Github Link: [Github](https://github.com/ganga-200514)

**Project Title:Predicting customer churn using machine learning to uncover hidden patterns**

**PHASE-2**

**Student Name:** Ganga.SS

**Register Number:** 712523104020

**Institution:** PPG Institute of Technology.

**Department:** BE Computer Science and Engineering.

**Date of Submission:** 05.05.25

### **1. Problem Statement**

### The goal is to predict customer churn in a subscription-based service using machine learning. Churn refers to customers who stop using the service within a certain period. Identifying such customers in advance can help businesses take proactive measures to retain them.

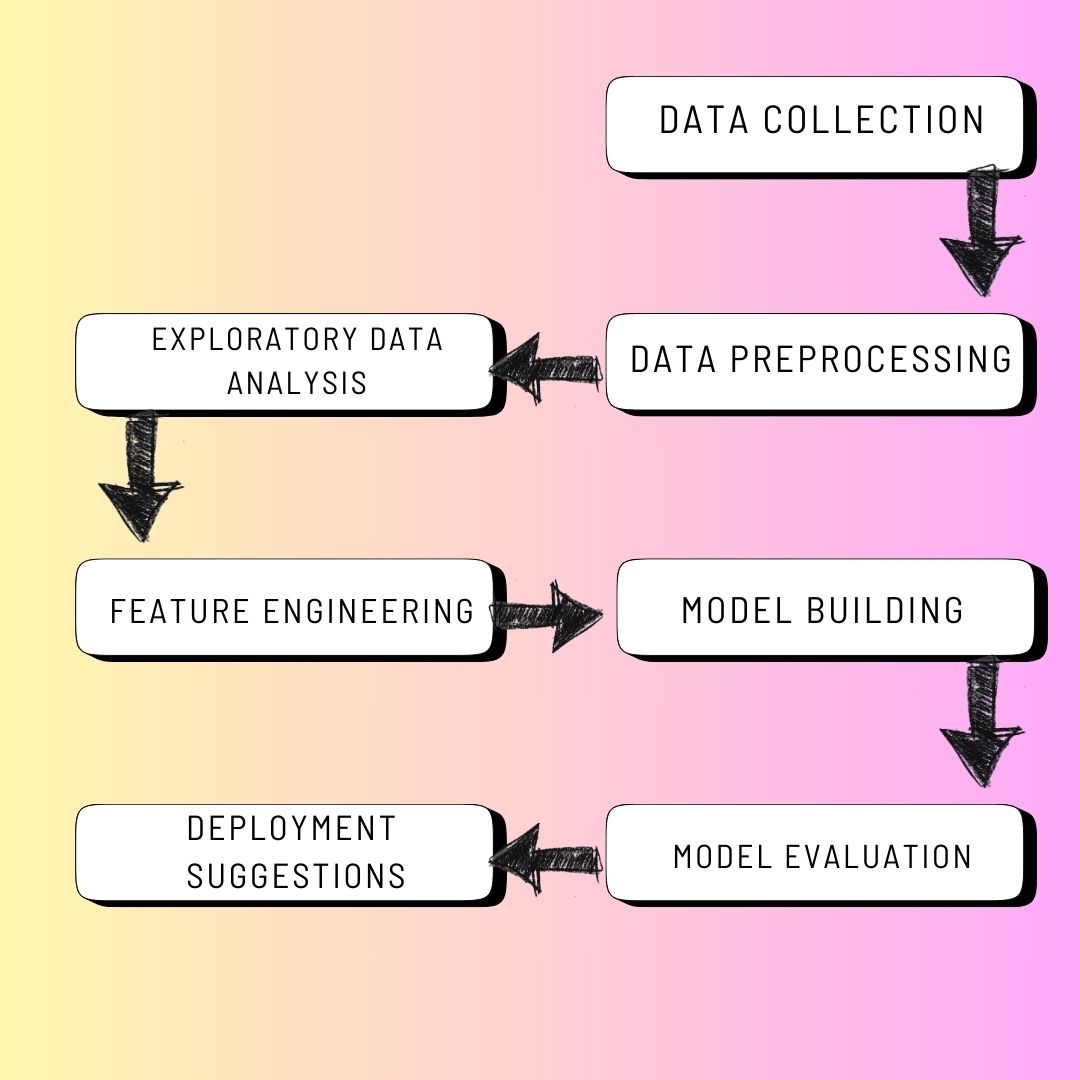
### **Problem Type:** Classification

### **Relevance:** Reducing churn can significantly improve revenue and customer lifetime value. Early detection of churn enables targeted marketing, personalized offers, and better resource allocation.

