DATA 607 – Project Proposal
Predicting Customer Churn Using the Iranian Telecom Dataset

Akinyemi Apampa | 30234456
David Fakolujo | 30273636
Joshua Ogunbo | 30272413
Ravin Jayasuriya | 30022788
Prince Oloma | 30263726

Dataset Description

For this project, we will analyze the Iranian Churn Dataset from the UCI Machine Learning Repository. It contains 3,150 customer records collected over 12 months by an Iranian telecom provider. Each record includes 13 variables related to customer usage patterns, account details, and demographic segments. Key features include subscription length, call failures, charge amount, SMS/call frequency, and customer value. The target variable is a binary churn indicator, showing whether a customer discontinued service during the final three months.

Research Questions

1. Which customer behavior patterns and demographic factors are the strongest predictors of churn in the Iranian telecom market?
2. Can we develop a classification model that prioritizes high recall to effectively identify at-risk customers and reduce churn?
3. How do different machine learning algorithms compare in their ability to generalize across the dataset, and what trade-offs exist between recall and overall performance?
4. How do dimensionality reduction and regularization techniques influence model complexity, overfitting, and predictive accuracy in churn classification?
5. How can insights from model results and feature design inform actionable customer retention strategies, such as targeted promotions or personalized plans?

Methods and Approach

We will begin with data preprocessing, including normalization of numeric features, encoding of categorical variables, and addressing class imbalance. Exploratory Data Analysis (EDA) will help uncover patterns and correlations relevant to churn.

The modeling phase will incorporate techniques from the course, including:

- Feature Engineering and Data Splitting using cross-validation strategies
- Dimensionality Reduction and Regularization
- Tree-based Models such as Decision Trees, Random Forest, and Gradient Boosting
- Deep Learning Models
- Traditional Models such as Logistic Regression, KNN, and SVM

Model performance will be assessed using cross-validation and metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. We will use insights from feature engineering and model performance to inform practical retention strategies such as targeted promotions and personalized service plans.