**Future Food Customer Needs**

|  |  |
| --- | --- |
|  |  |
| **Lecture** | Data Science  summer semester 2020 |
|  |  |
| **Participants** | Rahman, Safie Ur (2572863), s8sarehm@stud.uni-saarland.de |
|  | Rizvi ,Syed Taqi Abbas (2577651), s8syrizv@stud.uni-saarland.de |
|  | Munir, Talha (Matr. Nr.), E-Mail address |
|  |  |
| **Submission date** | 07-17-20 |
| **Chair** | Univ.-Prof. Dr.-Ing. Wolfgang Maaß  Chair in Information and Service Systems,  Campus A5 4, 66123 Saarbrücken |
|  |  |
|  |  |
|  |  |

**Executive Summary**

Text…

**Contents**

[**Introduction**](#_30j0zll) **6**

[**Data Set**](#_1fob9te) **8**

[**Procedure and Analysis**](#_3znysh7) **8**

[**Individual Approach**](#_8oja8z3212yr) **8**

[**Example Outline Level 2**](#_2et92p0) **9**

[**Example Outline level 3**](#_tyjcwt) **9**

[**Example Outline Level 2**](#_3dy6vkm) **9**

[**Results and Discussion**](#_1t3h5sf) **9**

**List of Abbreviation**

Please insert a list of abbreviations here, if necessary. Otherwise please delete this item.

**List of Figures**

Please insert a list of figures here, if necessary. Otherwise please delete this item.

**List of Tables**

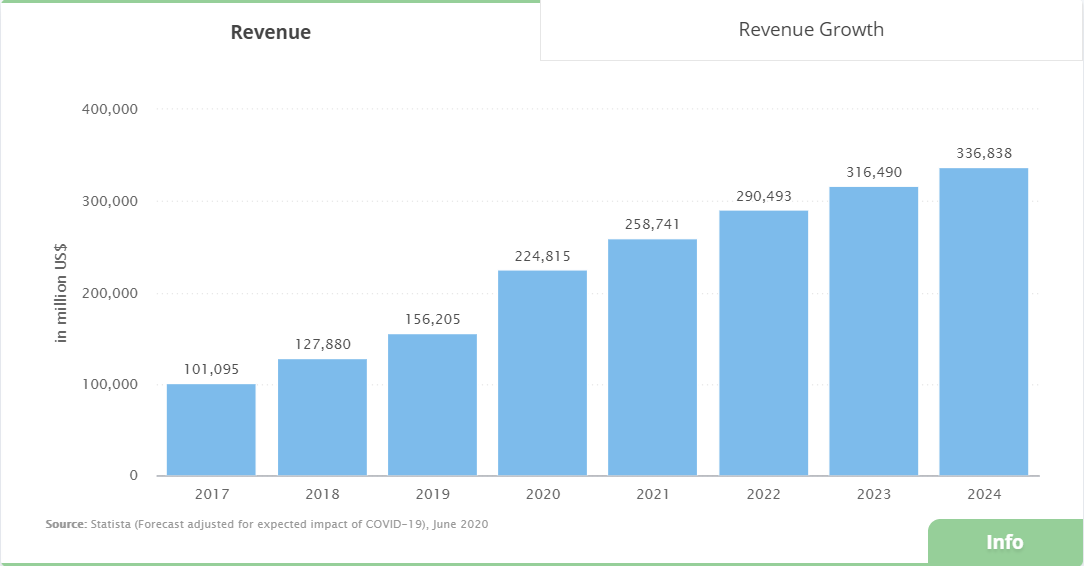
Please insert a list of tables here, if necessary. Otherwise please delete this item.

# Introduction

Please write in whole sentences and use the following key points as assistance.

* Problem statement.
* Current state of research on the topic under discussion.
* Research gaps.
* Please insert more information if necessary.

The following report presents the results from a data driven service to the task of finding the food trends in the next generation which can be utilized by the food producers to develop/update their product portfolio. The Food and Beverage industry as reported by Statistica[[1]](#footnote-1) is projected to reach a whopping US$224,815m in 2020 and is expected to show an annual growth rate of 10.6%.



