

**University Tunku Abdul Rahman**

**Faculty Information and Communication Technology**

UCCD2063 Artificial Intelligence Techniques

Practical Assignment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Name** | **Student ID** | **Program** | **Contribution** |
| 1. | ANG JIAN WEI | 19ACB00952 | CS | 33.33% |
| 2. | FOO YUE YI | 19ACB04735 | CS | 33.33% |
| 3. | SAMUEL TAN JOO WOON | 19ACB07010 | CS | 33.33% |

**Introduction**

Flags are national symbols, it a show that we belong to a community or nation and that we partake similar beliefs, goals, rules and regulations. Every country has got a specific flag as their national symbol. A country will spend a lot of time and plenty of money on their flag’s design because a flag is a country’s way of representing itself to the rest of the world. There are many characteristics in a flag which portray and define a country. There are many factors which affect the design of a national flag.

Each country in the world has its own combination of colours and symbols on their flags which are unique and distinguishable. The design of a national flag is also influenced by its religion. Most flags by their very nature adopt religious significance. In many countries, civil religion mingles with patriotism in some sort of veneration of the flag. So does the characteristic of a flag really determine its religion? How accurate is it if we use certain aspect of a flag to assume its religion status?

In this study, we focused on these aspects and made an analysis on different national flags to find out the relation between characteristics of the flag and the country’s religion. We have used feature extraction method, after that we applied different machine algorithms to predict religion of the country. We also presented the correlations of certain components that are exist on a country flag such as colours, bars or stripes, symbols such as sunstars, crosses, and saltires.

We will concentrate on constructing a machine learning method that can be utilized to predict the religion status of a country based on certain features on its national flag. In this assignment, we will retrieve data, explore data, process the data, select and train model, fine tuning, selecting features, selecting which model is the most accurate and efficient, evaluate those model and apply it to our analysis. We will finally implement it to perform classification of religion based on characteristics of a national flag.

**Chapter 2 Method and Experiments**

**2.1 Dataset collection, data exploration, visualization**

**2.1.1Data Collection:**

Flag Data set was collected primarily from the “Collins Gem Guide to Flags”. Collins Publishers (1986). This data file contains details of various nations, religion and their flags.

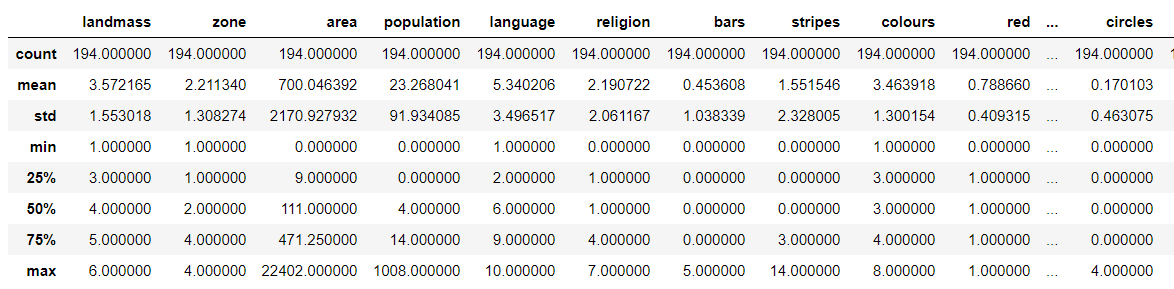
**2.1.2 Data Attribute**

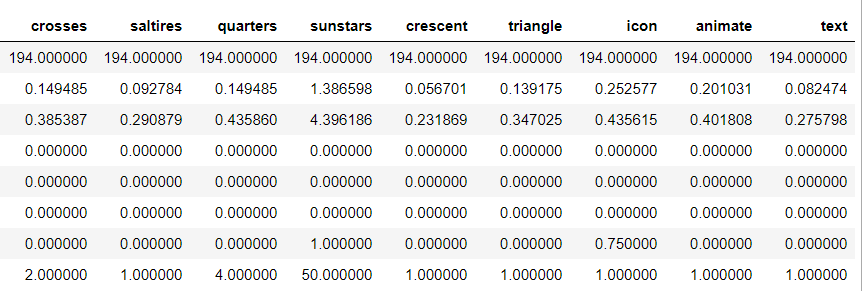
The data set characteristic is multivariate, and have 194 instances and 30 attributes. There is no any missing values in the column. There is 10 attributes are numeric-value, remainder are either Boolean or nominal valued.

Attributes Info

1. Name
2. Landmass – 1=N.America, 2 = S.America, 3= Europe, 4=Africa, 5= Asia, 6=Oceania
3. Zone – 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 absent, 1 present
12. Green - 0 if absent, 1 present
13. Blue - 0 if absent, 1 present
14. Gold/Yellow - 0 if absent, 1 present
15. White - 0 if absent, 1 present
16. Black - 0 if absent, 1 present
17. Orange - 0 if absent, 1 present
18. Mainhue – predominant colour in the flag
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. Sun stars – Number of sun or star symbols
24. Crescent – 1 if a crescent moon symbol present, else 0
25. Triangle – 1 if any triangles present , else 0
26. Icon – 1 if any inanimate image present , else 0
27. Animate – 1 if an animate image ,else 0
28. Text – 1 if any letters or writing on the flag , else 0
29. Top left – colour in the top-left corner (moving right to decide tie- breaks)
30. Bot Right – colour in the bottom-left corner (moving left to decide tie-breaks)

Continuous Attributes:



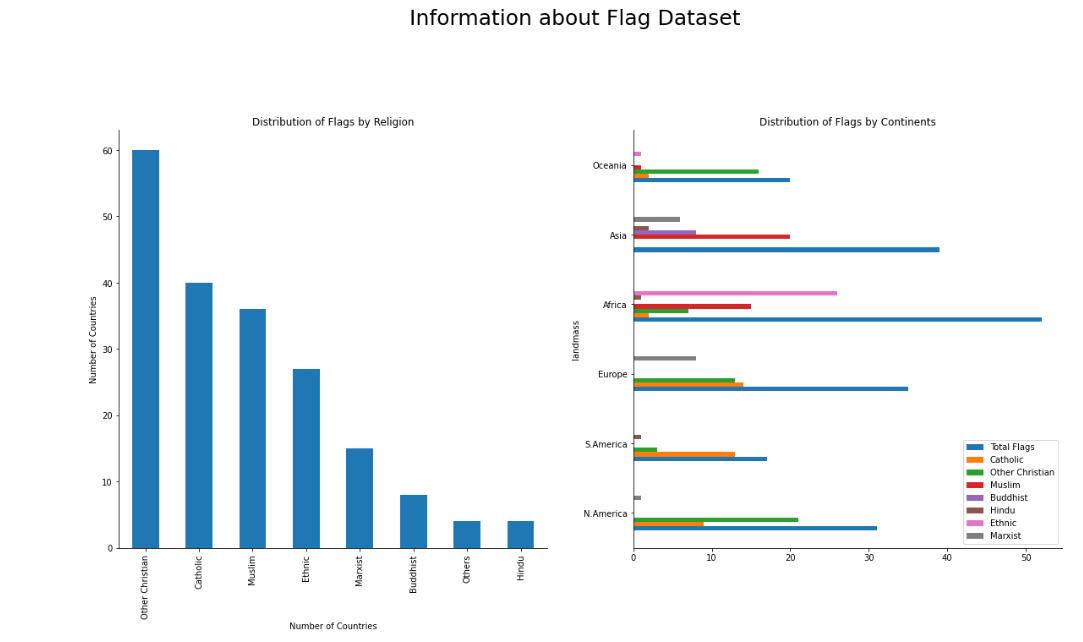


**2.1.3 Data Exploration and visualization**

Before a model is create an examination of the features was done. During the examination a number of necessary operations were done to the dataset to prepare for fitting in a model. Included

