NUT - Nutritionist Utility Tool

Team info:

Barel Mishal, id: 305109639, mail: barel.mishal@mail.huji.ac.il, cs id: barel.mishal
Yoav Orenbach, id: 208847749, mail: yoav.orenbach@mail.huji.ac.il, cs id: gingsley
Sapir Shapira, id: 204165732, mail: sapir.shapira@mail.huji.ac.il, cs id: sapir.shapira
Hen Emuna, id: 203285515, mail: hen.emuna@mail.huji.ac.il, cs id: hen.emuna

The following writeup is also available in our streamlit app. We recommend reading there for a more interactive experience.

Motivation and problem description:

Food and Nutrition Science is a newly developed field critical for health and development. It affects our body's homeostasis in general, specifically our immune system, sugar balance, and endocrine system. As a young research field, there is a need to deepen the field's knowledge further. Moreover, we cannot partially leverage existing knowledge, as it contains many myths and biases, some of which are incorrect. An excellent example of this bias is the relation to fat, which is incorrectly considered unhealthy. In fact, it is the vast intake of sugar that comes from this thinking that causes many of the diseases today. As a result, Nutritionists must use both reliable information and scientifically supported tools to know the exact composition of food. This knowledge can be used both for personal diet recommendation as well as interdisciplinary research that may link the different nutritions to metabolic processes and a wide range of non-communicable diseases: type 2 diabetes, developmental issues, etc [1].

A critical aspect of daily nutritionist's work is to suggest dietary alternatives for patients, based on their specific health profile, using previously defined food groups (s.a., Meat, milk, vegetables, fruits, and sweets). The rationale behind the clustering to food groups, based on known macronutrients (carbohydrates, fats, and proteins) and an estimation of the equivalent micronutrients (Vitamins, and minerals), is to enable the creation of reliable tools, which would

provide nutritionists and their patients the ability to create dietary alternatives and control regarding their food consumption. Nowadays, this clustering relies mainly on assumptions and lists of products memorized by Nutritionists.

Data:

We gathered from the Israeli government database catalog a table (10MBs) with 4650 records of 74 nutritional components: protein, fats, carbohydrates, amino acids, fatty acids, vitamins, and minerals. The table was created from a JSON file containing records and metadata, which was pre-processed (units conversion, missing fields, dropping non-helpful features like psolet, Supplements, and Recipes) and saved as a CSV. In addition, since the Israeli data is quite small we also use the U.S department of agriculture Food Data Central (FDC) dataset (135MBs) which the Israeli data relies on, in order to attempt to improve some of our results using more data. After using the FDC API key to extract information (with web scraping), we built a table of 7600 records of 149 nutritional components containing the components of the 1st table.

Comparing the two, the FDC dataset has more records than the Israeli data as well as additional features per record (for example, FDC uses 3 types of vitamin K, whereas the Israeli data uses just one type based on those three types). Therefore, while these datasets are connected, their features can't be mapped easily, in terms of features and item ids, requiring preprocessing to match them to the best of our knowledge.

- *The Israeli nutrition data source (https://data.gov.il/dataset/nutrition-database) from the ministry of health.*
- *Food Data Central source (https://fdc.nal.usda.gov/)*

Solution:

Our goal is to improve the existing assumptions and biases used in the Israeli nutrition community, thus helping nutritionists offer better alternatives for their patience. In particular, we intend to make improvements in three areas:

1. ### Clustering of food items to food groups.

