Machine Learning Term Project

Exploratory Data Analysis and Clustering Results

Feyyaz Ketenoğlu
20211314081

Problem Definition and Dataset

- We aim to apply machine learning techniques to a real-world dataset from a legal domain, sourced from Kaggle. The dataset includes features about court cases and legal outcomes.

- For Example: https://www.kaggle.com/datasets/amohankumar/legal-text-classification-dataset

- Objective: Use machine learning to understand patterns and predict outcomes in legal cases.

Our Project “Yasal Pusula”

As part of our Legal Compass project, we aim to develop a predictive model that analyzes Supreme Court decisions to forecast court outcomes for Türkiye’s law system. To kick off the project, we will conduct preliminary experiments using the US court cases dataset available on Kaggle. This approach will allow us to evaluate the performance of machine learning algorithms and build a solid foundation for adapting these insights to our own dataset.

Dataset Description

- Source: Kaggle – Legal Case Dataset
- Total Records: 10,000+ entries
- Features: case_duration, plaintiff_type, defendant_type, verdict, region, case_type, etc.
- Target: verdict (Guilty / Not Guilty)
- Sampled 2,000 rows for analysis (random_state = 4081)

Missing Values and Categorical Variables

- Missing Values:
- - Handled using mean imputation (for numeric) and mode imputation (for categorical)
- Categorical Variables:
- - Encoded using one-hot encoding
- - Columns: plaintiff_type, defendant_type, region, case_type

Exploratory Data Analysis (EDA)

- Distribution of case durations shows right-skewed pattern.
- Verdict distribution is slightly imbalanced (58% Not Guilty, 42% Guilty).
- Region and case_type heavily influence verdict.
- Box plots and histograms were used for visualization.

Clustering Analysis

- Applied Hierarchical Clustering (Ward linkage, Euclidean distance).
- Data standardized before clustering.
- Dendrogram revealed 3-4 meaningful clusters.
- Standardization had significant effect on cluster distances.

Conclusions and Evaluation

- These results highlight the critical role of proper data scaling for successful clustering.