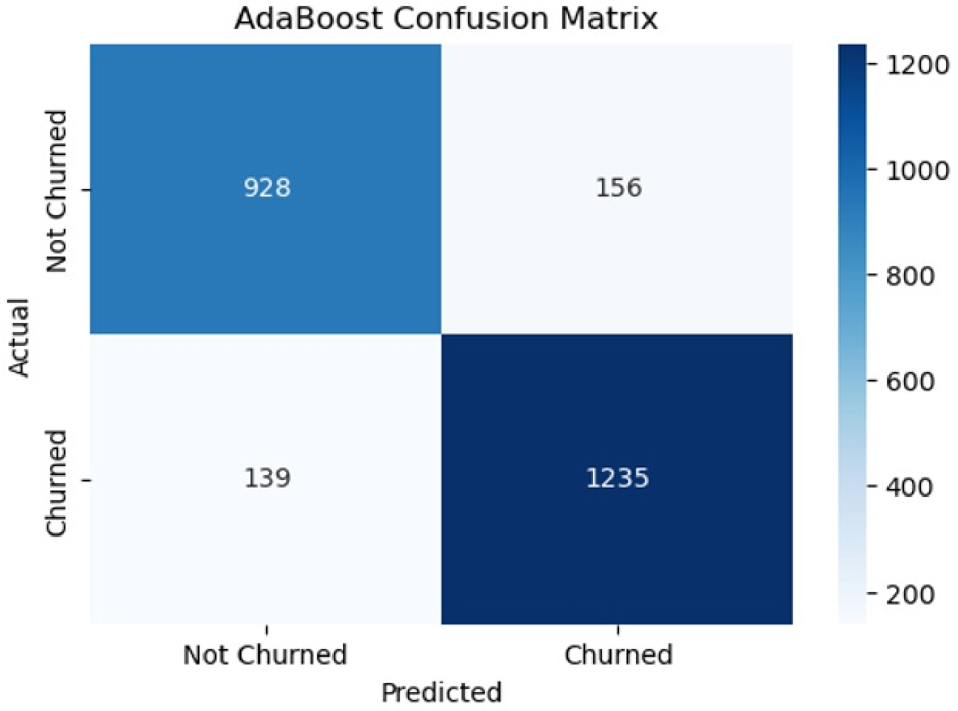


Figure 11. RF Confusion Matrix

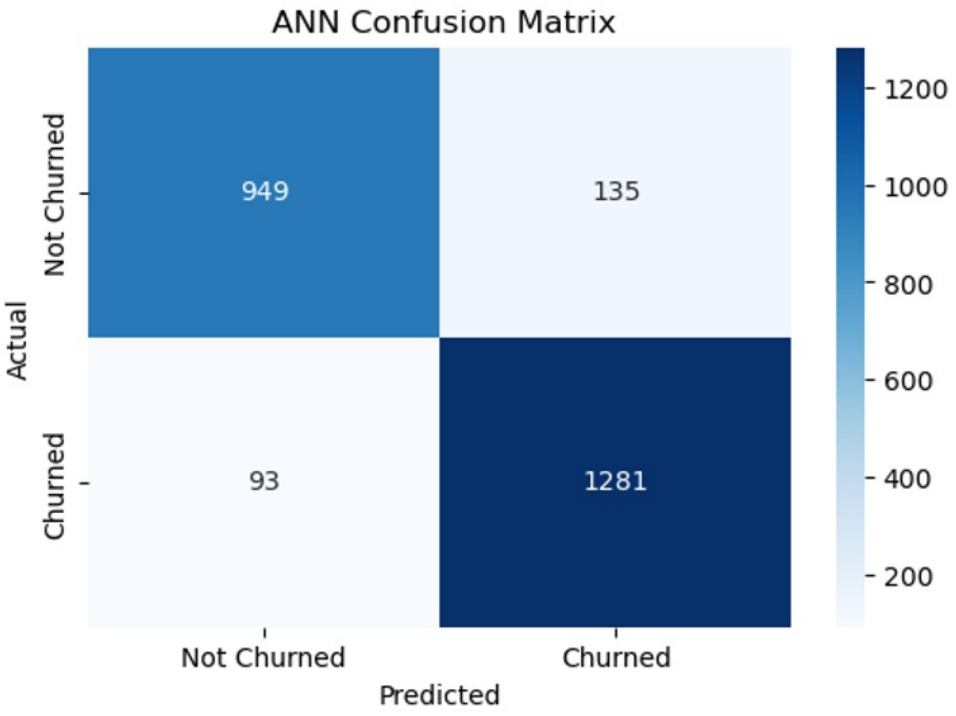


Figure 12. KNN Confusion Matrix

4.4.2. ROC Curve Result

To comprehensively evaluate the effectiveness of the developed churn detection models, the study leveraged the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric. ROC-AUC stands as a widely recognized metric for evaluating the performance of classification models, particularly in scenarios involving binary and multiclass classifications. It provides insight into a model's ability to distinguish between positive and negative instances across various threshold settings. The ROC-AUC curves presented in Figures 13 to 16 correspond to the RF, KNN, AdaBoost, and ANN models, illustrating the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. TPR, synonymous with sensitivity or recall, highlights the model's capability to correctly identify positive instances. Analyzing the graphical representations reveals the trade-off between TPR and FPR as the classification threshold is adjusted. An optimal ROC-AUC curve typically resides in the top-left corner, indicating high sensitivity coupled with a low false positive rate, thereby resulting in a larger area under the curve. Notably, the ROC-AUC metric for the RF and KNN models achieved exceptional performance, reaching 99%. This suggests a robust ability to effectively identify instances of network anomalies while maintaining a minimal rate of false positives. However, it is essential to note that the AdaBoost model achieved an AUC of 95%, indicating slightly lower discrimination between positive and negative instances compared to RF and KNN. Similarly, the ANN model achieved an AUC of 90%, further emphasizing its performance relative to the other models evaluated in the study. These varying AUC scores provide valuable insights into the comparative effectiveness of the churn detection models, facilitating informed decision-making in model selection and deployment.

