**Abstract:**

This report discussing the functions performed on the data which is taken from a dataset chosen from the UCL (University College London) repository (<https://archive.ics.uci.edu/ml/datasets.php>) called *Flags dataset*. Then descriptive statistics were performed. Ten columns were chosen for that and then I tried to discover relationships by taking some pairs of columns. Further, a question was proposed for explorative data analysis. Then in the next part we did data modelling where we chose two classification methods and data was split into training and test sets. Then several parameters were analysed for the same. Later, a comparison was sone between two classification methods.

**Introduction:**

In this project we used the dataset ‘Flags Data Set’ (<https://archive.ics.uci.edu/ml/datasets/Flags>).

All the tasks are performed on this dataset. It is taken from the UCL repository. Task 1 is about problem formulation, data acquisition and preparation and task 2 is about data exploration and task 3 is about data modelling.

**Tasks:**

Task 1: Problem Formulation, Data Acquisition and Preparation:

From the UCL repository we chose a dataset which was supposed to fulfil all the criteria as mentioned. I chose Flags dataset as I was intrigued about it that how flags and associated information can determine so many things. Also, the dataset matched the criteria as mentioned in the assignment.

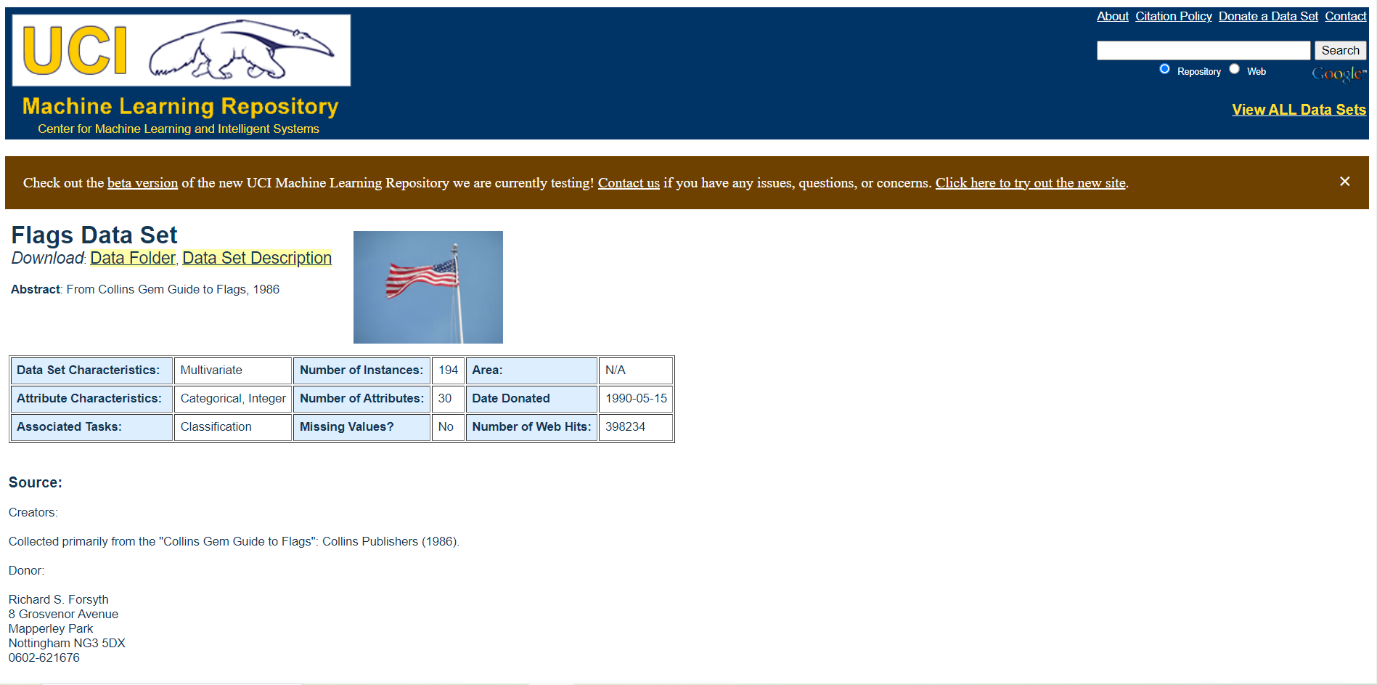


Figure 1: Flags Dataset(<https://archive.ics.uci.edu/ml/datasets/Flags>)

