**Project Report: Machine Learning Analysis on Telecommunication Data**

**1. Introduction** The rapid advancement of machine learning has enabled industries to analyze vast datasets and derive meaningful insights. This project focuses on a telecommunication dataset with the goal of classifying customer contracts using various machine learning techniques. The dataset includes multiple service-related attributes and billing details, which can be used to predict customer contract types. By preprocessing the data and applying machine learning models, this project aims to find the most effective method for classifying customer contracts, which can help telecommunication companies optimize their services.

**2. Objectives** The primary objectives of this project are as follows:

* To clean and preprocess the dataset to ensure data quality.
* To encode categorical variables for better compatibility with machine learning models.
* To train and evaluate multiple classification models.
* To compare model performances using various evaluation metrics.
* To determine the best model for predicting contract types.

**3. Dataset Description** The dataset consists of multiple features related to telecommunication services and customer billing information. The key attributes include:

* **Service-related Features:**
  + MultipleLines: Indicates whether the customer has multiple phone lines.
  + InternetService: Type of internet service subscribed (e.g., DSL, Fiber Optic, None).
  + OnlineSecurity: Whether online security services are included.
  + OnlineBackup: Availability of online backup services.
  + DeviceProtection: Presence of device protection services.
  + TechSupport: Availability of technical support.
  + StreamingTV: Whether streaming TV services are included.
  + StreamingMovies: Subscription to streaming movie services.
* **Billing Information:**
  + MonthlyCharges: The monthly bill amount for the customer.
  + TotalCharges: The total amount charged to the customer so far.
* **Target Variable:**
  + Contract: The type of customer contract ("Two year," "Month-to-month").

**4. Data Preprocessing** To ensure optimal performance of machine learning models, the dataset underwent several preprocessing steps:

* **Handling Missing Values:**
  + Numerical columns with missing values were replaced with the mean of their respective columns.
  + Categorical columns with missing values were imputed using the mode (most frequent value).
  + Columns with excessive missing values (greater than 50%) were removed to avoid data inconsistencies.
* **Encoding Categorical Variables:**
  + One-hot encoding was applied to categorical features with low cardinality.
  + Label encoding was used for high-cardinality categorical features to ensure machine learning compatibility.
* **Feature Selection:**
  + Irrelevant and redundant features were removed to improve model efficiency.
  + Correlation analysis was conducted to identify highly correlated features.

**5. Machine Learning Models Applied** Three classification models were implemented to predict customer contract types:

* **K-Nearest Neighbors (KNN):**
  + A distance-based algorithm that classifies data points based on their nearest neighbors.
  + Requires feature scaling for optimal performance.
  + Works best for datasets with balanced distributions.
* **Random Forest Classifier:**
  + An ensemble learning method that constructs multiple decision trees.
  + Handles both categorical and numerical data efficiently.
  + Reduces overfitting compared to single decision trees.
* **Naïve Bayes Classifier:**
  + A probabilistic model based on Bayes' theorem.
  + Performs well for categorical data and assumes independence between features.
  + Limited effectiveness for numerical data distributions.

**6. Model Evaluation** The models were evaluated using the following performance metrics:

* **Accuracy:** Measures the proportion of correctly classified instances.
* **Precision, Recall, and F1-score:** Evaluates the balance between false positives and false negatives.
* **Confusion Matrix:** Visualizes correct and incorrect classifications.

**7. Results and Discussion** Each model's performance was analyzed based on the evaluation metrics:

* **KNN Model:**
  + Provided moderate accuracy but showed sensitivity to dataset size and feature scaling.
  + Performance improved with optimized k-values.
* **Random Forest Model:**
  + Outperformed other models with high accuracy and robust feature importance.
  + Reduced overfitting by averaging multiple decision trees.
* **Naïve Bayes Model:**
  + Worked well for categorical variables but had limitations in handling continuous numerical data.
  + Performed best when features had clear probabilistic relationships.

**8. Conclusion** This project successfully applied machine learning techniques to classify telecommunication customer contracts. Among the models tested, the Random Forest classifier emerged as the most effective, providing high accuracy and stable performance. This result demonstrates the suitability of ensemble learning methods for structured datasets with categorical and numerical variables.

**9. Future Work** Several improvements can be made to enhance model performance and expand the study:

* **Hyperparameter tuning:** Optimize model parameters to improve accuracy and reduce computational complexity.
* **Advanced feature engineering:** Explore new feature transformations and interactions.
* **Deep learning approaches:** Investigate neural networks for more complex patterns in data.
* **Customer churn prediction:** Extend the analysis to predict customer retention trends based on service usage patterns.