Phase-1 Submission

Student Name	**:** A.Manju
Register Number	**:** 422723106024
Institution	**:** V.R.S.College of Engineering and Technology
Department	**:** ECE
Date of Submission	**:** 24:04:2025

1.Problem Statement

Customer churn poses a significant challenge for businesses, as losing existing customers can be more costly than acquiring new ones. This project focuses on leveraging machine learning techniques to predict churn by analyzing customer demographics, usage behavior, and interaction history. The objective is to discover subtle patterns and predictors that are not immediately apparent through conventional analysis, enabling data-driven strategies to retain at-risk customers and enhance business performance.

2.Objectives of the Project

The objective of this project is to build a machine learning-based predictive model that accurately identifies customers who are likely to churn. By analyzing historical customer data—including demographics, usage patterns, and engagement history—the model aims to uncover hidden trends and risk factors, enabling the business to implement targeted retention strategies and reduce customer attrition.

3.Scope of the Project

The scope of this project is to develop a machine learning model to predict customer churn by analyzing historical customer data and uncovering hidden behavioral patterns. The project will focus on key features such as usage metrics, customer demographics, and service interactions. Limitations include the use of only publicly available or company-approved datasets, reliance on scikit-learn and other open-source tools, and constraints on real-time deployment in production environments.

4.Data Sources

The dataset used for this project is the Telco Customer Churn dataset, which is publicly available on Kaggle. It contains customer-level information from a telecom company and is widely used for churn prediction modeling in supervised learning tasks.

- **Dataset Name and Origin**:
  Telco Customer Churn Dataset – Sourced from Kaggle
- **Type of Data**:
  Structured tabular data
- **Static or Dynamic Dataset**:
  The dataset is **static**, meaning it was collected at a single point in time and does not change or update in real time.

5.High-Level Methodology

- **Data Collection**: The dataset for customer churn prediction will be obtained from publicly available sources such as Kaggle or telecom datasets. If needed, synthetic data may be generated to simulate specific churn scenarios.

- **Data Cleaning**: We will address issues such as missing values by imputing with statistical methods or removing incomplete records. Duplicates will be eliminated, and inconsistent formats (e.g., date/time or categorical labels) will be standardized.
- **Exploratory Data Analysis (EDA)**: Techniques such as correlation heatmaps, box plots, histograms, and bar charts will be used to identify trends, customer segments, and relationships between features and churn behavior.
- **Feature Engineering**: New features may be derived from existing ones, such as tenure groups or customer activity levels. Categorical variables will be encoded, and features will be scaled or normalized as needed.
- **Model Building**: We will experiment with models such as Logistic Regression, Random Forest, XGBoost, and Support Vector Machines, chosen for their effectiveness in classification tasks and ability to handle complex patterns.
- **Model Evaluation**: Model performance will be evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation will be employed to ensure model generalizability.
- **Visualization & Interpretation**: Insights and predictions will be presented using graphs, confusion matrices, and feature importance charts, possibly through dashboards or interactive visualizations in Jupyter Notebooks.
- **Deployment**: The final model and findings may be deployed as an interactive web app using tools like Streamlit or Flask, allowing users to input data and receive churn predictions in real-time.

6.Tools and Technologies Us ed

In this phase of the customer churn prediction project, the following tools and technologies were utilized for data analysis, model building, and visualization.

1. Programming Language

- **Python**: The primary programming language used throughout the project for data manipulation, model building, and visualization. Python is widely used in data science due to its rich ecosystem of libraries and ease of use.

2. IDE / Notebook

- **Jupyter Notebook**: The primary development environment used for interactive coding, data exploration, and visualizations. Jupyter Notebooks allow us to document the process while running Python code, making it ideal for iterative data science projects.
- **Google Colab**: An online, cloud-based alternative to Jupyter Notebook, which provides free access to GPU/TPU resources for faster model training and experimentation.

3. Libraries

- **pandas**: Used for data manipulation, cleaning, and processing. It allows easy handling of structured data in the form of dataframes.
- **numpy**: Used for numerical operations and working with arrays.
- **scikit-learn**: A powerful library for machine learning, which includes models (e.g., Logistic Regression, Random Forest), preprocessing tools, and evaluation metrics.
- **seaborn**: Built on top of matplotlib, used for creating attractive and informative statistical graphics such as heatmaps, pair plots, and distribution plots.
- **matplotlib**: A low-level library for plotting in Python, often used for creating custom visualizations and charts.

- **xgboost** (optional, not used here but can be integrated in future): A gradient boosting framework that is efficient for high-performance and large datasets, often used to improve model accuracy.

4. Visualization Tools

- **matplotlib**: Used to generate basic plots such as histograms, boxplots, and line charts for visualizing the dataset and model performance.
- **seaborn**: Used in combination with matplotlib for more complex statistical plots, such as correlation heatmaps and pair plots.
- **Plotly** (optional, not used here but can be integrated in future): A library for creating interactive plots, often used for dashboards and web-based visualizations.
- **Tableau** / **PowerBI** (optional, if used for business reporting): Business intelligence tools that allow the creation of rich, interactive dashboards for presenting the final results to stakeholders.

5. Other Tools (if applicable)

- **Git**: For version control, enabling team collaboration and management of different versions of the project.
- **Anaconda**: A Python distribution that simplifies package management and deployment. It is useful for managing environments in data science projects.

7.Team Members and Contributions

1. Data Cleaning

- **Team Member 1:A.Manju**
  - **Responsibilities**:
    - Handled missing values by performing imputation and removal where necessary.
    - Removed duplicate records from the dataset.
    - Identified and treated outliers to ensure data quality.
    - Ensured consistency of data types across features.

2. Exploratory Data Analysis (EDA)

- **Team Member 2: A.Mahalakshmi**
  - **Responsibilities**:
    - Performed univariate analysis to examine the distribution of each feature.
    - Conducted bivariate and multivariate analysis to identify correlations and relationships between features and the target variable.
    - Generated various visualizations such as histograms, boxplots, and correlation matrices to derive insights.
    - Summarized findings, highlighting key trends and relationships in the dataset.

3. Feature Engineering

- **Team Member 3: K.Kiruthiga**
- **Responsibilities**:
  - Created new features based on domain knowledge and EDA insights, such as tenure groups, total services count, and monthly charge bins.

        - Applied techniques such as binning, categorization, and dimensionality reduction.
        - Justified and documented each new feature or transformation and its potential impact on the model.

4. Model Development

- **Team Member 4: R.Bhuvana**
    - **Responsibilities**:
        - Built and trained machine learning models (Logistic Regression and Random Forest) for churn prediction.
        - Tuned model parameters and evaluated model performance using appropriate metrics like accuracy, precision, recall, and F1-score.
        - Performed model comparisons and selected the best-performing model (Random Forest).

5. Documentation and Reporting

- **Team Member 5: Gayatri elangovan**
    - **Responsibilities**:
        - Compiled all the project stages, from data cleaning to model evaluation, into a comprehensive report.
        - Created visualizations for the final report, such as confusion matrices, ROC curves, and feature importance plots.
        - Wrote detailed explanations for each model, visualization, and analysis step.
        - Ensured the report was well-organized, clear, and aligned with project goals.