**Phase-2**

**Student Name:SARAVANAPRIYAN V**

**Register Number: 421023106301**

**Institution:** E.S COLLEGE OF ENGENEERING AND TECHNOLOGY

**Department:** ELECTRONIC AND COMMUNICATION

**Date of Submission: 11/05/2025**

**Github Repository Link:**

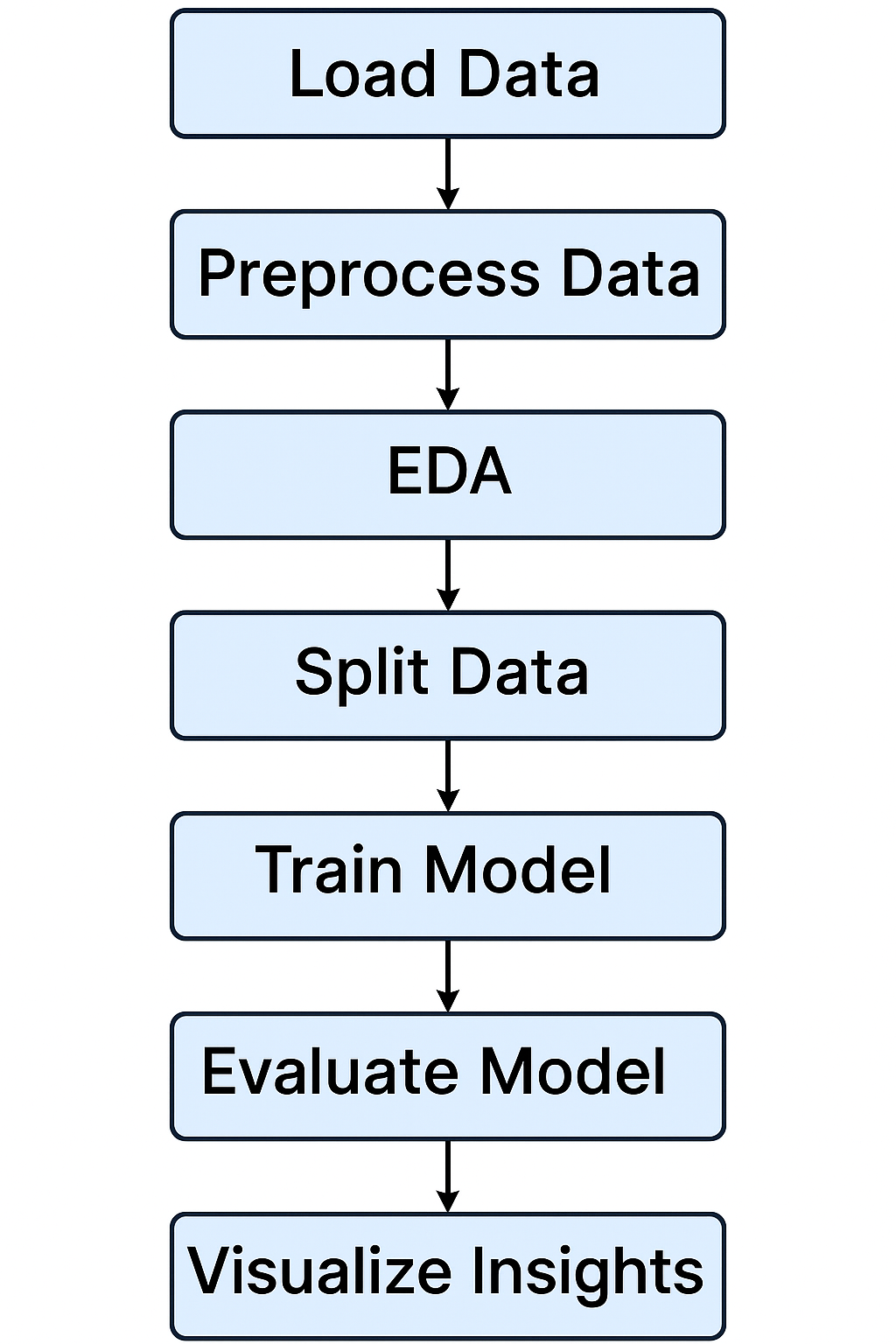
# Problem Statement

The goal is to predict customer churn based on historical data using machine learning. Accurately identifying customers likely to leave helps businesses proactively retain them, improving revenue and customer satisfaction. Churn prediction is a critical real-world business problem across industries like telecom, banking, and SaaS.

# Project Objectives

* + Develop a robust classification model to predict churn with high accuracy and generalizability.
  + Identify the most important features contributing to customer churn.
  + Interpret and visualize model results for actionable business insights.
  + Apply end-to-end data science workflow including preprocessing, EDA, feature selection, model training, and evaluation.

# Flowchart of the Project Workflow



# Data Description

* + Dataset Source: [Provide Source - e.g., Kaggle Telco Churn Dataset]*.*
  + Type of Data: Structured, Tabular*.*
  + Size: [Rows x Columns] – e.g., 7043 records with 21 features (update based on actual file)
  + Static or Dynamic: Static
  + Target Variable: Churn

# Data Preprocessing

* + Missing Values: Handled using forward-fill (fillna(method='ffill'))
  + Categorical Variables: Encoded using LabelEncoder
  + Duplicates: [Assumed none or not handled – optional improvement]
  + Normalization/Scaling: Not performed (Random Forest doesn't require feature scaling)
  + Outliers: Not explicitly handled

# Exploratory Data Analysis (EDA)

* + *Univariate Analysis:*
    - Code not shown in script, can be added for histograms/count plots*.*
  + *Multivariate Analysis:*
    - Correlation Heatmap used to identify relationships among features.
  + *Insights Summary:*
    - Heatmap helps identify redundant or highly correlated features*.*
    - Understanding these can guide feature selection and engineering

# Feature Engineering

* + No new features were created.
  + Label encoding used for categorical columns.
  + Further feature engineering could include interaction terms or derived features (e.g., tenure buckets).

# Model Building

* + Models Used: Random Forest Classifier
  + Justification: Random Forest handles categorical data, is robust to overfitting, and provides feature importance.
  + Train/Test Split: 80% Train, 20% Test
  + Evaluation Metrics Used:
    - Accuracy
    - Precision, Recall, F1-score
    - Confusion Matrix

# Visualization of Results & Model Insights

* + Confusion Matrix: Visualized using sns.heatmap
  + Feature Importance: Plotted to show most influential variables in predicting churn.
  + These visualizations help explain model decisions and support business interpretation*.*

# Tools and Technologies Used

* + Programming Language: Python
  + IDE/Notebook: Jupyter Notebook
  + Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn
  + Visualization Tools: seaborn, matplotlib

# Team Members and Contributions

* **Mohan Babu :** 
  + Data Collection and Data Cleaning
  + Handling missing values and encoding categorical variables
  + Training Random Forest model and evaluating it using confusion matrix and classification report
* **Rajesh :** 
  + Exploratory Data Analysis (EDA)
  + Generating correlation heatmaps and feature relationship visualizations
* **Saravana Priyan** 
  + Feature Engineering
  + Model Building and Evaluation
  + Preparing final datasets for model training
* **Vignesh:** 
  + Documentation and Reporting
  + Preparing final project report and visualization plots for presentation