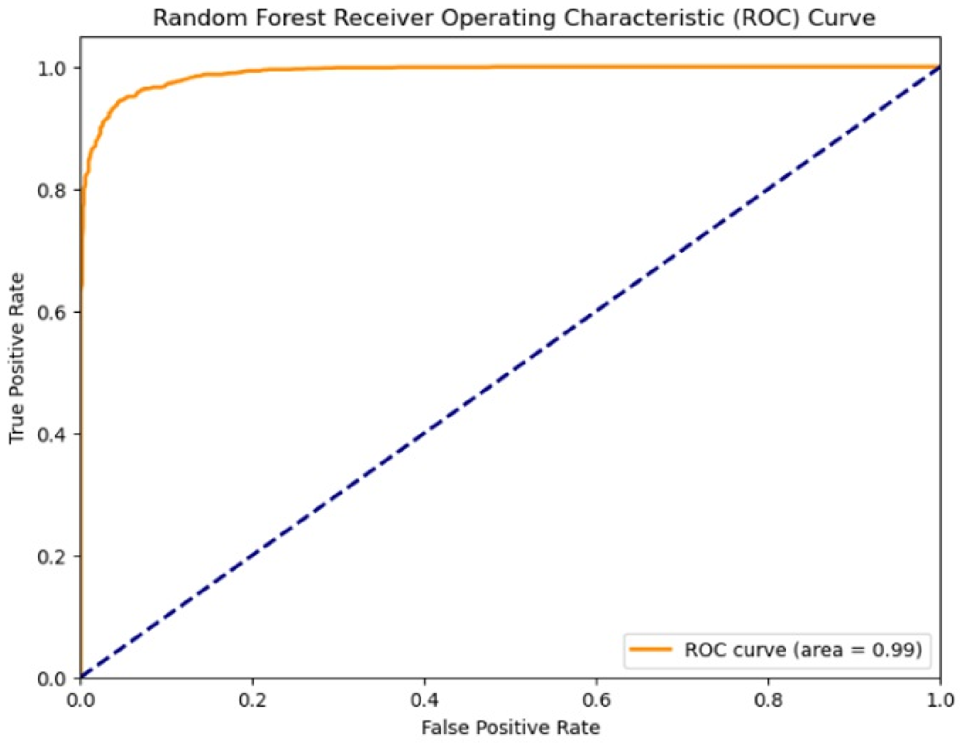


Figure 13. RF ROC

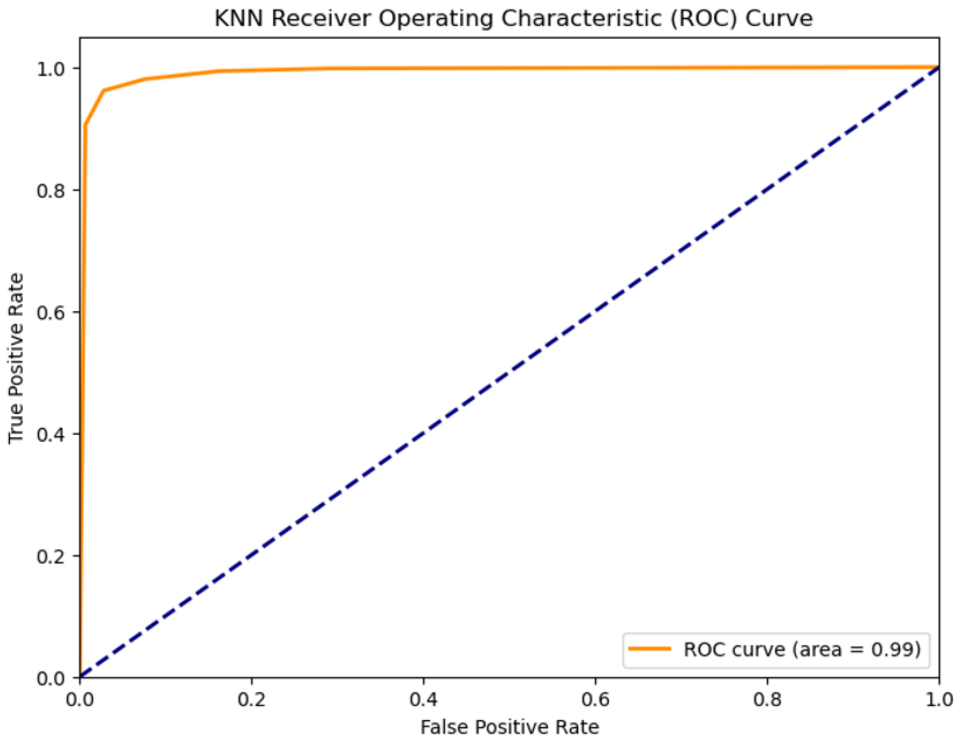


Figure 14. KNN ROC

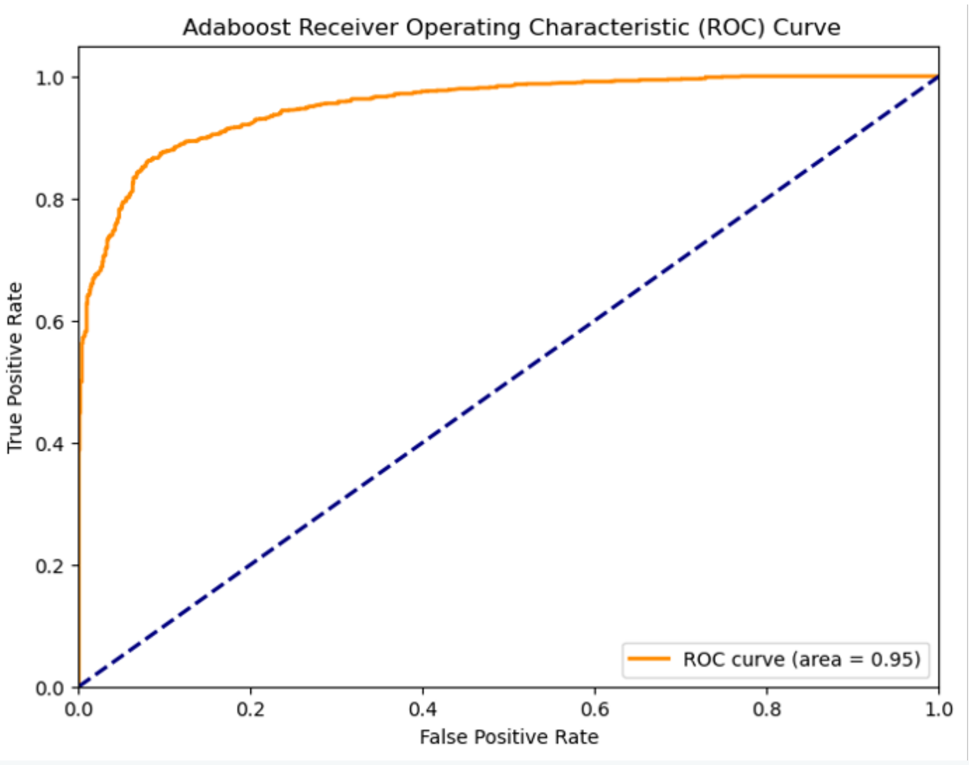


Figure 15. AdaBoost ROC

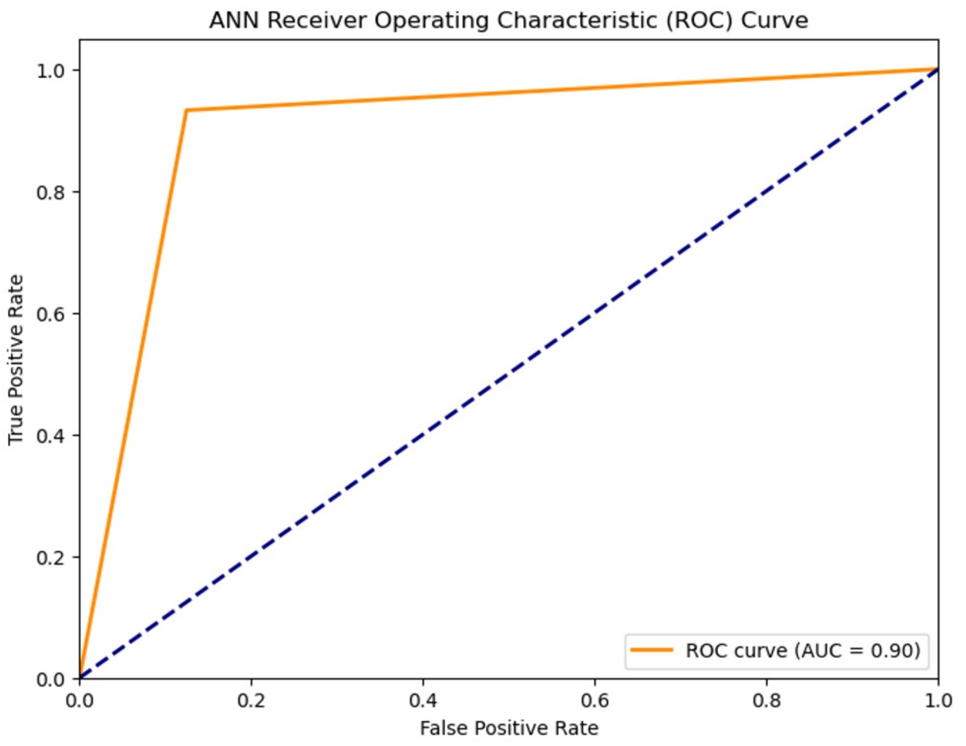


Figure 16. ANN ROC

4.5. Comparison with Existing Studies