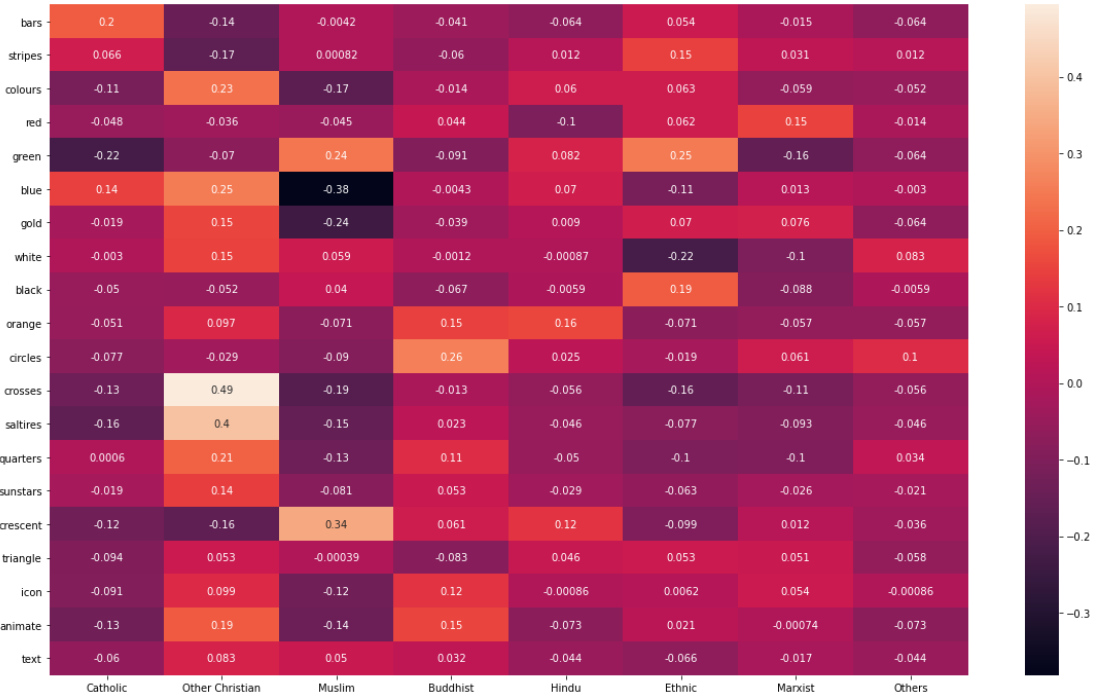


These features were removed as even though the features might have a high correlation with the country’s religion, this information will not be available in future usage of the model. Once the irrelevant data was removed, further exploring of the data occurred by using visualizations.



From the data, we can see that there is quite a high bias towards Christian flags. The charts also show that the other dominant flags are Christian, Catholic and Muslim. However, interestingly the Ethnic flags have a high bias in African countries. Due to the high bias, it could potential indicate an identifiable pattern that could be found among the flags despite the low amount of data collected.

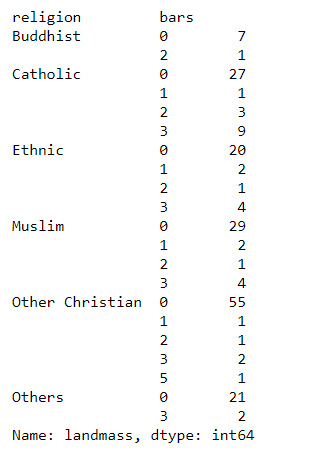
Further feature exploration will be conducted to see the correlation between the features and each religion.



From the data, we can that certain features correlate with certain religion more that others.

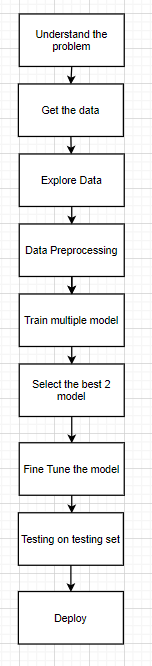
1. Catholic : number of bars
2. Christian : Crosses, Saltires, quarters
3. Muslim : Crescents, Green
4. Buddhist : Circles, Orange
5. Ethnic : Green

However, due to the limited data and therefore no recognisable patterns with flags belonging to Marxist, Hindu or others, a decision was made to combine the Marxist and Hindu flags into the others category. We will then examine the data to see the identified patterns that been discovered by the exploration while also further investigating the new 'other' section to see if a pattern can be found within it.



From the table above, we can that the number of bars does offer us some value in our prediction model. However, the number of bars on the flag could correlate more with its continent than its religion. As a lot of Catholic countries are in Europe and there are a lot of European flags that have 3 bars. This could be problematic as there is a high number of Catholic countries in South America, which may not have the same historic flag design as European countries.

**2.2 Flow Chart**

****

**2.3 Preprocessing Step**

**2.3.1 Preprocessing**

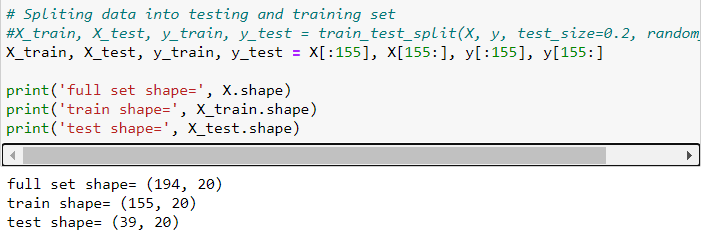
Before we start to run the training set and testing set, we had separate the output(Y) and input(X). For our data set, Y is the based on our flag design, to determine the country religion, and X is the design which is bars, stripes, colors, red, green, blue, gold, white, black, orange, circles, crosses, saltires, quarters, sun & stars, crescent, triangle, icon, animate and text.

1. Create the input Matrix x column from selected\_cols
2. Output vector Y created by selected the religion



Split the dataset into training and testing set

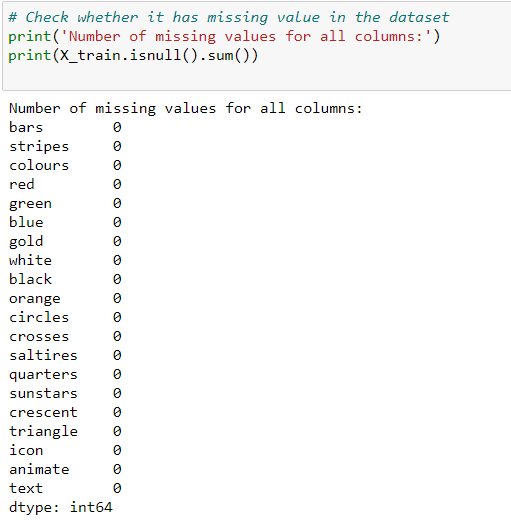
* Create X\_train, X\_test, y\_train, y\_test = X[:155], X[155:] ,y[:155], y[:155]
* Or X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=30)



Check Missing values

We had checked the data set, there is no missing values in each column in the dataset.

