Banking customer churn prediction

OBJECTIVES

The primary goal of our analysis is to leverage banking data to predict customer behavior regarding their likelihood of switching to another bank. To achieve this, we implemented and compared multiple supervised learning algorithms, evaluating their performance in terms of predictive accuracy.

The motivation behind this research lies in its strategic relevance for financial institutions: such predictive models serve as a key informational asset, supporting the design and implementation of customer retention strategies.

1. DATASET AND EXPLORATORY ANALYSIS (Source: Kaggle)

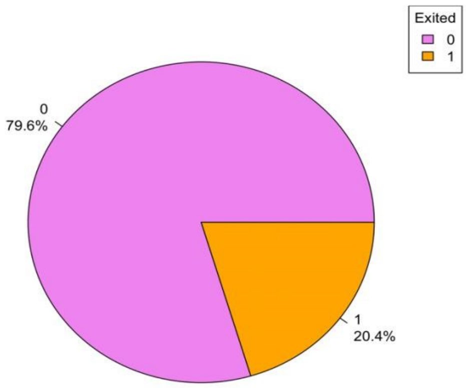
The dataset comprises 10,000 observations related to bank customers from three geographical regions — France, Germany, and Spain. It includes information across 14 variables, of which we selected 11 deemed most relevant for our predictive modeling and analytical objectives.

- **CreditScore:** the customer's credit rating.
- **Geography:** the customer's country of residence.
- **Gender:** the customer's gender.
- **Age:** the customer's age.
- **Tenure:** the number of years the customer has been with the bank.
- **Balance:** the customer's current account balance.
- **NumOfProducts:** the number of banking products held by the customer.
- **HasCrCard:** indicates whether the customer holds a credit card (binary: yes/ no).
- **IsActiveMember:** indicates whether the customer is considered an active member (binary: yes/no).
- **EstimatedSalary:** the customer's estimated annual income.
- **Exited:** indicates whether the customer has left the bank (binary: yes/no).

There are no missing values or duplicate records in the dataset.

1.1. Target Variable Frequencies

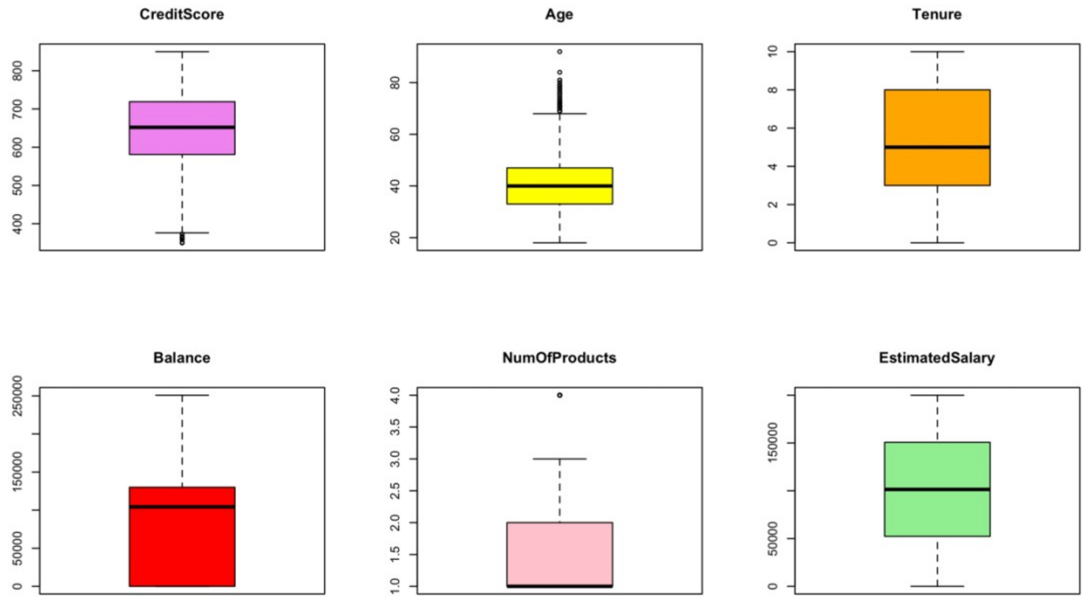
Since our focus is on the Exited variable, the first step was to examine the relative frequency distribution in order to understand how observations are distributed with respect to this target variable.



We observed an imbalance in favor of customers leaving the bank. Consequently, an undersampling technique was applied to rebalance the dataset. The resulting dataset consists of 4,140 records, with 50.8% representing customers who stayed with the bank and 49.2% representing those who exited.

