10/27/2020

STATEMENT OF WORK-1002

MARCOS BITTENCOURT

Roshandeep Singh Saini
100766638

Contents

1.1 Executive Summary

An analysis is conducted to predict whether a given dish is vegetarian or not, depending on the ingredients used in the dish. In simpler terms it is expected that that a dish with same ingredients, would result in a similar dish. Predictive modelling techniques would be used to find a solution which best classifies the dish as vegetarian or non-vegetarian in terms of diet.

1.2 Problem Statement

The problem statement is "To classify a dish as vegetarian and non-vegetarian based on the ingredients".

Based on the ingredients used to prepare an Indian dish, it can be classified as a part of a vegetarian diet or a non-vegetarian diet. To gain insights on how dishes with similar ingredients can be classified similarly or differently based on region, flavor, and course.

1.3 Analytics Rationale Statement

The rationale of the analysis is to accurately predict whether a dish can be called a vegetarian dish or a non-vegetarian dish depending on the contents, flavor profile, or the course of meal. Dishes with similar ingredients can be profiled as a different course of meal as per Indian cuisine. Hence, the aim is to quantify the similarity and disparity.

1.4 Data

The data has been acquired from Kaggle open datasets. It is a raw dataset named "indian_food", which represents entirety of the testing data for August 5th to October 5th, 2020. It contains details about Indian food, the features used to classify them.
It has 255 records. These variables are: -

- Independent Variables- Actual measurement parameters of an Indian Dish which are name, ingredients, prep_time, cook_time, flavor_profile, course, state, region.

- Dependent Variable- Classification of dish as part of the diet(vegetarian and non-vegetarian).

1.5 Output Variable Structure

The output variable will have data divided in 2 classes “Vegetarian” and “Non-Vegetarian” as this is a binomial classification problem. The hypothesis is that the ingredients contained in Vegetarian and Non-Vegetarian dishes would have close cosine similarity.

1.6 Data Analysis Approach

The methodology used will be to develop a classification model using machine learning approach to segregate Indian dishes into “vegetarian” and “non-vegetarian”. It is a scenario of Supervised Learning. This would require classification algorithms SVM, Decision Tree and an advanced modelling through TensorFlow(Keras) etc. depending on the complexity and the type of data received. This would be determined in the EDA and modelling phase of the project. The software tools used will be: -

- Python – for EDA (Exploratory Data Analysis), Data Cleaning, Model Building and Testing
- Jupyter Notebook – IDE for development
- Tableau/Power BI – for detailed visualizations

1.7 Assumptions And Constraints

The assumptions made for the given problem statement are: -

- The dataset acquired from Kaggle is a valid dataset, it is assumed that the records correspond to actual features and accurate measurements/description of the Indian dishes that were considered.

- All the features in the dataset corresponding to “indian_food”, acquired from Kaggle are required for the analysis and the features are independent of each other and are collectively required for tackling the problem statement.

- It is an assumption that the model is restricted to a binomial problem(Vegetarian and Non-Vegetarian) rather than multi class. Thus, this model can only be used to this specific problem statement.

The constraint in the entire scenario are: -

- No more data can be obtained. For the analysis and model building is limited to the given data only. This is a hard constraint that we have no control over.
- The above analysis is limited to the number of features available in the dataset, no more features can be obtained or used for the modelling purpose. This is a hard constraint that we have no control over.
- The classification model being built is a binomial classification problem, thus it cannot be used for any other study. The entire data set is classified in terms of “Vegetarian” and “Non-Vegetarian” Indian dish, so the model answers specifically to the above problem statement. This is a soft constraint, but it also depends on the problem statement.

1.8 Analytical Scorecard

The metrics that will be used to evaluate the models are accuracy and AUC/ROC curve.

Accuracy, because in this classification problem the model needs to perform good in the overall classification. It is a measure of all the correctly identified cases and is required when all the classes (Vegetarian and Non-Vegetarian) are equally important.

AUC/ROC, The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values. Therefore, we would get a balance evaluation for the model performance.

1.9 Exploratory Data Analysis Action Plan

The EDA was done on the cannabis dataset according to the stated EDA Action Plan and steps were taken to clean and transform the data in corresponding to the problem statement and assumptions

1.9.1 EDA Summary

Label Conversion - Initial steps correspond to transforming the data to fit the problem statement, tokenization and vectorization of “ingredients” feature, binary data encoding for the target variable “diet” as Class 0 for “Non-Vegetarian”, Class 1 for “Vegetarian”.

