**The FoodBase Corpus**

**Analyzing the Predictability of Meal’s Type by its Ingredients and Characteristics**

**Patrick Aquino and Christopher Fiaschetti**

**Introduction**

Coupled with water and air, food completes the trifecta of the quintessential components of sustained life. More and more is being learned about it every day and the necessity for insights from a data science perspective is ever growing. As such, Popovski, et. al, have composed the FoodBase corpus containing thousands of recipes for analysis. With this data in mind, the team came up with a few research questions to direct our study:

1. Can a reliable model be built with this dataset to classify a recipe as an Appetizer and Snack, Breakfast and Lunch, Dessert, Dinner, or Drink?
2. Did Popovski, et al. tag the recipes up to the level that can be told by the ingredients in the recipes?
3. What meal categories are most similar? Most distinguishable?

Aside from personal interest from the researchers, these questions are worth asking because they can help both corporations, restaurants and individuals in a number of ways. This data could be utilized crafted into an application where individuals type in ingredients they have and want to make a meal out of. Coupled with the intended meal type and individuals could discover new recipes they didn’t know about and broaden their culinary horizons. Restaurants can see what people are searching for in their area and try to make similar dishes or even have a competition to see who can make the best dish in the town. Corporations would be interested because they could see what ingredients are searched for in certain areas and provide more of them in those areas as well as figure out what isn’t being searched for, could be added to that area and increase their profits. All of this relies on accurate models being created, which is what was attempted in this project.

In terms of splitting up the project, Patrick handled the EDA and poster creation while Christopher handled the modeling and report write-up. Those were not mutually exclusive, as there was much overlap and collaboration throughout the entire project between group members.

**Data**

The original dataset is composed of thousands of recipes. However, Popovski and his team has sampled and curated a 1,000 of the entire dataset. They established semantic tags while curating. The curated dataset is composed of 1,000 entries, as mentioned, and evenly selected 200 of each class (“Appetizers and Snacks”, “Breakfast and Lunch”, “Dessert”, “Dinner”, and “Drinks”). Based on our experiences, we both have never dealt with an xml file. Thus, this posed as our first problem. The first process in this project is to convert the xml file into a workable dataframe. In order to accomplish this, regular expressions were utilized. From the xml file, we were able to extract the recipe name, class, ingredients, instructions and tags. This process took most of the cleaning and data preparation time. Once achieved, we were able to conduct exploratory data analysis (EDA) and proceed to model simulations. For EDA, there were not a lot of things to discover as the curated data was not randomly sampled. Each class has ⅕ representation of the overall data. Therefore, most of the EDA has been more of discovering what specific words show more often to specific classes. These words were extracted from both the recipe instructions and ingredients. Word Cloud generation became too helpful to quickly pinpoint frequent words. Sample word cloud is provided below. 

However, this does not answer our question of which ingredients are mostly tied to a specific meal/class. Therefore, we used tf-idf to indicate and provide values for distinguishable ingredients. From the visualization below, it has become clear which ingredients fall more heavily on a designated class.



Lastly further EDA, we have noticed that the classes, Dinners and Appetizers and Snacks, are mostly correlated--meaning they share the most identical ingredients.

**Methods & Procedure**

In order to build the best model, a number of different models needed to be constructed and evaluated. As such, we developed K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes (NB) and Decision Trees (DT) algorithms to perform multiclass classification on our dataset. Being that our recipes dataset contained Tags and Ingredients, models were made for both. One-hot encoding was performed on each parameter individually to create two different datasets. The datasets were then matched with the Categories parameter of our Recipes dataset in order to have evaluation of our models. We were effectively evaluating the tagging ability of Popovick et al versus the actual ingredients themselves. Each dataset was split into train and test datasets based on the 80-20 rule, respectively. Once this was completed, The aforementioned classification algorithms were trained and then tested for both datasets