Such a massive impact also invites detailed research to predict the future needs and every now and then, research is carried out to fore-see the future trends in general as reported by Fraunhofer ISI[[2]](#footnote-2) or specific to an age bracket as for 13-36 year olds as reported by YPulse[[3]](#footnote-3) among several other researches. It is important to note that the researches that we encountered during our literature survey in this short span of time just laid out the trends that could happen without explaining the reasons or factors contributing to that change. In this study, we would give recommendations considering these underlying causes based on the given data that may establish confidence in the results returned by other researches as mentioned before. We will overview the Data Set in Section 2, Procedure and Analysis in Section 3 and finally present the Results and Discuss them in Section 4.

# Data Set

Please write in whole sentences and use the following key points as assistance.

* Origin of the data.
* Data record description.
* Data availability.
* Data preparation.
* Please insert more information if necessary.

Two data sets were used for the service, both available on Kaggle[[4]](#footnote-4). One[[5]](#footnote-5) lists the food choices and preferences of students of Mercyhurst University and spans 126 uncleaned and raw responses in 61 columns from students on food choices, nutrition, preferences, childhood favorites, and other miscellaneous information. The other dataset[[6]](#footnote-6) is more generalized and records 288 unique responses in 8 columns from participants from different countries and demographics.

# Procedure and Analysis

Please write in whole sentences and use the following key points as assistance.

Your individual approach to the problem.

Description of the procedure and the methods used.

Please insert more information if necessary.

## Individual Approach

The task to predict the future food trends specific to a food producer was pretty challenging for us with the provided raw and unstructured data. What added more to our misery was that one of our group mates, “Khizar Razzaque” left us uninformed pretty early in the project. We still carried on to gather data and overview how the trends are categorized as previewed in Section 1. Once we did this, each of us worked on defining Use Cases to visualize what would be expected of the service, The inputs that we’ll feed and output we expect. Following a brainstorming meeting, we agreed on a Use Case (more on this as we proceed) and decided to cluster the agreed features through K-Means clustering and Polynomial Regression.

Finding dependencies Part (to be added) and others!!

## Example Outline Level 2

This is an example of how to insert subchapters. Please delete if not needed.

### Example Outline level 3

This is an example of how to insert subchapters. Please delete if not needed.

## Example Outline Level 2

This is an example of how to insert subchapters. Please delete if not needed.

# Results and Discussion

Please write in whole sentences and use the following key points as assistance.

* Results and Discussion of your analysis.
* Short outlook.
* Please insert more information if necessary.

- Simple Clustering+ Word Cloud perform at the start, result & insights shown here

- Clustering results and insights shown here

**Key Features of the Predictive model**

Prediction algorithm has been implemented in order to find the most important factors that affect the likeness of each food.

The table attached below describes the top 5 features for each food and how one unit change in

these features would effect the probability of the likeness of food.

|  |  |  |  |
| --- | --- | --- | --- |
| Italian | Asian | American | Spanish |
| Income  (+0.4) | Tortilla Calories  (-0.3) | Coffee  (-0.34) | Favorite food  (-0.23) |
| Cuisine  (+0.3) | Ethnic Food  (+0.26) | Ethnic Food  (-0.33) | Comfort food  (+0.28) |
| Tortilla Cuisine  (+0.29) | Exercise  (+0.26) | Thai Food  (-0.25) | On\_off\_campus  (+0.27) |
| On\_off\_campus  (-0.22) | Comfort-food  (-0.19) | Income  (-0.21) | Waffle Calories  (+0.27) |
| Exercise  (-0.19) | Waffle Calories  (-0.18) | Favorite Food  (+0.21) | Coffee  (+0.23) |

It is observed from the table that every food has different features that would affect their attractiveness. These weights importance’s will tell us how much increase in one unit of any feature would increase the likelihood of the particular food.

There occurs very interesting pattern in these features. If an income of student is increased by one bin according to our binning of income data, probability that student food like Italian food is increased by 40% where probability of American Food is decreased by 21%. Similarly if a student has better exercise routine, likelihood of student choosing exercise would be increased by 26%.

If the data of the customers is already available, food producers could start producing foods by looking at these factors. For example, if the food producer is operating in the particular vicinity and he has some representative data of the students residing there, then based on these factor, he could decide whether to produce expensive Italian food or cheap American food etc.

