Endometriosis versus food

Creating a prototype for a recommendation system.

Sara Galatro – April 2024

Abstract

The goal of this project is to analyze the relationship between food and people diagnosed with endometriosis to create a **recommendation system** where users are the patients.

The data were collected through a Google form shared on various Facebook groups and then it was elaborated through two methods: **normalized K-Nearest Neighbour** and **Singular Vector Decomposition**.

The results show the utility such a system would have and that, with the necessary changes, it is possible to create such a software.

Motivation

Endometriosis is a chronic illness which deeply affects the lives of the people diagnosed with it, both physically and psychologically. To this day, **no cure** has been found.

One of the main therapies suggested is linked to specific diet guidelines based on chemical and biological analysis of foods. However, many patients signal how **these diets do not work** and, in some cases, even worsen their symptoms.

Therefore, we propose to shift the focus from the eats to the people and to analyze their characteristics and reactions to what they eat to create an auxiliary system that will help more patients create a **new balanced lifestyle** with endometriosis.

Dataset

To gather the data, we created a **Google form** that we later sent out through social media and word of mouth. Patients were asked to

- Give their age;
- Say how long they have been suffering from endometriosis;
- Evaluate 19 pre-selected (by us) foods using numbers from 1 to 5 based on personal experience, where 1 codes “safe food” and 5 “harmful food”.

We collected a total of **492 answers**, which we then analyzed and formatted into the desired dataframe layout to use the Surprise Python library. We also created a training set and a test set with a ratio of 80%-20%.

From a preliminary run, we were able to infer two important notions:

1. The score does not depend on **age** nor on the **period of time** the patient has been dealing with endometriosis;
2. There is **no clear-cut between harmful and safe foods**, proving that the proposed diet is not as effective as theorized.

Used Methods

Normalized K-Nearest Neighbour

The first test was to associate a new user, with missing food scores, to the other users with whom they share the most similar experiences. Once the new user is placed in a cluster, each of the missing scores are estimated as the mean from its neighbours:

$$\hat{r}_{ui} = \mu_u + \sigma_u \frac{\sum_{v \in N_i^k(u)} \text{sim}(u, v) \cdot (r_{vi} - \mu_v)/\sigma_v}{\sum_{v \in N_i^k(u)} \text{sim}(u, v)}$$

Singular Vector Decomposition

The second test is less intuitive but more performant. The goal is to decompose an input matrix with missing values as a product, thus inferring the missing numbers. This is achieved minimizing the following cost function through a stochastic gradient descent:

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(\|q_i\|^2 + \|p_i\|^2)$$

Used Metrics

We used two accuracy metrics for both models:

- **Mean Absolute Error**:

$$mae = \frac{1}{m} \sum_{1 \leq k \leq m} |\hat{r_k} - r_k|$$

- **Fraction of Concordant Pairs**:

$$fcp = \frac{n_c}{n_c + n_d} \qquad \text{with} \qquad n_c = \sum_u n_c^u$$

where “concordant pairs” identifies the pairs whose estimations respect the original score order, i.e.

$$n_u^c = |\{(i, j) : \hat{r}_{ui} < \hat{r}_{uj} \text{ and } r_{ui} < r_{uj}\}|$$

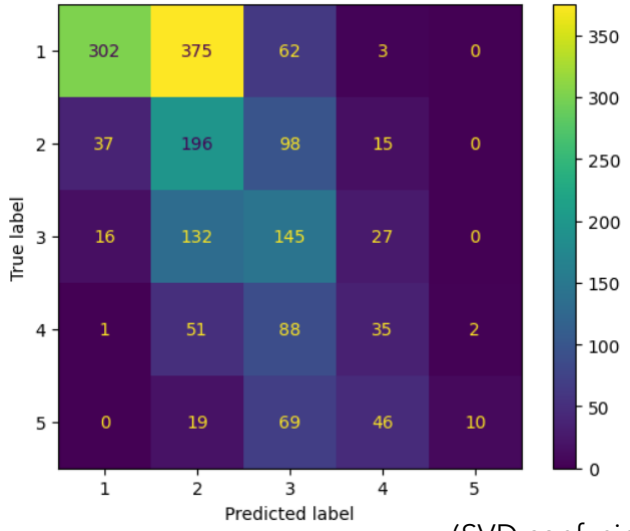
Results

Score: 1 or 2 Safe food	Score: 3 Safe if eaten every once in a while	Score: 4 or 5 Harmful food

Picturing a website with which users will interact, the system should be able to identify **three classes** based on the safety of that food, giving as the output a number that tells the user if it is safe to eat or not.

Hence, we focused on "**serious mistakes**", i.e. when the system assigned the wrong class to the food for that specific user.

With this idea in mind, the K-NN model made 14.75% serious mistakes, while the SVD made 13.65% serious mistakes.



