Customer Churn Prediction Project Summary

# 1. Introduction

The goal of this project was to develop a supervised machine learning model to predict customer churn using the Telco Customer Churn dataset. The project involved building a complete local ETL (Extract, Transform, Load) pipeline, followed by model training, evaluation, and interpretation. Churn prediction is critical in many sectors, especially telecom, where retaining customers is significantly more cost-effective than acquiring new ones.

# 2. Data Preprocessing and Transformation

The dataset was loaded and explored for missing values and data inconsistencies. Categorical variables were encoded using both label encoding and one-hot encoding. Numerical features were scaled using StandardScaler to ensure even weight distribution during model training. Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance and improve model performance on the minority class (churned customers).

# 3. Feature Engineering

Additional features were engineered to enhance model learning. These included calculated metrics such as tenure, monthly charges, and total charges. Irrelevant or redundant features were removed to streamline the dataset and reduce noise in the learning process.

# 4. Model Building and Evaluation

The dataset was split into training and testing subsets. Three supervised classification algorithms were implemented: Logistic Regression, Naive Bayes, and Decision Tree. Each model was trained and evaluated using accuracy, precision, recall, and F1-score. The models showed varied performance, with Logistic Regression and Decision Tree demonstrating a good balance between precision and recall.

# 5. Results and Conclusion

All three models were able to predict customer churn with reasonable accuracy. Feature scaling and SMOTE significantly improved the model’s ability to correctly identify churned customers. Logistic Regression emerged as the most interpretable model, while Decision Tree provided strong predictive performance. Further improvement could include hyperparameter tuning, feature selection, and using advanced models like Random Forest or XGBoost. The ETL pipeline and modeling process successfully fulfilled the project's objectives.