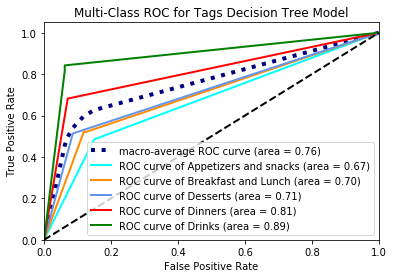
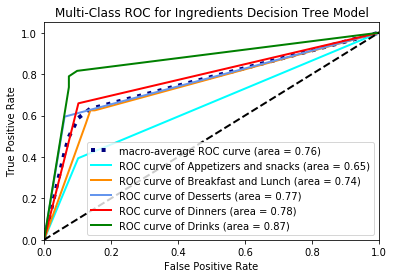
**Results**

Being that we care most about the positive class of our models’ classification, each model was primarily evaluated on F1 Score, Precision and Recall. ROC AUC was also taken into consideration but the primary factors in analyzing the performance of our algorithms were F1 Score, Precision and Recall.

The first algorithm developed was a Decision Tree. As mentioned, each Decision Tree was based off of their own dataset. Results from each of those decision trees can be seen below. The left side corresponds to models based off of Ingredients, while the right side corresponds to models based off of Tags.

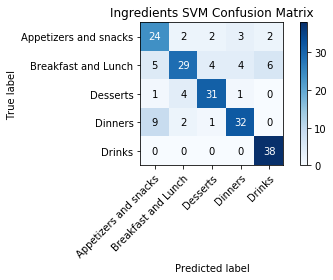
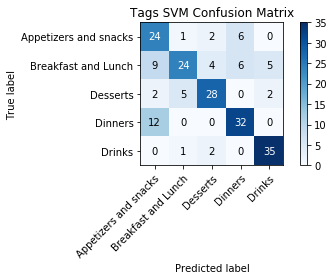
** **

*F1 Score:* 0.62 | *Precision:* 62% | *Recall:* 62% |||| *F1 Score:* 0.59 | *Precision:* 60% | *Recall:* 60%

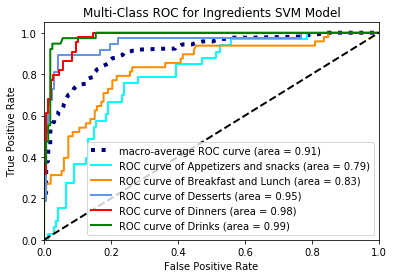
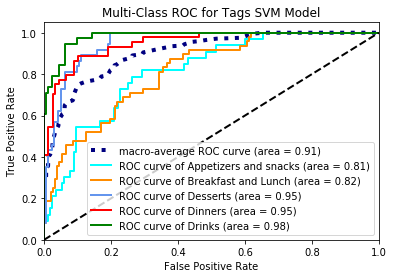
****

With an F1 score of 0.62, the ingredients decision tree model outperforms the tags decision tree model, which had an F1 score of 0.59. The same notion held true in terms of precision and recall, with Ingredients outperforming Tags. It can also be seen in the confusion matrices that Dinners and Appetizers and Snacks were confused with each other more than any other pairing of categories by both models. Furthermore, Breakfast and Lunch was the only category that had predicted labels in each of the other categories for both models. Our secondary metric, ROC AUC (marco-average), scored each tree the same at 0.76. The Tags model struggled more with Appetizers and Snacks, Breakfast and Lunch, and Desserts more than the Ingredients model did. Altogether, the Ingredients Decision Tree was better than the Tags Decision Tree.

The next set of models that was developed was SVM models. As was the case with the Decision Trees, two models were created. One was for Ingredients and the other was for Tags.

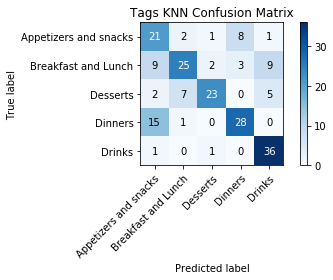
** **

*F1 Score:* 0.77 | *Precision:* 77% | *Recall:* 78% |||| *F1 Score:* 0.72 | *Precision:* 72% | *Recall:* 73%