1.2. Boxplot

Subsequently, we constructed boxplots for the variables identified as numerical.



We observed the presence of outliers in the Age variable. However, since there are numerous observations with values between 60 and 80 years, we decided to also analyze the behavior of this customer segment.

Regarding the outlier in NumOfProducts, being a single occurrence, it does not significantly affect the variable's distribution.

Some variables exhibit high variability, such as Tenure and EstimatedSalary, while others are asymmetrically distributed, like Balance and NumOfProducts. Many observations are concentrated between 100,000 and 130,000, whereas below these values there is greater dispersion.

Additionally, all customers hold at least one banking product.

1.3. Summary Statistics

The Gender variable has a mean of 0.51, indicating that there are slightly more males than females in our dataset.

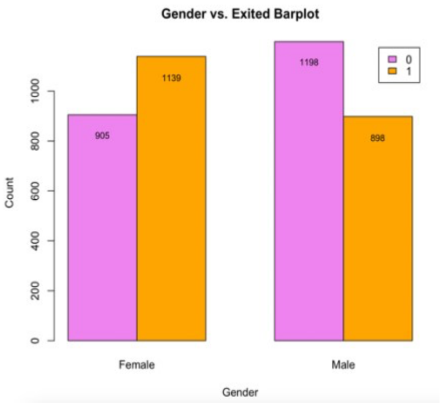
The HasCrCard variable has an average value of approximately 0.71, meaning that more than half of the customers hold a credit card.

The IsActiveMember variable has a mean of 0.46, indicating that the majority of customers are not considered active members of the bank.

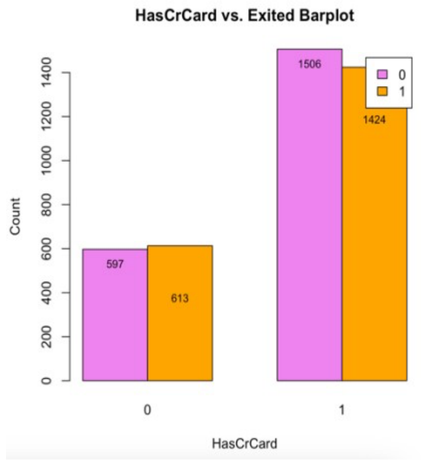
The Exited variable has a mean of 0.49, showing that the proportion of customers who remain with or leave the bank is roughly equal.

1.4. Visualizations in Relation to the Target Variable

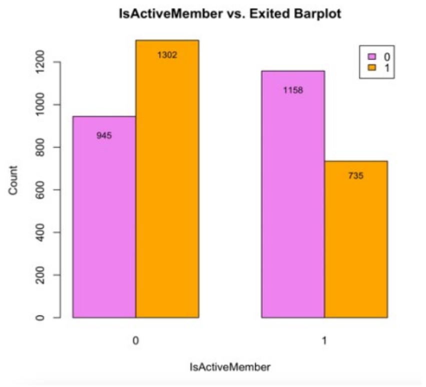
Another key aspect of the exploratory analysis is examining individual variables in relation to the target variable Exited. For categorical variables, we used barplots as the method of representation.



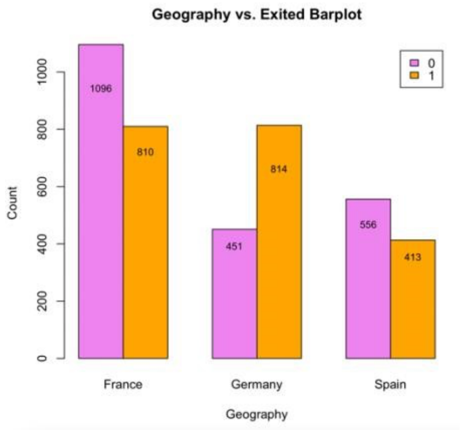
Regarding the Gender variable, females are more likely to exit the bank compared to males.



The HasCrCard variable does not appear to be significant in distinguishing between customers who stay or leave the bank. When deciding whether to switch banks, this factor may not have had an influence, as a customer could hold a credit card without actively using it, or simply may not have considered the possession of a credit card when making the decision to change banks.

