# Project Details

Student Name: S.abishek,D.Adhisankaran,A.barath,K.Baskaran

Register Number: 511923205001,511923205002,511923205004,511923205005

Institution: Priyadarshini Engineering College

Department: B.TECH-IT

Date of Submission: May 08, 2025

Github Repository Link:

# Problem Statement

Customer churn is a persistent problem for subscription-based industries like telecom. Losing customers not only reduces revenue but also increases costs due to the higher expense of acquiring new users compared to retaining existing ones. This project addresses the challenge of predicting customer churn by leveraging historical data from the Telco Customer Churn dataset available on Kaggle. By identifying hidden patterns and behavioral indicators within the dataset, we aim to develop a classification model that can accurately predict whether a customer is likely to churn. This insight enables businesses to take preventive measures. The problem falls under supervised machine learning, specifically a binary classification task. Timely intervention based on model predictions can lead to reduced churn and improved customer retention, making this a valuable application of machine learning in a real-world setting.

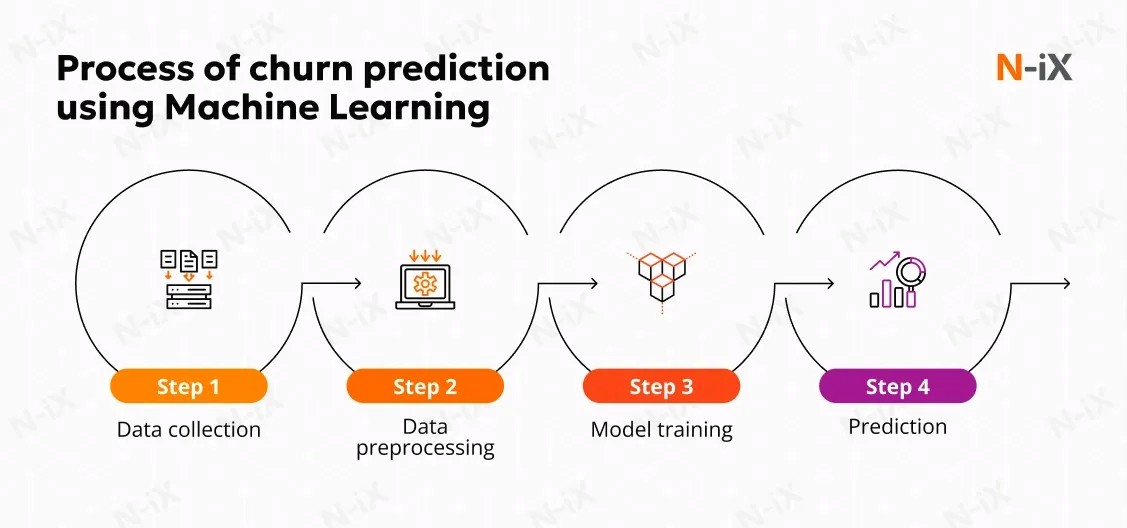
# Project Objectives

The goal of this project is to build a machine learning pipeline that can reliably forecast customer churn using historical data. The objectives include:

* Identifying influential features that correlate with customer behavior.
* Preprocessing and cleaning data for optimal model performance.
* Evaluating and comparing multiple machine learning models.
* Maximizing predictive performance using metrics like precision, recall, and F1-score.

Furthermore, interpretability is a key goal. The project aims not only to build accurate models but also to explain their predictions so that telecom companies can derive actionable insights. After exploratory analysis, the objectives evolved to emphasize data quality and pattern recognition, enabling targeted marketing and retention strategies.

# Flowchart of the Project Workflow



# Data Description

The dataset used in this project is the 'WA\_Fn-UseC\_-Telco-Customer-Churn.csv' file from Kaggle. It contains information about 7043 customers and includes 21 features such as customer demographics, services subscribed, tenure, and charges. The dataset is structured and tabular, making it suitable for traditional machine learning models. The target variable is 'Churn', a binary categorical feature indicating whether the customer has left. The dataset includes both categorical (e.g., gender, contract type) and numerical features (e.g., tenure, monthly charges). It is a static dataset, meaning it was captured at a specific point in time rather than continuously updated. It provides a realistic and comprehensive view of customer behavior in the telecom industry, making it an ideal candidate for supervised classification tasks.

