DOCUMENTATION OF COURSEWORK

Discipline: Fuzzy Sets and Applications

Topic: *"Fuzzy clustering and rule extraction"*

Date: 05.02.2026, Winter semester, 2025/2026

Prepared by:

- Gabriel Tuparov, 3MI3400841, special. Information Retrieval and Knowledge Discovery

1. Introduction to the problem and purpose of the project

In the modern craft beer industry, there is a huge variety of flavors and styles. Traditional classifications are often subjective or marketing-oriented, making it difficult for consumers to find products with specific characteristics. The main challenge is the transformation of raw taste data (alcohol, bitterness, density) into human-understandable knowledge.

The main goal of the project is to develop a smart model for Knowledge Discovery through fuzzy clustering. The project aims to:

- Identifies sustainable beer archetypes using the Fuzzy C-Means (FCM) algorithm.
- Extract linguistic IF-THEN rules that describe these archetypes
- Validates the knowledge gained through quantitative analysis (MAE) and comparison with official beer classifications.

2. Theoretical statement and algorithm used

2.1. Fuzzy-C-Means

In contrast to hard clustering, fuzzy clustering (Fuzzy C Means) allows each object to belong to several sets with a degree of belonging (membership) in the interval [0,1]. In our case, the objects are the beers, the sets are beer archetypes. The algorithm iteratively minimizes the target function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} ||x_i - c_j||^2$$

Where *m* is the fuzzy parameter. Centroids are calculated according to the formula:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

After calculating the new centroids, the degrees of affiliation for each site are updated according to the following equation: $u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}}$

To avoid local lows, a strategy with 20 independent starts has been implemented, choosing the model with the lowest end value of J_m.

2.2. Adaptive sigmas for affiliation functions

To define fuzzy rules, Gaussian membership functions are used. Because different characteristics have different variations, the project applies adaptive sigmas to each characteristic for each archetype, calculated as a weighted standard deviation:

$$\sigma_{jk} = \sqrt{\frac{\sum_{i=1}^{N} u_{ij}^{m} \left(x_{ik} - c_{jk}\right)^{2}}{\sum_{i=1}^{N} u_{ij}^{m}}}$$

A small value for sigma indicates that the characteristic is strictly defining for the archetype, while a large value allows for a wider tolerance of deviation.

2.3. Aggregation and Protection from Underflow

To implement the logical operation "AND" in the rules, the multiplicative T- is used. The common activation for an archetype is:

$$\mu_j(x) = \prod_{k=1}^{7} \exp\left(-\frac{\left(x_k - c_{jk}\right)^2}{2\sigma_{jk}^2}\right)$$

A technical optimization has also been applied: Due to the multiplication of many small numbers, there is a risk of numerical errors (underflow). In the code, this is solved by summing the exponents before applying the function:exp

$$\ln\left(\mu_j(x)\right) = \sum_{k=1}^{7} \left(-\frac{\left(x_k - c_{jk}\right)^2}{2\sigma_{jk}^2}\right)$$

After calculating the activation for each of the five archetypes, a final normalization takes place. The final degree of belonging is obtained by dividing the activation of a given archetype by the sum of the activations of all archetypes:

$$\mu_j'(x) = \frac{\mu_j(x)}{\sum_{k=1}^{C} \mu_k(x)}$$

2.4. Quantitative Assessment (MAE)

To validate the rules, a Mean Absolute Error is calculated between the original matrix U^{FCM} and the activations from the extracted rules U^{Rules}:

$$MAE = \frac{1}{N \cdot C} \sum_{i=1}^{N} \sum_{j=1}^{C} \left| u_{ij}^{FCM} - u_{ij}^{Rules} \right|$$

Additionally, Pearson correlation is used to measure the linear dependency between FCM memberships and rule-based activations to evaluate model fidelity.

2.5 Extracting fuzzy rules

A key stage in the project is the transformation of numerical centroids into logical rules. For this purpose, 5 linguistic categories are defined for each feature: *Very Low, Low, Average, High, Very High*.

