ML SS 2023
Task 0

Vladimir Panin (12238686) Reema George Dass (12144026)
Valentin Schweitzer (51829840)

March 2023

1 Introduction

The aim of this task is to analyze two different data sets: the Spambase dataset and the Flag dataset.

The Spambase dataset contains a total of 4,601 instances, with 57 features capturing various attributes of each email. We chose this dataset because it provides an interesting challenge in classification and because spam filtering is a ubiquitous and important problem in today's digital world. As for the number of instances used, this is considered to be the bigger dataset.

The Flag dataset, the smaller dataset in comparison, is made up of 194 instances with 30 attributes. It contains data, categorical as well as integer attribute characteristics, on 194 nations. The classification task could consists in predicting the religion of its' population or the colours of the associated flag. The decision to choose this dataset was taken in order to learn about the limits of machine learning and investigate the difference between causality and correlation, as the assumption would be that there is no causality. We are curious to see how well a Machine Learning model can be trained on this data.

2 Data Description

2.1 Spambase Dataset

The Spambase dataset was created by Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt at Hewlett-Packard Labs. The collection of e-mails was collected from their postmasters and other individuals. The dataset contains personal as well as work e-mails. In total, the data set consists of 58 column, with 57 attributes that can be used, to predict, whether the element should be considered spam or not, which is a categorical feature.

In this dataset, there are 48 continuous real attributes by the form word_freq_Word, which represent the percentage of words in the e-mail that match WORD. Additionally, there are are six continuous real attributes of type char_freq_CHAR, which state the percentage of characters in the e-mail that match CHAR. Furthermore, there are the continuous real attributes capital_run_length_average, capital_run_length_longest, and capital_run_length_total, which state the average length, the longest length and the total length of uninterrupted sequences of capital letters respectively. Additionally, there is a class attribute, which signals whether the e-mail was considered spam or not. Finally, there are not missing values and 39.4% of e-mail were actually spam and hence 60.6% were non-spam e-mails.

2.2 Flag Dataset

This dataset, titled "Flag database," was collected primarily from the "Collins Gem Guide to Flags" by Collins Publishers in 1986, and contains details of various nations and their flags. 10 of the

attributes of the dataset are numeric and the rest being Boolean or nominal. The attributes include information on the country like for example name, landmass, zone, area, population, language, religion, number of bars, stripes, colours. There are no missing values in the dataset.

3 Data Preprocessing

3.1 Spambase Dataset

After loading the data and creating the correlation matrix, the most correlated and uncorrelated attributes were identified by examining the matrix and selecting the corresponding pairs of attributes. The results, including the correlation matrix and the most correlated attributes, were visualized in the two figures 1 2 . Finally, the most correlated attributes were used as input features for the RandomForestClassifier model to predict the importance of the feature variables.

The correlation matrix in 1 shows that the majority of attributes are uncorrelated. However, it can be pointed out that some attributes are highly correlated like for example "word_freq_telnet" and "word_freq_data".

Figure 2 shows the importance of the feature variables in a ranked horizontal bar chart. The ranked feature importance increases steadily and most important attributes have a feature importance of multiple times the average feature importance. It is to point out, that the most used attributes are "char_freq_!", "char_freq_$" and "word_freq_remove"

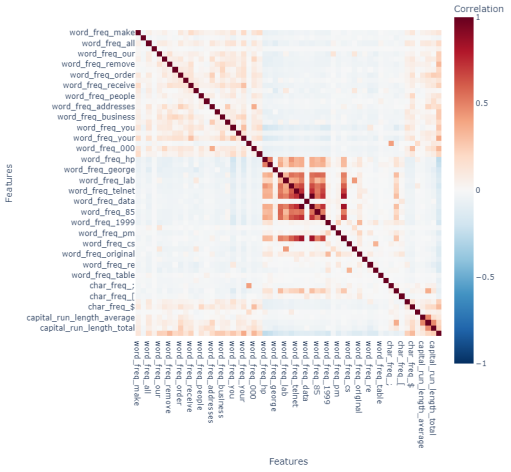

Figure 1: Correlation matrix for all attributes

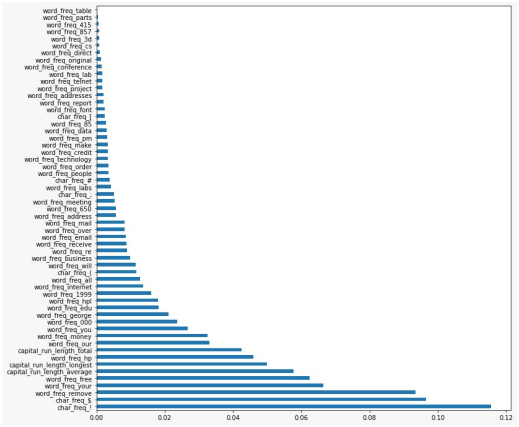

Figure 2: Ranked importance of features for the prediction

Figure 3: Plots from Spambase dataset

3.2 Flag Dataset

After loading the dataset from the CSV file two dictionaries for the attributes "language" and "religion" are created. The next step is to plot features of the dataset. To show the features of the dataset two plots are created. Figure 4 is a heatmap that shows the fraction of flags with each main hue by religion/ideology. It is to note, that most of the correlations between the main hue and the dominant religion or ideology are quite low, with a few notable example as for the correlation between the main hue color "white" and "Others" for ideology or religion or between the main hue color "red" and being "Marxist".

The Figure 5 is a bar plot of the 20 most populous countries and their associated religion/ideology. The plot shows that the 3 most populous countries in 1986 were either Marxist or Hindu. Countries with a smaller population have a higher variety in terms of belonging to a religion or an ideology.

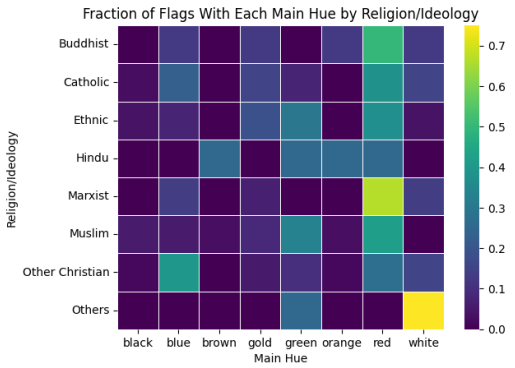

Figure 4: Main hue by religion

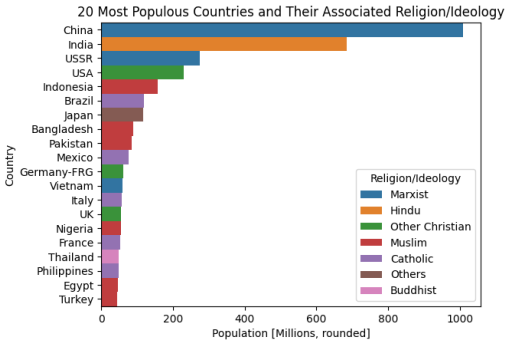

Figure 5: Religion by country and population

Figure 6: Plots from Flag dataset

After the plots are created, the code replaces the numerical values in the "language" and "religion" columns with their corresponding strings using dictionaries. Dummy variables for categorical features are created using the "get_dummies" function in pandas. Finally, the numeric columns are normalized by dividing each column by its maximum value. The resulting dataframe is stored in "normalized_df".