**ML MINI PROJECT**

**NAME: Priyal Kamdar (60004210111)**
**Shashwat Shah (60004220126)**
**Falguni Parmar (60004220130)**

**TITLE :**

A data-driven approach to predict the success of bank telemarketing.

**KEYWORDS :**

Bank telemarketing, Predictive analytics, Random Forest, Features selection, SVM, Classification , KNN.

**ABSTRACT :**

This research paper presents the development and evaluation of a machine learning model designed to predict the effectiveness of bank marketing campaigns using a dataset obtained from a Portuguese banking institution's direct marketing efforts. The model incorporates various customer demographic and behavioral attributes to make predictions, employing multiple algorithms and evaluating its performance through metrics including accuracy, F1-score, precision, recall, and roc_auc_score. The study provides insights into the efficacy of different machine learning techniques in optimizing marketing strategies for financial institutions, contributing to the advancement of data-driven decision-making in the banking sector.

**INTRODUCTION :**

Classification and prediction algorithms serve as the cornerstone of predictive analytics, offering a systematic approach to uncovering hidden patterns and trends in data. Decision Trees, for instance, employ a hierarchical structure to recursively partition the data based on different attributes, making them well-suited for both classification and regression tasks. Similarly, k-Nearest Neighbors (kNN) relies on the principle of proximity, where data points are classified based on the majority vote of their nearest neighbors. This algorithm is particularly effective for handling noisy data and does not require explicit model training, making it suitable for real-time applications.

On the other hand, Support Vector Machines (SVM) excel in finding the optimal hyperplane that separates data points into different classes with the maximum margin. SVMs are highly versatile and can handle both linear and nonlinear relationships, making them a preferred choice for complex classification tasks. Additionally, Random Forest (RF) algorithms utilize an ensemble of decision trees to improve prediction accuracy and mitigate overfitting. By aggregating the

predictions of multiple trees, Random Forests offer robust performance and can handle high-dimensional datasets efficiently. Overall, the diverse array of classification and prediction algorithms in machine learning empowers data scientists and analysts to extract valuable insights and drive informed decision-making in a wide range of applications.

This project focuses on utilizing machine learning techniques to predict the effectiveness of bank marketing campaigns. The data used in the project is provided by a Portuguese banking institution and includes input variables such as age, job, marital status, education, and balance etc. The goal of this project is to develop a classification model that can accurately predict the effectiveness of bank marketing campaigns. Through the use of machine learning algorithms and techniques, the model will be able to classify a client's response to a campaign as either positive or negative.

This project will provide insight into how different input variables can affect the effectiveness of bank marketing campaigns and help banks better target their customers. The data set contained details about bank marketing campaigns. Descriptive statistics were computed for each variable as part of the analysis, and visualizations were made to investigate the relationships between the various variables. We created a number of graphs, such as a distplot, count plot, bar plot, pair plot, heatmap, and boxplot, to gain insight from the dataset. There are 45211 observations in this dataset and 17 columns with the following names: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, and y (target variable).

There are no duplicate values in this dataset. The 10 categorical variables in this dataset are: employment, marital, education, default, housing, loan, contact, month, poutcome, and y. This dataset contains seven numerical variables: age, balance, day, duration, campaign, pdays, and prior. For the variables job, education, contact, and poutcome, the number of unknown tagged values are 288; 1857; 13020; and 36959, respectively. Unknown tagged values can be treated as null since they are not defined and can be taken out of features by treatment. Replaced null values with their respective modes for features like contact, education, and job. Moreover, features with more than 50% null values were eliminated because they were useless and negatively impacted model performance. Outliers are treated using the interquartile range for the variables age, balance, duration, campaign, p-days, and previous.

This project utilizes machine learning techniques to predict the effectiveness of bank marketing campaigns using data from a Portuguese banking institution. With input variables such as age, job, marital status, and education, the goal is to develop a classification model to classify clients' responses as positive or negative. Descriptive statistics and visualizations, including distplot, count plot, and pair plot, were employed to analyze relationships between variables. The dataset

comprises 45211 observations and 17 columns, with categorical variables like employment, marital status, and education, as well as numerical variables such as age and balance. Treatment of unknown tagged values as null and handling of outliers using interquartile range were essential steps in data preprocessing. Features with over 50% null values were eliminated to enhance model performance.