### **2. Project Objectives**

* Build a model that accurately predicts whether a customer will churn.
* Achieve high precision and recall to minimize false positives and false negatives.
* Provide insights into the key features influencing churn.
* Develop a robust, interpretable solution suitable for real-world deployment.

### **3. Flowchart of the Project Workflow**



### **4. Data Description**

* **Dataset Name:** Telco Customer Churn
* **Source:** Kaggle
* **Type:** Structured tabular data
* **Records & Features:** ~7,000 records, 20+ features
* **Nature:** Static
* **Target Variable:** Churn (Yes/No)

### **Data set link:** [**https://www.kaggle.com/**](https://www.kaggle.com/)

### **5. Data Preprocessing**

* Handled missing values using mode imputation for categorical and mean for numerical columns.
* Removed duplicate entries based on customer ID.
* Detected outliers using IQR and handled extreme values.
* Converted data types (e.g., TotalCharges from object to float).
* Encoded categorical variables using One-Hot Encoding*.*

### **6. Exploratory Data Analysis (EDA)**

* **Univariate Analysis:**
  + Histograms and boxplots for tenure, MonthlyCharges, and TotalChargesCount plots for churn distribution
* **Bivariate/Multivariate Analysis:**  
  + Correlation heatmap
  + Churn vs contract type, payment method, and tenure plots.
* **Insights Summary:**
  + Month-to-month contracts have higher churn
  + Customers with fiber internet tend to churn more

### **7. Feature Engineering**

* Created a new feature AvgMonthlySpend = TotalCharges / tenure.
* Extracted tenure groups using binning.
* Removed less impactful or redundant features (e.g., customerID).
* Tried polynomial interaction features but retained only those improving model accuracy.
* Considered PCA but skipped due to good interpretability without dimensionality reduction.

### 

### **8. Model Building**

### **Models Used**: Logistic Regression,Random Forest Classifier

### and interpretability

### Random Forest for handling non-linear relationships and feature importance

### **Training/Test Split:** 80/20 with stratification

### **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score

### **Logistic Regression:** F1 = 0.79

### **Random Forest**: F1 = 0.86

### 

### **9. Visualization of Results & Model Insights**

### **Confusion Matrix:** Showed a balanced true positive/negative rate

### **ROC Curve:** AUC = 0.90 for Random Forest

### **Feature Importance:** Tenure, Contract Type, Monthly Charges were top predictors

### **Insights:**

### Long-tenure customers less likely to churn

### Auto-pay customers have lower churn rates

### **10. Tools and Technologies Used**

* Programming Language: Python.
* IDE/Notebook: Google Colab, VS.
* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost.
* Visualization Tools: Matplotlib, Seaborn, Plotly.

### 

### **11. Team Members and Contributions**

|  |  |  |
| --- | --- | --- |
| **NAME** | **ROLE** | **DESCRIPTION** |
| Krisha.A.P | Project Manager | Oversees project planning, timelines, and deliverables, Coordinates communication among team members, Ensures the project aligns with goals and deadlines |
| Nivitharasri.V | Analyst & Visualizer | Identifies business goals and interprets model insights,Translates technical findings into actionable business strategies,Works on visualization and presenting results to stakeholders |
| Ganga.SS | Machine Learning Developer | Develops and trains machine learning models, Conducts feature engineering and model evaluation, Tunes hyperparameters and selects the best model. |
| Hirudhanya.S | Data Engineer | Responsible for data collection, cleaning, and preprocessing, Builds and maintains data pipelines, Ensures data quality and integrity. |
| Vijay.S | Deployment | Develops a user interface or dashboard for model outputs, Handles integration and deployment of the model, Ensures scalability, maintainability, and user accessibility. |