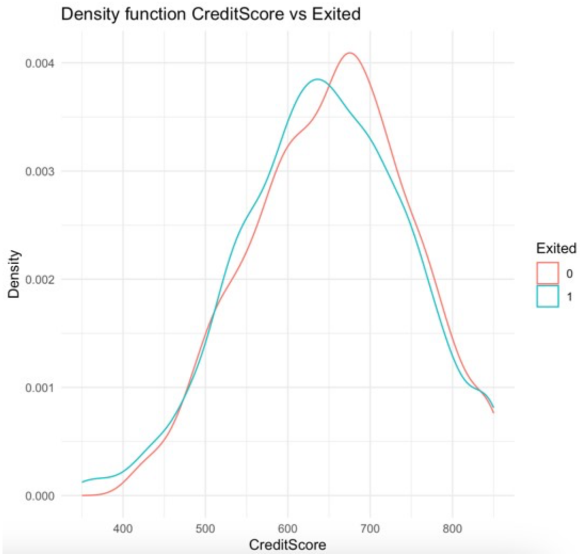


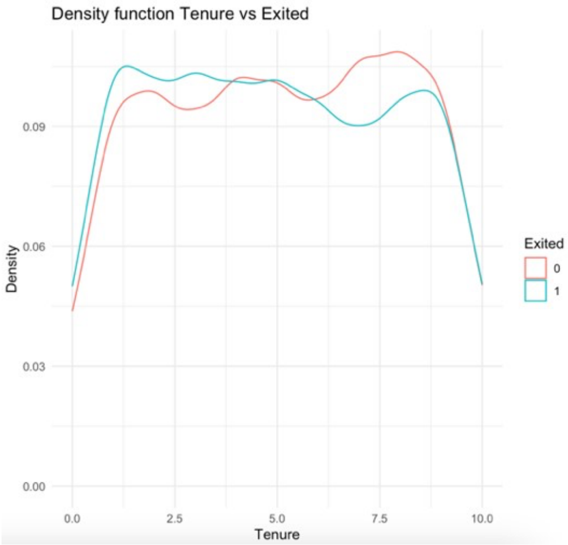
It is observed that more active members tend to leave the bank less frequently, and vice versa. A possible explanation is that they may be more inclined to stay due to the personalized services offered by the bank.



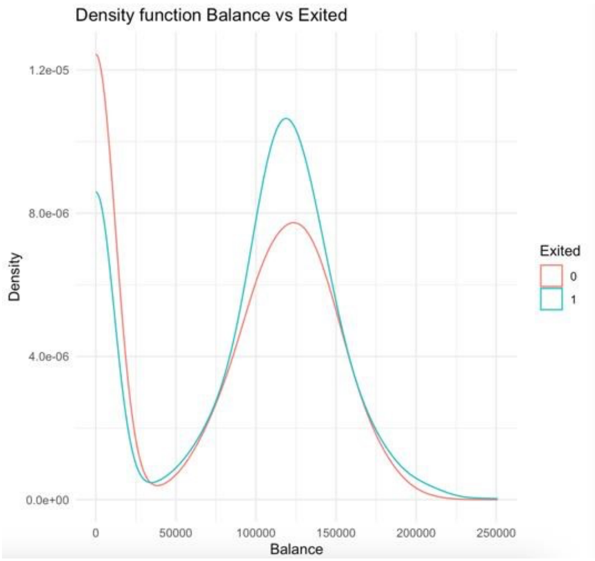
In France, there is a stronger tendency for customers to remain with their bank, and a similar pattern is observed in Spain.
In contrast, in Germany the opposite trend is more pronounced. The more consumer-oriented German culture leads customers to actively seek the best offers and compare financial products.

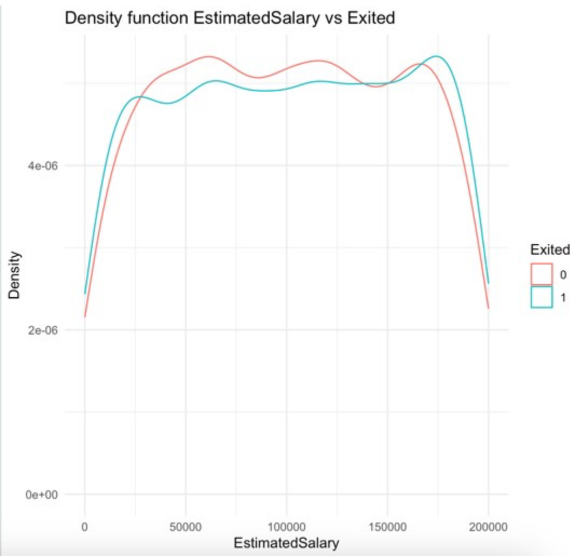
For the numerical variables, we used density plots as the method of representation, again analyzed in relation to the target variable.



