📊 Telco Customer Churn Prediction

Project Overview

The **Telco Customer Churn Prediction** project aims to predict whether a customer will leave a telecom service provider. Predicting churn helps businesses **retain customers**, **optimise marketing strategies**, and **enhance customer satisfaction**.
This project leverages **machine learning classification models** to categorise customers as either **churned** or **non-churned** based on their demographics and usage patterns.

Dataset

The dataset contains detailed customer-level information collected by a telecom company. It includes **demographics, account information, and service usage details**.

Key Columns:

- `gender` – Male/Female
- `SeniorCitizen` – Indicates if the customer is a senior citizen (0 or 1)
- `tenure` – Number of months the customer has stayed with the company
- `MonthlyCharges` – The amount charged to the customer monthly
- `TotalCharges` – Total amount charged to the customer
- `HasInternetService` – Whether the customer has internet service (Yes/No)
- `Churn` – Target variable (Yes = churned, No = not churned)

Data Preprocessing

Steps performed before model training:

1. **Handling Missing Values** – Checked for and handled missing or inconsistent data.
2. **Encoding Categorical Variables** – Converted categorical features like `gender`, `HasInternetService`, etc., into numeric representations using Label Encoding or

One-Hot Encoding.

3. **Feature Scaling** – Standardised numerical columns (`tenure`, `MonthlyCharges`, `TotalCharges`) to normalise the feature range.

4. **Train-Test Split** – Split the dataset into training and testing sets (typically 80:20 ratio) for model evaluation.

Machine Learning Models

Several classification algorithms were applied and compared:

- **Logistic Regression** – A linear baseline model for binary classification.
- **Multinomial Naive Bayes** – A probabilistic model suitable for categorical input features.
- **Support Vector Classifier (SVC)** – Effective in high-dimensional spaces and for non-linear decision boundaries.
- **Random Forest Classifier** – An ensemble model combining multiple decision trees for higher accuracy and robustness.

Model Evaluation Metrics

The following metrics were used to assess performance:

- **Accuracy** – Measures the overall correctness of the model.
- **Precision** – Percentage of correctly predicted churn cases among all predicted churns.
- **Recall** – Percentage of actual churn cases correctly predicted by the model.
- **F1-Score** – Harmonic mean of precision and recall.
- **Confusion Matrix** – A Visual representation of true vs. predicted classifications.

Insights

Key insights derived from the analysis and models:

- Customers with **longer tenure** and **higher engagement** are more likely to stay.
- **Senior citizens** show a lower tendency to churn compared to younger customers.
- **Internet service** and **monthly charges** are strong predictors of churn behaviour.
- **Logistic Regression** and **Random Forest** provided the best trade-off between accuracy and interpretability.

Skills Learned

During this project, the following key skills and techniques were developed:

- Data cleaning and preprocessing
- Handling categorical and numerical data
- Applying and comparing multiple ML models
- Model evaluation using classification metrics
- Extracting and interpreting business insights from ML outputs

---

Technologies Used

- **Python** (pandas, numpy, scikit-learn, matplotlib, seaborn)
- **Jupyter Notebook / Google Colab**
- **Machine Learning Algorithms** (Logistic Regression, SVC, Naive Bayes, Random Forest)