**Name: Bibakananda Roy Shuvo**

**ID:212-15-14747**

**Training Name:** **Enhancing Digital Government and Economy (EDGE)**

**Project Report: Machine Learning Analysis on EDGE Dataset**

**1. Introduction** Machine learning techniques are widely used for analyzing structured datasets to derive meaningful insights. This project focuses on the classification of customer-related data using machine learning models. The dataset contains service-related features, billing details, and contract types, aiming to predict contract categories effectively. The project involves data preprocessing, feature engineering, and model evaluation to determine the best classification approach.

**2. Objectives**

* To preprocess and clean the dataset to ensure data consistency.
* To handle missing values and encode categorical variables appropriately.
* To implement multiple classification algorithms and compare their performance.
* To evaluate models based on accuracy, precision, recall, and other metrics.
* To identify the best machine learning model for contract classification.

**3. Dataset Description** The dataset consists of customer service features, billing information, and contract details. The key attributes include:

* **Customer Information:**
  + User ID: A unique identifier for each customer.
  + Demographic attributes (if available).
* **Service-related Features:**
  + MultipleLines: Indicates whether the customer has multiple service lines.
  + InternetService: Type of internet service used (DSL, Fiber optic, No service).
  + OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies: Status of additional services subscribed by the customer.
* **Billing Information:**
  + MonthlyCharges: Monthly payment amount.
  + TotalCharges: Total cumulative charges for the customer.
* **Target Variable:**
  + Contract type (Month-to-Month, One Year, Two Year).

**4. Data Preprocessing** Preprocessing steps included:

* **Handling Missing Values:**
  + 'TotalCharges' was converted to numeric format, with missing values imputed using the median.
* **Feature Encoding:**
  + One-hot encoding was applied to categorical variables such as Contract type.
  + Label encoding was used for InternetService and other categorical fields.
* **Feature Scaling:**
  + Min-Max scaling was applied to MonthlyCharges and TotalCharges to normalize values for better model performance.

**5. Machine Learning Models Implemented** The following classification models were implemented:

* **Logistic Regression:** A baseline linear model for binary/multi-class classification.
* **K-Nearest Neighbors (KNN):** A distance-based classifier requiring proper feature scaling.
* **Random Forest Classifier:** An ensemble learning method using multiple decision trees for robust classification.
* **Naïve Bayes Classifier:** A probabilistic classifier that performs well for categorical data but may be less effective for continuous variables.
* **Support Vector Machine (SVM):** A powerful classifier that maps data to a high-dimensional space to find an optimal decision boundary.

**6. Model Evaluation** To assess the models' effectiveness, the following evaluation metrics were used:

* **Accuracy:** Measures the proportion of correctly classified instances.
* **Precision, Recall, and F1-score:** Evaluates the balance between false positives and false negatives.
* **Confusion Matrix:** Provides an overview of correct vs. incorrect classifications.
* **ROC Curve and AUC Score:** Measures model discrimination capability.

**7. Results and Discussion**

* **Logistic Regression:** Performed well as a baseline model but lacked flexibility for complex relationships.
* **KNN:** Showed moderate accuracy, with performance varying based on the choice of k.
* **Random Forest:** Achieved the highest accuracy due to its ensemble learning approach, handling both categorical and numerical features effectively.
* **Naïve Bayes:** Worked well for categorical data but struggled with numerical attributes.
* **SVM:** Provided high accuracy but required careful parameter tuning for optimal performance.

**8. Conclusion** The study successfully classified customer contract types using multiple machine learning techniques. Among the evaluated models, the **Random Forest Classifier** exhibited the best performance in terms of accuracy and robustness. This model is suitable for contract prediction and can assist businesses in optimizing customer retention strategies.

**9. Future Work** To further improve the analysis, the following enhancements are recommended:

* **Hyperparameter tuning:** Optimize model parameters for improved accuracy.
* **Feature engineering:** Introduce additional derived features to enhance classification.
* **Deep learning models:** Explore neural networks for advanced predictive capabilities.
* **Customer behavior analysis:** Expand the study to include churn prediction based on historical usage patterns.