**LITERATURE REVIEW :**

In Selecting critical features for data classification based on machine learning methods by Rung-Ching Chen, Christine Dewi, Su-Wen Huang and Rezzy Eko Caraka [1] The significance of feature selection in high-dimensional datasets is the main topic of this study, especially with regard to machine learning applications like classification. The study uses three different datasets for its feature selection: Bank Marketing, Car Evaluation Database, and Human Activity Recognition Using Smartphones. Its primary algorithm is Random Forest. It assesses the effectiveness of several machine learning models, highlighting the influence of feature selection on classification performance and accuracy. These models include Random Forest, Support Vector Machines, K-Nearest Neighbors, and Linear Discriminant Analysis. Through an analysis of several feature selection techniques, including varImp(), Boruta, and Recursive Feature Elimination (RFE), the study emphasizes how important critical feature selection is to achieving better classification results. By evaluating feature relevance, contrasting machine learning models, and offering suggestions for further research, the study makes a contribution.

In Predicting the Success of Bank Telemarketing using various Classification Algorithms by Muneeb Asif [2] In conjunction with feature selection strategies like best-subset Logistic Regression, LASSO, and Random Forest, this thesis investigates the effectiveness of several classification techniques, such as Support Vector Machine, Decision Trees, Random Forest, and Artificial Neural Network. The goal is to minimize dimensionality without sacrificing prediction performance and accuracy. The application context is telemarketing in the banking industry, which highlights the significance of efficient decision-making aided by data mining and Decision Support Systems (DSS). In supervised learning for outcome prediction, in particular, machine learning (ML) is essential. Authentic data is taken from the University of California, Irvine database, and statistical methods are applied to feature selection. The findings show that the accuracy of a smaller set of variables chosen by Random Forest is similar to the accuracy of the entire model across classification techniques, with Random Forest outperforming the others. Notably, Random Forest found that age, balance, and duration are the most influential characteristics.

In Application of Selected Supervised Classification methods to Bank Marketing Campaign by Danieal Grzonka, Grazyna Suchaka and Barbara Borowik [3] This paper explores the use of

classification models, particularly decision trees and ensemble methods like bagging, boosting, and random forests, to predict the effectiveness of marketing campaigns, focusing on a telemarketing campaign by a Portuguese bank. By analyzing real campaign data, the study identifies key attributes influencing client decisions. Despite omitting certain parameters known only post-campaign, the models demonstrate significant predictive capability, with random forests yielding the best results. However, the randomness inherent in the modeling process highlights the need for careful interpretation. Nonetheless, the findings underscore the potential of decision tree-based methods in optimizing bank marketing strategies.

In Research Paper Classification using Supervised Machine Learning Techniques by Shovan Chowdhury and Marco P. Schoen [4] The problem of effectively classifying research papers into different fields, like business, social science, and science, is the focus of this study. The study attempts to automate this process, thereby relieving the laborious task of manual classification by utilizing supervised machine learning techniques and text classification. Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbor, and Decision Tree are the four classification algorithms that are used after a balanced dataset consisting of abstracts from different research domains is created. Along with vector representation techniques such as Term Frequency Inverse Document Frequency (TF-IDF) and Bag of Words, text preprocessing techniques like tokenization and stemming are applied. SVM performs better than the other methods, according to the results, though KNN and Naïve Bayes also do fairly well. Overall, the research highlights machine learning's potential.

Machine Learning: Algorithms, Real-World Applications and Research Directions by Iqbal H. Sarker [5] In the context of the Fourth Industrial Revolution (4IR), large and diverse datasets are frequently analyzed. This paper investigates the importance of machine learning (ML) techniques in this regard. It covers a range of machine learning algorithms, such as supervised, unsupervised, semi-supervised, and reinforcement learning, emphasizing how they can be used in real-world settings like e-commerce, cybersecurity, smart cities, healthcare, and agriculture. Emphasis is placed on classification, a crucial component of supervised learning, covering binary, multiclass, and multi-label classification scenarios. The function of popular classification algorithms like Random Forests, Decision Trees, Naive Bayes, and Linear Discriminant Analysis in creating data-driven systems is explained. The study also emphasizes how crucial it is to comprehend the fundamentals and subtleties of various machine learning algorithms in order to effectively handle a range of application requirements in the context of Industry 4.0.