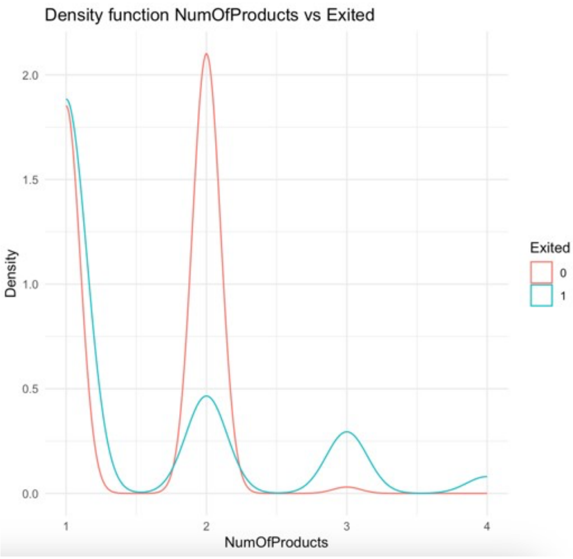


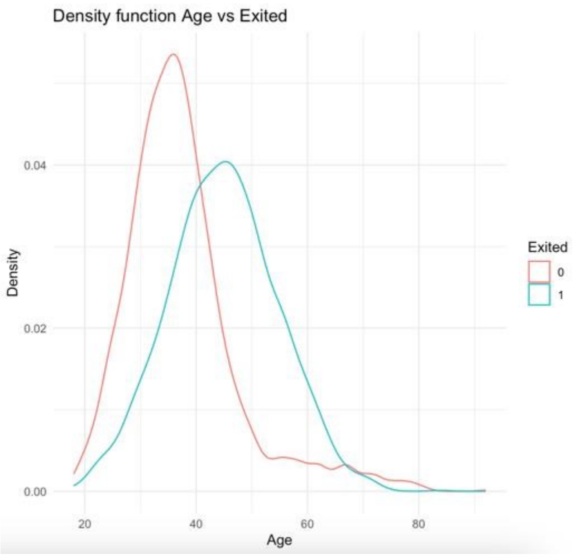
Whether a customer stays with or leaves the bank does not appear to be strongly influenced by CreditScore**;** indeed, this parameter is primarily used internally by banks to assess a customer's creditworthiness.
It is observed that customers tend to leave the bank during the first four years of tenure, whereas those with more than six years tend to stay, likely due to trust built over time in the relationship with the bank.





Individuals with a Balance between approximately 75,000 and 150,000 tend to be more likely to leave the bank, which could be influenced by offers from other financial institutions. However, this represents a very small percentage of customers.
The EstimatedSalary variable does not appear to be particularly relevant in predicting customer exit, suggesting that the decision to leave is made independently of personal income.