Predictive algorithm goes hand in hand with the clustering we have done above as Clustering will be used to understand the different cluster of the students and find distinctive behaviors using multiple attributes within each cluster while predictive algorithm will provide information about each attribute of the student and how one attribute affect the student preference of food.

**Trend Tracking**

This prediction algorithm is also used to track the trends in the data. So if the accuracy of the model is significantly drops with the passage of time, it could be inferred that the behavior of the customers have been changed and we need to update our model.

We could also design the **self-learning pipeline** which could automatically check whether the trends have been changed based on some evaluation metric like accuracy. If there is change in trend, we could initialize the self-learning pipeline which would retrain the model as well as do the clustering again in order to give the updated behavior of our customer

* Bibliography

Images of all the results and jupyter notebook screenshot for data exploration, data cleaning , results and insights shown here. Or we could refer some of these image available in jupyter notebook(In this case , we need to mention jupyter cell no where we mention the result or insight in the report)

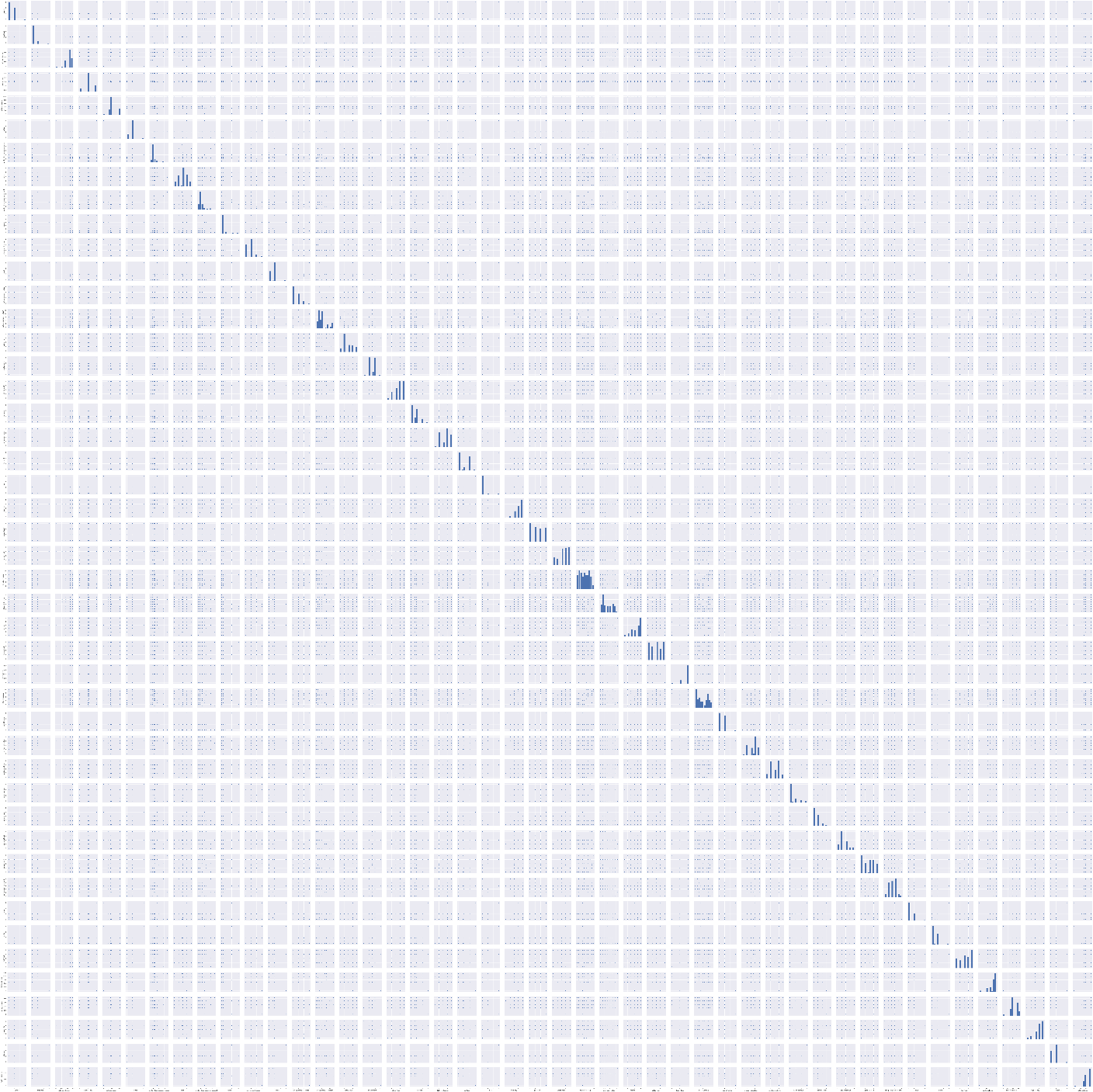
