1. Introduction

This project focuses on predicting customer churn for a telecommunications company. Customer churn refers to the rate at which customers stop doing business with a service provider. By leveraging machine learning algorithms, this project aims to classify whether a customer is likely to churn, allowing companies to take proactive actions to retain valuable customers.

2. Dataset Description

The dataset used in this project is sourced from the **Telco Customer Churn** dataset, which contains various customer attributes that help in determining the churn rate. The dataset includes the following columns:

- **customerID**: Unique customer identifier.
- **gender**: Customer gender (Male/Female).
- **SeniorCitizen**: Whether the customer is a senior citizen (1: Yes, 0: No).
- **Partner**: Whether the customer has a partner (Yes/No).
- **Dependents**: Whether the customer has dependents (Yes/No).
- **tenure**: Number of months the customer has stayed with the company.
- **PhoneService**: Whether the customer has phone service (Yes/No).
- **MultipleLines**: Whether the customer has multiple lines (Yes/No).
- Other attributes include information about internet services, payment methods, contract types, and monthly charges.

The target variable is **Churn**, which indicates if the customer has churned (Yes/No).

3. Project Workflow

The project is divided into the following major steps:

1. **Data Loading**:
    - The dataset is loaded using the Pandas library for further manipulation.
2. **Data Cleaning and Preprocessing**:
    - Missing values are handled.
    - Encoding of categorical variables.
    - Scaling of numerical variables where necessary.
3. **Exploratory Data Analysis (EDA)**:
    - Visualization of the dataset to understand the relationships between variables.
    - Important metrics such as churn rates across various demographics are analyzed.
4. **Model Building**:
    - Machine learning models such as Logistic Regression, Decision Trees, and Random Forest are applied to predict churn.
5. **Model Evaluation**:
    - The models are evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

6. **Visualization**:
    - Several visualizations are used to depict data distribution, churn correlations, and model performance.

4. Modeling

In this project, various machine learning algorithms are employed to predict customer churn:

- **Logistic Regression**: A baseline model to check the linear relationship between features and churn.
- **Random Forest**: A robust ensemble model that captures non-linear relationships and interactions between variables.
- **Neural Networks (Keras)**: A deep learning approach using Keras to model more complex patterns in the data.

Each model is fine-tuned using hyperparameter optimization, and performance is compared using classification metrics.

5. Results and Visualizations

Key results from the project include:

- **Churn Rate**: Percentage of customers who churned.
- **Feature Importance**: Ranking of the most important features contributing to customer churn.
- **Confusion Matrix**: A visualization of true positives, false positives, etc., for each model.
- **ROC Curve**: A plot comparing the performance of models using the AUC-ROC metric.

6. Future Improvements

Potential improvements for the project:

- Implementing advanced feature engineering techniques.
- Trying more complex models like XGBoost or CatBoost.
- Performing cross-validation to better generalize the model.
- Exploring other customer datasets to further improve model performance.