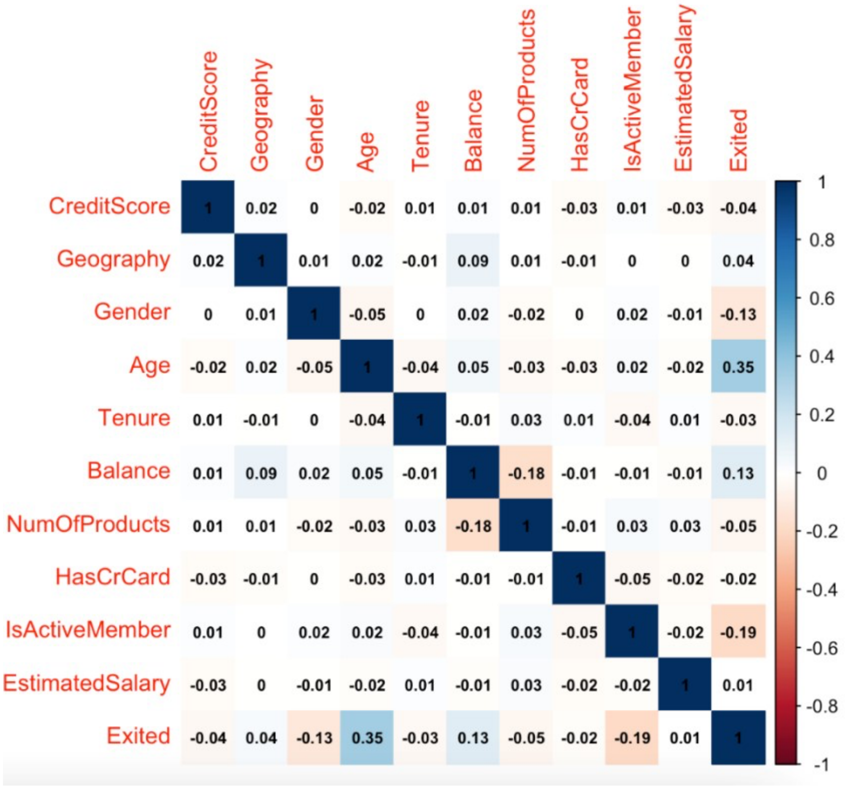




An individual holding two products is more likely to remain with the bank. For one product, the outcome is indifferent, whereas for three or four products, the likelihood of leaving the bank slightly increases. This may be due to the fact that customers interested in holding more than the standard products (one or two) might seek better deals from other banks.
For the Age variable, customers between 40 and 70 years old are more likely to leave the bank, whereas those under 40 tend to stay, as younger customers are generally offered more incentives. Additionally, older customers may have accumulated sufficient savings, reducing their reliance on banking services.

1.5. Correlation Matrix

Per costruire la matrice di correlazione tra le variabili del dataset prima di tutto abbiamo trasformato Gender e Geography in factor.



	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1	0.02	0	-0.02	0.01	0.01	0.01	-0.03	0.01	-0.03	-0.04
Geography	0.02	1	0.01	0.02	-0.01	0.09	0.01	-0.01	0	0	0.04
Gender	0	0.01	1	-0.05	0	0.02	-0.02	0	0.02	-0.01	-0.13
Age	-0.02	0.02	-0.05	1	-0.04	0.05	-0.03	-0.03	0.02	-0.02	0.35
Tenure	0.01	-0.01	0	-0.04	1	-0.01	0.03	0.01	-0.04	0.01	-0.03
Balance	0.01	0.09	0.02	0.05	-0.01	1	-0.18	-0.01	-0.01	-0.01	0.13
NumOfProducts	0.01	0.01	-0.02	-0.03	0.03	-0.18	1	-0.01	0.03	0.03	-0.05
HasCrCard	-0.03	-0.01	0	-0.03	0.01	-0.01	-0.01	1	-0.05	-0.02	-0.02
IsActiveMember	0.01	0	0.02	0.02	-0.04	-0.01	0.03	-0.05	1	-0.02	-0.19
EstimatedSalary	-0.03	0	-0.01	-0.02	0.01	-0.01	0.03	-0.02	-0.02	1	0.01
Exited	-0.04	0.04	-0.13	0.35	-0.03	0.13	-0.05	-0.02	-0.19	0.01	1

From the correlation matrix, it is evident that no pair of variables is strongly correlated. Slight correlations are observed for:

- Age and Exited**:** a positive relationship, as seen in the density plots, occurring mainly between 40 and 70 years of age.
- **I**sActiveMember and Exited**:** a negative relationship, also confirmed by the barplots; the more active a customer is, the less likely they are to leave the bank.
- NumOfProducts and Balance: a negative relationship, since the more a customer invests in banking products, the lower their account liquidity.

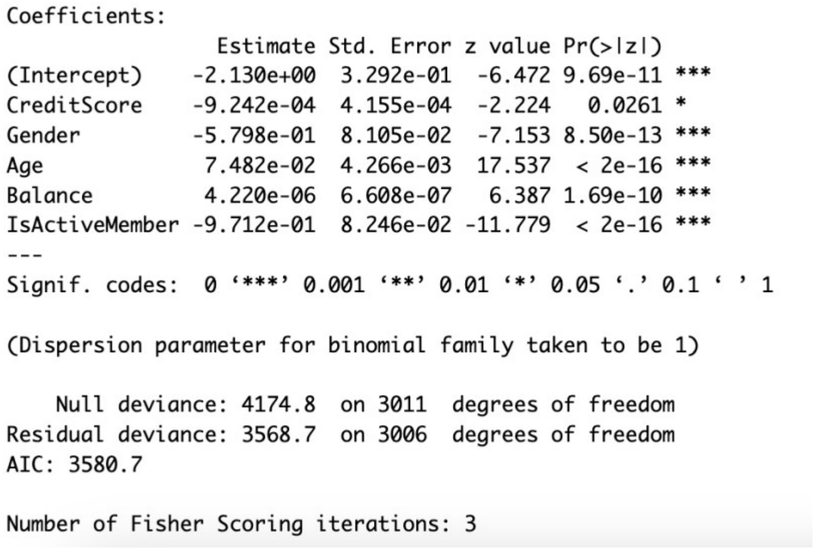
The absence of multicollinearity suggests that applying PCA for dimensionality reduction would have limited benefit, as 75% of the variance is explained only by the eighth principal component.