The boundaries of these categories are not fixed arbitrarily, but instead are derived by quantile division (20%, 40%, 60%, 80%) of the entire data set. The rule generation process (in generate_linguistic_rules_denorm) works like this:

1. The centroid of the archetype in its normalized form is taken.
2. The value is denormalized to the real units (e.g. ABV as a percentage).
3. This real value is compared with the quantile limits to determine the relevant linguistic label.

3. Description of the data, pre-process processing

Data with over 3000 unique records were used. The profile of each beer is defined by 7 numerical dimensions: ABV, Body, Bitter, Sweet, Sour, Hoppy, Malty.

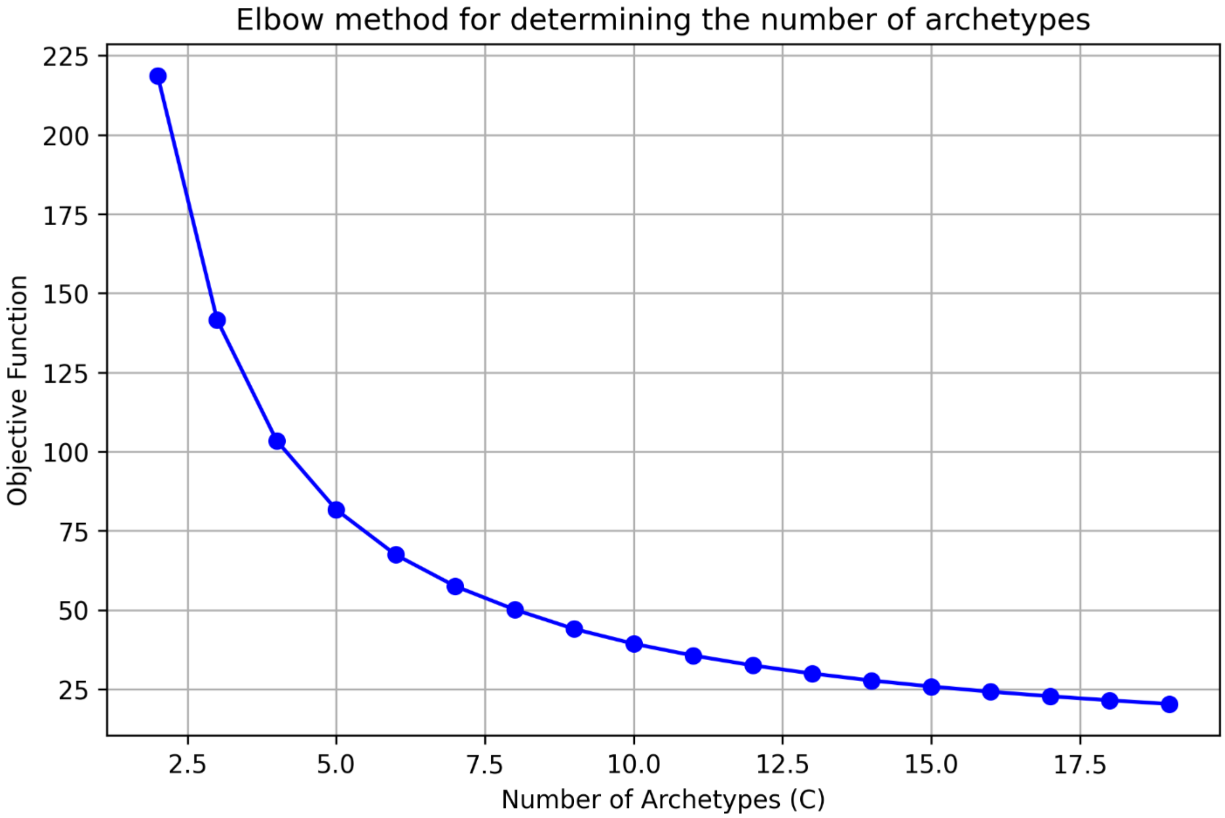
As for preprocessing, there are two:

1. **Cleaning:** Remove incomplete entries via dropna to preserve the integrity of the 7D space.
2. **Normalization:** Min-Max normalization is applied in the interval to neutralize the difference between the different units of measurement of the characteristics and their corresponding different ranges of values.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

To determine the verbal limits of taste characteristics (the *categories Very Low, Low, Average, High* and *Very High*), quintiles defined by the 20th, 40th, 60th and 80th percentile of the actual distribution of data were used.