The data file is having information about several countries and their respective flags. You can explore the data and try things like predicting the religion of a country by analysing its size and the colours in its flag. Out of 30, 10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued attributes (<https://archive.ics.uci.edu/ml/datasets/Flags>).

Attribute information of the dataset taken from the website itself(<https://archive.ics.uci.edu/ml/datasets/Flags>) -

1. name: Name of the country concerned  
2. landmass: 1=N.America, 2=S.America, 3=Europe, 4=Africa, 5=Asia, 6=Oceania  
3. zone: Geographic quadrant, based on Greenwich and the Equator; 1=NE, 2=SE, 3=SW, 4=NW  
4. area: in thousands of square km  
5. population: in round millions  
6. 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  
7. religion: 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others  
8. bars: Number of vertical bars in the flag  
9. stripes: Number of horizontal stripes in the flag  
10. colours: Number of different colours in the flag  
11. red: 0 if red absent, 1 if red present in the flag  
12. green: same for green  
13. blue: same for blue  
14. gold: same for gold (also yellow)  
15. white: same for white  
16. black: same for black  
17. orange: same for orange (also brown)  
18. mainhue: predominant colour in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)  
19. circles: Number of circles in the flag  
20. crosses: Number of (upright) crosses  
21. saltires: Number of diagonal crosses  
22. quarters: Number of quartered sections  
23. sunstars: Number of sun or star symbols  
24. crescent: 1 if a crescent moon symbol present, else 0  
25. triangle: 1 if any triangles present, 0 otherwise  
26. icon: 1 if an inanimate image present (e.g., a boat), otherwise 0  
27. animate: 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise  
28. text: 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise  
29. topleft: colour in the top-left corner (moving right to decide tie-breaks)  
30. botright: Colour in the bottom-left corner (moving left to decide tie-breaks)

First, we import the dataset and convert it into desired format and when it is in tabular form, we are going to use the dataset.

Task 2: Problem Formulation, Data Acquisition and Preparation:

I chose ten columns out of 30 which were a mixture of numerical and categorical attributes. They are landmass, zone, area, population, language, religion, bars, stripes, colours, and main hue. Descriptive statistics were performed on these columns, and they were visualized, keeping in mind the numerical and categorical attributes.

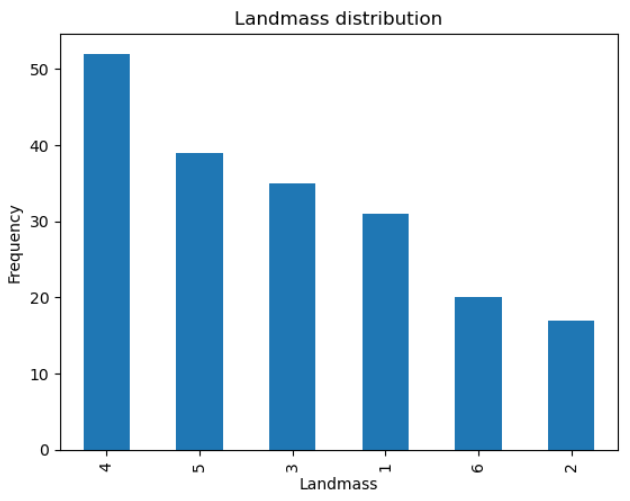


Figure 2: Landmass

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 3: Zone

A picture containing text, screenshot, number, diagram

Description automatically generated

Figure 4: Area

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure 5: Population

A picture containing text, screenshot, font, diagram

Description automatically generated

Figure 6: Language

A picture containing text, screenshot, diagram, font

Description automatically generated

Figure 7: Religion

A picture containing text, screenshot, number, plot

Description automatically generated

Figure 8: Bars

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 9: Stripes

A picture containing text, screenshot, font, plot

Description automatically generated

Figure 10: Colours

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 11: Main Hue

Relationship between some columns:

* Area and population: Do more area means more population or area does not impact population significantly? Here, we can see that when the area was highest population was not that significant (20000 and 250 approximately) while, when area was around 10000, population was the most.