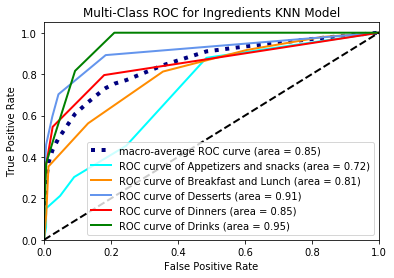
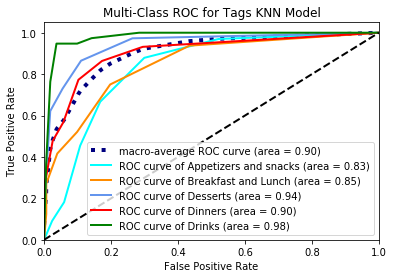
 

Once again, the Ingredients model outperforms its counterpart in F1 Score, Precision, and Recall. A larger gap between each of the metrics can be seen here than there was for the Decision Tree models (about 0.04 for each metric). As was seen in Decision Trees, Dinners and Appetizers and Snacks are the most confused for one another. Moreover, Breakfast and Lunch was yet again the only category to have misclassifications in each category for both models. Another trend that exhibited some development was Drinks being very easily distinguishable from the rest. Its own ROC AUC was north of 0.97 in both SVM models and also had the largest ROC AUC for the DT models. The final notion of Ingredients models being better than Tags models also persists.

Subsequently, KNN models were developed for both datasets. Here, we set the nearest neighbor number to seven. Results from the models can be seen below.

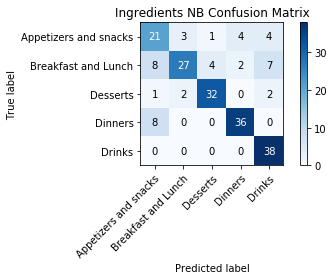
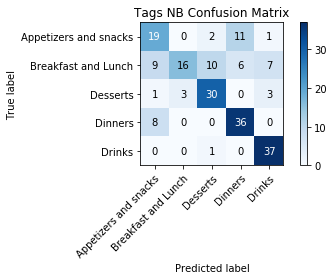
 

*F1 Score:* 0.55| *Precision:* 64% | *Recall:* 56% |||| *F1 Score:* 0.66 | *Precision:* 69% | *Recall:* 67%

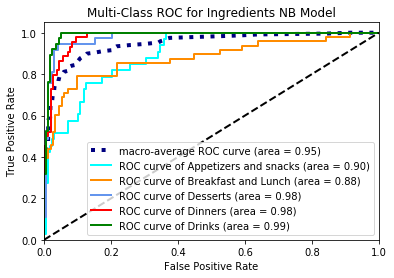
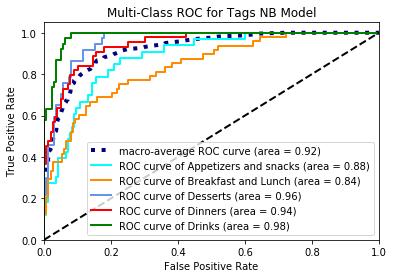
 

In KNN, the first time where the Tags model performs better than the Ingredients model can be seen. The difference between the models is drastic, as there is a difference in .11 in F1 Score, 0.05 in Precision and 0.11 in Recall all in favor of the Tags model. Other notions seen in the DT and SVM models also do not persist here. Breakfast and Lunch does not maintain its distinction as the only category with misclassifications in all other categories. No category falls into that bucket. Additionally, Ingredients is outperformed by Tags in the secondary metric of ROC AUC, as a difference of 0.05 separates the two. What does maintain for these models is that Drinks is the most easily distinguished model and that Dinners and Appetizers and Snacks are confused with each other. For the latter discovery, this falls more on Dinners being misclassified than Appetizers and snacks.

The final set of models developed were Naive Bayes models. Being that the dataset being worked with was a binary one for each column after one-hot encoding, Bernoulli Naive Bayes models were developed.