2. METHODOLOGIES: SUPERVISED ALGORITHMS

Given a target variable to predict, we employed supervised algorithms. In our case, since the target is categorical, the task is a classification rather than regression. For this purpose, the dataset was split into training and testing sets.

2.1. Logistic Regression

For the logistic regression analysis, we initially considered the full model and gradually removed the less significant variables, ultimately explaining customer churn through Credit Score, Gender, Age, Account Balance**,** and Active Membership**.** These five variables were retained for subsequent models as well.
The final model selection was based on a significance level of 0.05**,** along with considerations of AIC and the Likelihood Ratio Test values.



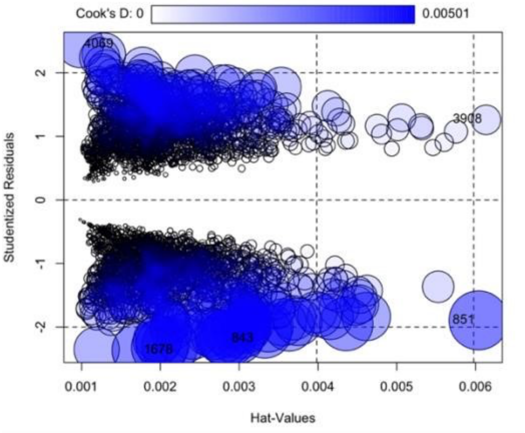
```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -2.130e+00  3.292e-01  -6.472 9.69e-11 ***
CreditScore    -9.242e-04  4.155e-04  -2.224   0.0261 *  
Gender         -5.798e-01  8.105e-02  -7.153 8.50e-13 ***
Age             7.482e-02  4.266e-03  17.537  < 2e-16 ***
Balance         4.220e-06  6.608e-07   6.387 1.69e-10 ***
IsActiveMember -9.712e-01  8.246e-02 -11.779  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4174.8  on 3011  degrees of freedom
Residual deviance: 3568.7  on 3006  degrees of freedom
AIC: 3580.7

Number of Fisher Scoring iterations: 3
```

Regarding the coefficients, the relationship with the target variable is as follows:

- As **Credit Score** increases, individuals tend to remain with the bank;
- **Females** tend to switch banks more than males;
- As **age** increases, individuals are more likely to leave the bank;
- As **Account Balance** increases, individuals tend to leave the bank;
- If an individual is an **active member** of the bank, they tend to stay.

Finally, we evaluated the model’s accuracy on both datasets to check for potential overfitting. With a value of 69.5% on the training set and **69.1%** on the testing set, we confirm that the model fits well for predicting both training and testing data. It should be noted that accuracy is considered better the higher it is; in our case, the result can be considered acceptable but not entirely reliable. Indeed, the misclassification error rate, representing the percentage of incorrectly predicted individuals, is 30.9% in the testing set.

To conclude the logistic regression model analysis, we examined the presence of observations with a significant influence on the regression model using Cook’s distances**.**



As shown in the chart on the left, there are some influential values. For example, observation no. 851 stands out as influential for the model due to its maximum values of Age and Credit Score and minimum Balance**,** when compared with the dataset summary.

2.2. Linear Discriminant Analysis

We decided to implement another technique to discriminate the target variable, LDA, which also allows for dimensionality reduction and imposes stricter assumptions, such as the multivariate normality of features. In our case, this assumption does not hold. We attempted to apply transformations to the variables, but these did not yield a positive result in the test. Therefore, we concluded that this analysis cannot be applied to our dataset.