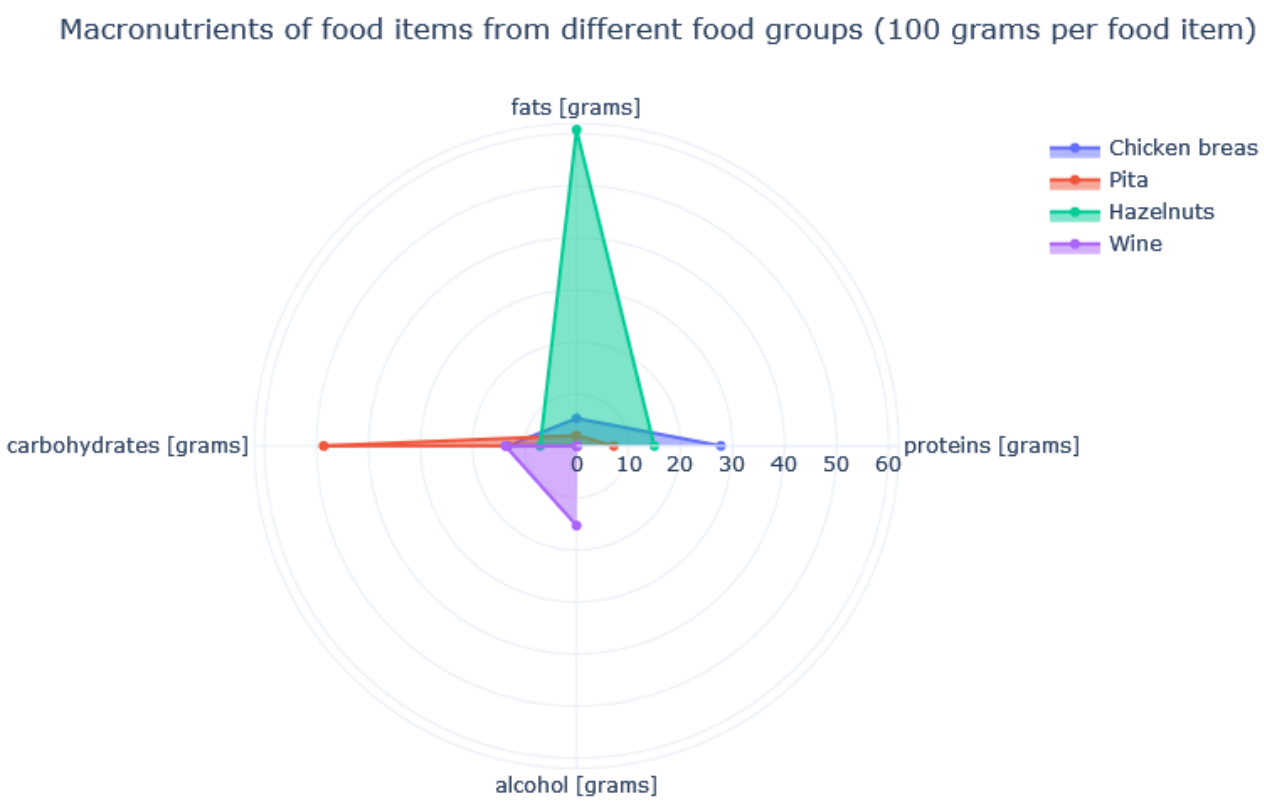
   There are no known food group labels for food items in the Israeli data (or the American data for that matter) since some food items cannot be classified into one food group but into several. For example, if we look at a food item - Pizza, it could fall into the Fat food group as it has high amounts of fat, however, it could also fall into the dairy food group depending on the amount of cheese on the pizza. Moreover, a pizza could be in the pastries food group because it contains bread. Many more examples follow, though there are some food items which we unequivocally know their food group, like milk, for instance.
   Therefore, our objective as a first step in the analysis is to cluster all food items in the Israeli data into their respective food group.
   To deal with the ambiguity in the food-groups clustering, we received a list from the ministry of health containing all the Israeli food groups with known food items in each group (containing 11 main food groups and 32 sub food groups). However, this list is

lacking, containing only 320 food items. To expand our food-groups estimation with this ground-truth information, we use a set of clustering algorithms to cluster the additional food items in our dataset, i.e. the missing food items in the given list, as close as we can to the known food items.

For our clustering algorithm, we compared the following (algorithms with varying un/even cluster size, different expected manifolds geometry, with/out outlier removal): K Means, Agglomerative, DBScan, and Spectral clustering.

As our features vector for the clustering, we chose the three macronutrients (proteins, carbohydrates, and fats) and alcohol, as these gave the best results based on our evaluation (shown at Fig 3). Fig 1 shows an illustration of the possible distinction between different food items based on these four features. As we can see, for the exemplar food items, there is a clear separation, matching their expected food groups.