In A novel ensemble approach for estimating the competency of bank telemarketing by WeiGuo, YaoYao, Lihua Liu & Tong Shen [6] In order to forecast bank customers' long-term deposit subscription status, this study explores the field of bank telemarketing. For this purpose, it assesses how well four metaheuristic algorithms—social ski-driver (SSD), harmony search algorithm (HSA), future search algorithm (FSA), and electromagnetic field optimization

(EFO)—train artificial neural networks (ANNs). In terms of prediction accuracy, the hybrid EFO-ANN model proves to be the best, outperforming traditional methods such as logistic regression, decision trees, and support vector machines (SVM). Furthermore, prior studies that utilized machine learning techniques, like ANN and naive Bayes, for comparable predictive tasks in bank telemarketing are referenced, emphasizing the importance of classification algorithms in enhancing marketing tactics and outreach to customers.

In Machine Learning Algorithm for Classification by Haoyuan Tan [7] This study emphasizes supervised learning while examining the performance of several machine learning classifiers in classifying tasks. Popular methods like Random Forest, XGBoost, Naive Bayes, Support Vector Machines (SVM), Gaussian Mixture Model (GMM), and Random Forest are covered, along with examples of how they are used in various datasets. The results show that while overall performance of all classifiers is good, the accuracy is highly dependent on the task's complexity. For example, SVM performs poorly in text classification while Random Forest achieves the best accuracy in remote sensing. Overall, the study emphasizes how important it is to select classifiers based on the particular classification tasks in order to achieve the best results.

In Bank Marketing Data Classification Using Machine Learning by Dr. Smitha Shekar B and Pooja A [8] This paper investigates the application of machine learning concepts, particularly supervised learning algorithms, in analyzing bank marketing data for term deposit subscription prediction. Utilizing Python as the programming language, the study focuses on building a predictive model using the Naïve Bayes classifier algorithm. The dataset, sourced from the UCI Machine Learning Repository and Kaggle, pertains to bank marketing campaigns conducted via phone calls. By assessing metrics like accuracy, precision, recall, and F1 score, the study aims to enhance the bank's targeting strategy towards potential customers who are more likely to respond positively to marketing calls, ultimately improving campaign effectiveness and customer response rates. Additionally, it discusses various machine learning techniques, including Random Forest and Naïve Bayes, emphasizing their roles in data analysis and decision-making in the banking sector.

In Enhancing bank marketing strategies with ensemble learning: Empirical analysis by Xing Tang and Yusi Zhu [9] This paper introduces a customer demand learning model for bank marketing, aiming to enhance market segmentation and competitiveness in the e-commerce banking sector. Leveraging ensemble learning techniques, the study compares the predictive performance of random forest and support vector machine (SVM) models. Results indicate that the ensemble learning model achieves higher accuracy (92%) compared to the SVM model (87%), leading to improved targeted marketing, customer relationship maintenance, and product marketing success. The research contributes valuable insights for bank marketing decision-making, emphasizing the significance of predictive modeling in optimizing marketing strategies and enhancing overall marketing capabilities in the banking industry. Additionally, the

study suggests avenues for future research to further refine the predictive model's accuracy and applicability.

In Classification and prediction‑based machine learning algorithms to predict students' low and high programming performance by Aykut Durak and Vahide Bulut [10] This study explores various estimation and classification techniques, including Decision Trees, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Naive Bayes, Logistic Regression (LR), and Random Forest (RF), to analyze programming performance factors among students. Decision trees demonstrate superior performance in predicting programming performance, while other classifiers exhibit differences in handling the dependent variable. Factors such as gender and educational level influence variables like computational identity, computational thinking perspective, and programming empowerment scores. Notably, these variables remain consistent across different levels of academic success. By extending existing models and incorporating programming-related variables, the study provides valuable insights into understanding and predicting programming performance among students.