The choice of the optimal number of archetypes is made through empirical analysis – the elbow (knee) method. Tests were carried out for values of 2 to 20.



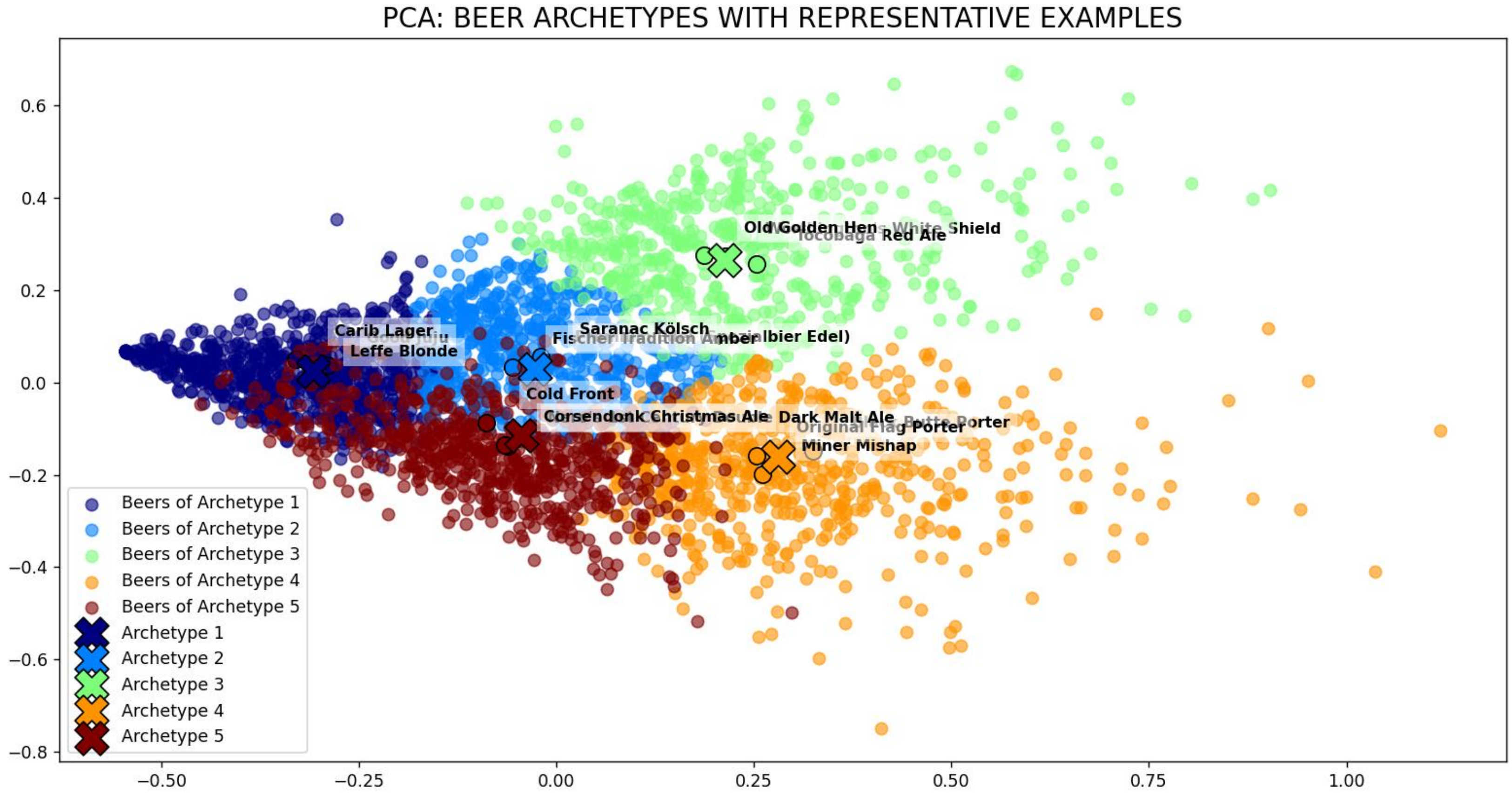
Although the value of the target function decreases as the number of clusters increases (which is normal behavior), after 5 clusters the curve begins to slow down its slope more significantly. Therefore, the choice of 5 is justified as a balance between the detail of the description and the value of the target function.

