Data Analysis and Machine Learning Report

This report summarizes the data preprocessing steps, exploratory data analysis (EDA), and machine learning models' performance on the given dataset. The dataset consists of customer information and service usage details to predict customer churn.

Data Preprocessing:
1. Converted 'TotalCharges' to numeric and filled missing values with the median.
2. Encoded categorical columns using LabelEncoder.

Exploratory Data Analysis (EDA):
- Correlation matrix revealed that 'tenure' and 'TotalCharges' are highly correlated.
- Distribution of churn: 26.5% churned, 73.5% did not churn.
- Key factors influencing churn: 'tenure', 'MonthlyCharges', and 'TotalCharges'.
- Analyzed relationships between categorical variables and churn.

EDA Observations:
1. The count plot and pie chart for churn show that a majority of customers did not churn.
2. Box plots and KDE plots for numerical features ('tenure', 'MonthlyCharges', 'TotalCharges') were analyzed to understand their distributions and relationships with churn.
3. Count plots for categorical variables like 'gender', 'Partner', 'Dependents', etc., were used to see how these features vary with churn.
4. Histograms and pie charts for customer demographics and services provided give insights into customer distributions.
5. A pair plot was generated to visualize relationships between different features and churn.

Machine Learning Models:
Three models were trained and evaluated:
1. Logistic Regression
2. Random Forest Classifier
3. Gradient Boosting Classifier

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	81.62%	67.92%	57.91%	62.52%
Random Forest	79.56%	65.92%	47.18%	55.00%
Gradient Boosting	80.55%	67.13%	52.01%	58.61%

Cross-Validation and Hyperparameter Tuning:
- Cross-validated F1 scores for Logistic Regression: [0.608, 0.617, 0.559, 0.602, 0.595]
- Mean CV F1 score: 0.596
- Best parameters from GridSearchCV: {'C': 100, 'solver': 'lbfgs'}
- Best cross-validated F1 score: 0.589
- Tuned model F1 score on test set: 0.624

Based on the evaluation metrics, Logistic Regression was found to be the most accurate model for this dataset.