Initial Document: Telecom Churn Prediction Project

---

1. Introduction:

Telecom companies face the challenge of retaining customers in a highly competitive market. **Customer churn**, when customers leave a service is a critical issue, as acquiring new customers is significantly more expensive than retaining existing ones.

This project aims to **predict customer churn** using machine learning models, helping companies identify high-risk customers and take preventive actions.

Core Concept:

- Using **historical customer data** to train models that predict churn
- Evaluating models using **Accuracy, Precision, Recall, F1 Score, Confusion Matrix, and ROC-AUC**

---

2. Objectives:

- **Predict customer churn using ML models** – SDG 9: Innovation & infrastructure.
- **Evaluate models with key metrics (Accuracy, Precision, Recall, F1, ROC-AUC)** – SDG 8: Economic growth & efficiency.
- **Implement visualization & inference pipeline for business use** – SDG 12: Reduce resource waste.
- **Compare ML models to select the most robust** – Supports data-driven innovation (SDG 9) & efficiency (SDG 8).

---

3. Technology and Business Benefits:

3.1 Technology Benefits

- Automation of churn prediction using ML pipelines.
- Visualizations for model evaluation and reporting.
- Scalable models using Python and scikit-learn.

3.2 Business Benefits

- Identify high-risk customers proactively.
- Reduce revenue loss due to churn.
- Improve targeted marketing and retention strategies.

---

4. Dataset Overview:

- **Rows (Customers):** 7032
- **Features:** 20+ (numeric, categorical, boolean)
- **Target Variable:** Churn (1 = Churn, 0 = Stay)

Key features:

- tenure: Number of months with the company
- MonthlyCharges: Customer monthly payment
- TotalCharges: Total charges for the customer
- Contract type, PaymentMethod, InternetService
- SeniorCitizen: Binary indicator

Data Preprocessing:

- Handling missing values
- Encoding categorical features using one-hot encoding
- Normalization/Scaling not required for tree-based models

---

5. Methodology:

5.1 Model Selection

We trained five supervised learning models:

1. **Logistic Regression** – Simple baseline classifier

2. **Decision Tree** – Captures nonlinear relationships
3. **Random Forest** – Ensemble of decision trees, reduces overfitting
4. **K-Nearest Neighbors (KNN)** – Distance-based classification
5. **Support Vector Machine (SVM)** – Maximizes class separation

Rationale:

- Tree-based models (Decision Tree, Random Forest) handle categorical features well and are interpretable.
- Logistic Regression provides a baseline probability estimate.
- KNN and SVM were tested for performance comparison.

5.2 Evaluation Metrics

Metric	**Purpose**
Accuracy	Overall correctness
Precision	Correctly predicted churn among predicted churns
Recall	Correctly predicted churn among actual churns
F1 Score	Harmonic mean of Precision and Recall
Confusion Matrix	Visual summary of true positives/negatives
ROC-AUC	Model's ability to distinguish classes

All metrics are computed using **5-fold cross-validation** and on the **test set** for robust performance evaluation.

---

6. Results and Visualizations:

Sample model performance (from evaluation.py):

Random Forest: (Best performing model)

- Accuracy: 0.83
- F1 Score: 0.82
- Confusion Matrix:
- ROC-AUC: 0.80
- Precision/Recall/F1 Bar Chart:

Summary of all models:

Model	**Accuracy**	**F1 Score**
Logistic Regression	0.78	0.78
Decision Tree	0.80	0.80
Random Forest	0.83	0.82
KNN	0.75	0.74
SVM	0.68	0.70

---

7. Project Timeline:

Task	**Hours**
Data Cleaning & Preprocessing	6
Model Training	4
Evaluation & Visualization	4
Inference Pipeline	2
Documentation & Presentation	4
Total	20

---

8. Tech Stack:

- **Programming Language:** Python 3.10
- **IDE/Environment:** Anaconda
- **Libraries:** scikit-learn, numpy, pandas, matplotlib, seaborn, pickle
- **OS:** Windows 10

---

9. Conclusion:

The project successfully predicts telecom customer churn using machine learning. Random Forest achieved the highest performance with **Accuracy 83%** and **F1 Score 0.82**, providing actionable insights for customer retention strategies.

- The inference pipeline allows quick predictions for new customers.
- Visualizations aid stakeholders in understanding model decisions.
- Core ML concepts like **cross-validation, feature importance, and metrics evaluation** were applied correctly.