**RESEARCH OBJECTIVE :**

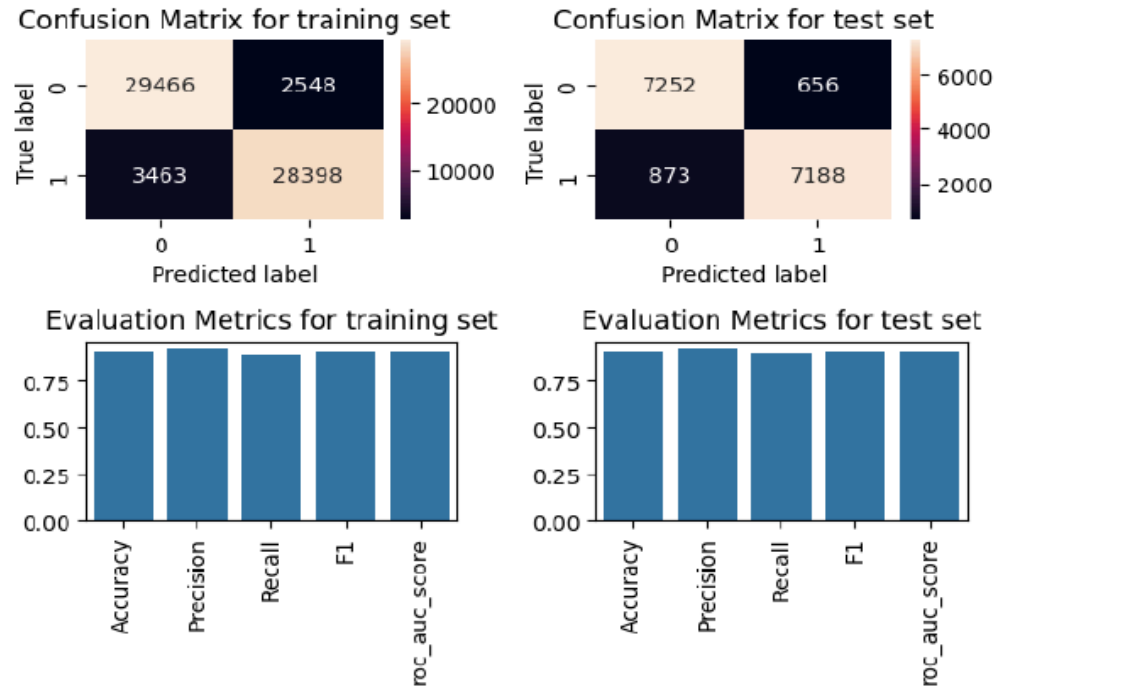
The aim of this study is to create and assess a machine learning-based method for estimating bank marketing campaigns' efficacy. By utilizing a dataset that was acquired through direct marketing initiatives of a Portuguese banking institution, the research endeavors to build predictive models that encompass various customer demographic and behavioral characteristics. Through the use of a variety of machine learning algorithms and performance metrics, including accuracy, precision, recall, F1-score, and roc_auc_score, the study aims to shed light on how best to optimize marketing tactics for the banking industry. The ultimate objective is to use data-driven decision-making processes to increase campaign effectiveness and customer engagement.

## RESEARCH METHODOLOGY :

### 1. LOGISTIC REGRESSION

```
# Import Logistic Regression algorithm in envirnoment
from sklearn.linear_model import LogisticRegression
# Fitting Logistic Regression model to training set
Logistic_regression=LogisticRegression(fit_intercept=True, max_iter=10000,random_state=0)
lr=classification_model(X_train, X_test, y_train, y_test, Logistic_regression)
```

```
================================================================================================
Training set evaluation result :

Confusion Matrix:
 [[29466  2548]
 [ 3463 28398]]
Accuracy:  0.905894324853229
Precision:  0.9176630259161119
Recall:  0.8913091240074071
F1 Score:  0.9042941073447226
roc_auc_score:  0.9058594723554246

--------------------------------

Test set evaluation result :

Confusion Matrix:
 [[7252  656]
 [ 873 7188]]
Accuracy:  0.9042519882271902
Precision:  0.9163691993880673
Recall:  0.8917007815407517
F1 Score:  0.9038667085822069
roc_auc_score:  0.9043734054390659
================================================================================================
```

