**Text Classification on Consumer Complaint Dataset Documentation**

**Table of Contents**

1 Introduction

2 Explanatory Data Analysis and Feature Engineering

2.1 Data Loading

2.2 Basic Statistics

2.3 Handling Missing Values

2.4 Class Distribution Visualization

2.5 Category Mapping

2.6 Feature Engineering

3 Text Pre-Processing

3.1 Data Splitting

3.2 TF-IDF Vectorization

4 Selection of Multi-Classification Model

4.1 Random Forest Classifier

5 Comparison of Model Performance

5.1 Model Training

5.2 Model Evaluation

6 Prediction

6.1 New Data Processing

6.2 Making Predictions

6.3 Decoding Predictions

**1. Introduction:**

* This document provides a step-by-step guide for performing text classification on the Consumer Complaint Dataset.
* The goal is to categorize consumer complaints into four predefined categories: Credit reporting, repair, or other (0), Debt collection (1), Consumer Loan (2), and Mortgage (3).

**2. Explanatory Data Analysis and Feature Engineering:**

* **2.1 Data Loading:**

Load the Consumer Complaint Dataset from the provided CSV and JSON files.

* **2.2 Basic Statistics**

Explore basic statistics of the dataset using df.info ().

* **2.3 Handling Missing Values**

Identify and handle missing values in the dataset.

* **2.4 Class Distribution Visualization**

Visualize the distribution of classes in the 'Product' column.

* **2.5 Category Mapping**

Map product categories to numerical values for classification.

* **2.6 Feature Engineering**

Extract relevant features ('Consumer complaint narrative' and 'category') for further analysis.

**3. Text Pre-Processing**

* **3.1 Data Splitting**

Split the dataset into training and testing sets.

* **3.2 TF-IDF Vectorization**

Perform text processing using TF-IDF vectorization.

**4. Selection of Multi-Classification Model**

* **4.1 Random Forest Classifier**

Train a Random Forest classifier using the TF-IDF vectorized data.

**5. Comparison of Model Performance**

* **5.1 Model Training**

Train the Random Forest classifier on the training set.

* **5.2 Model Evaluation**

Evaluate the model's performance on the testing set using accuracy and classification report.

**6. Prediction**

* **6.1 New Data Processing**

Process new data for making predictions.

* **6.2 Making Predictions**

Use the trained model to make predictions on new data.

* **6.3 Decoding Predictions**

Decode numerical predictions back to category labels for interpretation.