# Phase-2

**Student Name:** Sudharsan V.S (2006)

**Register Number:** 410723106033

**Institution:** Dhanalakshmi College of Engineering

**Department:** Electronics and Communication Engineering

**Date of Submission:** 07-05-2025

## Github Repository Link:

https://github.com/sudharsan2121/NM_sudharshan_ds

---

## 1. Problem Statement

Customer churn is a critical issue for businesses, especially in subscription-based industries like telecom, banking, and SaaS. Losing customers leads to revenue decline and increased customer acquisition costs. By analyzing customer behavior and demographics, machine learning can help predict which customers are likely to leave a service, enabling proactive retention strategies.

- **Refined Problem Type**: **Classification problem**

- **Business Relevance**: Early churn prediction allows companies to implement targeted interventions (e.g., discounts, customer support outreach) to retain customers, directly impacting profitability and customer satisfaction.

- **Updated Understanding**: After exploring the dataset, behavioral indicators such as service usage and support interaction
  are significant churn predictors.

## 2. Project Objectives

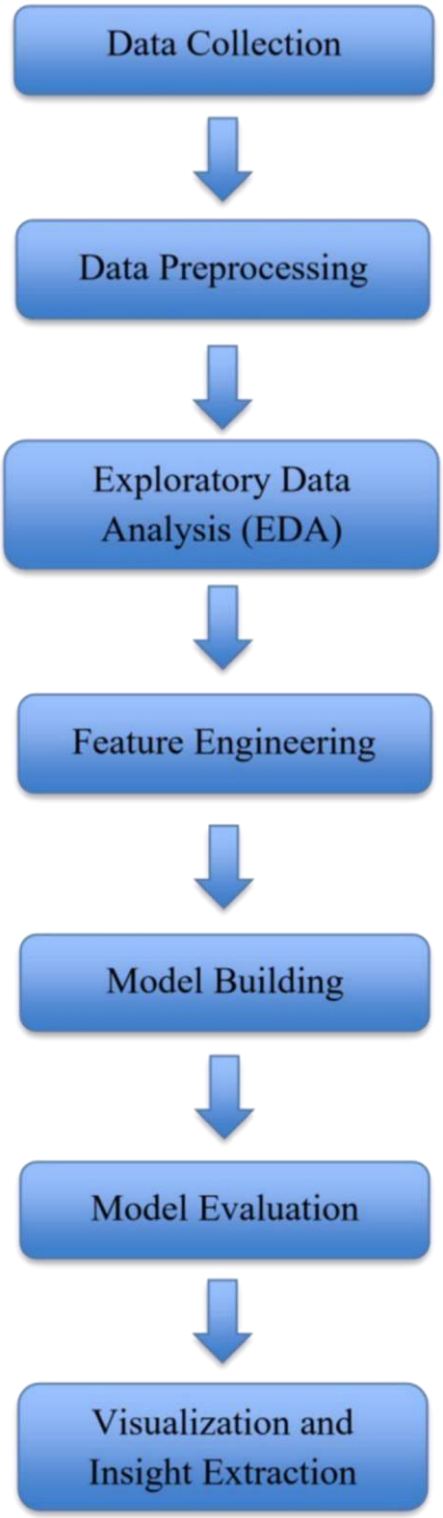
**Technical Objectives**:

- Build a machine learning model to predict whether a customer will churn.
- Achieve high accuracy while maintaining interpretability.
- Compare multiple models to determine the best performer.
- Generate feature importance insights for actionable business decisions.

**Updated Goals**:

- Focus has shifted from purely maximizing accuracy to balancing performance and interpretability.
- Emphasis added on extracting business-relevant insights from feature importance.

## 3.Flowchart of the Project Workflow



## 4.Data Description

**Dataset Name**: Telco Customer Churn

**Data Type**: Structured tabular data

**Records**: ~7,000 customer entries

**Features**: 20+ features (gender, tenure, internet service type, etc.)

**Static or Dynamic**: Static dataset

**Target Variable**: **Churn (Yes/No)**

**Data Set Link** : https://www.kaggle.com/datasets/bhuviranga/customer-churn-data

## 5. Data Preprocessing

- **Missing Values**:
  - TotalCharges had empty strings – converted to NaN and imputed with median.