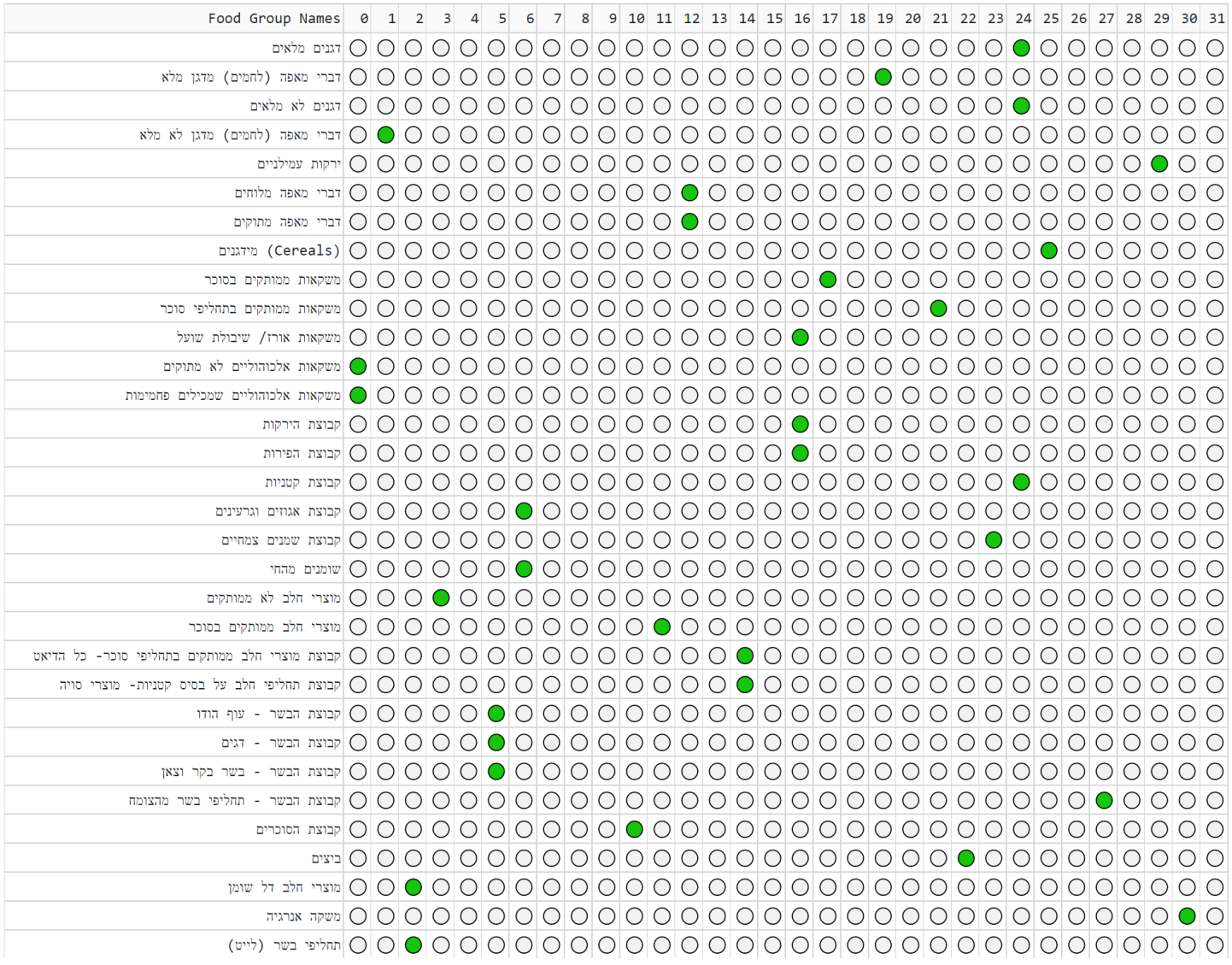
Fig 1



Comparing the proportion of the three macronutrients (proteins, carbohydrates, and fats) and alcohol for different food items; Chicken breast (blue) belongs to Meat group for chicken and turkey, Pita (red) belongs to Pastry group for breads, Hazelnut (green) belongs to Nuts and seeds group, and Wine (purple) belongs to Alcoholic beverages.

