**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Product Classification using their Ingredient**

**Week 4 Report**

| **Name** | **Enrolment Number** |
| --- | --- |
| Aharnish Pithva | AU2040022 |
| Jevin Jivani | AU2040051 |
| Astha Bhalodiya | AU2040067 |
| Yug Patel | AU2040181 |

**1) Tasks Performed in the week.**

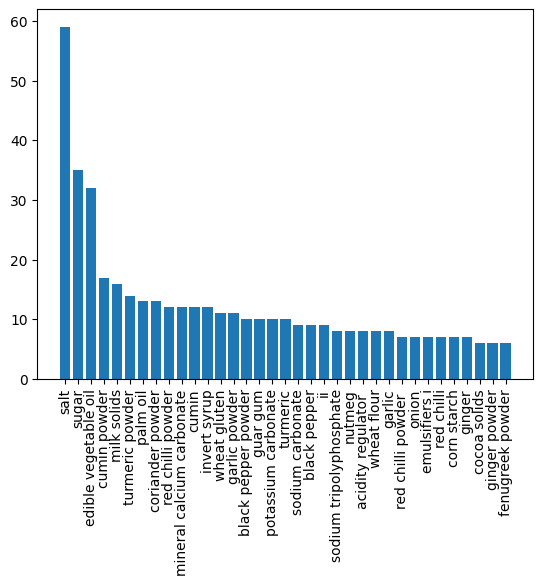
Data was gathered for the category of food and beauty products. Afterwards, we conducted data cleansing and analysis specifically for the food data to isolate the list of ingredients and eliminate any extraneous or irrelevant values. This process allowed us to refine the data and obtain a more accurate and useful representation of the relevant information.

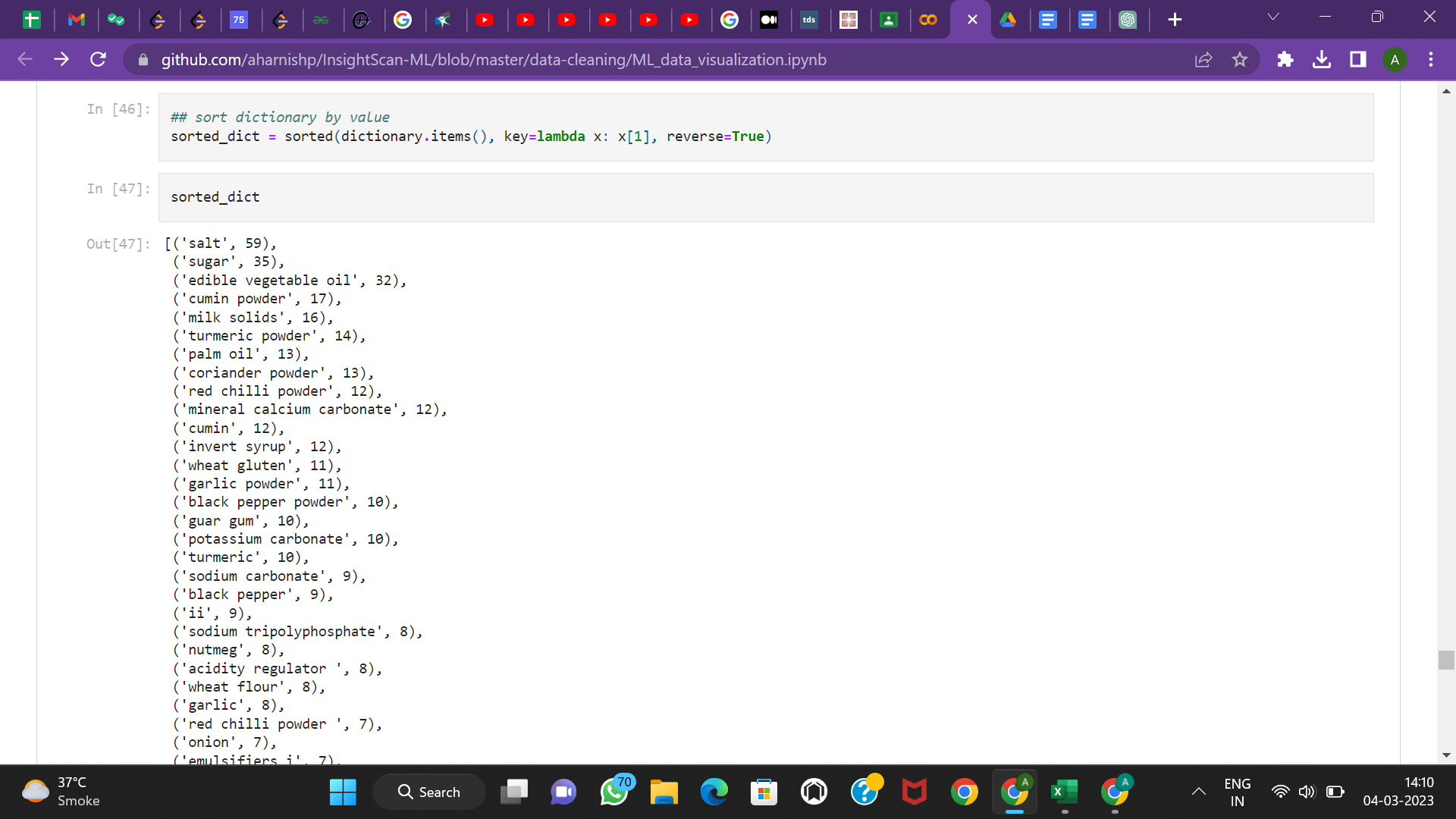
Steps of Data Analysis:

* Begin by uploading the dataset to the collab file.
* Create a list of all the ingredients (by checking if it has another character, then space and a new line)
* Separate the ingredients in the list by splitting them from ','.
* Store all the separated ingredients in a CSV file.
* Count the occurrence of each ingredient in the dataset.
* Sort the output in descending order, with the ingredients having the most occurrence coming first.
* Create a histogram with all the ingredients on the x-axis and the number of occurrences on the y-axis.
* Create another histogram with the top 35 ingredients on the x-axis and the number of occurrences on the y-axis.

**2) Outcomes of the tasks performed.**

We obtained a clean dataset containing information about food ingredients, free from any irrelevant or incorrect data. Additionally, we analysed the training dataset to determine the frequency of each ingredient and presented our findings through two histograms. The first histogram displays the total count of all ingredients, while the second focuses on the top 35 most frequently occurring ingredients.





*Google Collab Link:* [*https://colab.research.google.com/drive/1zrkQWYC-mCPeJHsH6-7QnELE390\_BTey?usp=sharing*](https://colab.research.google.com/drive/1zrkQWYC-mCPeJHsH6-7QnELE390_BTey?usp=sharing)

**3) Tasks to be performed in the upcoming week.**

We intend to apply various algorithms such as logistic regression, decision tree, naive base, SVM, KNN, and Random Forest to the analysed data. We will then evaluate the accuracy of each algorithm to determine which one yields the best results to use for the project further.