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 12: Area and Population

* Mainhue and colour: Mainhue tells the most predominant colour in a flag. Colours are the range of colures present in the flag. Here we can see that the colours mentioned in the dataset are most common colours found. Blue and orange are leading colours and brown is the least common among these eight.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 13: Main Hue - Colours

* Stripes and bars: If a flag is having stripes, bars or both, which is actually very common feature in a flag. They are not found together in a flag.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 14: Stripes- Bars

* Landmass and zone: A particular landmass (1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania) falls into which zone (1=NE, 2=SE, 3=SW, 4=NW).

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

Figure 15: Landmass-Zone

* Landmass and language: What language is prominent in a particular landscape.

A picture containing text, screenshot, colorfulness, diagram

Description automatically generated

Figure 16: Language-Landmass

Then, a question was supposed to be proposed, I was in the description. What is the religion distribution among different landmasses as mentioned in the dataset description?

* I grouped the data by landmass and religion.
* Frequency of each religion was counted within each landmass.
* Create a bar chart which is stacked showing the distribution of what religion is followed by the masses within each landscape.
* It can be inferred that all land masses provided that are, 1=N.America, 2=S.America, 3=Europe, 4=Africa, 5=Asia, 6=Oceania, there is diversity in terms of religion (0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others) in every land. In North America, there are prominently Catholics and other Christians. In south America, the situation is also same. However, population on which it is accessed is less.
* The same goes for Europe as well however there is Marxism as well with the otherb two.
* In Africa, there are various religions practiced, Ethnic being most common.
* Same situation is in Asia but here Islam is practiced the most. In Oceania, other Christians take the lead.

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

Figure 17: Religion - landmass

Task 3: Data Modelling:

Data modelling was done with ‘scikit-learn’ package specifically and various libraries were imported from it. I chose two classification methods they were KNN and Decision Trees. K-Nearest Neighbours (KNN) and decision trees are supervised machine learning algorithms used for classification and regression functions. Next, the data is to split data into training and test sets respectively. However, this will be done in three suites. I chose ‘language’ as the target variable. First suite splitting into 50% training and 50% testing, second splitting into 60% training and 40% testing and third one, 80% training and 20% testing.

For each suite, first the models were trained accordingly and then Confusion matrix, Classification accuracy, Precision, Recall, F1 score are the parameters that were used to evaluate the performances of the models. Classification reports were generated for the same.

Then both methods, KNN and decision trees were compared, and graphs were plotted with respect to the confusion matrices (10 attributes were considered). Here, confusion matrix tells us how a classification algorithm performs.

This was the result:

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 18: Comparison between two methods

K-Nearest Neighbours:

Accuracy: 0.23076923076923078

Precision: 0.3143719026071967

Recall: 0.23076923076923078

F1 Score: 0.1956315289648623

Decision Tree:

Accuracy: 0.5128205128205128

Precision: 0.6162393162393163

Recall: 0.5128205128205128

F1 Score: 0.5050505050505051

(Both recall and F1 Score are integral parameters for evaluating the performance of a classification model, especially when positive instances need to be correctly identified.)

The above comparison shows that ‘Decision tree’ was a better choice as compared to ‘KNN’ method.

Conclusion:

Here, we also used ‘scikit-learn’ package. We ventured into the basics of machine learning. We chose ‘Flags’ dataset from the UCL repository. Data exploration was performed followed by data modelling. Further we found out that ‘Decision tree’ is a better classification method as compared to KNN.