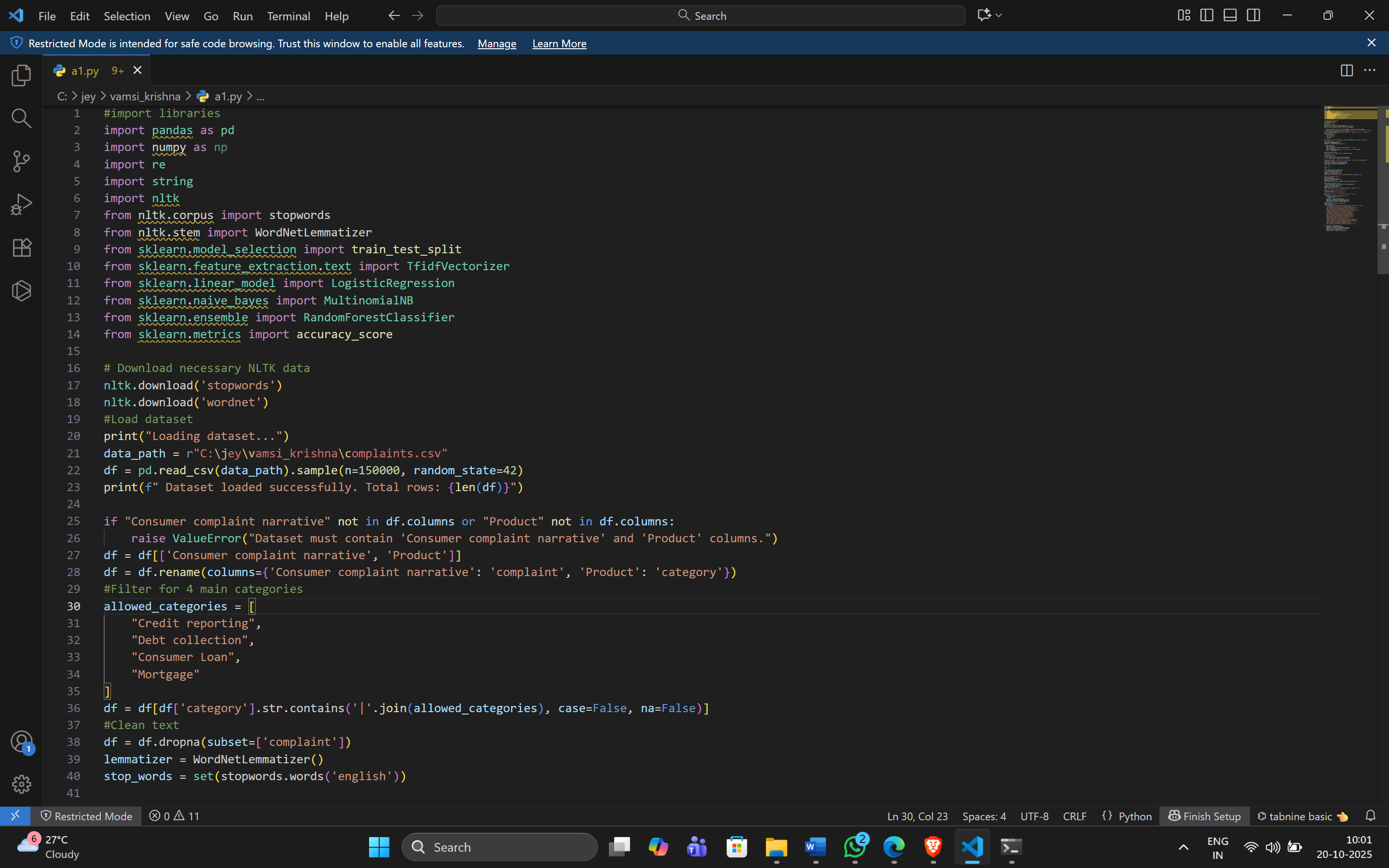
