# Master in Data Science

Module: Data Science Languages

Assignment Report: Customer Churn Prediction – Alternate Simulation

## 📊 Objective

This simulation replicates the logic of a submitted Jupyter notebook aimed at predicting banking customer churn. Due to memory constraints, a mock dataset was used to evaluate the model logic and produce comparable metrics.

## 📂 Simulated Dataset Overview

- Size: 10,000 records  
- Target Classes: 0 = No Churn (97%), 1 = Churn (3%)  
- Features: Simulated numerical and binary features replicating the original schema

## ⚖️ Methodology Summary

1. One-hot encoding and imputation of missing values  
2. Train-test split with stratified sampling  
3. Downsampling majority class for balanced training (5:1 ratio)  
4. Models trained:  
 - Decision Tree (min\_samples\_leaf=50)  
 - Random Forest (balanced class weight)  
 - Logistic Regression (balanced class weight)

## 📊 Simulated Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Recall (Churn) | Precision (Churn) | F1-Score | AUC Score |
| Decision Tree | 0.00 | 0.00 | 0.00 | 0.523 |
| Random Forest | 0.00 | 0.00 | 0.00 | 0.571 |
| Logistic Regression | 0.49 | 0.03 | 0.06 | 0.530 |

## 🚀 Conclusion

Despite efforts to balance the training set, only Logistic Regression successfully identified churn cases. Decision Tree and Random Forest models defaulted to predicting non-churn due to the severe class imbalance. This simulation highlights the importance of balancing techniques and model choice when handling imbalanced datasets. Further improvements could include SMOTE, cost-sensitive learning, or tuning decision thresholds.