2.3. K-Nearest Neighbor

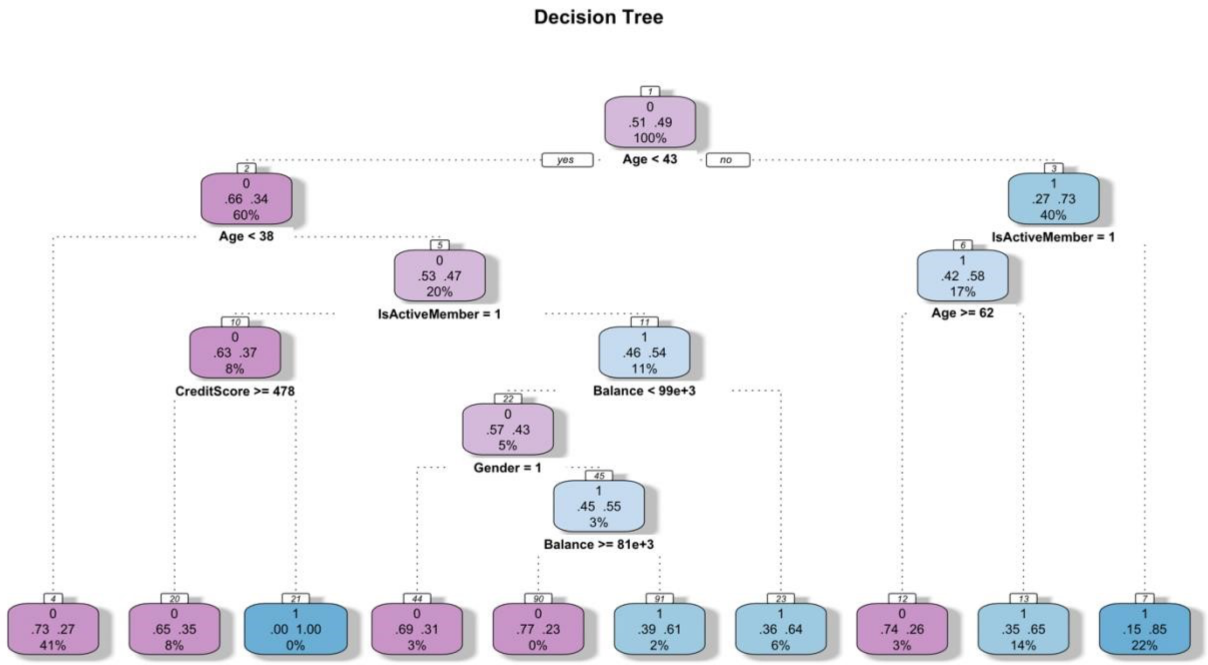
Next, we applied a simple and intuitive classification method, kNN, which makes predictions based on the training data. In our case, the model with the highest average accuracy used 1 nearest neighbor. After implementation, we evaluated the model’s accuracy on both datasets and observed overfitting on the training set. Specifically, the training set achieved a correct prediction rate of 99.9%, while the testing set dropped to **65.9%.** This phenomenon can be attributed to the high computational cost of the algorithm and the low number of neighbors, which may cause the model to be overly influenced by points in that specific training set and result in poor generalization to new data, in our case the testing set. For these reasons, the kNN algorithm’s performance is less satisfactory compared to logistic regression.

2.4. Support Vector Machine

Subsequently, we implemented another classification algorithm that identifies the optimal hyperplane separating observations linearly in the feature space. The results obtained are slightly better than those from previous analyses, likely because SVM can also be applied to data that are not linearly separable through the kernel function (in our case, linear). The accuracy on the training set was 70.28%, while on the testing set it was **70.03%.** We can conclude that, in this case, there is no overfitting.

2.5. Decision Tree

Continuing our analysis, we also adopted the decision tree technique, which allowed us to identify classification rules based on the values that observations take on the variables.
The following image shows the tree configuration, which minimizes the complexity parameter with 10 terminal nodes (9 levels).



We note that the results obtained are largely consistent with the conclusions from the exploratory analysis. Indeed, as age increases, we would expect a higher probability of leaving the bank, and observing the tree, this occurs in most cases. Additionally, another key feature identified in the exploratory analysis is the negative relationship between leaving the bank and being an active member. Again, this relationship is reflected in the structure of the tree.