* Check (X\_train.isnull().sum())

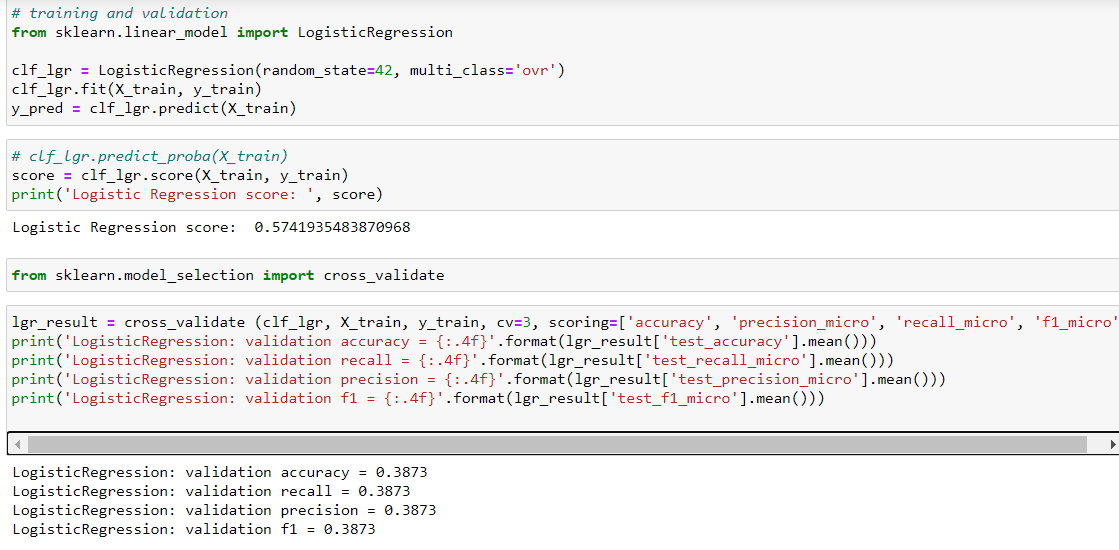


No missing value, so proceed to the next step.

**2.3.2 Model Training**

We had selected 4 algorithm to train the model which is Logistic Regression, Decision tree classifier with max depth(8) and max depth(5), SVM classifier and KNN classifier.

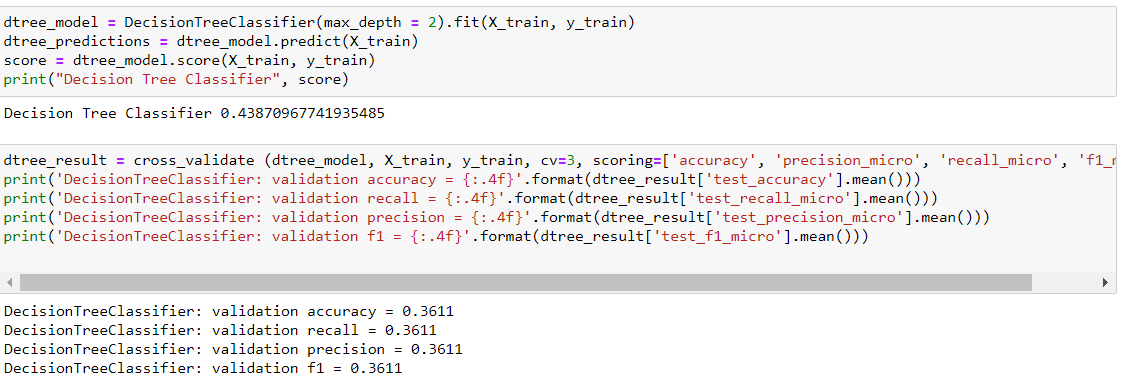
**2.3.2.1 Logistic Regression**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logistic regression) is estimating the parameters of a logistic model (a form of binary regression). 

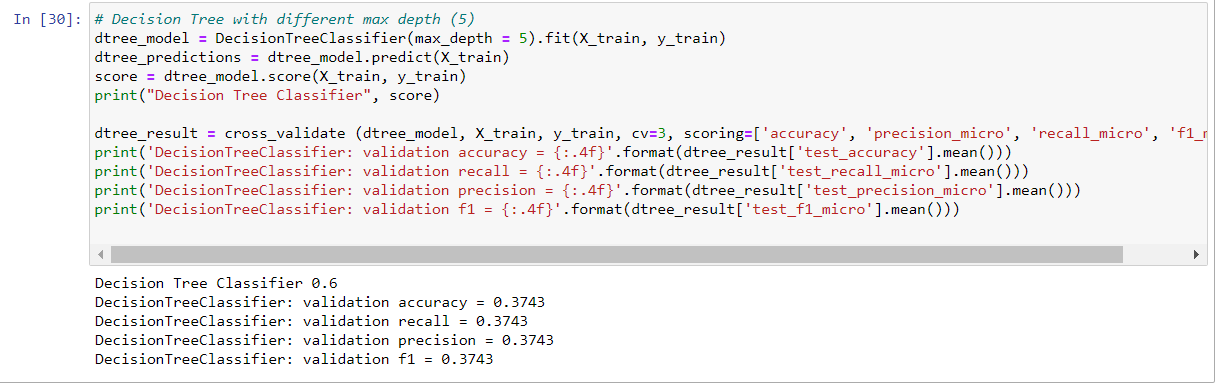
**2.3.2.2 Decision Tree Classifier**

Decision Tree Algorithm is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees in these tree structures, leaves represent class labels and branches represent conjunction of features that lead to those class labels. Different type of depth were using in this training set which normal test, with max depth (5) and with max depth (8), and the depth will affect the classifier.