******

Figure 1: Scatter Plot

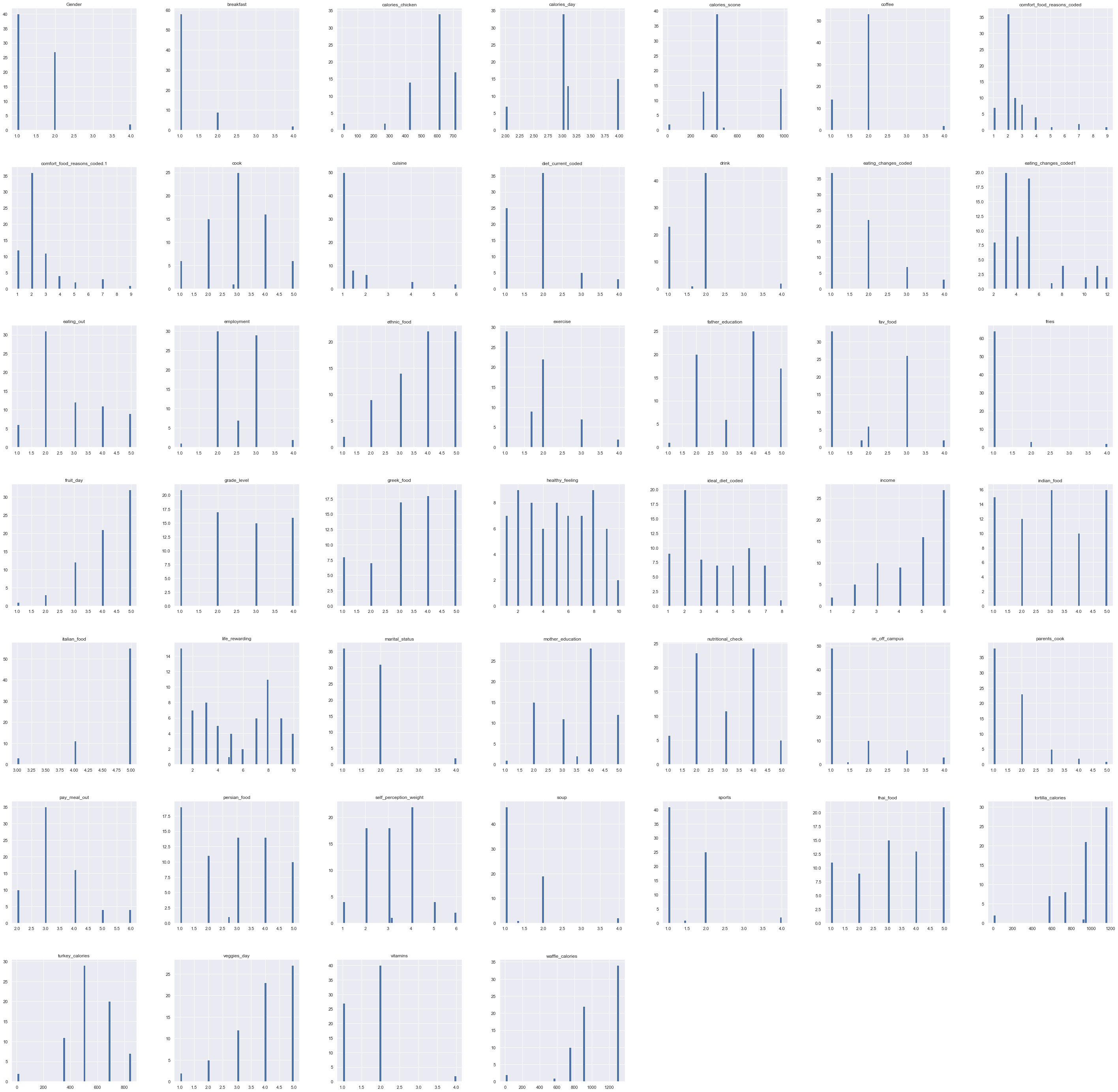
******

Figure 2: Histogram

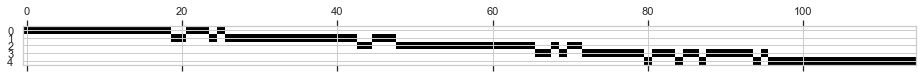
******

Figure 3: K Fold Stratified Cross Validation

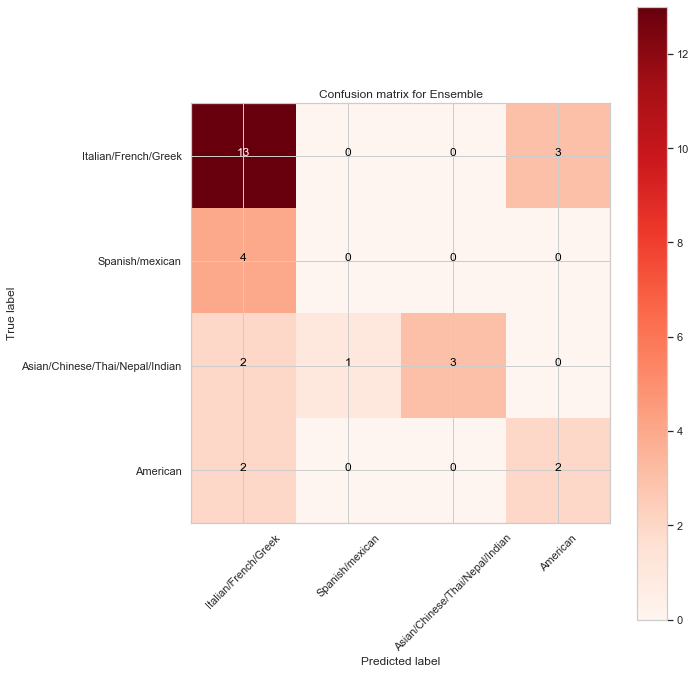
******

Figure 4: Multi Class Confusion Matrix

**Appendix**

Please insert here your **Description of the Code**.

Please write in **whole sentences** and use the following key points as assistance.

* Detailed description of the data organization (like cleaning and preparation of the data).
* Description of architecture of the data.
* Please insert more information if necessary.

Business Question

* We first define 3 questions(email question) by looking at the data. Decide the answers of these 3 question and then we move forward.
* Need to describe the column we are making prediction and doing clustering

Data exploration

After setting the business objectives, efforts were directed towards real understanding of the data and its semantic. Due to large number of available columns, mind map was developed to cluster similar column in the same place. It help in understanding different dimensions of the available data. Mind map image has been attached in the appendix.

After creating mind map, Histograms were made in order to check the skewness in the data. It gave an idea where most of value of each column lies. Image has been added in the appendix.

In the next step, Scatter plot and correlation matrix were visualized to check what type of different relationships are present among columns and check if there are any outlier which could affect. It assisted us in understanding the numerical and categorical data. Image has been added in the appendix. (Add library or function to create these in scikit)

Word Clouds are developed based on Tf-idf to understand the semantic of string data. It helped us in learning the most important terms used in this type of columns. This word cloud also proves beneficial in performing clustering at the later stage.

All of these plots and techniques proved useful in comprehending the underlying patterns in the data and gave us good overview of data we have.

Data Cleaning and Preparation:

Data Cleaning is performed to identifying and correcting the irregularities in the data. For this purpose, we have replaced the missing numerical value in a column with the mean value using ‘SimpleImputer’ class in the scikit-learn. For the string columns, we have replaced it with the relevant terms as comes out using tf-idf vectors.

In the next step, standardization is performed in order to smooth the data. It is also needed because for the prediction algorithm, we need to check the relative importance of each variable with the output variable which requires all the data columns standardized and normalized.