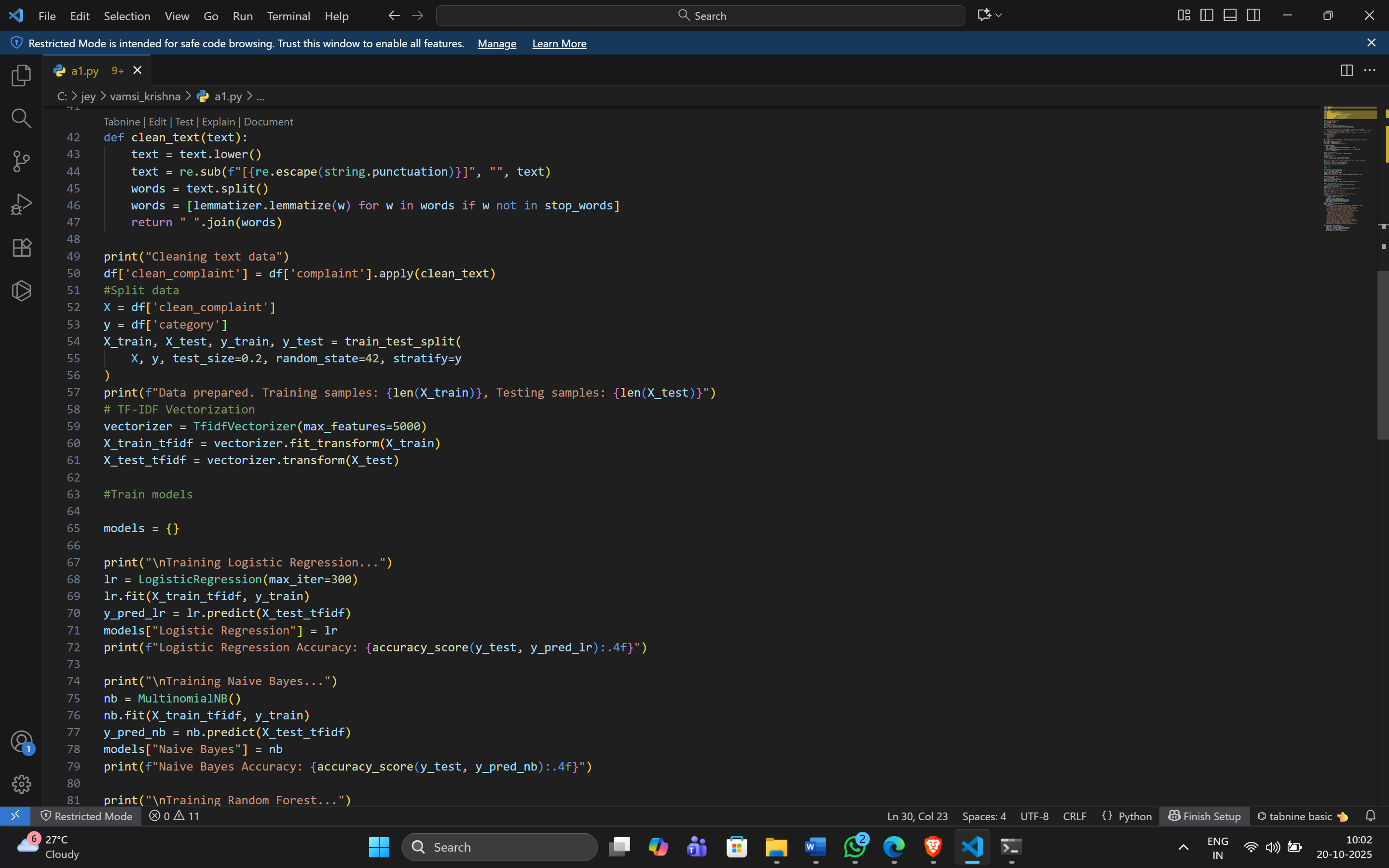