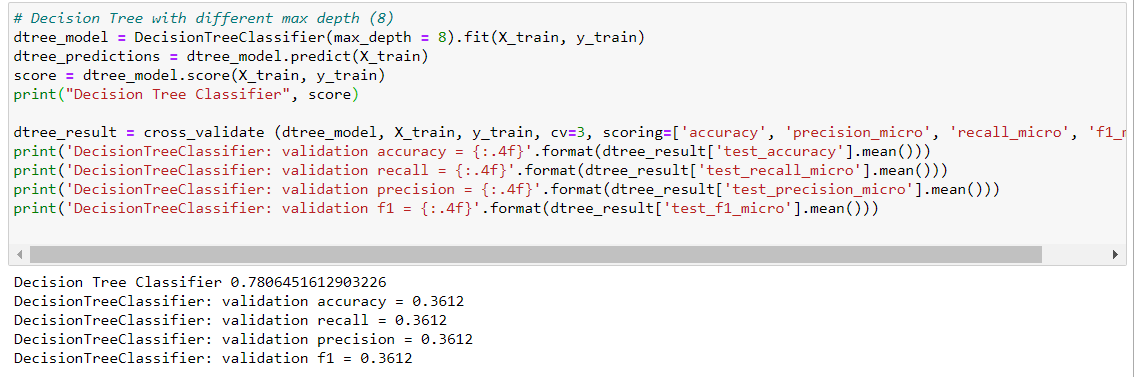
**No Max Depth**



**Max depth (5)**

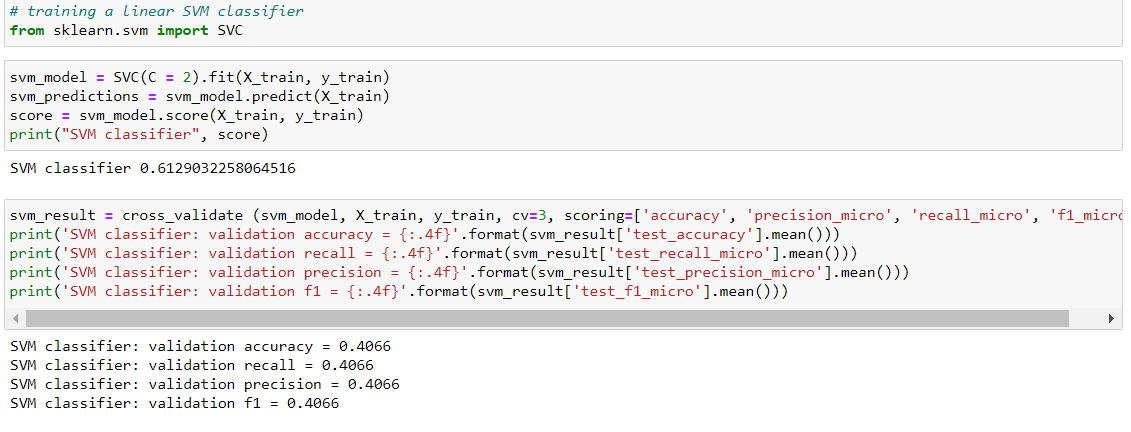


**Max depth (8)**



**2.3.2.3 SVM classifier**

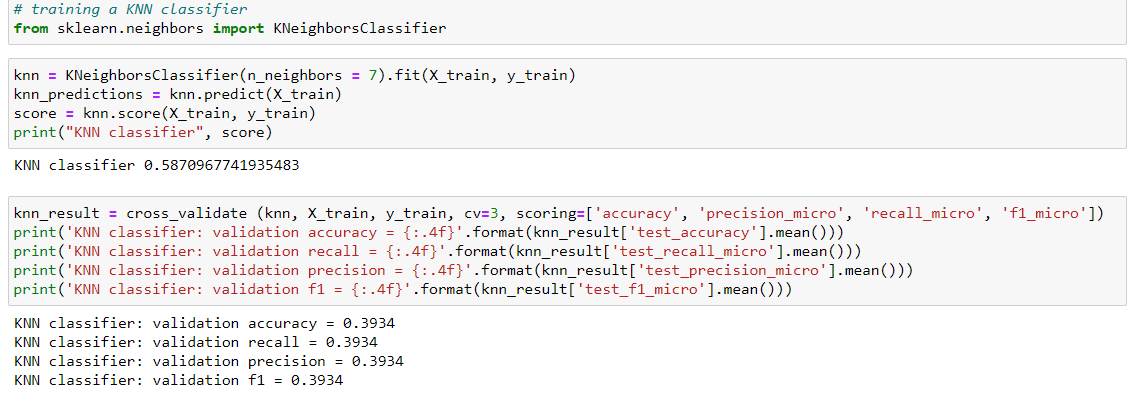
A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.



**2.3.2.4 KNN Classifier**

A k-nearest-neighbor algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

The k-nearest-neighbor is an example of a "lazy learner" algorithm, meaning that it does not build a model using the training set until a query of the data set is performed.



**2.3.3 Cross Validation**

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation

**2.3.3.1 Performance measure: Accuracy, recall, precision and F1 score**

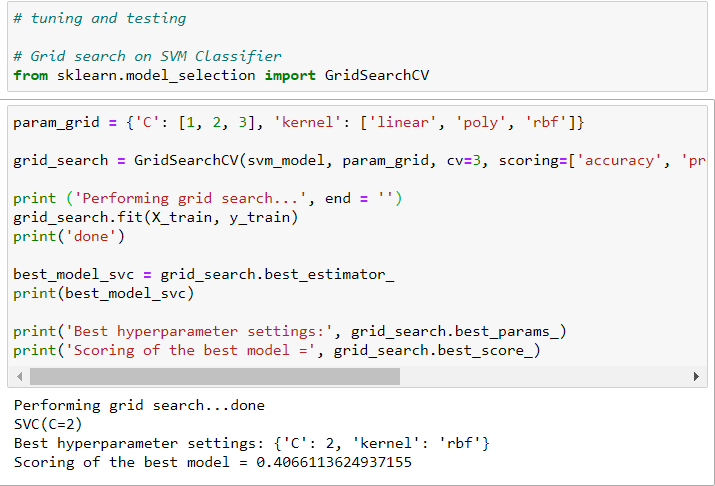
For all the training, we choose cross validation to get the validation accuracy, validation recall, validation precision and validation f1.

1. **Accuracy:** the most intuitive performance measure, and defined as the ration of the number of correctly classified objects to the total number of objects evaluated.
2. **Precision:** it is simply ratio of correctly predicted positive data objects to the total predicted positive data object.
3. **Recall:** it is defined by the number of correct positive results divided by the total number of relevant samples (all samples that should have been identified as positive).
4. **F1-score:** combines precision and recall into a single metric. It is the harmonic mean of the two measures.

**2.3.4 Tuning and Testing**

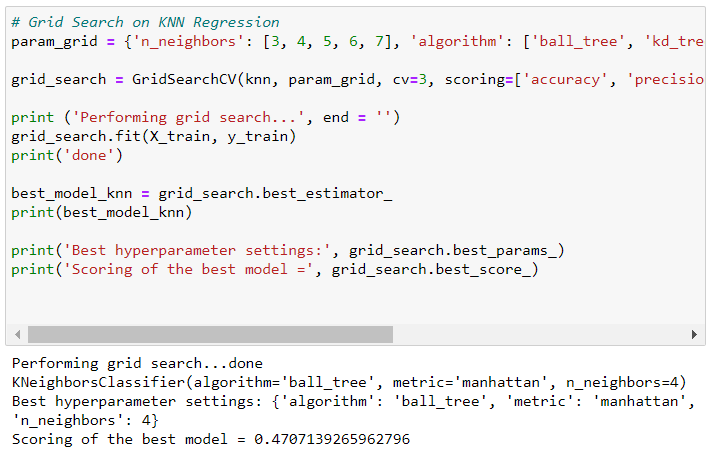
In tuning and testing step, we choose 2 highest validation accuracy after did the training algorithm. Based on the training result, we choose SVC classifier and KNN Classifier as our grid search.