Visualization of our clustering result can be seen at the following [Heroku dashboard site] (might take a few seconds to load), allowing a qualitative evaluation of the clustering, for each of the chosen algorithms (a quantitative evaluation is expanded under ‘evaluation’ section).
Our clustering results are verified between the different algorithms used. Fig 2 shows an example of a Spectral clustering result, where we map each cluster to the matching food-group, using the known food-items list mentioned above.
As we can see, there are cases where the clustering result is not ideal, meaning we don’t get a one-to-one mapping between each cluster and its respective food group, but rather several food-groups (at most 3 food groups assigned).

Fig 2



Mapping food groups to the clustering algorithm`s labels. Green checkmark signifies a mapping between a food group and a label. An ideal clustering would have a one-to-one mapping between the food groups and the labels. As we can see there are mismatches in this mapping, for example cluster #24 has two contradicting food groups, and cluster #8 does not match any known food groups.

2. **Predicting food's micronutrients based on their macronutrients.**

We attempt to predict the values of the most well known vitamins (vitamin A IU, vitamin A RE, vitamin E, vitamin C, thiamin, riboflavin, niacin, vitamin B6, folate, folate dfe, vitamin B12, carotene, vitamin K, vitamine D, and choline) and minerals (calcium, iron, magnesium, phosphorus, potassium, sodium, zinc, copper, manganese, and selenium), since the connection between macronutrients and micronutrients are unknown and not trivial.

Our input contains the macronutrients – proteins, carbohydrates, and fats, while our output contains the micronutrients – vitamins and minerals.

We applied various machine-learning algorithms:

- Linear regression (compared with variations s.a. ridge and kernels).
- K-nearest neighbors.
- Gaussian-process regression.
- Tree-based regressors: Decision tree, Random forest, XGBoost.
- Neural network (multi-layer perceptron)

As part of our preprocessing for the prediction, we had to deal with the existence of NaN values across the different features. In our data there are many macronutrients with NaN values, partly because they were not tested for some food items and partly because it is known that they are equal to zero so there was no need to test. Due to this ambiguity we decided to drop NaN values in our training and testing, to later predict those values for our next step. Evaluation of the predictions are shown at Fig 5.

3. **Recommend food alternatives based on similar products.**

In this section our goal was to provide alternatives for a specific food item within the same food group.
Since there are no ratings for each food item, we turn to a content-based recommender system. One of the downsides of using content-based recommenders is that finding the appropriate features for each food item profile is hard. Therefore, we used three different feature vectors to test three different recommendations: Food items names, Food items macronutrients, Food items macronutrients + micronutrients.

As for the prediction heuristic, given a food item and any of the item's profiles, we compute the cosine similarity between them and return the top ten most similar food items (based on that profile).
In order to compute the cosine similarity between food items names, we first use the Term frequency-Inverse Document Frequency (TF-IDF) with an hebrew tokenizer (to transform the hebrew text to a meaningful representation) and then we can apply the cosine similarity. Computing the cosine similarity for food items macronutrients and micronutrients is straightforward by applying it on its values.
In addition, since many micronutrients values are missing from the data and we wish to