4. ***Experimental/simulation results***

The analysis process is organized in three phases.

Phase 1: Geometric verification via PCA

To assess the quality of clustering, the 7-dimensional space is designed in two dimensions through PCA (Principal Component Analysis). The visualization confirms the presence of 5 clearly distinct archetypes. The archetypes do not overlap chaotically, but rather occupy different zones corresponding to finite flavor profiles. The most representative beers for the respective archetypes are also included.



Phase 2: Functions of Belonging and Quintile Denormalization

The graphs of the membership functions of all archetypes are presented, as well as the boundaries that define the quintiles. These boundaries are used to determine verbal categogies - Very Low, Low, Average, High, Very High. For the sake of easier interpretation of the graphs, red and blue lights alternate for the lines of the quintiles. The data is denormalized. At the same time, the verbal rules for all archetypes are displayed on the console. It describes how every feature falls into one of the five linguistic categories. The value with the maximum membership rate is also shown. An example of how a single rule looks is given:

RULE 1: IF ABV is Low (approximately 5.5)

AND Body is Very Low (approximately 22.5)

AND Bitter is Very Low (approximately 13.4)

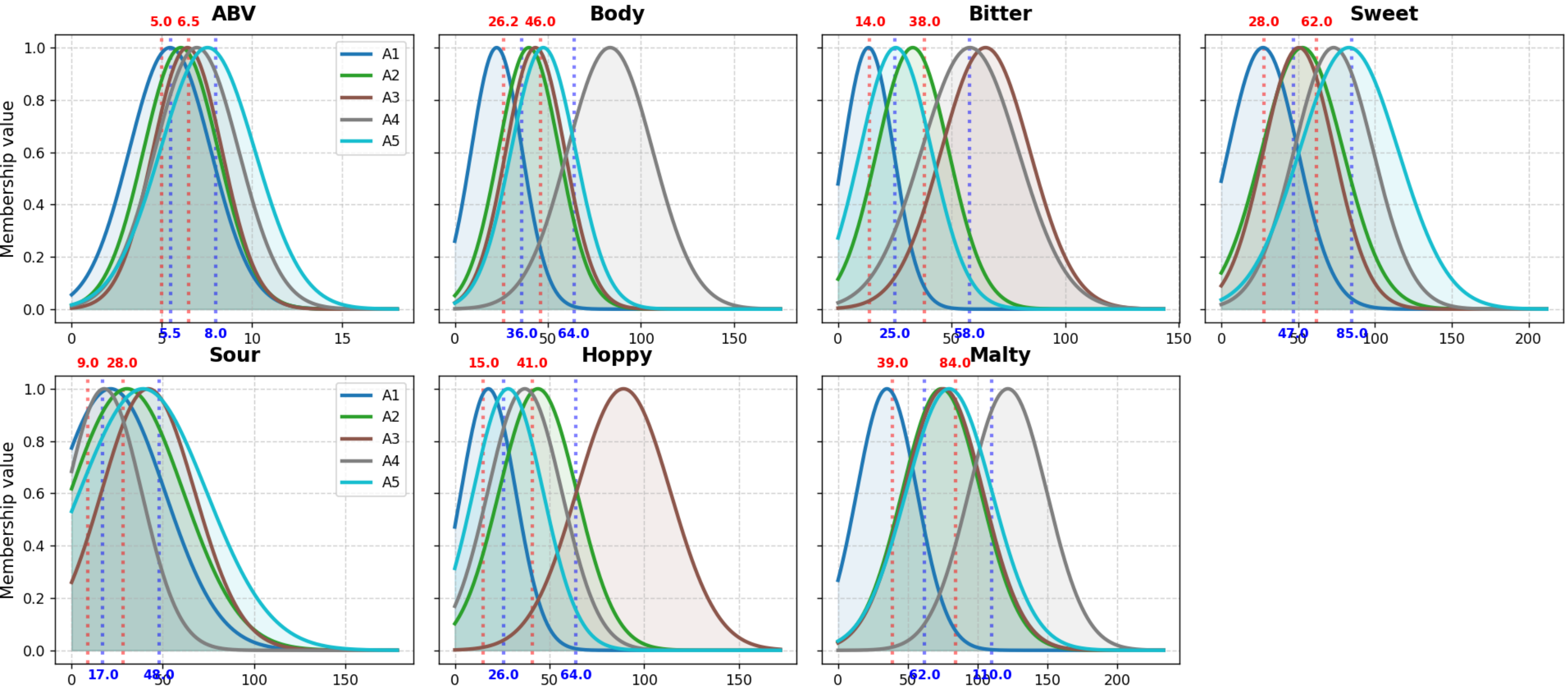
AND Sweet is Very Low (approximately 27.2)

