Name – Binay Borun Mann

Reg. No – 554

1. Data Format and Features:

The dataset is loaded from a CSV file, and it contains information about Telco customers. Here are the key details:

- **Data Format:**
    - The dataset is stored in a CSV file.
    - Each row corresponds to a Telco customer.
    - Various features capture customer information.
- **Features in the Data:**
    - The dataset includes a mix of numerical and categorical features.
    - Some of the features include 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'OnlineSecurity', 'OnlineBackup', 'TechSupport', 'PaperlessBilling', 'StreamingTV', and 'Churn'.
    - The 'Churn' feature serves as the target variable, indicating whether a customer has churned ('Yes' or 'No').

2. Use of Entropy for Features:

- **Entropy Calculation:**
    - Entropy is calculated for each feature using the formula:
    - Entropy=$-\sum_{i=1}^{n} P_i * \log_2(P_i)$, where *Pi* is the probability of each unique value in the feature.
    - Entropy is a measure of impurity or disorder in a set of data.
    - Features with higher entropy have more diverse values, providing more information for predictive modeling.
    - So , we have selected the features with highest probability like Total charges, Monthly charges , tenure etc.
- **Skipping Features with Similar Domain Knowledge:**
    - Certain features like 'StreamingTV' and 'StreamingMovies' may have similar information about customers' streaming preferences.
    - In some cases, domain knowledge or feature importance analysis may lead to the exclusion of redundant or less informative features to avoid multicollinearity or overfitting.

3. Use of GradientBoostingClassifier:

- **Data Preprocessing:**

    - Unnecessary columns, such as 'customerID', 'gender', etc., are dropped from the dataset.
    - Remaining columns are one-hot encoded to handle categorical variables.
- **Target Encoding:**
    - The target variable 'Churn' is encoded into binary values (1 for 'Yes' and 0 for 'No').
- **Data Splitting and Standard Scaling:**
    - The dataset is split into training and testing sets (80% training, 20% testing).
    - Standard scaling is applied to the features to ensure consistent units and improve model performance.
- **GradientBoostingClassifier:**
    - Gradient Boosting is a machine learning technique that builds a series of weak learners (usually decision trees) to create a strong predictive model.
    - The **GradientBoostingClassifier** from scikit-learn is employed, initialized with a random state for reproducibility.
    - The model is fitted to the training data using the **fit** method.
- **Prediction and Evaluation:**
    - Predictions are made on the test set using the trained model.
    - The accuracy of the model is evaluated using the **accuracy_score** function from scikit-learn.
    - Additional evaluation metrics, such as precision, recall, and F1-score, can be obtained using the **classification_report**.

In summary, the report outlines the data format, justifies the use of entropy for feature analysis, and describes the application of the GradientBoostingClassifier for predicting customer churn based on preprocessed data.