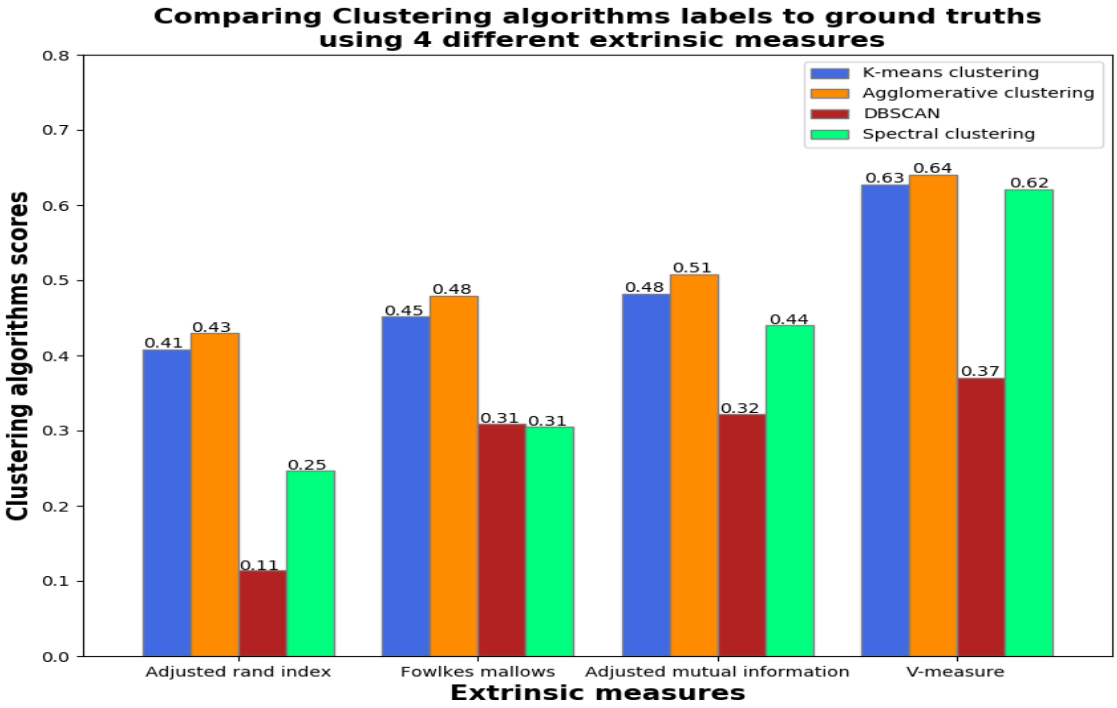
use them for our third feature vector, we use our second step – the micronutrient prediction to fill any missing values.
A demo for the recommendations can be found on the streamlit app, where one can enter a food item in Hebrew and see 3 different kinds of recommendations based on the three item profiles.

Evaluation:

1. Evaluating food groups clusters:

The main issue with the visual evaluation (as shown in Fig 1 and Fig 2) is that we would prefer a quantification using our ground truth labels. Thus, in order to see a numerical evaluation of our clustering, we created a test set containing the known food items to food groups (from the ministry of health list, 320 food items). We applied the following extrinsic measures (metrics used for comparing two clustering labels): Adjusted Rand Index score, Fowlkes Mellows score, Adjusted Mutual Information score and the V-measure score. Using these metrics we could further optimize our clustering algorithms (for instance, we decided to also use alcohol as part of our input as it also helped us increase our scores).

Fig 3



We can see that the Agglomerative clustering achieves the best scores on all metrics (the orange bar is the highest across all measures). Some sub food groups have very similar macronutrient breakdown, so it is quite hard to cluster them perfectly when the inputs are very similar, though we can see promising results with Agglomerative and with more ground truths we believe it can be improved.

To further visualize and qualitatively evaluate our clustering, we use a word-could to show similarities between the food-groups. These figures show the food-items text (without stopwords) as well as their attached food groups written as their titles. Clusters without a mapping to a known food group are titled as 'לא סווג'. All word clouds can be seen in the streamlit app. The following example shows 4 correctly defined clusters. As we can see the upper word clouds are food items with labels that were mapped to a single food group, while the bottom word clouds are food items that were mapped into two food groups.