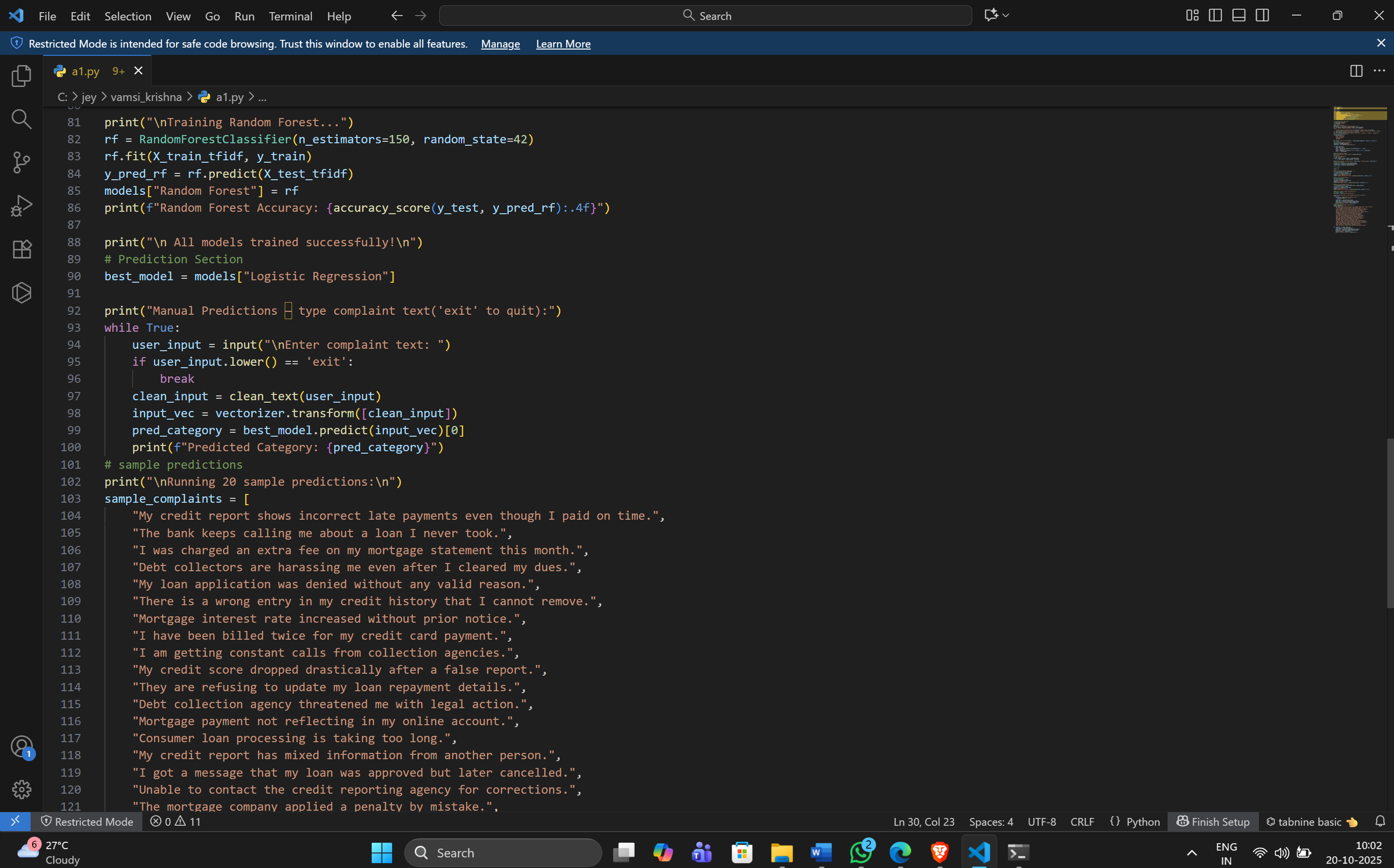
**KAIBURR-TASK-5**

