

## PROJECT REPORT

**Course Title: Data Mining & Machine Learning**

**Course Code: CSE3**

**Submitted to: Ms. Fariha Jahan**

**Group Member’s Name:**

Name: Bidyut Sharma (Id: 191-15-12720)

Name: Zannatun Nayem Vasha

(Id: 191-15-12939)

Name: Israt Jahan Esha (Id: 191-15-12921)

Name: Mosa. Rabeya (Id: 191-15-12145)

## GitHub link:

Flags Dataset

## Tittle:

**Abstract:**

This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colours in its flag.  
  
10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.

# Introduction:

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This

is important because it allows trends and patterns to be more easily seen. Data visualization analyst: job description. Data visualization analysts make large data sets useful and meaningful by presenting key information in a variety of ways. Data visualization analysts deliver data in useful and appealing ways to users.

We use a dataset (flags dataset) for data visualization. The ensign may be a image or emblem of a rustic, and therefore it represents a rustic. every country within the world has its own combination of coolers, shapes and symbols, but conventionally most national flags are rectangular. Every nation withiin the world decides their own style of flag. Maybe, this is the explanation why each ensign provides info about nation that they're belong to. Certainly, there are several fetors taking part in role on style of a ensign. during this study, we centered on these factors And worked an analysis on different national flags. Flag dataset’s UCI repository (https://archive.ics.uci.edu/ml/datasets/Flags). The UCI repository is a collection of databases, domain theories, and data generators. First of all, we import this dataset into database and also import in google colaboratory.Then we have done data preprocessing. After that, we connected dataset and google colaboratory.

# Methodology:

**Dataset:** This dataset has 194 rows and 30 columns. The columns are name, landmass, zone, area, population ,language, religion, bars, stripes, colors, red, green, blue, gold, white, black, orange, mainhue, circles, crosses, saltires, quarters, sunstars, crescent, triangle, icon, animate, text,topleft,botright.

## Pandas:

Pandas is a fast, powerful, flexible and easy to opensource data analysis and manipulation tool, built on top of the python programing language.

## Advantages of Pandas Library

* 1.1. Data representation.Pandas provide extremely streamlined forms of data representation. ...
* 1.2. Less writing and more work done. ...
* 1.3. An extensive set of features. ...
* 1.4. Efficiently handles large data. ...
* 1.5. Makes data flexible and customizable. ...
* 1.6. Made for Python.

## Matplotlib:

Matplotlib pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

## Numpy:

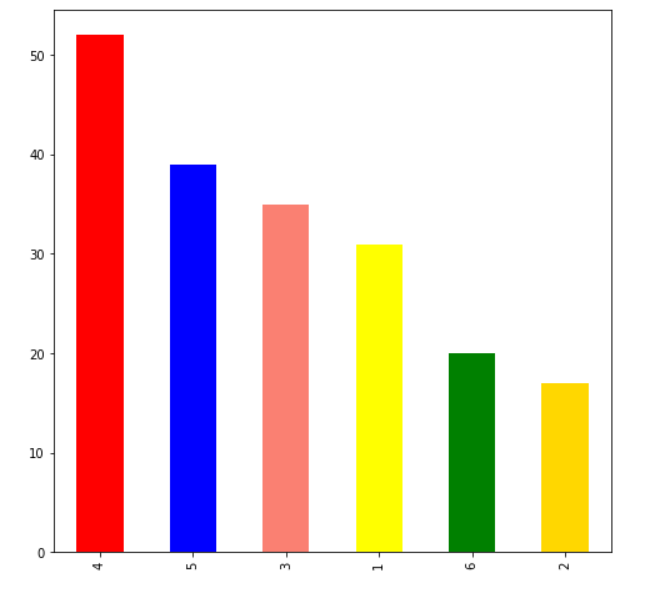
NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an opensource project and you can use it

freely. NumPy stands for Numerical Python.

**Seaborn:**

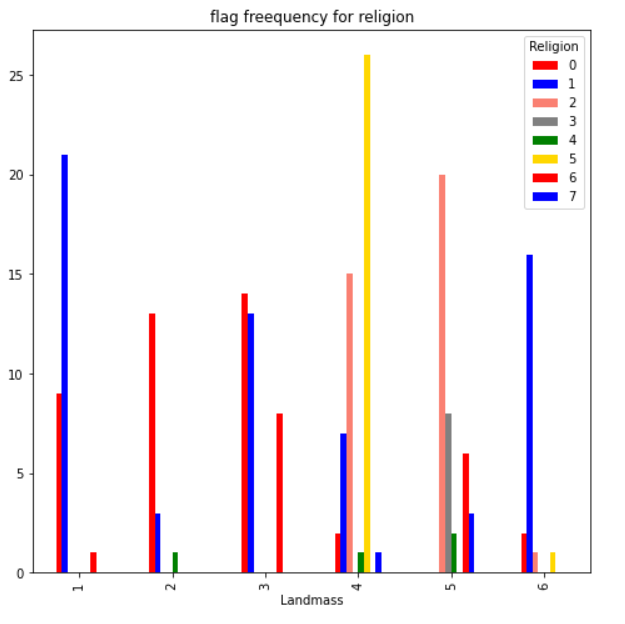
Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

