Data Preprocessing Report

1. Introduction

This report details the preprocessing and exploratory data analysis (EDA) steps performed on the dataset for the term project: "Predicting Outcomes of Turkish Constitutional Court Decisions Using Explainable Artificial Intelligence." The aim is to prepare the dataset for binary classification tasks while ensuring data quality and integrity.

---

2. Dataset Overview

The dataset consists of 13,676 individual application decisions from the Turkish Constitutional Court. These decisions were downloaded from the publicly available decision search engine and focus on individual applications evaluated on admissibility and merits.

---

3. Data Cleaning and Preprocessing

3.1 Initial Cleaning

- **Objective:** Create a binary classification dataset with two labels: "violation" and "no violation."
- Decisions containing both "violation" and "rejection" were excluded to ensure clarity in classification.
- Decisions with multiple outcomes were filtered to retain only those with consistent results for the first two rights most relevant to the case.
- **Outcome:** Final dataset reduced to **13,286 decisions** after cleaning.

3.2 Text Preprocessing

- Conversion of text to lowercase for normalization.
- Removal of stopwords, punctuation, and irrelevant symbols to improve model performance.
- Tokenization of the decision texts for feature engineering.
- Lemmatization for reducing words to their base forms.

---

4. Exploratory Data Analysis (EDA)

4.1 Label Distribution

- **Findings:**
    - The dataset is imbalanced with a higher number of "no violation" labels compared to "violation" labels.
    - Addressing this imbalance will require techniques such as oversampling, undersampling, or weighted loss functions in model training.

4.2 Word Frequency Analysis

- Common terms were analyzed to identify key patterns.
- High frequency of legal terms such as *"violation," "admissibility,"* and *"rights"* indicates their importance in predictive modeling.

4.3 Distribution of Word Count

- Most decision texts ranged between **500–8000 words**.
- Longer texts were identified as outliers, and a threshold was considered to exclude excessively verbose decisions.

4.4 Frequency of Right Types

- Certain types of rights (e.g., *freedom of expression* and *right to a fair trial*) were more prevalent in the dataset.
- Their frequency distribution indicates the potential influence of specific rights on court decisions.

---

5. Feature Selection and Engineering

5.1 Feature Selection

- **Rationale:** Focused on preserving linguistic features most relevant to classification.
- Selected features include:
    - N-grams (bigrams and trigrams) to capture context.
    - Term Frequency-Inverse Document Frequency (TF-IDF) for text representation.
    - For Transformer architecture we will use tokenizer of the chosen pretrained models.

5.2 Feature Engineering

- Creation of new features such as:
    - Length of decision text.
    - Average word length.
    - Count of legal-specific terms.

---

6. Visualizations

The following visualizations were created as part of EDA:

1. **Label Distribution Chart:** Highlights the balanced dataset
2. **Word Cloud:** Displays frequently occurring words in decision texts.
3. **Histogram of Word Count:** Illustrates the distribution of text lengths.

4. **Bar Chart of Right Type Frequencies:** Shows the prevalence of rights involved in decisions.

---

7. Conclusion

The data preprocessing and EDA steps outlined above were crucial for understanding the dataset and preparing it for machine learning tasks. Future steps will involve addressing class imbalance, experimenting with different feature representations, and optimizing models for explainability.