**SVM Grid Search**



We get the best scoring of the model in SCV is 0.4066

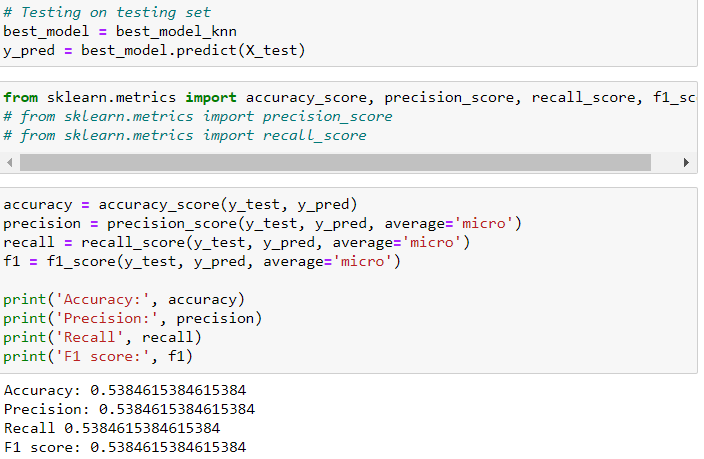
**KNN Grid Search**



We get the best scoring of the model in KNN is 0.4707

After compare the scoring of the model in SVM and KNN we decided to choose KNN classifier as our testing set because KNN give us the most scoring of the model.