AND Sour is Average (approximately 21.4)

AND Hoppy is Low (approximately 17.8)

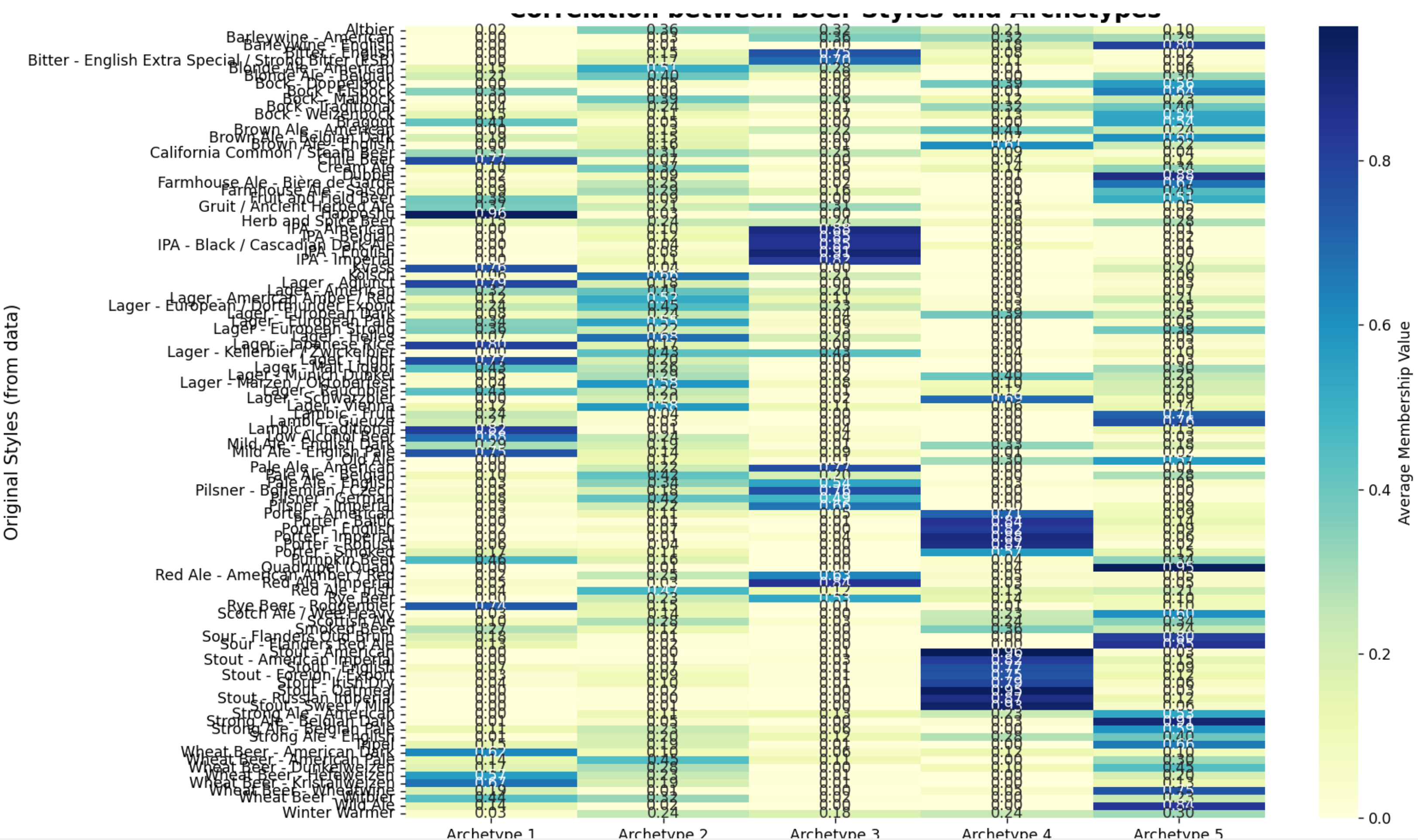
AND Malty is Very Low (approximately 35.0)