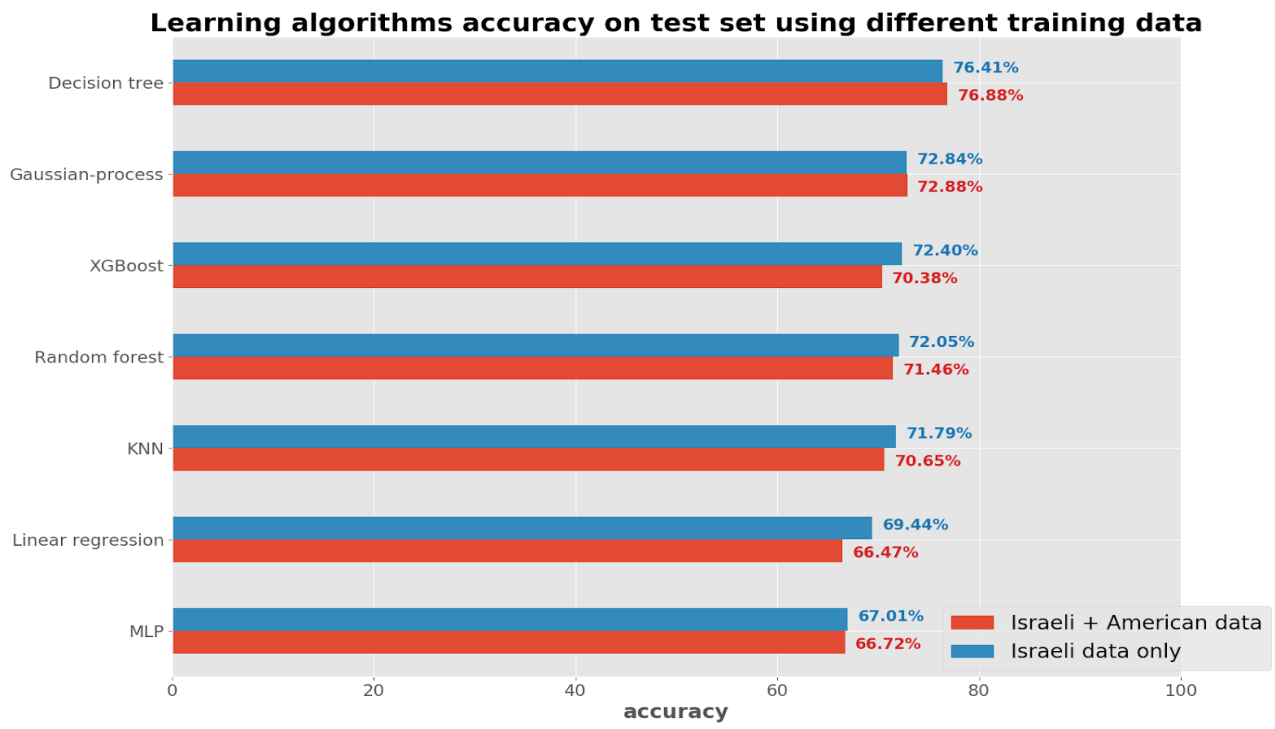
Fig 4

2. Evaluating micronutrients prediction:

The prediction task is a multi-output regression problem, thus we can evaluate it by splitting our data into train and test sets and computing our accuracy on the test set. To count for the wide range of numeric values (our prediction is in the Milligram or Microgram ranges), when calculating the accuracy we decided to give a small error range of one milligram and we can consider our predictive model to be successful if it is correct with an error on one milligram. To deal with the lack of data in the Israeli records (only 4560 food items), for training we used the FDC data, while our test set contains only the Israeli data.

From Fig 5 we can see that the FDC data does not help much, probably due to overfitting or because it is still not enough to properly learn (even combined, the amount of food items is just a little over 10,000). Secondly, we can see that the Decision Tree algorithm achieved the highest accuracy with about 76% accuracy on average, where ensemble methods only reduce the prediction accuracy. Trying to apply the same algorithm to a single micronutrient at a time (instead of a multi-output) improved the accuracy for some. For instance, we saw that for vitamin b6 the Linear regression algorithm performs best with 98% accuracy, hence we assume that this vitamin has a linear relation to the macronutrients, while other vitamins and minerals have different relations. To conclude, on average the decision tree is the most suitable for the prediction task, with possibility to improve for specific micronutrients.