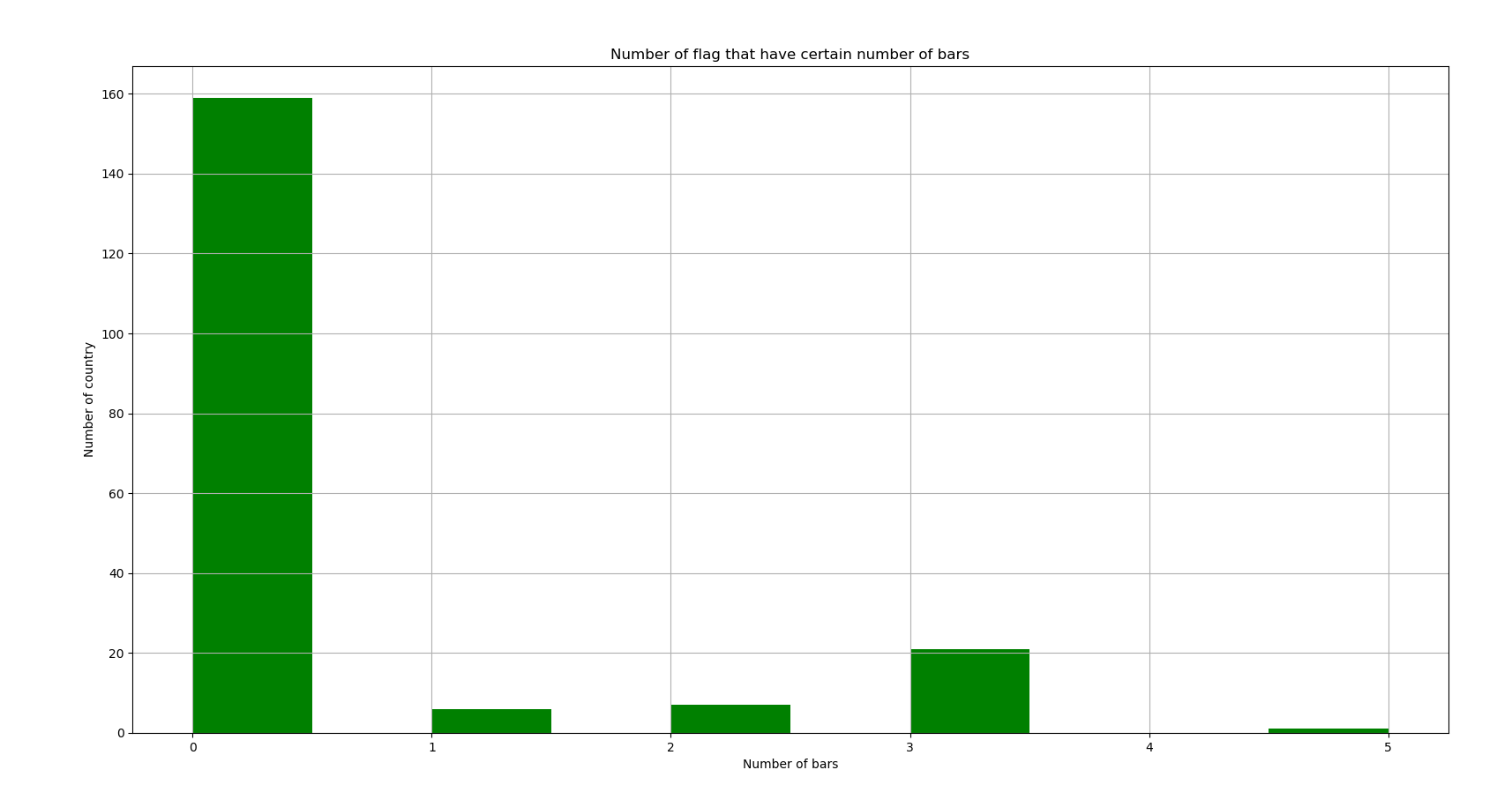
**Testing algorithm**

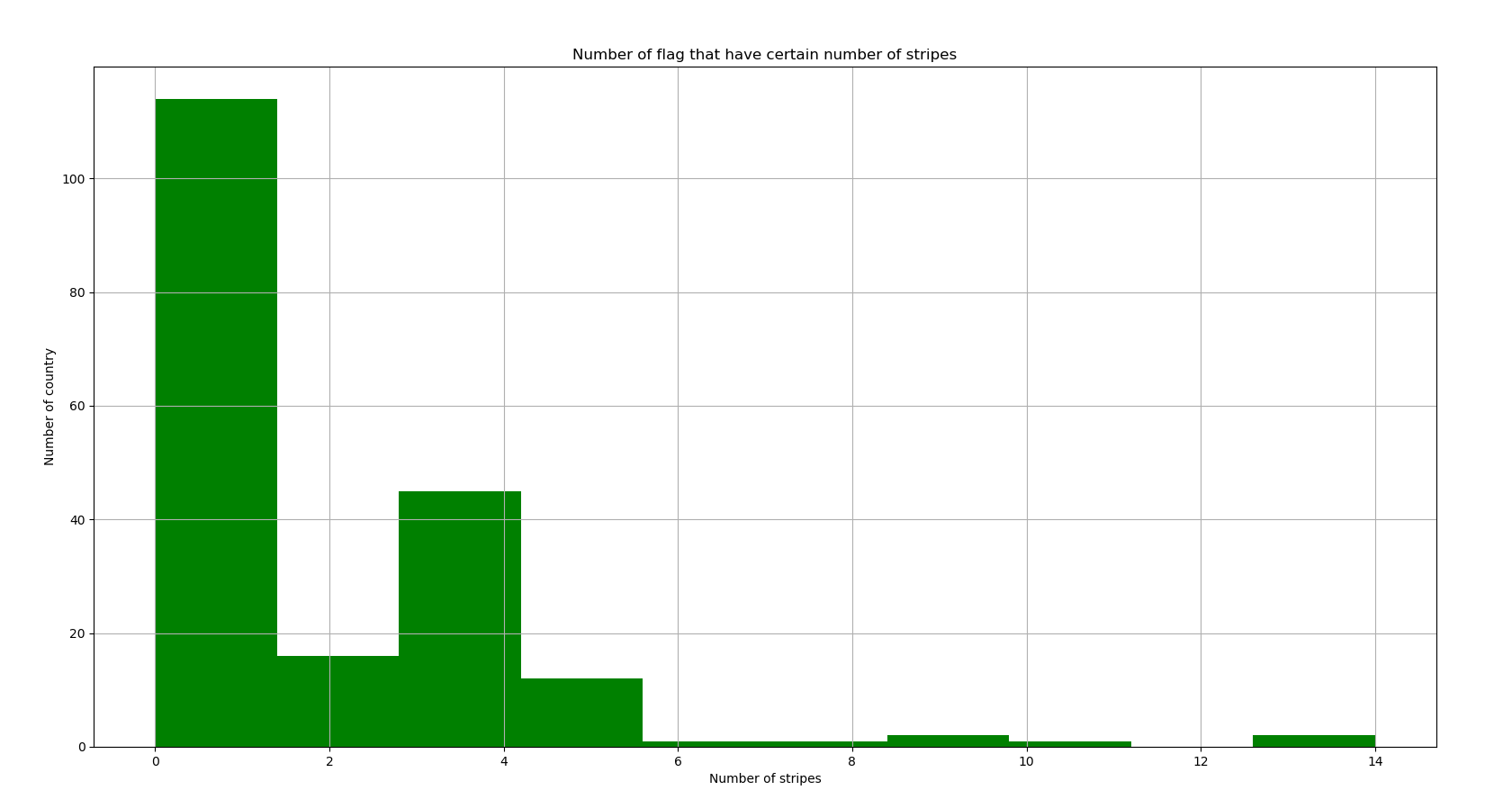


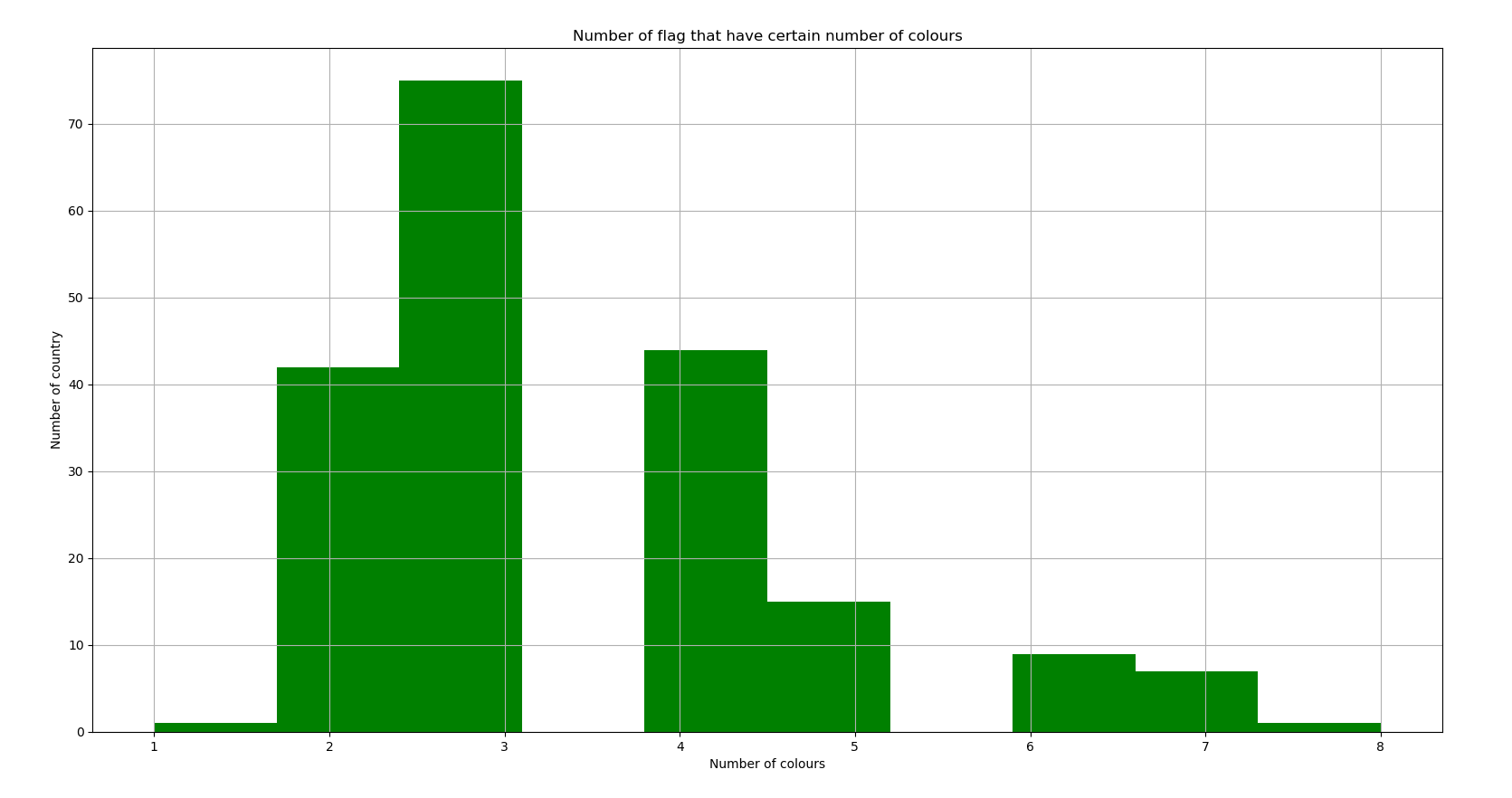
**Chapter 3: Result and Discussion**

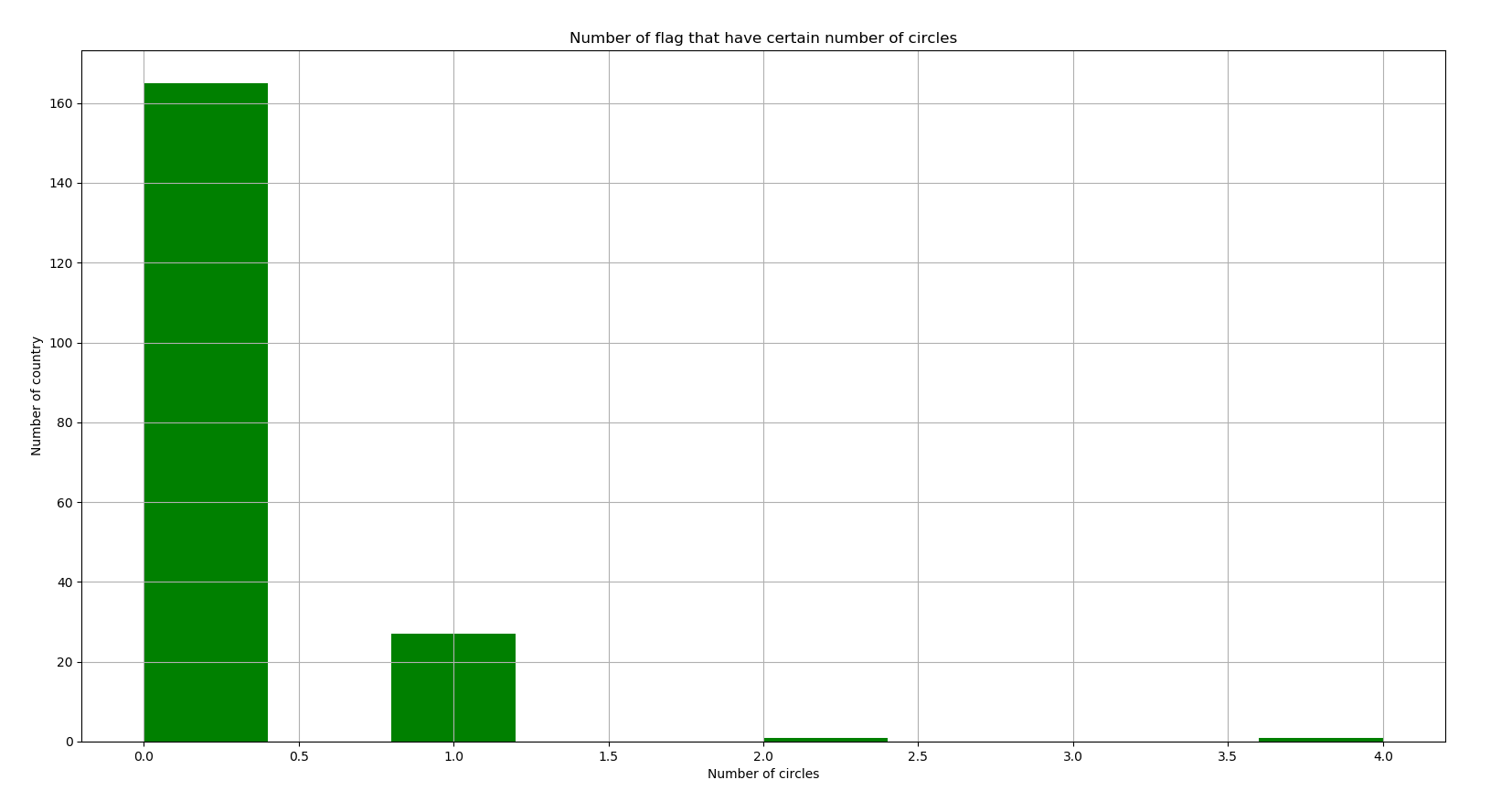
**3.1 Data Preprocessing**

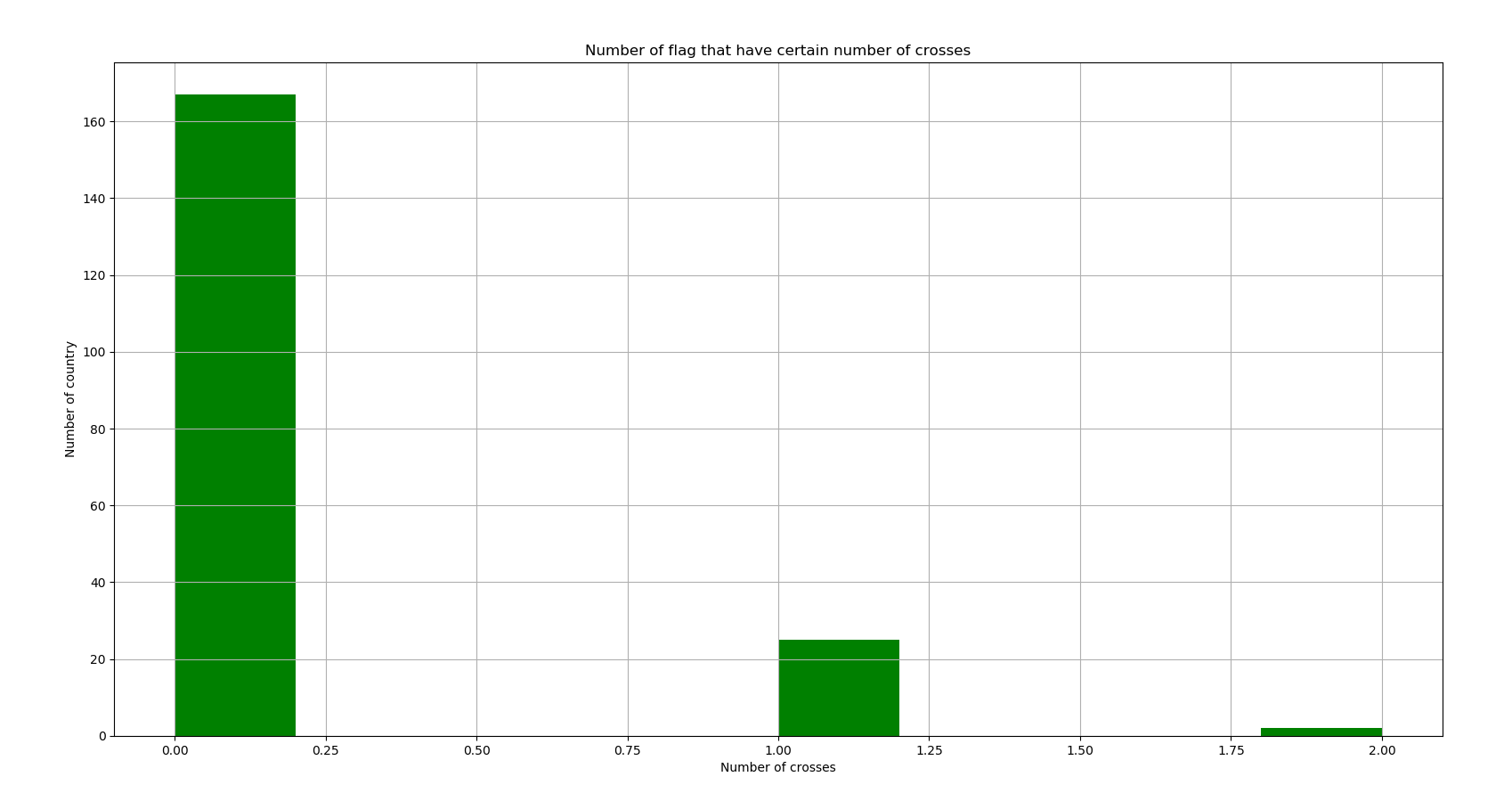
The dataset has been split into 80/20 of training and testing set, which means 80% of the data is under training set and 20% of data is under testing set. By the total 194 of country, 155 country is categorized under training set. The missing value had been checks before proceeds to training step. Various graphs had been plotted to show number of flags that have certain feature, such as bar, stripes, colour, circle, crosses, saltires, quarters and sunstars.