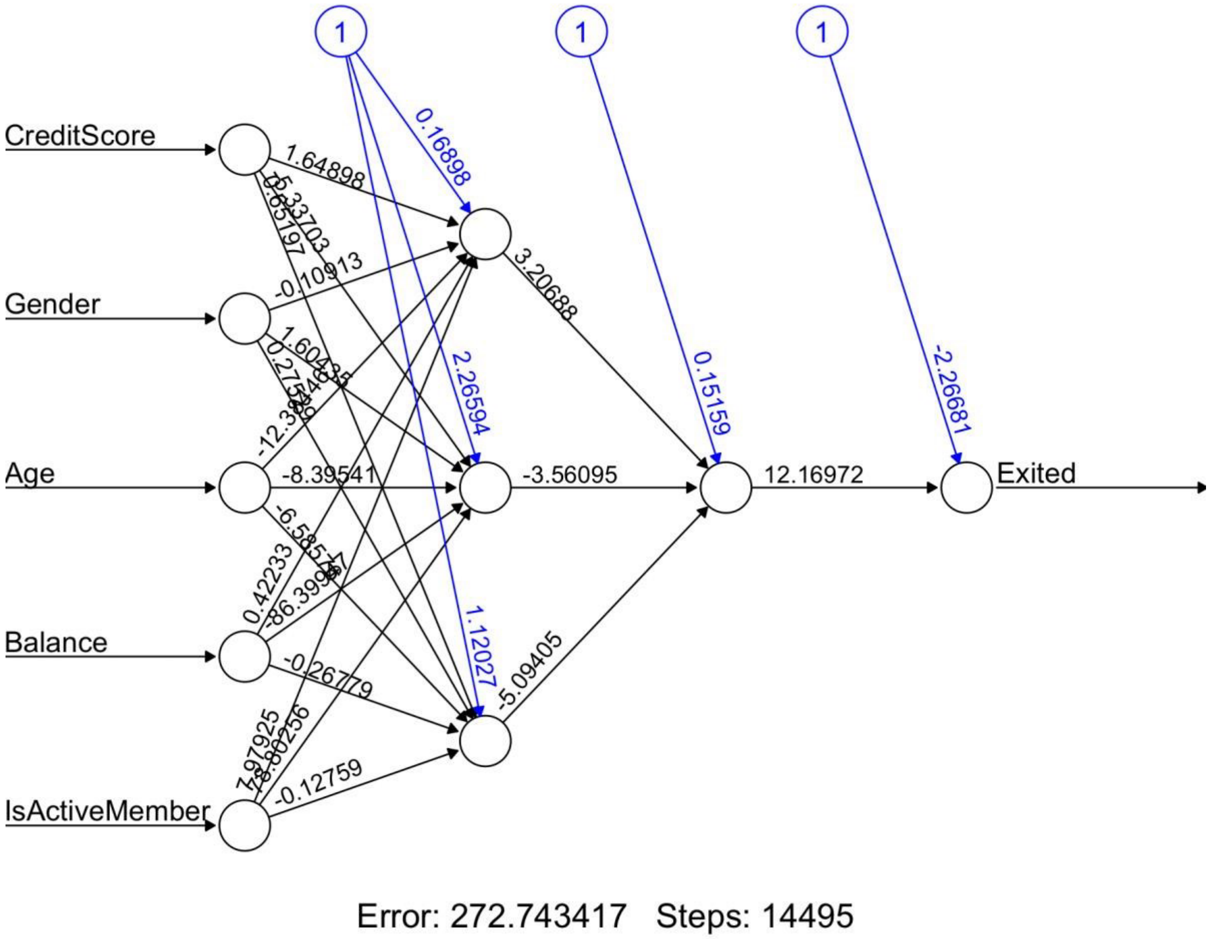
With an accuracy of 73.04% on the training set and **71.28%** on the testing set, the decision tree outperforms the previous algorithms in terms of predictive accuracy. Furthermore, the absence of overfitting is confirmed by the comparable performance between training and testing sets.

2.6. Random Forest

The decision tree technique is often followed by random forest**,** because by aggregating multiple learning models, i.e., multiple decision trees, the overall classification accuracy can be further improved. In our results, however, with an accuracy of 70.42% on the training set and **70.57%** on the testing set, this improvement does not occur, even though the model continues to fit both the training and testing sets well. This is likely due to the presence of some binary variables, which make the dataset relatively homogeneous.

2.7. Neural Network

Neural networks are models inspired by the human brain and are used for predictions because they can learn from data through training and produce specific outputs. In our case, the neural network was applied for the same classification purpose, and its accuracy was evaluated, resulting in the lowest performance among all analyses. Despite the absence of overfitting, with an accuracy of 56.34% on the training set and **58.86%** on the testing set, it cannot be considered a reliable predictive model. The chosen configuration, with two hidden layers containing 3 nodes in the first layer and 1 node in the second, proved to be the best among the various trials conducted, considering both model error and accuracy. We conclude that the poor predictive performance may be due to the fact that neural networks' complexity requires a large amount of data to be properly trained on the training set.



3. CONCLUSIONS

The exploratory analysis revealed critical factors influencing customer churn, such as age, account balance, and whether the customer is an active member. For example, younger customers tend to leave the bank less frequently, likely due to lower activation costs. Conversely, customers maintaining a balance above 75,000 are more likely to exit, potentially attracted by competitive offers from other banks. Additionally, more active members tend to remain with the bank, incentivized by the personalized services offered.
Below is a summary of the accuracy of the various models implemented.

Methodology	Accuracy (test)
Regressione logistica	69.1%
Support Vector Machine	70.03%
Decision Tree	71.28%
Random Forest	70.57%
Neural Network	58.86%

By comparing the accuracy of the various algorithms, we concluded that the most effective model for making predictions on our dataset is the decision tree, which also provides results consistent with the exploratory analysis.
Banks can leverage this information to improve customer retention by implementing targeted strategies and mitigating the risk of churn.