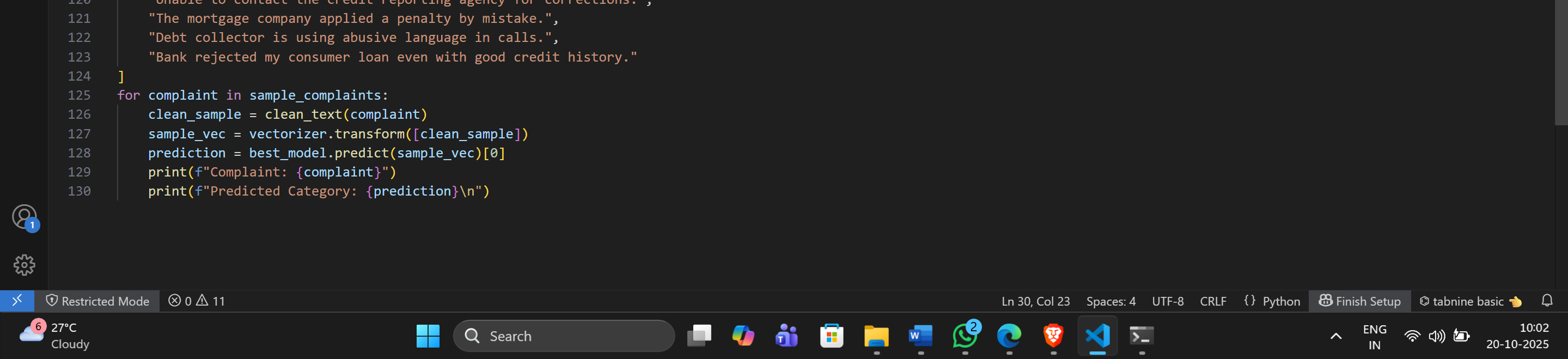
**DATASET LINK:** [**https://catalog.data.gov/dataset/consumer-complaint-database**](https://catalog.data.gov/dataset/consumer-complaint-database)

