Customer Churn Prediction Using Supervised Machine Learning on Telco Data

## Aims and Objectives

This project aims to develop a predictive model that identifies customers likely to leave a telecommunications company. The analysis is based on the Telco Customer Churn dataset, which contains information on customers who left the company in the last month, along with details such as services signed up for, account information, contract type, and payment methods. Customer churn significantly affects business revenue and customer retention strategies. As such, accurately predicting churn is essential for proactive engagement.

Dataset Description  
The dataset used in this study is the Telco Customer Churn dataset, publicly available on OpenML. It contains **7,043 rows** representing individual customers of a telecom provider. Each row includes **21 variables** that describe demographic attributes, services used (such as internet, phone, and streaming services), contract details, payment methods, tenure length, and billing information.

The key target variable in the dataset is **‘Churn’**, which indicates whether the customer **left the company in the previous month** (Yes) or remained (No).

Important columns include:

* **Demographics**: gender, senior citizen status, partner and dependent information.
* **Service details**: phone service, internet service type (DSL, Fiber), online security, tech support, streaming features.
* **Contract and billing**: contract type (monthly, yearly), paperless billing, payment method, monthly and total charges.
* **Customer behaviour**: tenure (in months), which is further grouped into bins for modelling.

**Dataset Source:**  
OpenML URL: <https://www.openml.org/search?type=data&id=46280>

## Aims and Objectives

This research aims to develop a robust and interpretable machine learning model for predicting customer churn in the telecommunications industry. The study focuses on practical applicability, data-driven insights, and model generalization under class imbalance.

The specific objectives are as follows:

1. **To perform exploratory data analysis (EDA)** to identify distributions, trends, and relationships among variables associated with customer churn, including tenure, contract type, billing method, and service subscriptions.
2. **To preprocess the dataset** through detection and treatment of missing values, transformation of categorical variables via encoding, and standardization of numerical features to ensure compatibility with machine learning models.
3. **To address class imbalance** in the churn target variable using the SMOTEENN technique, which combines oversampling of the minority class and removal of noisy samples from the majority class to improve model sensitivity and precision.
4. **To train and evaluate three supervised learning models** Decision Tree, Random Forest, and Logistic Regression on both the original and resampled data for comparative analysis.
5. **To evaluate model performance** using classification metrics such as accuracy, precision, recall, and F1-score, emphasizing minority-class (churn) performance.
6. **To interpret model behavior** through feature importance plots to identify key predictors of churn.
7. **To finalize the optimal model** to find the best model which is suitable for potential use in customer retention strategies.

## Research Questions

To support the development and evaluation of the proposed churn prediction framework, the study is structured around the following key research questions:

* **Which customer-specific features are most influential** in determining churn behavior within the telecommunications domain?
* **Can supervised machine learning models** specifically Decision Tree, Random Forest, and Logistic Regression reliably classify customer churn in the presence of significant class imbalance?
* **How does the application of the SMOTEENN resampling technique** affect model performance in terms of precision, recall, and F1-score, particularly for the minority churn class?
* **Which classification model offers the best trade-off** between predictive accuracy, generalization, and interpretability for deployment in real-world customer retention systems?

These questions form the analytical foundation for the research, guiding the exploration of algorithmic performance, feature relevance, and the practical feasibility of deploying a predictive machine learning system within a commercial telecommunications environment.

## Project Relevance to Skills and Success Criteria

This project fits well with my current knowledge and skills. I have already worked on classification problems and have good experience using Python libraries for data analysis and machine learning. I’m comfortable with cleaning data, transforming it, training models, and interpreting the results.

By using SMOTEENN (Synthetic Minority Oversampling Technique + Edited Nearest Neighbours), I will apply advanced techniques for imbalanced datasets, which is common in real-world problems.

Because the project follows a step-by-step process from data understanding to model deployment, it supports my learning goals and will help me grow further in this field.

## Expected Outcomes

The outcomes of this research are intended to bridge the gap between academic data science methodology and real-world business application in the context of customer churn prediction.

The project is expected to yield the following deliverables:

* **A comparative performance evaluation** of supervised machine learning classifiers—Decision Tree, Random Forest, and Logistic Regression—under conditions of class imbalance.
* **Enhanced sensitivity to the minority churn class**, as demonstrated by improvements in recall and F1-score following the application of the SMOTEENN hybrid resampling technique.
* **Robust evaluation using standard classification metrics**, including accuracy, precision, recall, F1-score, and confusion matrix analysis, highlighting the practical viability of the selected model.
* **Feature importance analysis** to identify and explain the most influential predictors of customer churn across all classifiers.
* **Visualization outputs**, including confusion matrices and importance plots, to support interpretability and understanding of model behavior.
* **Selection of a finalized model** finding the optimal model based on both empirical performance and interpretability criteria.

Conclusion  
In summary, this project will allow me to:

* Explore a real dataset related to customer behaviour
* Build, test, and compare machine learning models
* Learn how to handle common issues like missing data and class imbalance
* And gain experience in preparing a production-ready model

This project not only matches my current skills but also helps me move one step closer to becoming a data scientist ready to work on practical business challenges.