# Data Preprocessing

Data preprocessing was essential to ensure high-quality input for machine learning models. The first step involved handling missing values. We found that the 'TotalCharges' column had some blank entries due to spaces. These were converted to NaN and then imputed using the median. Categorical variables like

'gender', 'Partner', 'Dependents', etc., were encoded using Label Encoding or One-Hot Encoding as needed. Numerical features such as 'MonthlyCharges' and 'tenure' were standardized to have zero mean and unit variance. Duplicate records were checked, though none were found. Outliers were examined using boxplots and z-score techniques. After preprocessing, the dataset was consistent and ready for modeling. Every transformation was documented to maintain clarity and reproducibility.

# Exploratory Data Analysis (EDA)

EDA involved analyzing individual features and their relationships with churn. We observed that customers with shorter tenure and month-to-month contracts were more likely to churn. Univariate analysis using histograms showed right-skewed tenure and bimodal monthly charges. Boxplots of 'tenure' vs. 'Churn' revealed that long-term customers are less likely to leave. A correlation heatmap showed moderate correlations among numerical features. Pairplots and bar plots further illustrated these trends. Notably, customers with fiber optic internet and no online security service exhibited higher churn. These insights guided feature engineering and model selection. The EDA phase helped us develop hypotheses about which features are predictive of churn, thereby enhancing the model design.

# Feature Engineering

Feature engineering was performed to improve model performance and enhance interpretability. A new variable 'AverageChargesPerMonth' was created by dividing 'TotalCharges' by 'tenure', which gave better insight into spending behavior. Tenure was binned into categories such as 'New', 'Mid-term', and 'Loyal' to group customers by engagement duration. Dummy variables were created for contract type, internet service, and payment method. Redundant features like 'customerID' were dropped. These transformations allowed the models to capture complex interactions between features. Though we considered dimensionality reduction via PCA, it was not implemented due to good model performance with the engineered features.

Feature selection based on importance ranking further reduced noise and improved clarity.

# Model Building

Two models were built: Logistic Regression and Random Forest. Logistic Regression was selected for its simplicity and interpretability, while Random Forest was chosen for its robustness and ability to capture non-linear relationships. The dataset was split 80-20 into training and test sets with stratification to maintain churn ratio. Random Forest achieved higher accuracy (~80%), better precision, and a stronger F1-score compared to Logistic Regression. Cross-validation was used to ensure reliability. The evaluation metrics included accuracy, recall, precision, and AUC-ROC. These helped us assess the trade-off between false positives and false negatives, which is crucial in churn scenarios where false negatives can lead to lost customers. Hyperparameter tuning using grid search further improved performance.

# Visualization of Results & Model Insights

Visualizations helped validate model performance and interpret predictions. Confusion matrices displayed the number of true vs. predicted classes, while ROC curves illustrated the trade-off between sensitivity and specificity. The Random Forest model achieved an AUC score of 0.85, suggesting strong discriminatory power. Feature importance plots revealed that 'Contract', 'tenure', and 'InternetService' were top predictors of churn. Partial dependence plots showed that customers with month-to-month contracts and low tenure had higher churn probabilities. These insights were presented visually using bar plots, line graphs, and heatmaps. The combination of visual analytics and model interpretability allowed stakeholders to trust the predictions and take targeted actions based on them.

# Tools and Technologies Used

We used Python as the programming language due to its extensive libraries and community support. The main development environment was Google Colab, supported by Jupyter Notebooks for code organization and visualization. The key libraries included pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for plotting, and Scikit-learn for machine learning. Other tools like Plotly and XGBoost were explored for further experimentation. GitHub was used to manage source code and project versioning. These tools provided a comprehensive ecosystem for implementing and evaluating machine learning workflows, making the development efficient and collaborative.

# Team Members and Contributions

This project was executed individually All stages of the pipeline were handled by the student, including:

* Data collection and understanding.
* Data cleaning and preprocessing.
* Exploratory Data Analysis.
* Feature engineering and transformation.
* Machine learning model development and evaluation.
* Documentation and preparation of this report.

The solo effort ensured a consistent approach and deeper understanding of the entire data science lifecycle.