**Result:**



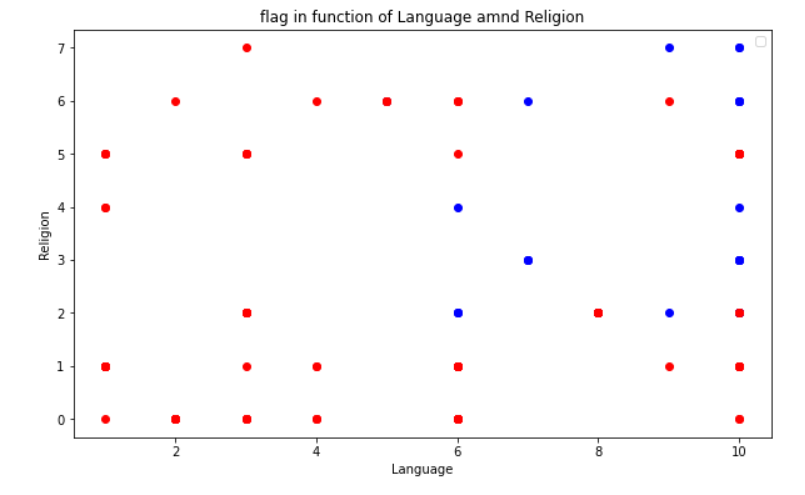
## Figure:01

Here, we can describe our plot through values of landmass column. We have taken the column of the landmass along with y-axis.



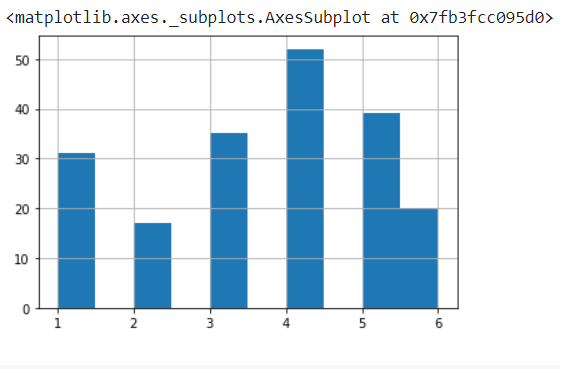
## Figure:02

Here landmass and religion are column name. we can describe our plot through values of landmass and religion column. We have taken the column of the landmass along with x-axis.



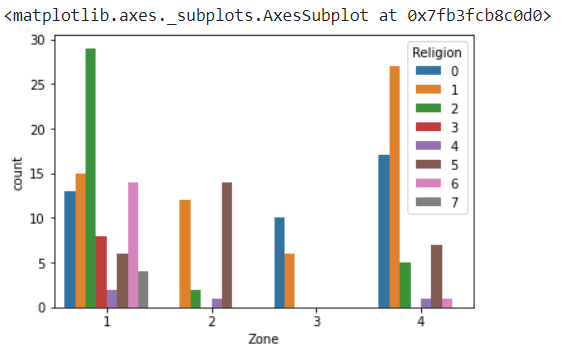
## Figure: 03

We can see that the graph above is a scatter. Here we used religion and language column and we set that the color of the column are red and blue . we can describe our bar scatter through values of religion column and language column. We have taken the column of the religion along with y- axis and column of the language along with x – axis.



## Figure: 04

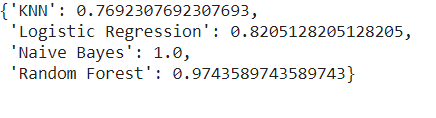
We can see that the graph above is a histogram. we can describe our histogram(hist()) through values of landmass column. We have taken the column of the landmass along the x-axis.



# Figure: 05

Here we use seaborn library function only for this graph. we use sns.countplot keyword for seaborn function. Here we use zone and religion column. we can describe our countplot through values of zone column and religion column. We have taken the column of the zone and the column of the religion along the x-axis and y-axis.

**Classification:**

****

# Figure: 06

# Conclusion:

We can predict the language of the country as well just the way as the religion is predicted. In future we can predict the change in economic status of a country based on the previous years’ financial data. The population growth of the country can also be predicted using any of the data mining techniques. The data will have more intra-cluster similarity and less inter-cluster similarity. We consider Naive Bayes, KNN and J48 algorithms for classification. And ROC graph the knowledge is for the abalone dataset the best classifier is j48 cause its accuracy 100% and J48 Classifier as TPR &FPR IS 1,0 ,so we say j48 is best classifiers