Group member: Diego Parra

Title: Customer Churn Prediction

Explanation: This project aims to predict customer churn using a historical dataset from a bank. Customer churn refers to the phenomenon of customers discontinuing a service or leaving a company. Identifying at-risk customers is a critical objective for many businesses, especially in highly competitive industries such as banking.

We used a dataset that includes various customer attributes such as credit score, geography, gender, age, account balance, and activity level, among others. The target variable is Exited, which indicates whether or not a customer has left the bank.

The goal was to explore the dataset, build predictive models, and draw conclusions that could help a business proactively address customer retention. We applied two machine learning classification models: K-Nearest Neighbors (KNN) and Logistic Regression. Both models were trained and tested using RapidMiner, and their performance was evaluated using accuracy, confusion matrix, and other classification metrics.

Additionally, we imported the dataset into MongoDB and performed queries to support real-time customer analysis in a NoSQL environment. This demonstrated how big data tools can be used to supplement machine learning workflows.

We concluded that Logistic Regression provided slightly more accurate results and clearer interpretability of feature importance, while KNN offered competitive performance with a simple structure. Features such as activity level, credit score, and account balance play significant roles in predicting churn.

Conclusion: Through this project, we gained valuable hands-on experience applying big data and machine learning techniques to a real-world business problem. By analyzing customer data and predicting churn using both K-Nearest Neighbors and Logistic Regression, we were able to identify patterns that could help a bank improve customer retention strategies. Logistic Regression proved to be slightly more effective in terms of accuracy and interpretability, while KNN provided a strong baseline with minimal configuration. Additionally, integrating MongoDB allowed us to experience how NoSQL databases can be used for flexible, large-scale data storage and retrieval. Overall, this project reinforced the importance of data preprocessing, model evaluation, and the role of data-driven insights in business decision-making.