Telco Customer Churn Prediction

Summary:

Objective

- Predict customer churn using IBM's Telco Customer Churn dataset.
- Build a full ML pipeline with preprocessing, tuning, and export.
- Make the code fully reusable and Colab-friendly.

Tools & Libraries Used

- pandas, numpy
- scikit-learn (Pipeline, ColumnTransformer, GridSearchCV)
- joblib for model export

What I Did

- Loaded dataset from a verified GitHub source (fixed Unicode/URL issues).
- Cleaned data:
  - Removed customerID
  - Converted TotalCharges to numeric
  - Filled missing values
- Encoded target variable (Churn: Yes → 1, No → 0)
- Separated features into:
  - Numerical: scaled with StandardScaler
  - Categorical: encoded with OneHotEncoder
- Built pipelines for:
  - Logistic Regression
  - Random Forest Classifier
- **Tuned hyperparameters** using GridSearchCV (5-fold CV)
- **Evaluated models** on test set using accuracy and classification report
- **Exported best pipeline** as a .pkl file using joblib

What I Learned

- How to build **clean ML pipelines** using Scikit-learn
- Importance of **modular preprocessing** with ColumnTransformer
- How to tune models with **GridSearchCV**
- Best practices for **data cleaning and type handling**
- How to make models **production-ready** using joblib

Next Steps

- Add model explainability (SHAP, LIME)
- Deploy via Flask, FastAPI, or Streamlit
- Visualize churn predictions in a dashboard
- Monitor model drift and retrain automatically