**Project Documentation: Drug Classification Prediction using Web Scraping and Machine Learning**

**1. Introduction**

In this project, I aimed to predict the drug class of pharmaceutical drugs using a combination of **web scraping** and **machine learning**. The data for this project was extracted from the RXList website, which provides detailed information about drugs, including their **brand names**, **generic names**, and **drug classes**. By scraping this data and leveraging machine learning algorithms, I developed a model capable of classifying new drugs based on their names.

The objectives of the project were:

* To scrape and extract drug information from the RXList website.
* To preprocess the data and train a machine learning model.
* To evaluate the model’s performance and make predictions.
* To save the trained model, vectorizer, and encoder for future predictions.

**2. Technologies Used**

* **Python**: The programming language used to implement the entire project, including web scraping, data processing, and machine learning.
* **BeautifulSoup**: A library used for scraping and parsing HTML web pages to extract drug information from the RXList website.
* **Requests**: A Python library used to send HTTP requests to retrieve web pages.
* **scikit-learn**: A machine learning library used for model training, data preprocessing, and evaluation. Specifically, it was used for text vectorization (TF-IDF), label encoding, and training the Random Forest Classifier.
* **joblib**: A library used for saving the trained machine learning model, vectorizer, and encoder for future use.

**3. Web Scraping with BeautifulSoup**

The data extraction process involved scraping the RXList website. The key steps were:

1. **Accessing Drug Pages**: I created a list of URLs for each drug category alphabetically (a to z) to scrape data on all available drugs.
2. **Extracting Drug Details**: For each drug page, I scraped the following information:
   * **Brand Name**: The official name of the drug.
   * **Generic Name**: The generic equivalent or active ingredient of the drug.
   * **Drug Classes**: The medical classification(s) of the drug (e.g., antibiotics, pain relievers, etc.).
3. **Storing Data**: The extracted drug data was stored in a structured format (e.g., a DataFrame) for further processing and machine learning.

**4. Data Preprocessing**

After scraping the data, it was cleaned and transformed to make it suitable for machine learning:

1. **Handling Missing Data**: I ensured that there were no missing values in the dataset, removing or imputing data as necessary.
2. **Combining Features**: I combined the **Brand Name** and **Generic Name** columns into a single text feature to capture more comprehensive information for classification.
3. **Text Vectorization**: The combined text data was converted into numerical form using **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization, which reflects the importance of each word in the context of the entire dataset.
4. **Encoding Target Variable**: The drug classes, which were categorical, were encoded into numerical labels using **LabelEncoder** for compatibility with machine learning algorithms.

**5. Machine Learning Model**

The core of this project involved training a machine learning model to classify drugs into different classes. The steps were as follows:

1. **Splitting the Data**: The dataset was split into training and testing sets, with 80% used for training and 20% for testing.
2. **Model Choice**: A **Random Forest Classifier** was selected due to its ability to handle high-dimensional data and its robustness in classification tasks.
3. **Training**: The Random Forest model was trained on the preprocessed text data (converted into numerical vectors using TF-IDF) and drug class labels.
4. **Prediction**: After training, the model was tested on the unseen data (test set) to evaluate its performance.

**6. Model Evaluation**

The model’s performance was evaluated by calculating the accuracy score, which is the percentage of correctly predicted drug classes on the test set. The accuracy of the model was **86.91%**, indicating that the model is capable of predicting the correct drug class for most drugs with high reliability.

**7. Saving the Model**

To enable future use and predictions, I saved the trained model, vectorizer, and encoder:

* **Model**: The Random Forest Classifier was saved so it can be used to make predictions on new data.
* **Vectorizer**: The TF-IDF vectorizer was saved to ensure that new drug names can be transformed in the same way as the training data.
* **Encoder**: The LabelEncoder was saved to decode the predicted drug classes into their original labels.

These components were saved as **.pkl** files using **joblib**, allowing for easy loading and future predictions.

**8. Conclusion**

This project demonstrates the integration of web scraping and machine learning to build a predictive model for drug classification. By scraping drug data from the RXList website and, I successfully trained a Random Forest model that predicts drug classes with **86.91% accuracy**.

The project allowed me to showcase skills in:

* **Web Scraping**: Extracting and cleaning data from a live website using BeautifulSoup.
* **Data Preprocessing**: Converting text data into numerical features and encoding categorical labels.
* **Machine Learning**: Training and evaluating a model using scikit-learn.
* **Model Deployment**: Saving the trained model, vectorizer, and encoder for future use.