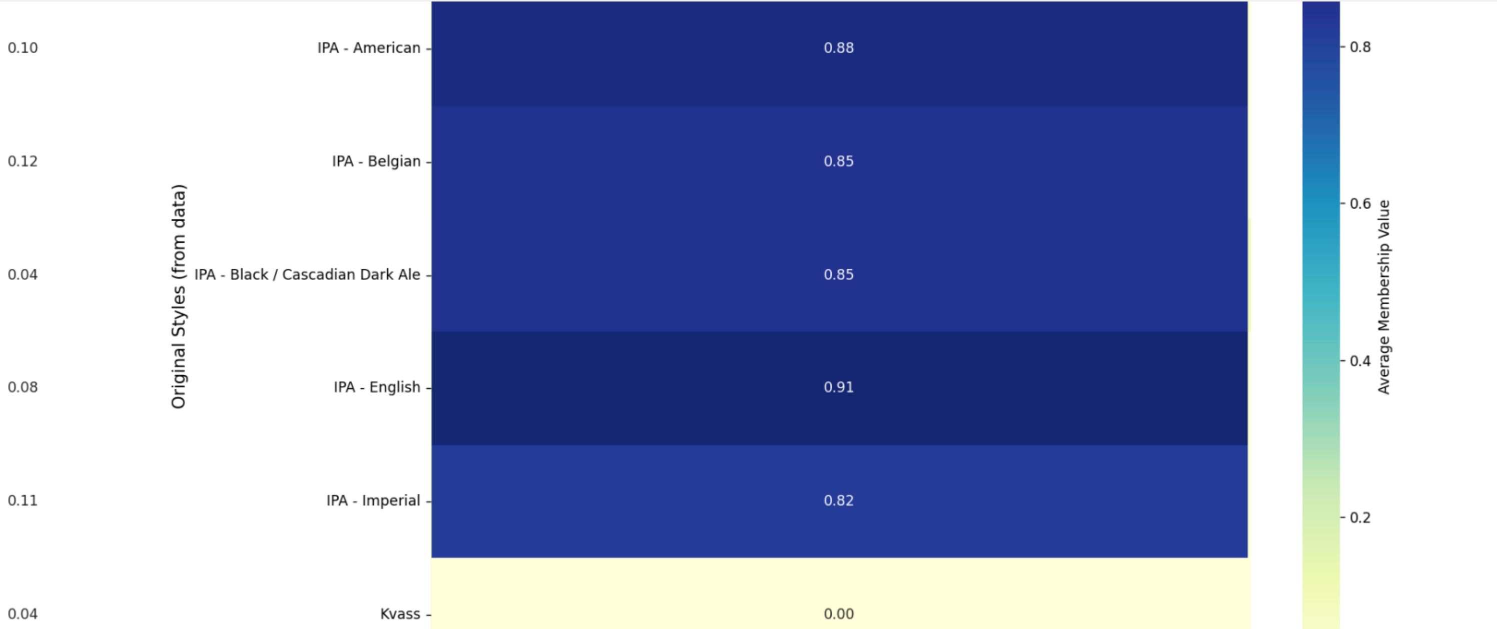
THEN Beer belongs to ARCHETYPE 1



Phase 3: Validation via Heatmap and Style Matches

The final step is to match the extracted archetypes to the formal beer styles from the dataset. For this purpose, a heat map was used, including all the beer styles from the database and all archetypes. It is observed that each archetype has at least one very typical beer for it. Other archetypes have more. An example of such is archetype 3, which is mainly defined by IPA-style beers.





At the end of the execution, the system calculates the Mean Absolute Error (MAE) and the Pearson correlation. The obtained values prove that the simplified linguistic rules are a good approximation of the complex FCM model. The bottom is the console output that describes this.

```
=== MODEL VALIDATION REPORT ===

Mean Absolute Error (MAE): 0.1318

Model Fidelity (Correlation): 90.19%

STATUS: SUCCESS. The fuzzy rules accurately approximate the FCM model.
```

5. *Main conclusions*

This project demonstrates the successful application of artificial intelligence methods and the fuzzy logic of data extraction. Through the integration of the Fuzzy C-Means algorithm and the concept of fuzzy rules, the raw statistical indicators of thousands of beer products have been transformed into understandable and mathematically based archetypes. The main conclusions are:

- **FCM Efficiency:** Fuzzy clustering demonstrates its quality as a taste data analysis approach. Compared to "hard" algorithms, it allows modeling the natural overlap between styles, which reflects the real world much more relevantly.
- **Automated interpretation:** The use of quintiles to define linguistic labels makes the system universal and independent of inaccurate "expert" opinions.