NaN Values – EDA shows there is no null values in the dataset but features region and state have a value “ -1 ” which is replaced NaN so that “-1” is not treated as a separate value but an

unknown value. It is not replaced with a synthetic value because it is a small dataset and even 1 value addition to a any group would create an unbalance.

Categorical Variables Encoding – Features like flavor_profile, course, state, and region are categorical variables which needed to be one-hot encoded before model building. A new binary variable is added for each unique integer value for model to work on numerical input where for categorical variables where no such ordinal relationship exists.

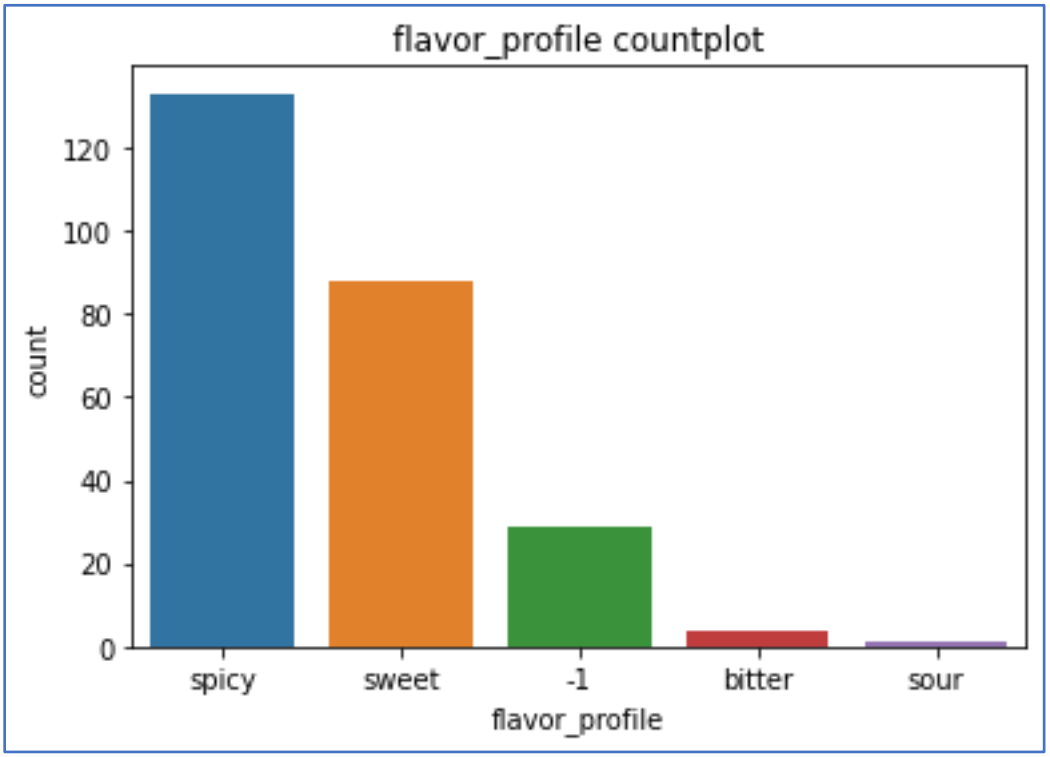

FIGURE 1 FLAVOR PROFILE DISTRIBUTION

Dependent Variable Labeling - The dependent variable "diet" was labelled and converted to numeric to be better used by the model Class 0 for "Non-Vegetarian", Class 1 for "Vegetarian".

Correlation - The correlation heatmap was used to check the correlation between the features and no highly correlated feature were indicated. All correlations are less than 0.8, indicating low correlation.

Duplicates – There were no duplicates in the dataset all the dishes were unique. It was comparatively a very clean dataset

Imbalance - The clean dataset has an imbalance of 226:29 for Vegetarian: Non-Vegetarian. This could have been handled by SMOTE technique before model building but because of the underlying architecture of SMOTE it would lead to overfitting of Non-vegetarian data points. The only solution for this is more data.