During data preparation step, we have simplified some assumption only for the predictive part of our project. We have been predicting the favorite food for the students. Due to vary small amount of data available, we have drop some rows in the dataset which belong to categories having less than 5 rows. It will help us in reducing the over fitting in the model.

(Check on the validation SET)

Data Splitting

In order to develop appropriate model according to the real world scenario and to avoid the risk of over fitting, we have split the data into train set, validation set and train set.

Stratification is used in splitting in order to avoid risk of clustering all rows of one type of pattern/class in one set and other types of rows in second set. It will also provide representative sampling of all the classes in every set.

Feature Selection

In the predictive module of the project, we have used different techniques to perform feature selection.

* Since most of the data is categorical in nature , so we have performed Chi-Square test to check the relative importance with the target variable
* Categorical data is ordinal in nature so correlation could also be used to see the pattern which gave good insights for the project.
* Random Forest is used for the feature selection as well because it performs well on our limited available dataset.
* Feature appears in Chi-Square and Random Forest both are also used to perform the modeling part.

Feature selected in all these techniques are shown in the jupyter notebook on the relevant cell.

Data Modeling, Hyper parameter Tuning and Optimization

The training set as gathered from the previous steps is then subjected to Linear Classification (Multinomial and OneVsAll), Gradient and Ada Boosting models in a Grid Search with multiple attributes in parameter grid to select the hyper parameters that best fit the data respectively.

The best parameters sought out are retained and used on the test set to get accuracies. Furthermore, the instances of the above models with best parameters through grid search are passed to Ensemble Voting Classifier that does a ‘soft’ classification to get the final model for the data.

The resultant model is also subjected to K Fold Stratified Cross Validation. This takes the combined set (training and testing) in k-folds and cross validates the model on the shuffled set to be sure that the selected model doesn’t over fit to a specific chunk of data.

The standard deviation that results from the K Fold Cross Validation approves of the fact the resultant model is generalized on the whole data and not over-fits.

Data Evaluation

For evaluation purposes, a confusion matrix is generated that lists down the precision, recall, f1-score, support, average, macro and weighted average values for the multi classes ‘Italian/French/Greek’, ‘Spanish/Mexican’, ‘Asian/Chinese/Thai/Nepal/Indian’ and ‘American.’

Moreover, we also generate the lifts for the multi-classes relevant to baselines.

Also we could create a pictorial representation of pipeline showing all these steps.

* **Declaration of Authorship**

I affirm that I have produced the work independently, that I have not used any aids other than those specified and that I have clearly marked all literal or analogous reproductions as such.

Location, Date: Saarbrucken, July 17, 2020

|  |
| --- |
| Rahman, Safie Ur |
| Rizvi, Syed Taqi Abbas |
| Munir, Talha |
|  |

1. "Food & Beverages - worldwide | Statista Market Forecast." <https://www.statista.com/outlook/253/100/food-beverages/worldwide>. Accessed 21 Jun. 2020. [↑](#footnote-ref-1)
2. "50 trends influencing Europe's food sector by 2035." <https://www.isi.fraunhofer.de/content/dam/isi/dokumente/ccv/2019/50-trends-influencing-Europes-food-sector.pdf>. Accessed 21 Jun. 2020. [↑](#footnote-ref-2)
3. "The 14 Biggest Food Trends Gen Z & Millennials Are ... - YPulse." 1 May. 2019, <https://www.ypulse.com/article/2019/05/01/the-14-biggest-food-trends-gen-z-millennial-are-interested-in/>. Accessed 21 Jun. 2020. [↑](#footnote-ref-3)
4. "Kaggle." <https://www.kaggle.com/>. Accessed 21 Jun. 2020. [↑](#footnote-ref-4)
5. "Food choices | Kaggle." 23 Apr. 2017, <https://www.kaggle.com/borapajo/food-choices>. Accessed 21 Jun. 2020. [↑](#footnote-ref-5)
6. "Food Preferences | Kaggle." 18 Jul. 2019, <https://www.kaggle.com/vijayashreer/food-preferences>. Accessed 21 Jun. 2020. [↑](#footnote-ref-6)