CSE453-Pattern Recognition
FINAL PROJECT REPORT

Berke Süslü
Computer Engineering
161044076
berke.suslu2016@gtu.edu.tr

I. Introduction

This project is about using flag attributes in order to guess countries religion.To achieve this, feature selection and classification methods are used.

II. Dataset

A. General Information

The dataset contains details of various nations and their flags. In the dataset, the fields are separated by commas. The dataset is created in 5/15/1990 and has 30 attributes and 194 instances. [1]

B. Attributes Information

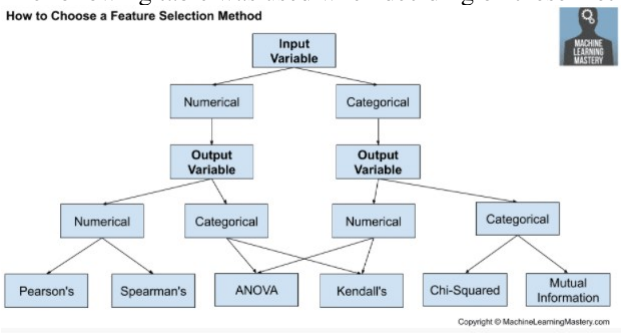
Atrribute name	Information about attributes
name	Name of the country concerned
landmass	1=N.America, 2=S.America, 3=Europe 4=Africa, 5=Asia, 6=Oceania
zone	Geographic quadrant, based on Greenwich and the Equator: 1=NE, 2=SE, 3=SW, 4=NW
area	in thousands of square km
population	in round millions
language	1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
religion	0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
bars	Number of vertical bars in the flag
stripes	Number of horizontal stripes in the flag
colours	Number of different colours in the flag
red	0 if red absent, 1 if red present in the flag
green	same for green
blue	same for blue
gold	same for gold(also yellow)
white	same for white
black	same for black
orange	same for orange
mainhue	predominant colour in the flag
circles	Number of circles in the flag
crosses	Number of (upright) crosses
saltires	Number of diagonal crosses
quarters	Number of quartered sections
sunstars	Number of sun or star symbols
crescent	1 if a crescent moon symbol present, else 0
triangle	1 if any triangles present, 0 otherwise
icon	1 if an inanimate image present (e.g., a boat), otherwise 0
animate	1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
text	1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
topleft	colour in the top-left corner
botright	Colour in the bottom-left corner

III. Feature Selection Methods

The following methods were used for feature selection.

a) ANOVA Test
b) Chi-squared Test
c) Recursive Feature Elimination

The following table was used when deciding on these methods.



[2]

A. ANOVA Test

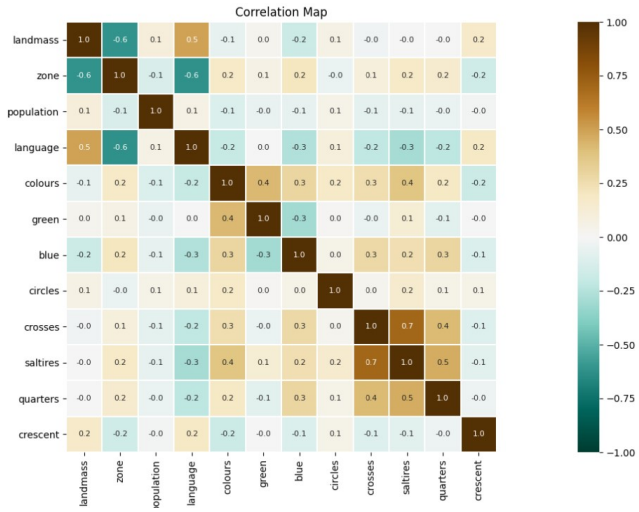
Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the ”variation” among and between groups) used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. [3] ANOVA test was used to obtain categorical outputs from numerical inputs. The test produced p-values and a significance level of 0.05 was set. The features below the significance level were selected. Selected features:”population”, ”colours”, ”circles”, ”crosses”, ”saltires”, ”quarters”.

B. Chi-squared Test

A chi-squared test, also written as X^2 test, is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. [4] Chi-squared test was used to obtain categorical outputs from categorical inputs. The test produced p-values and a significance level of 0.05 was set. The features below the significance level were selected. Selected features:"landmass", "zone", "language", "green", "blue", "crescent".

C. Recursive Feature Elimination

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. [5] The Recursive Filter Elimination method was used to combine the results from the two feature selection methods. Decision Tree Classifier was used as the estimator. The best 10 features are selected.



IV. Classification

A. Decision Tree Classifier

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. [6] Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model. Entropy is used as a classification criteria. The obtained features were loaded into the Decision Tree model and the training data were placed in this model(%80 training data, %20 test data). After the model was created, predictions were made with the test values and compared with the actual values. Example of a decision tree model:

V. Conclusion

The test and training data were shuffled so that the results were different each time. As a result, the accuracy rate varies between %50 and %80. This means that the probability of finding the right religion from the predictions is higher than %50.

References

[1] UCI Machine Learning Repository: Flags Data Set
George H. John and Ron Kohavi and Karl Pfleger.
Irrelevant Features and the Subset Selection Problem. ICML. 1994.
https://archive.ics.uci.edu/ml/datasets/Flags

[2] How to Choose a Feature Selection Method For Machine Learning
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

[3] Wikipedia - Analysis of Variance
https://en.wikipedia.org/wiki/Analysis_of_variance

[4] Wikipedia - Chi-squared test
https://en.wikipedia.org/wiki/Chi-squared_test

[5] SKLearn Manual - RFE
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[6] SKLearn - Decision Trees
https://scikit-learn.org/stable/modules/tree.html