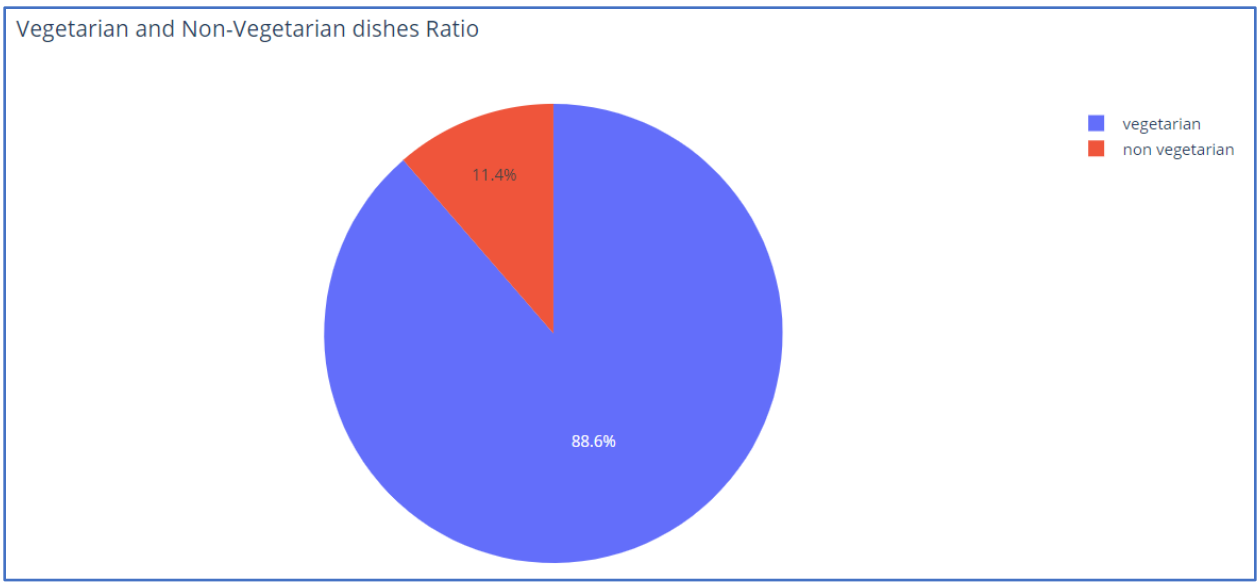

FIGURE 2 IMBALANCE IN DIET DISTRIBUTION

Vectorization - The ingredients need to be tokenized for further analysis. This is the most important feature of the model where the diet is predicted mainly using the ingredients. The food ingredients are taken, and vectors are created for each and every dish. This is similar to label encoding. This is done so that the algorithm can process the contents of the dish in the prediction process. The vectors are of the shape (255, 337) or (no of dishes, no of total ingredients)

Cosine Similarity - Ingredient vectors are used to check the cosine similarity between 2 dishes. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two vectors are far apart by the Euclidean distance, chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