A comparison was conducted between the proposed model and existing models in the domain. The proposed model achieved an accuracy of 96%, maintaining identical precision and F-measure values, as illustrated in Table 6. The proposed approach outperformed some of the studies in the literature except the work of Seyed Mohammad Sina Mirabdolbaghi and Babak Amiri (2022) that yield the same result with our approach. Our comparison is based on those authors that carried out the experiments on the bank churn, so as to maintain a level plain ground.

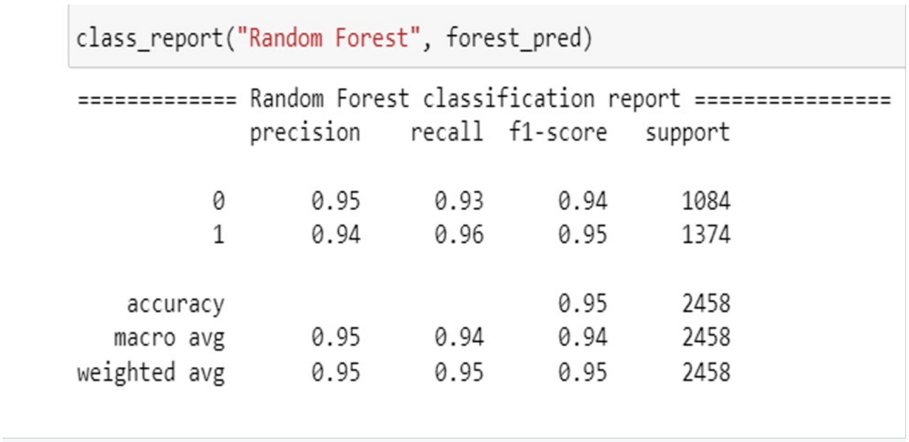
Table 6. Comparison with Previous Studies

