Investigating Key Factors Influencing the Primary Color (Mainhue) of National Flags

Web page with results: link

Data Preprocessing: For the initial phase, we streamlined our dataset by omitting attributes such as 'name', 'topleft', and 'botright', as these were the only nominal variables apart from our target, 'mainhue'. To ensure uniformity, we normalized all numeric attributes to a range between 0 and 1, utilizing the MinMaxScaler. Furthermore, we employed the LabelEncoder to convert our target variable into a numeric format, enhancing its compatibility with our analytical methods.

Utilized Libraries: We used sklearn library, paired with pandas dataframes, to facilitate both data preprocessing and modeling.

Adopted Methodologies:

1. **Attribute Selection through Iteration:** Our primary approach involved iteratively evaluating every possible combination of 'n' attributes. For each value of 'n' (1 to 4), we exhaustively examined each attribute subset of that size. By training a classification model on each subset, we assessed its training accuracy. We chose training accuracy as our metric, acknowledging its limitations for generalization, as our objective was solely to identify the most relevant features for our training dataset, without intending to apply the model to unseen data.

   We selected a decision tree as our classification model. This choice was made because of the predominance of Boolean attributes in our dataset, which serve as effective decision points within the tree structure.

2. **Experimenting with Regularization:** We explored regularization techniques (both L1 and L2), in conjunction with logistic regression and SVM, fine-tuning hyperparameters manually and via grid search. However, these models ended up being not that relevant with training accuracy of approximately 70%. This limitation could be caused by our atypical dataset with combination of numerical and majority of Boolean features.

3. **Unsupervised Learning with t-SNE:** We applied the t-SNE method for clustering. The resulting visualizations in 2 or 3 dimensions, however, revealed predominantly overlapping clusters. This overlap might come from the high proportion of Boolean attributes, leading to a sparse feature space that complicates the t-SNE model's ability to discern clear structures. Our attribute subset trials indicated that numeric factors such as area, language, and landmass have a more significant impact on classification, suggesting the relative insignificance of many Boolean attributes for clustering, potentially posing challenges for t-SNE analysis.