Data Mining Project

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Fuzzy Clustering Implementation

Group 20

Lucas Campos Ferreira, 20250448

João Paulo de Avila, 20250436

Gonçalo Torrão, 20250365

Maria Leonor Ribeiro, 20221898

Fall/Spring Semester 2025-2026

TABLE OF CONTENTS

1. Executive Summary

Our customers can be organized into two practical groups that support day-to-day targeting and personalization. Most customers align clearly with one group, so segmentation remains actionable for campaign design. However, a meaningful minority—about 11% (1,760 out of ~16,390)—shows mixed behavior, meaning they sit between the two groups rather than fitting cleanly into one. Treating these "in-between" customers as purely one type can weaken performance because the message may feel incomplete or misaligned. The recommended approach is to keep the two-group segmentation as the foundation, while adding a simple execution rule: use group-specific messaging for clear-fit customers, and use hybrid messaging plus testing for the mixed segment to learn which framing works best at the moment.

2. Methodology

We implemented Fuzzy C-Means (FCM) with scikit-fuzzy as a complementary analysis to the main MeanShift segmentation. The model was trained on the preprocessed feature set (excluding the identifier column, Loyalty#), using the standard FCM input format (features × customers). We tuned two hyperparameters—number of clusters (c = 2 to 9) and fuzziness (m = 1.2, 1.5, 2.0, 2.5)—and selected the best configuration using the Fuzzy Partition Coefficient (FPC), where higher values indicate clearer separation. The best-performing setup was c = 2 and m = 1.2, and the final model achieved FPC = 0.9011, indicating strong overall separation into two broad groups under the fuzzy framework.

3. Analysis of Results

FCM produces a membership matrix u, assigning each customer a degree of belonging to each cluster, with memberships summing to 1. To compare "soft" vs. "hard" interpretations, we derived a hard label by taking the maximum membership (argmax), which mimics what a traditional hard clustering assignment would produce. The added value of the fuzzy view is that it provides a confidence signal: two customers can share the same hard label but differ substantially in certainty. A customer with memberships like (0.95, 0.05) is a strong fit for one group, while (0.55, 0.45) indicates a borderline profile.

To quantify overlap, we used a strict high-confidence rule: customers whose maximum membership is below 0.80 are flagged as overlap customers. Under this definition, the notebook identifies 1,760 overlap customers, which is roughly 11% of the population. This corrects the "100% hybrid" narrative: the data does not suggest that segmentation is invalid. Instead, it shows segmentation is broadly stable for most customers, with a meaningful boundary region where customers should be handled more flexibly.

4. Business Recommendations & Strategy

This fuzzy layer is intended to complement MeanShift rather than replace it. MeanShift provides the primary segmentation structure for operational targeting, while fuzzy memberships add a reliability flag that improves execution at the edges. For customers with high membership confidence (≥ 0.80), segment-specific messaging is appropriate because alignment is clear. For the overlap segment (< 0.80), avoid rigid or exclusionary targeting based on a single label, since these customers still have meaningful affinity with both groups. The most defensible approach is to use hybrid offers that combine the strongest value points of both groups and to apply A/B testing to determine which framing drives better response, then scale the winner. In practice, this reduces mis-targeting at the boundary and increases relevance for mixed customers without undermining the main segmentation strategy.

5. Conclusion

The fuzzy clustering component meets the required goals by applying fuzzy c-means, analyzing membership degrees, contrasting hard vs. soft interpretations, quantifying overlap, and translating uncertainty into actionable guidance. The correct takeaway is not that "distinct groups do not exist," but that most customers align clearly with one of two groups while ~11% sit near the boundary. We chose this enhancement because it addresses a key limitation of hard clustering used in the main work: hard methods provide a single label but not how reliable that label is, especially for customers near cluster boundaries. FCM contributes a practical interpretability layer by measuring how strongly each customer aligns with each group, enabling more careful targeting where uncertainty is highest. This makes the overall segmentation strategy more robust and execution-ready while keeping MeanShift as the primary operational segmentation.