Authors	Approach	Accuracy	Precision	F- measure
Rahman & Kumar, (2020)	KNN, SVM, DT, RF	95.74	-	-
Leung & Chung, (2020)	LR, RF GBM	94.78	24.80	-
Domingos et al., (2021)	DNN, SGD, MLP, AdaGrad	86.45	-	-
Sagala & Permai, (2021)	XGBoost, LightGBM, and CatBoost	91.4%,	87.7%	-
Tékouabou et al., (2022)	SMOTE-RF	86	-	86
Muneer et al., (2022)	SMOTE with RF, Adaboost & SVM	88.7	-	91.90
Mirabdolbaghi & Amiri, (2022)	PCA, LDA, T-SNE. GBM	**96**	**96**	
Chen et al., (2023)	logistic regression, SVM, GBDT, random forest and AdaBoost	0.811	0.514	0.599
(Hon et al., 2023)	*KNN, SVM, NB, DT, RF, and XGBoost*	84.76		56.95
Proposed model	SMOTE-KNN	**96.00**	**96.00**	**96.00**

5. Conclusion and Future Work

The objective of this research is to aid the banking industry in devising a model to enhance their profitability. It is evident that predicting customer churn is crucial for boosting revenue in banking companies. Therefore, this research endeavors to formulate a predictive model for forecasting customer attrition in the banking sector. The increasing emphasis on developing an accurate and efficient churn prediction model poses a notable research challenge for both academics and industry experts. This study suggests that employing machine learning (ML) techniques holds considerable promise in addressing the issue of customer churn management, enabling the establishment of an early-warning system tailored to the dynamic customer landscape. The comprehensive evaluation of the model proposed in this study reveals an impressive overall accuracy of 96% in forecasting customer churn. Notably, the study did not analyze the characteristics of anticipated customer churns, although such insights could be invaluable for organizations in deciding whether to retain or release individual customers. Therefore, future research will concentrate on examining customer churn characteristics, as they may offer greater long-term value to the organization.

Appendix



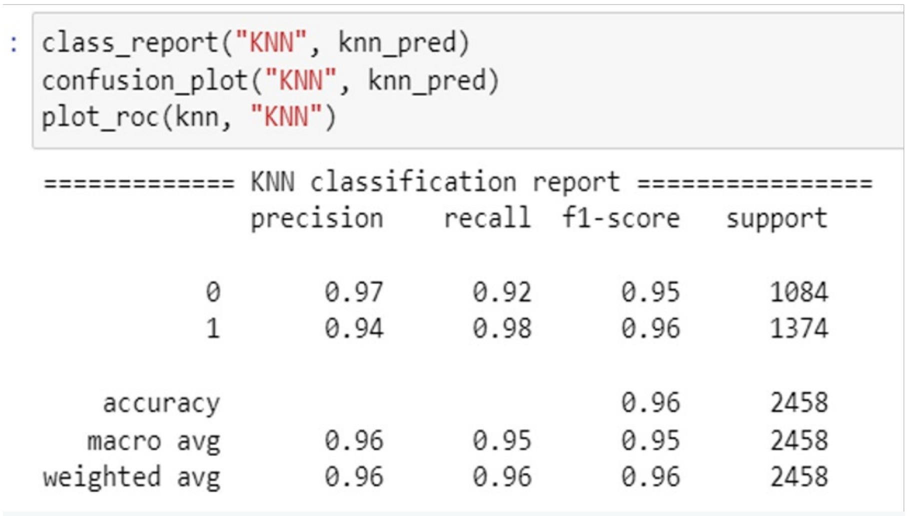
```
class_report("Random Forest", forest_pred)
```

```
============= Random Forest classification report ================
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      1084
           1       0.94      0.96      0.95      1374

    accuracy                           0.95      2458
   macro avg       0.95      0.94      0.94      2458
weighted avg       0.95      0.95      0.95      2458
```