CODE:

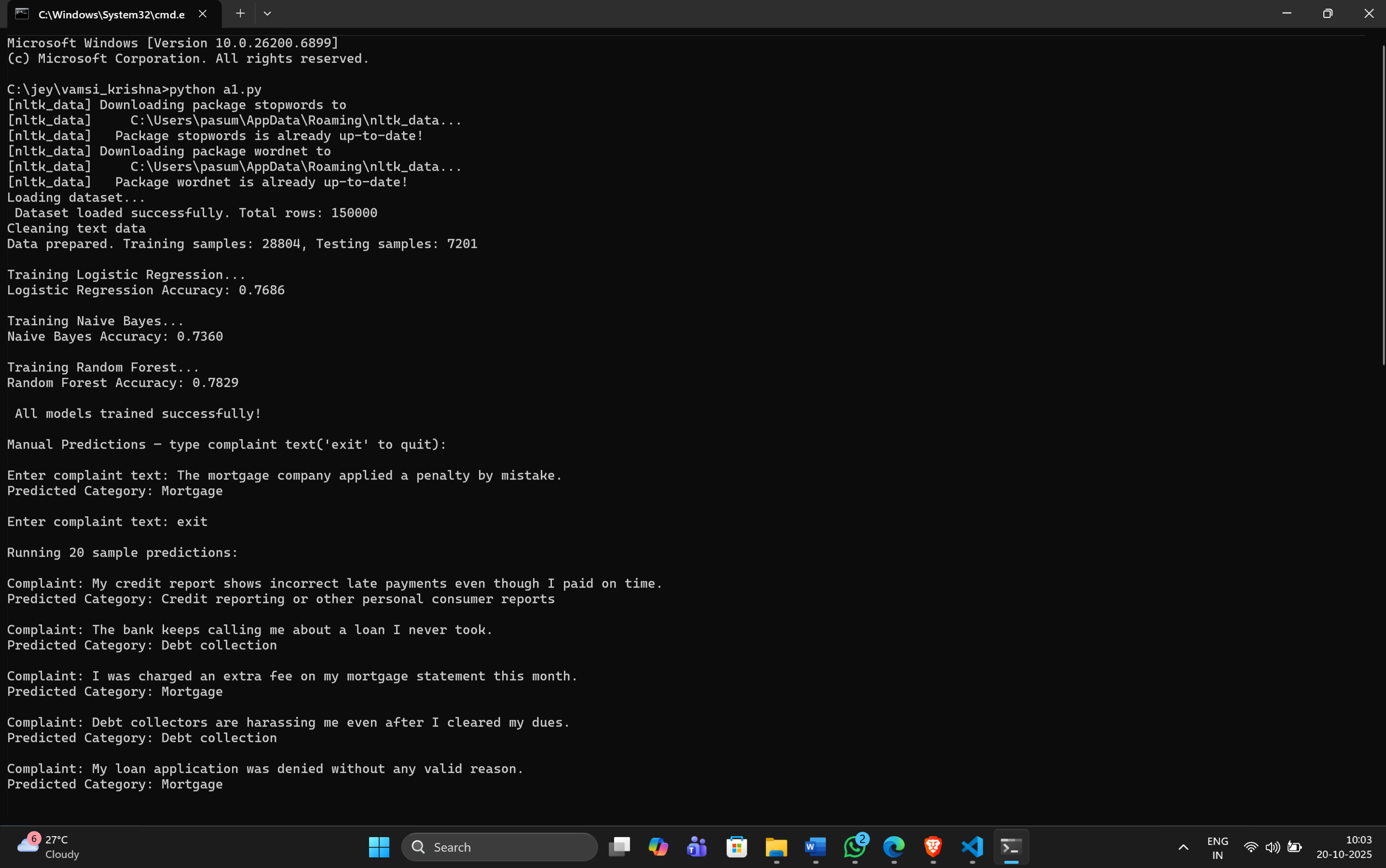








**Output:**



**Explanation:**

The first step of the entire procedure is loading the dataset with the dataset consisting of approximately 150,000 customer complaints regarding financial products and services. The data first undergoes a verification process to see if it has the appropriate columns which are primarily complaint text and product category. After that rows containing unnecessary data and those with missing values are deleted. In order to emphasize on the main issues, only four core categories are retained namely, Credit Reporting, Debt Collection, Consumer Loan, and Mortgage. This filtering process not only simplifies the classification task but also makes it more efficient.

The next step is the clutter clearing and text preprocessing phase. The complaint texts are all changed to lowercase, the characters are stripped off, and the loquaacious terms (known as stop words) like "the", "and", or "is" are removed. Further, the technique of lemmatization is applied to reduce words to their root (for instance, "running" turns into "run"). Consequently, it is ensured that all complaint texts are clean, consistent, and prepared for the next stage of analysis. This is a significant part of NLP that enhances the model accuracy by concentrating on the significant words.

Once the text was cleaned, it was then converted to numerical format by means of a TF-IDF vectorizer (Term Frequency–Inverse Document Frequency). This system translates the words into numbers depending on how crucial they are in each complaint as compared to the whole dataset. The purified complaints are then divided into two parts — one that is for training (80%) and the other that is for testing (20%). The training data works with the models and helps them learn while the testing data measures how accurately the models can predict new, unseen data.

Following this, the process continues with the training of different machine learning models such as Logistic Regression, Naive Bayes, and Random Forest. Through this learning, each of the models gets to know the interrelationship between the complaint text and its category. When the training ends, the models are tested to measure their accuracy. The results show that Logistic Regression achieved around **76.86%**, Naive Bayes **73.60%**, and Random Forest **78.29%** accuracy. Random Forest performed the best, making it the most reliable model for predicting complaint categories.

Finally, the trained model is used for real-time prediction. When a user types a new complaint, the system cleans and processes it just like the training data, then predicts which category it belongs to. For example, if someone enters *“The bank keeps calling me for a loan I never took,”* the model predicts it as Debt Collection. This approach helps automatically identify complaint types and can be used by companies to sort and address customer issues faster and more accurately.