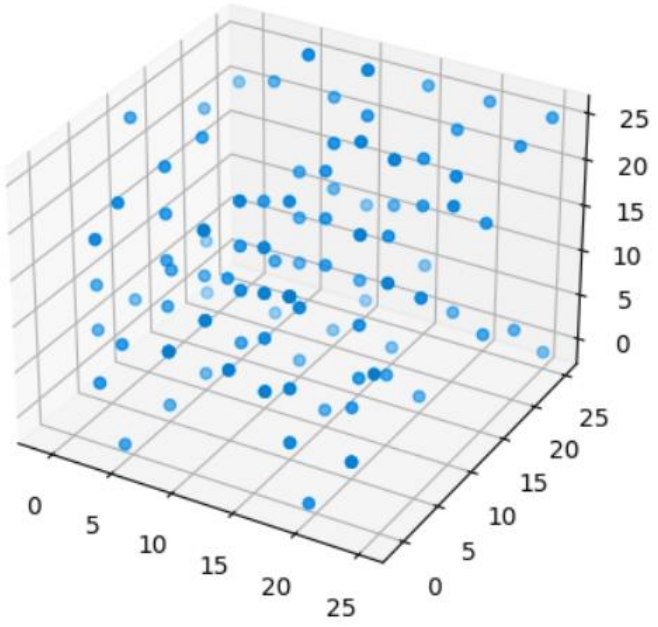
(SVD confusion matrix)

Future Work: new data

One of the main difficulties encountered is the **dimension of the food space in relation to the scores** (19 versus 5). This disproportion makes it so that the users are condensed into a relatively small volume, into which they are uniformly distributed-ish.

Furthermore, the data may contain some **noise** due to some food categories grouping more than one kind of eats.

It could also be interesting to implement an algorithm that focuses on the **estimations' ordering** rather than their actual value (e.g. OrdRec).

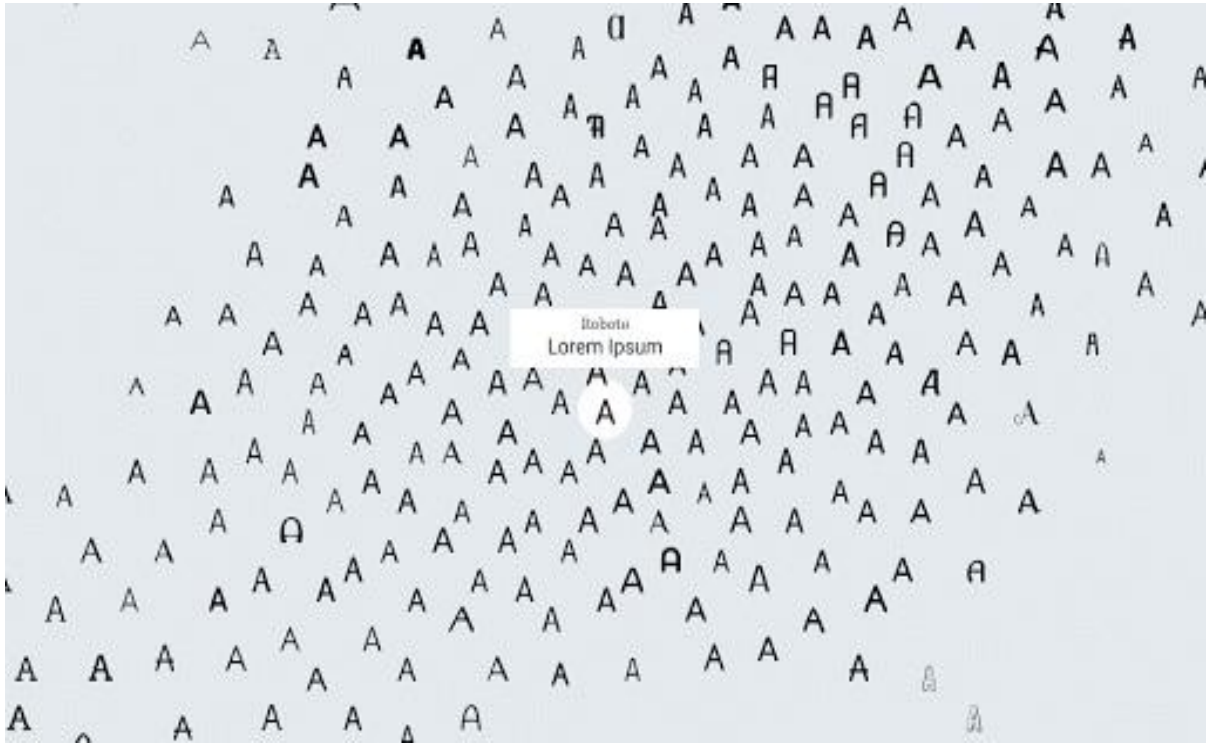


Conclusions

The results of this prototype bode well for **future implementations**, as no current medical strategy performs safely enough.

The idea, in the long run, is to create an **interactive system** that maps foods in a space explorable both by doctors and patients, complete with links to specialists to contact, recipes and online stores.

The road to fully understand **endometriosis** may still be long, but we can try and make it more bearable for those who suffer from it.



©Font Map, by Kevin Ho

Main References

[Les+14] Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, “Mining of Massive Datasets”. Cambridge University Press, 2014

[Ho17] Font Map – Using machine learning to surface new relationships between fonts. July 2017. url: experiments.withgoogle.com/font-map

[KS11] Yehuda Koren and Joseph Sill, “Collaborative Filtering on Ordinal User Feedback”. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2011.

[IBM23] IBM Notebooks - Quantum Machine Learning, 2023. url: github.com/Qiskit/textbook/QML

[Raj+11] Rajdeep Kumar Nath and Himanshu Thapliyal and Travis S. Humble, “A Review of Machine Learning Classification Using Quantum Annealing for Real-world Applications”, 2021.

Thank you for your attention!