- **Duplicates**: No duplicate rows found.
- **Outliers**:
  - Detected high MonthlyCharges – retained based on business context.
- **Type Conversion**: ₀ Converted TotalCharges to float.
- **Categorical Encoding**:
  - Label encoding for binary columns.
  - One-hot encoding for multi-category features.
- **Normalization**:
  - StandardScaler applied to continuous variables (tenure, MonthlyCharges, etc.).

```
# Handle missing values
df['TotalCharges'].fillna(df['TotalCharges'].median(), inplace=True)

# Encode categorical variables
df = pd.get_dummies(df, drop_first=True)
```

## 6. Exploratory Data Analysis (EDA)

**Univariate Analysis**:

- Histograms of numerical features show right-skewed distribution.
- Countplots show higher churn among Fiber Optic users.

- **Bivariate/Multivariate Analysis**:
  - Correlation matrix reveals strong relationship between MonthlyCharges and Churn
  - Churn more common among customers with short tenure and no contract.
- **Insights**:
  - Contract type, payment method, and service add-ons significantly impact churn.
  - Customers with electronic checks and no security features tend to churn more.

```
# Correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(), cmap='coolwarm', annot=False)

# Bivariate analysis
sns.boxplot(x='Churn_Yes', y='MonthlyCharges', data=df)
plt.title('Monthly Charges vs Churn')
```

## 7. Feature Engineering

- **Univariate Analysis**:
  - Histograms of numerical features show right-skewed distribution.
  - Countplots show higher churn among Fiber Optic users.
- **Bivariate/Multivariate Analysis**:
  - Correlation matrix reveals strong relationship between MonthlyCharges and Churn. ○ Churn more common among customers with short tenure and no contract.
- **Insights**:
  - Contract type, payment method, and service add-ons significantly impact churn.
  - Customers with electronic checks and no security features tend to churn more.

```
# Create a new feature: average monthly spend
df['AvgMonthlySpend'] = df['TotalCharges'] / (df['tenure'] + 1)
```

## 8. Model Building

**Models Used**:

- Logistic Regression (baseline, interpretable) ○ Random Forest Classifier (nonlinear, robust)

**Model Selection Justification**:

- Logistic Regression: Simplicity and interpretability ○ Random Forest: Handles categorical data well, provides feature importance

- **Data Split**:
  80/20 train-test split with stratification on target (Churn)

- **Performance Metrics**:
  Accuracy, Precision, Recall, F1-score

Random Forest performed better with balanced F1 and precision

```
# Split data
X = df.drop('Churn_Yes', axis=1)
y = df['Churn_Yes']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
stratify=y, random_state=42)

# Logistic Regression
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred_log = log_model.predict(X_test)
```

## 9. Visualization of Results & Model Insights

- **Confusion Matrix**: Visualized true/false positives/negatives for both models
- **ROC Curve**:
  AUC ~0.85 for Random Forest
- **Feature Importance**:
  Contract Type, MonthlyCharges, tenure are top predictors
- **Plot Types Used**:
  Bar plots, ROC curves, confusion matrix
- **Interpretation**:
  Customers on month-to-month contracts with high charges are more likely to churn

## 10. Tools and Technologies Used

- **Programming Language**: Python

  **IDE/Notebook**: Google Colab

  **Libraries**: pandas, numpy,

  seaborn, matplotlib pandas,

  numpy, seaborn, matplotlib (EDA,

  preprocessing)

- scikit-learn (modeling)
- xgboost (for extended experiments)
- **Visualization Tools**: Plotly, Seaborn

## 11. Team Members and Contributions

| S.NO | NAMES | ROLE | RESPOSIBILITY |
|---|---|---|---|
| 1. | Sudharsan V.S | Member | Visualization & Cleaning |
| 2. | Aswin P.G | Member | Exploratory data analysis(EDA)<br>Feature engineering |
| 3. | Sudharsan V.S | Member | Model building, model Evaluation |
| 4. | Yogesh J | Leader | Data Colletion, Data Cleaning |