*F1 Score:* 0.76| *Precision:* 77% | *Recall:* 78% |||| *F1 Score:* 0.67 | *Precision:* 70% | *Recall:* 70%

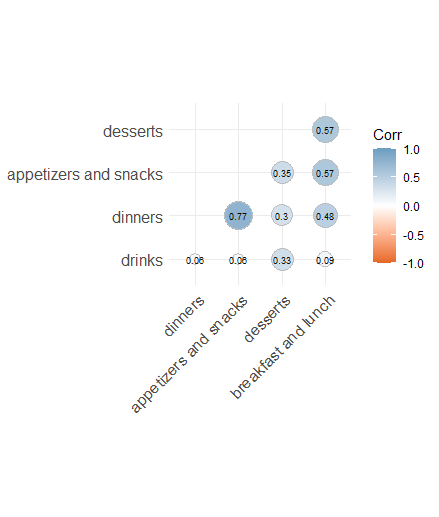
 

After results differed from the majority with the KNN modeling, they fell back into line with NB modeling. The Ingredients model once again is the better of the two by F1 Score (0.76 vs 0.67), Precision (77% vs 70%) , Recall (78% vs 70%) and the secondary metric, AUC ROC (0.95 vs 0.92). Other trends persist here as well. Dinners and Appetizers and Snacks are misclassified for one another more than any other pairing. Breakfast and Lunch is the only category with misclassifications in all the other categories. Drinks are once again the most distinguishable category. Finally, the best model is once again the Ingredients model.

**Analysis**

With three out of the four types models developed showing in favor of ingredients by way of F1 Score, Precision, Recall, and ROC AUC, it is clear that developing models for predicting the category of the recipe is better done by the ingredients of the recipe rather than the tags given to each recipe. The implications of the tags not doing as well as the ingredients are that the tags can be improved. This was something mentioned by Popovski et al. in their work and is a potential focus point as FoodBase continues to be improved upon.

Two models that were developed stood out among the rest - the SVM Ingredients model and the NB Ingredients model. The models had almost identical lineups in terms of F1 Score (0.77 vs 0.76), Precision (77% vs 77%) and Recally (78% vs 78%). To distinguish between the two, ROC AUC, the secondary metric, needed to be called into question. With that now impacting the decision, the NB model pulled ahead, as it outscored the SVM model 0.95 to 0.91. With both of these models being almost identical, either can be looked at as the most reliable for categorizing a given recipe.

Other insights were also able to glean from the dataset. Dinners and Appetizers and Snacks were often confused, for one. As can be seen in the graphic to the right, they had high correlation to begin with. This notion was only further confirmed during model building as for almost each model constructed, those two categories were confused most often. The misclassification in question accounted for 26% of misclassifications when it came to the two best models. At its peak, that pairing accounted for over 34% of misclassifications (Tags KNN model).

Drinks was consistently the most distinguishable category, as anticipated prior to analyzing. Its ROC AUC persisted as the top score among individual ROC AUCs. That number never dropped below 0.87 and a mainstay in the mid-to-upper 0.90’s. Conversely, Breakfast and Lunch was the least distinguishable of all, as it consistently was misclassified in every other category. Its persistent correlation to the other categories, other than drinks, also exemplifies this point.

Breakfast and Lunch being so indistinguishable leads to a further piece of work for this corpus. As it continues to be developed Breakfast should be separated from Lunch and insights should be gathered distinctly. Additional future works include more preprocessing of the dataset in general and attaching food “ethnicities” to each recipe. That being the case, how each area of the world cooks could be looked into. Questions such as “How do cooking styles across the globe differ and relate?” and “Is there a lack of food types in certain areas of the world that can be either made more accessible by a charity organization or capitalized upon for profit by a corporation?” can be answered.

**References**

Title: FoodBase corpus: a new resource of annotated food entities

Author(s): [Gorjan Popovski](https://www.ncbi.nlm.nih.gov/pubmed/?term=Popovski%20G%5BAuthor%5D&cauthor=true&cauthor_uid=31682732), [Barbara Koroušić Seljak](https://www.ncbi.nlm.nih.gov/pubmed/?term=Seljak%20BK%5BAuthor%5D&cauthor=true&cauthor_uid=31682732), and [Tome Eftimov](https://www.ncbi.nlm.nih.gov/pubmed/?term=Eftimov%20T%5BAuthor%5D&cauthor=true&cauthor_uid=31682732)

Link:<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6827550/>