- **Common features of different styles of beer:** The heat map reveals a high correspondence in taste extremes (e.g. IPA and Imperial Stout), but at the same time registers a "diffusion" in the more balanced styles. This discrepancy is evidence that many beers with different marketing labels share identical or very similar flavor profiles, making archetypes a more objective classification criterion than traditional brewing categories.

Possible applications and future work

The developed algorithm can serve as a basis for:

- **Intelligent Recommender Systems:** Offering new products to consumers based on belonging to a particular archetype.
- **Quality control:** Breweries can use the model to check if a new product falls within the desired linguistic parameters of a style.

In the future, the model can be expanded by adding additional features such as color, price, or user rating.

6. *Literature and data used*

- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.
- Zadeh, L. A. (1965). *Fuzzy Sets*. Information and Control, Vol. 8, pp. 338-353.
- Ross, T. J. (2010). *Fuzzy Logic with Engineering Applications*. John Wiley & Sons.
- Pedrycz, W. (1996). *Fuzzy Control and Fuzzy Systems*. Research Studies Press Ltd.
- Python Software Foundation. Scikit-learn Documentation. Достъпно на: https://scikit-learn.org/stable/documentation.html
- Python Software Foundation. Matplotlib Documentation. Достъпно на: https://matplotlib.org/stable/contents.html
- Python Software Foundation. Seaborn Documentation. Достъпно на: https://seaborn.pydata.org/tutorial.html
- Python Software Foundation. *Pandas Documentation*. Достъпно на: https://pandas.pydata.org/docs/
- Python Software Foundation. *Numpy Documentation*. Достъпно на: https://numpy.org/doc/
- The used data is from Kaggle. Name of the database: „Beer Profile and Ratings Data Set“. Link: https://www.kaggle.com/datasets/ruthgn/beer-profile-and-ratings-data-set?select=beer_profile_and_ratings.csv

As of 05.02.2026, the above links are up-to-date.