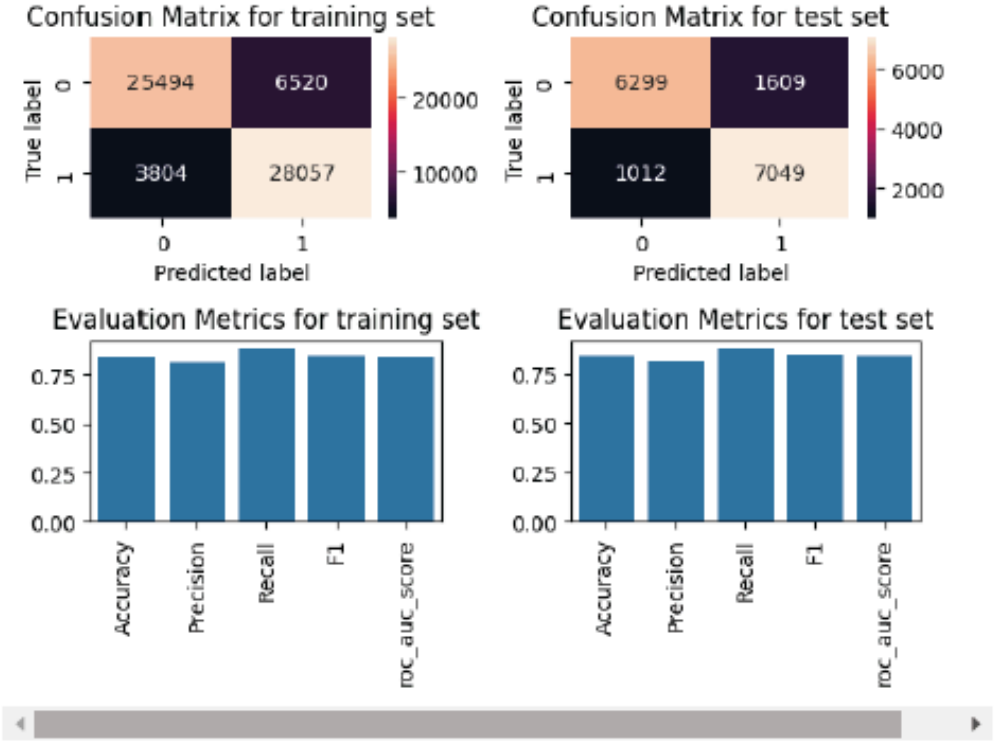



```
================================================================================================
```

## 2. DECISION TREE:

```
# Import Decision Tree algoritham in envirnoment
from sklearn.tree import DecisionTreeClassifier
# Fitting Decision Tree model to training set
classifier_dt = DecisionTreeClassifier(criterion='entropy', max_leaf_nodes=10, random_state=0)
dt=classification_model(X_train, X_test, y_train, y_test, classifier_dt)
```

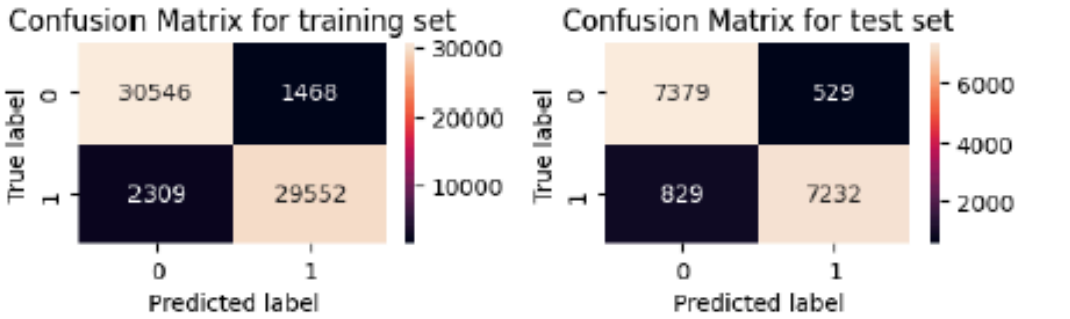
```
DecisionTreeClassifier(criterion='entropy', max_leaf_nodes=10, random_state=0)
================================================================================
Training set evaluation result :

Confusion Matrix:
 [[25494  6520]
 [ 3804 28057]]
Accuracy:  0.8383718199608611
Precision:  0.811435347196113
Recall:  0.8806063839804149
F1 Score:  0.8446070020169181
roc_auc_score:  0.838472742811723

------------------------------

Test set evaluation result :

Confusion Matrix:
 [[6299 1609]
 [1012 7049]]
Accuracy:  0.8358694971507296
Precision:  0.8141603141603142
Recall:  0.8744572633668279
F1 Score:  0.8432322507326994
roc_auc_score:  0.8354962088204904
================================================================================
```