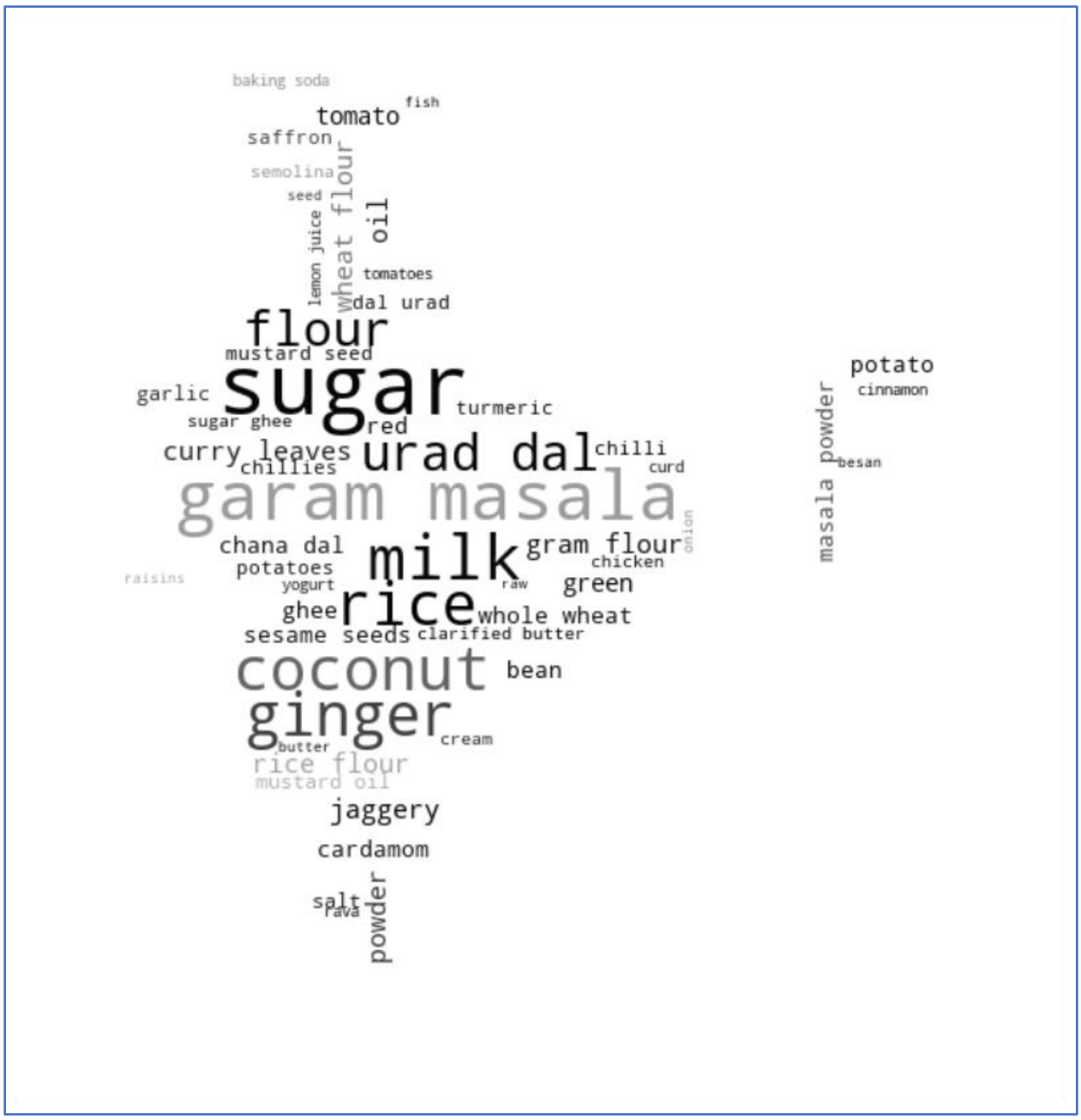

FIGURE 3 INGREDIENTS USED IN INDIAN DISHES

1.9.2 EDA Insights

The three key insights from the EDA process are following:

- No high correlation is seen between the independent variables, but on model building PCA might be required. There is little or no multicollinearity within the independent

variables. Logistic Regression can be used as a base model to get an approximation of outputs.

- In the Heatmap 0 - 66th (these numbers correspond to index of "data frame") ingredients vectors have high cosine similarity each other. Cosine similarity to calculate similarity of ingredient vectors. If cosine similarity between two foods is high, it can be inferred that dishes are similar.
- There is a large imbalance between the two classes (Vegetarian and Non-Vegetarian) which might affect the model.
- As seen in the EDA dishes at location 9, 14, and 15 are sweet dishes and have very similar ingredients. Therefore, the cosine similarity between the ingredient vectors is high, indicating actual closeness between the dishes. Also seen in the EDA dish at location 30 is a savory dish compared to a sweet dish. Therefore, having a small cosine similarity between them, Indicating no actual closeness between the contents of the dishes.

1.10 Project Plan

The table below contains the project tasks and their estimated completion dates.

Phase	**Task**	**Deliverables**	**Delivery Date**
Project Organization	Setup	Setting up a GitHub repository for the project deliverables. Setting up the project structure.	October 28, 2020
Business Understanding and Problem Discovery	Statement of Work(V1)	Developing a problem statement and a rational statement for the business problem. Refining the problem, identifying the accurate requirements, assumptions, constraints for the business problem. Acquiring the data sources relevant for solving the business problem, analysing the data source for restrictions and limitations. Defining parameters to test the results and setting up a benchmark for the analysis.	November 1, 2020

Data Acquisition and Understanding	Data Acquisition and Understanding, Statement of Work(V2)	Performing preliminary EDA on the data set for basic statistics. Identification of the steps required for feature engineering, normalization, scaling, cleaning, outlier removal, conversion to numerical features from strings and categories, correlation. Data preparation for a base model prediction. Selection of models required for advance classification. Detailed documentation of the complete EDA process.	November 23, 2020
ML Modelling and Evaluation	Advanced Predictive Modelling	Developing the base model and analyzing the results. Evaluation of base results. Development of architecture and pipelines for advance classification algorithms. Using optimization techniques for further improvement and evaluation of results. Prototyping the model, tweaking the parameter to attaining the proposed results for the evaluation metric selected.	November 23, 2020
Delivery and Acceptance	Deployment of Software Pipeline	Development of a software pipeline to consume the results from the model solution. Providing and end-to-end solution for the final user to be consumed. To submit the results, solution, and a detailed presentation of the entire analysis to the instructor.	December 18, 2020