Fig 5

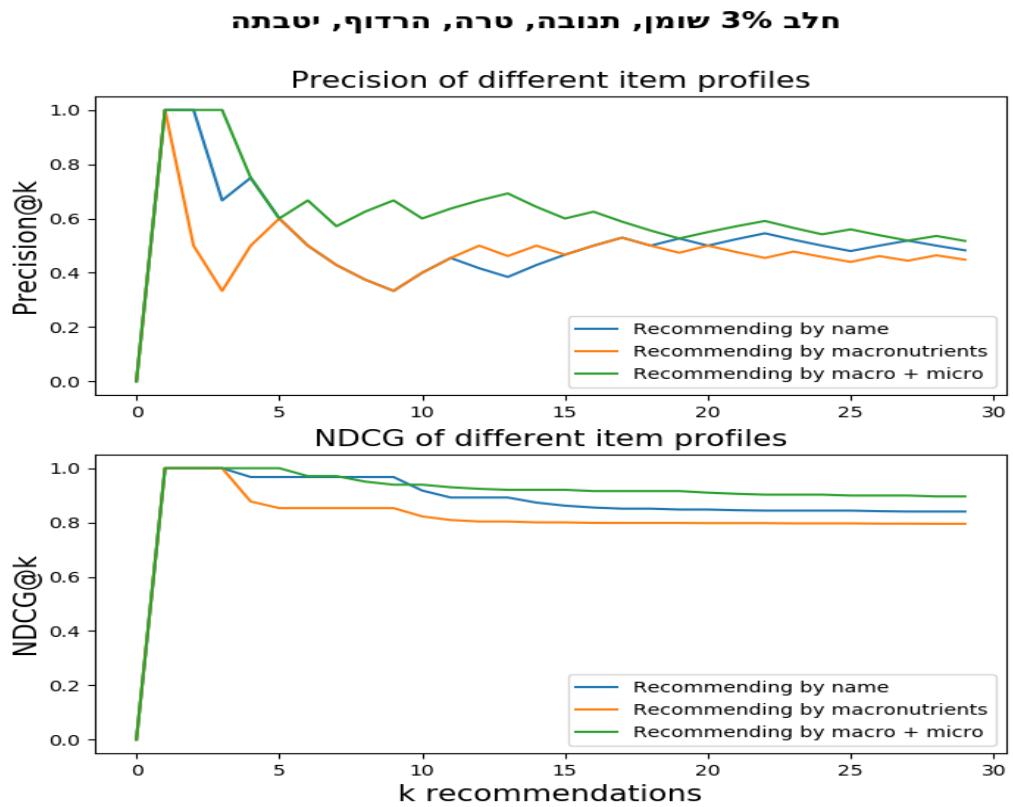


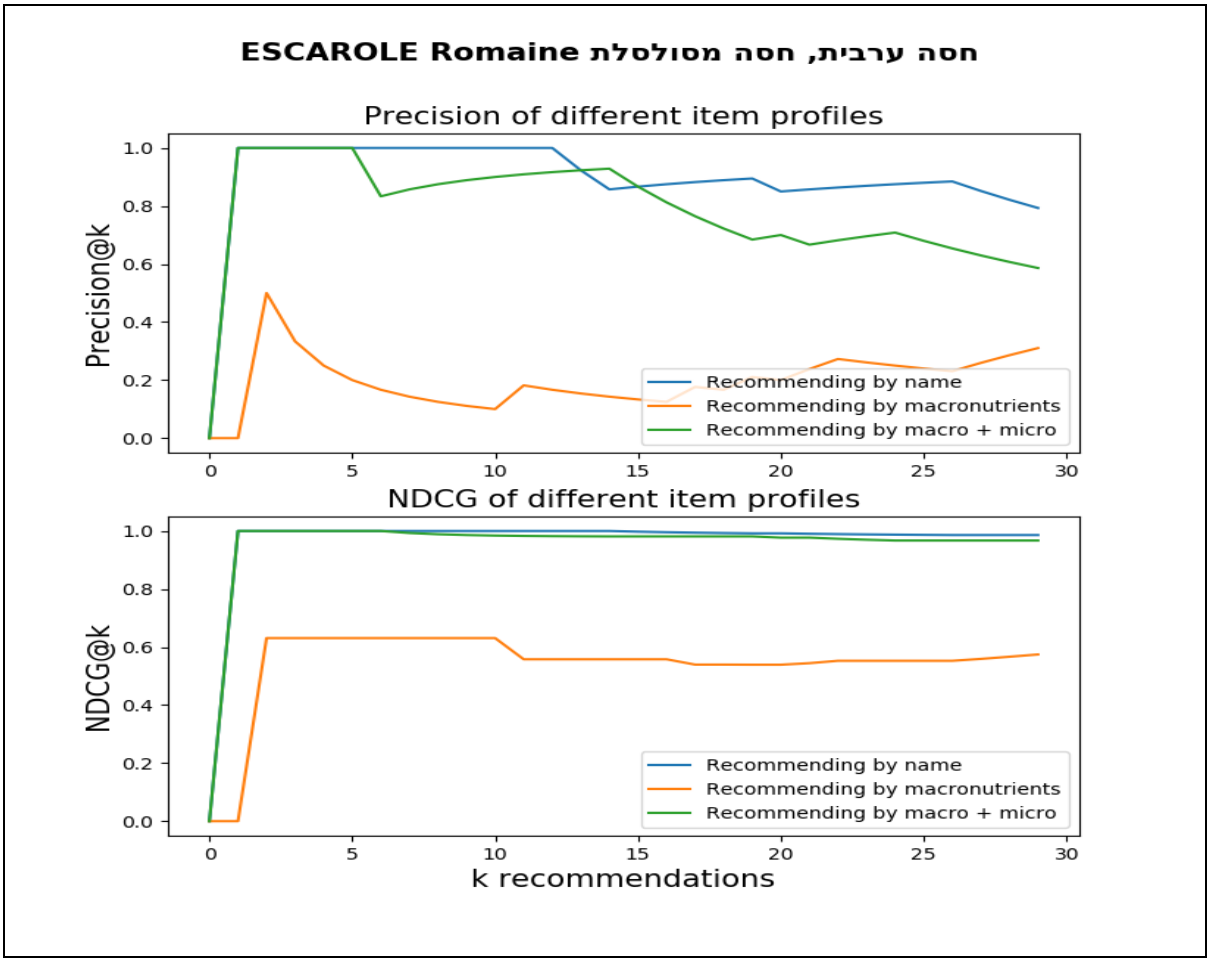
3. Evaluating food alternatives recommendation:

Since there are no ratings of food items we can use, we cannot evaluate our recommender system by creating a test set and computing the root-mean-square error. However, we can know if our recommender is successful by using the Precision and the Normalized Discounted cumulative gain (NDCG) evaluation metrics. Given a food item, for each feature vector we recommend K items which we consider as positives. To calculate Precision and NDCG, we use our food groups clustering as ground truth. Since we want to offer alternative food items within the same food group, we compare the food group label of the given food item to the food group label of the recommended items. For the K positive recommendations, true positive is when the labels match, otherwise it is a false positive. With that in mind, we can compare the different feature vectors of the recommendations on different food items to see which has the best results, and find the proper value of K to see how many items we should recommend. Inspecting varios results (not shown) on average K=10 has the best scores for both Precision and NDCG.

Fig 6 shows results for 2 food items (milk and lettuce). We can see that for milk we achieve a very high NDCG for both macronutrient and micronutrient feature vector, while for lettuce we achieve higher NDCG for the name feature vector rather than the macronutrient and micronutrient feature vector.

Fig 6





Future Work:

In the future, an improved clustering can be achieved by better filtering outliers and noise (food items that should not be clustered to any food group like 'similak'). While we tried our best to filter any noise, we believe that better constructed data can help in the clustering. Moreover, further investigation regarding the relation of every micronutrient to the macronutrients can help in deciding which algorithm should be used for every micronutrient and improve the prediction accuracy. Finally, by collecting user data on the Israeli food items, the recommendation can be improved significantly and with enough ratings, a collaborative recommender system could be used.

Conclusion:

In this project we experimented with different areas of Israeli nutrition, attempting to help nutritionists better classify the food groups of food items, predict unknown micronutrients, and hopefully offer better food item alternatives for their patients. We hope our findings can help the nutrition field and act as a basis for further improvements.

Bibliography:

1. Mozaffarian D, Rosenberg I, Uauy R. History of modern nutrition science—implications for current research, dietary guidelines, and food policy BMJ 2018; 361 :k2392
2. Williams, P. Nutritional composition of red meat. Nutr. Diet. 2007, 64, S113–S119.