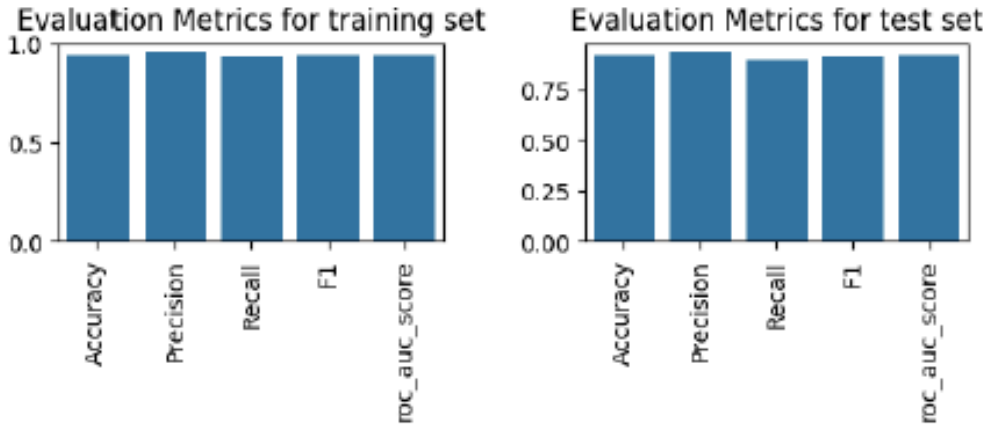



## 3. K-NEAREST NEIGHBOUR

```
# Import KNN algoritham in envirnoment
from sklearn.neighbors import KNeighborsClassifier
# Fitting model to training set
classifier_knn = KNeighborsClassifier(n_neighbors=5)
knn=classification_model(X_train, X_test, y_train, y_test, classifier_knn)
```

```
================================================================================
Training set evaluation result :

Confusion Matrix:
 [[30546  1468]
 [ 2309 29552]]
Accuracy:  0.9408688845401174
Precision:  0.952675693101225
Recall:  0.9275289538934748
F1 Score:  0.9399341613523958
roc_auc_score:  0.9408370077145266

--------------------------------

Test set evaluation result :

Confusion Matrix:
 [[7379  529]
 [ 829 7232]]
Accuracy:  0.9149602354561964
Precision:  0.9318386805823992
Recall:  0.8971591613943679
F1 Score:  0.9141701428390847
roc_auc_score:  0.9151324385626367
================================================================================
```




**RESULT:**

Based on the performance metrics provided for the three classification algorithms (Logistic Regression, Decision Tree, and K-Nearest Neighbors), we can make several observations to determine the most suitable model for the bank dataset:
Accuracy: KNN has the highest accuracy of approximately 91.50%, followed by Logistic Regression with around 90.43%, and Decision Tree with about 83.59%.
Precision: KNN has the highest precision of approximately 93.18%, followed by Logistic Regression with around 91.64%, and Decision Tree with about 81.42%.
Recall (Sensitivity): KNN has a recall of approximately 89.72%, followed by Logistic Regression with around 89.17%, and Decision Tree with about 87.45%.
F1 Score: KNN has the highest F1 score of approximately 91.42%, followed by Logistic Regression with around 90.39%, and Decision Tree with about 84.32%.
ROC AUC Score: KNN has the highest ROC AUC score of approximately 91.51%, followed by Logistic Regression with around 90.44%, and Decision Tree with about 83.55%.
Based on these observations, the K-Nearest Neighbors (KNN) model appears to be the most suitable for the bank dataset among the three algorithms. It consistently outperforms Logistic Regression and Decision Tree across all key performance metrics such as accuracy, precision, recall, F1 score, and ROC AUC score. The higher values in these metrics indicate better overall performance and predictive capability of the model.