a. RF Classification Report



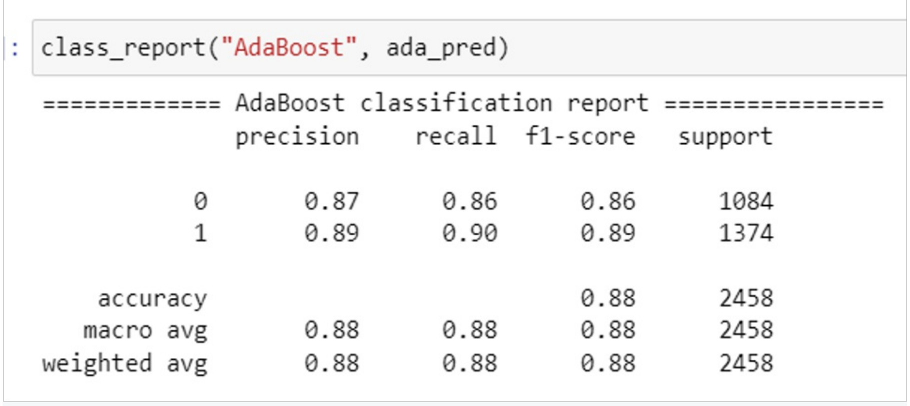
```
class_report("KNN", knn_pred)
confusion_plot("KNN", knn_pred)
plot_roc(knn, "KNN")
```

```
============= KNN classification report ================
              precision    recall  f1-score   support

           0       0.97      0.92      0.95      1084
           1       0.94      0.98      0.96      1374

    accuracy                           0.96      2458
   macro avg       0.96      0.95      0.95      2458
weighted avg       0.96      0.96      0.96      2458
```

b. KNN Classification Report



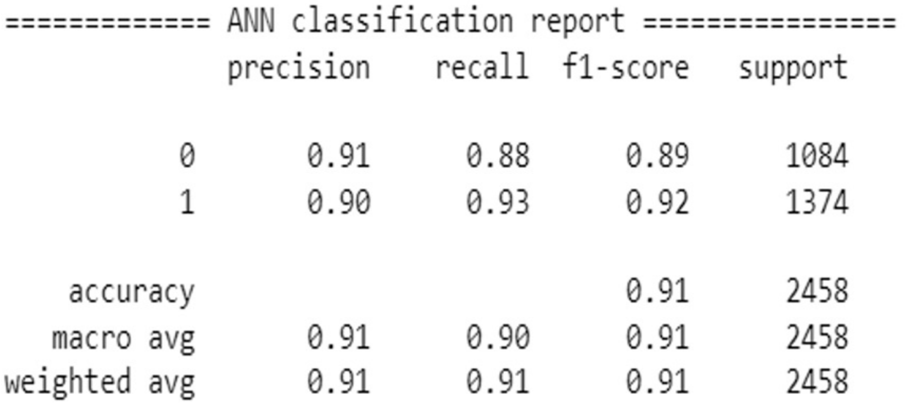
```
class_report("AdaBoost", ada_pred)
```

```
============= AdaBoost classification report ================
              precision    recall  f1-score   support

           0       0.87      0.86      0.86      1084
           1       0.89      0.90      0.89      1374

    accuracy                           0.88      2458
   macro avg       0.88      0.88      0.88      2458
weighted avg       0.88      0.88      0.88      2458
```

c. Adaboost Classification Report



```
============= ANN classification report ================
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      1084
           1       0.90      0.93      0.92      1374

    accuracy                           0.91      2458
   macro avg       0.91      0.90      0.91      2458
weighted avg       0.91      0.91      0.91      2458
```

d. ANN Classification Report

References

- Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, 1-24.
- Rahman, M., & Kumar, V. (2020). Machine learning-based customer churn prediction in banking. In 2020 4th International Conference on Electronics, communication and Aerospace Technology (ICECA) (pp. 1196-1201). IEEE.
- Upreti, K., Verma, M., Agrawal, M., Garg, J., Kaushik, R., Agrawal, C.,... & Narayanasamy, R. (2022). Prediction of mechanical strength by using an artificial neural network and random forest algorithm. Journal of Nanomaterials, 2022.
- Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. Remote Sensing, 12(10), 1687.
- Abiodun, O.J., and Wreford, A.I. (2023) Stroke Prediction Using Smote for Data Balancing, XGBoost and KNN Ensemble Algorithms. Journal of Applied Physical Science International, 15 (1). pp. 42-53. ISSN 2395-5279.
- Zhou, C., Li, M., & Yu, S. (2022). Intelligent Grouping Method of Science and Technology Projects Based on Data Augmentation and SMOTE. Applied Artificial Intelligence, 36(1), 2145637.