**RESEARCH GAP :**
While the presented research provides valuable insights into the development and evaluation of a machine learning model for predicting the effectiveness of bank marketing campaigns, there exist several potential avenues for further exploration and research. Specifically, the following research gaps could be addressed:

1. Exploration of Additional Features: The study mentions incorporating various customer demographic and behavioral attributes into the predictive model. However, there might be other relevant features that could enhance the model's predictive performance. Future research could focus on exploring additional features, such as socio-economic indicators, customer preferences, or external economic factors, to further refine the predictive capabilities of the model.

2. Model Comparison and Selection: The research mentions employing multiple algorithms for prediction but does not delve deeply into the comparative analysis of these algorithms. Future studies could focus on systematically comparing different machine learning algorithms, including ensemble methods and deep learning approaches, to identify the most effective model for predicting marketing campaign effectiveness in the banking sector.

4. Generalizability and External Validation: The research focuses on a specific dataset from a Portuguese banking institution, raising questions about the generalizability of the findings to other banking contexts or geographic regions. Future studies could address this limitation by validating the predictive model using data from multiple banking institutions or conducting cross-validation across diverse datasets to assess its external validity and applicability in different settings.

5. Ethical and Privacy Considerations: The study does not explicitly address ethical or privacy considerations related to the use of customer data for predictive modeling in the banking sector. Future research could explore the ethical implications of deploying machine learning models in marketing campaigns, including issues related to data privacy, fairness, transparency, and the potential impact on consumer trust and autonomy.

Addressing these research gaps could contribute to a deeper understanding of the challenges and opportunities associated with leveraging machine learning for optimizing marketing strategies in the banking sector, ultimately advancing the field of data-driven decision-making in financial institutions.

**CONCLUSION :**

The K-Nearest Neighbors (KNN) algorithm performs the best among the three models, consistently demonstrating higher accuracy, precision, recall, F1 score, and ROC AUC score.
KNN achieves an accuracy of approximately 91.50%, precision of 93.18%, recall of 89.72%, F1 score of 91.42%, and ROC AUC score of 91.51%.
Logistic Regression and Decision Tree algorithms also show competitive performance but are slightly outperformed by KNN across all metrics.
Therefore, based on the evaluation results, KNN is deemed the most suitable model for the bank dataset, showcasing superior predictive capability and overall effectiveness in classification tasks.

**REFERENCES :**

**1)** Selecting critical features for data classification based on machine learning methods by Rung-Ching Chen, Christine Dewi, Su-Wen Huang and Rezzy Eko Caraka.

**2)** Predicting the Success of Bank Telemarketing using various Classification Algorithms by Muneeb Asif.

**3)** Application of Selected Supervised Classification methods to Bank Marketing Campaign by Danieal Grzonka, Grazyna Suchaka and Barbara Borowik.

**4)** Research Paper Classification using Supervised Machine Learning Techniques by Shovan Chowdhury and Marco P. Schoen**.**

**5)**Machine Learning: Algorithms, Real-World Applications and Research Directions by Iqbal H. Sarker.

**6)** A novel ensemble approach for estimating the competency of bank telemarketing by WeiGuo, YaoYao, Lihua Liu & Tong Shen.

**7)** Machine Learning Algorithm for Classification by Haoyuan Tan.

**8)** Bank Marketing Data Classification Using Machine Learning by Dr. Smitha Shekar B and Pooja A.

**9)** Enhancing bank marketing strategies with ensemble learning: Empirical analysis by Xing Tang and Yusi Zhu.

**10)** Classification and prediction-based machine learning algorithms to predict students' low and high programming performance by Aykut Durak and Vahide Bulut.