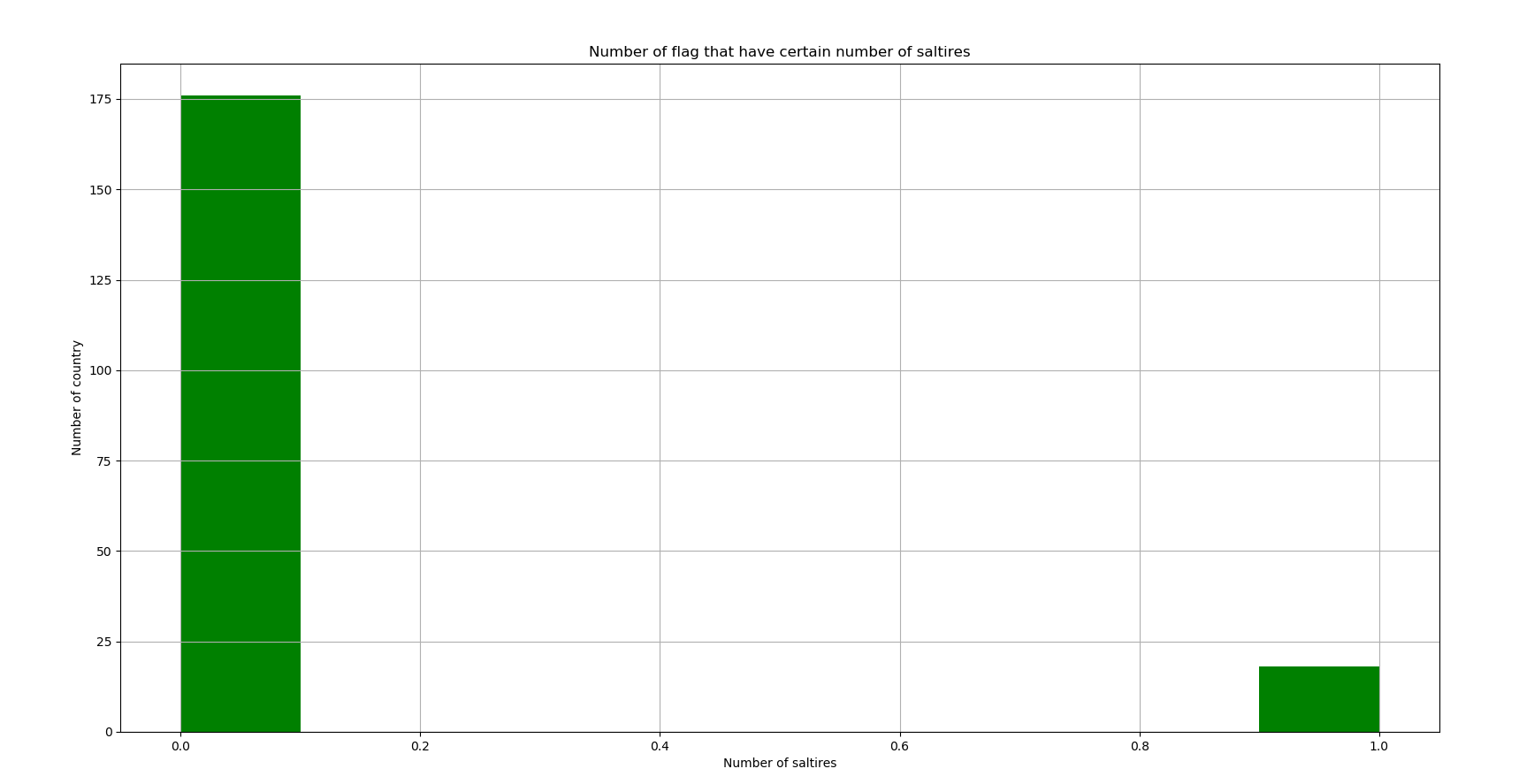


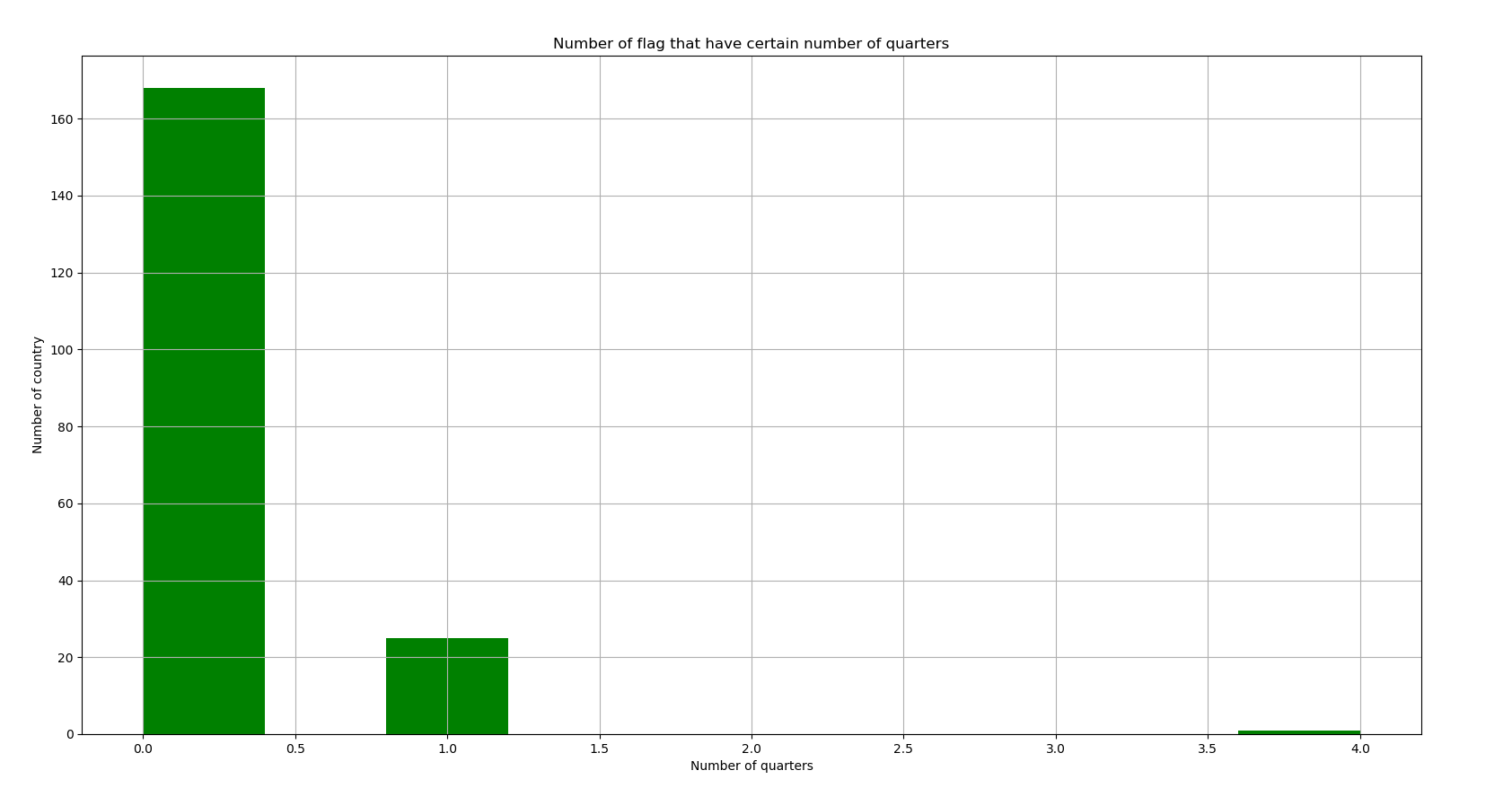


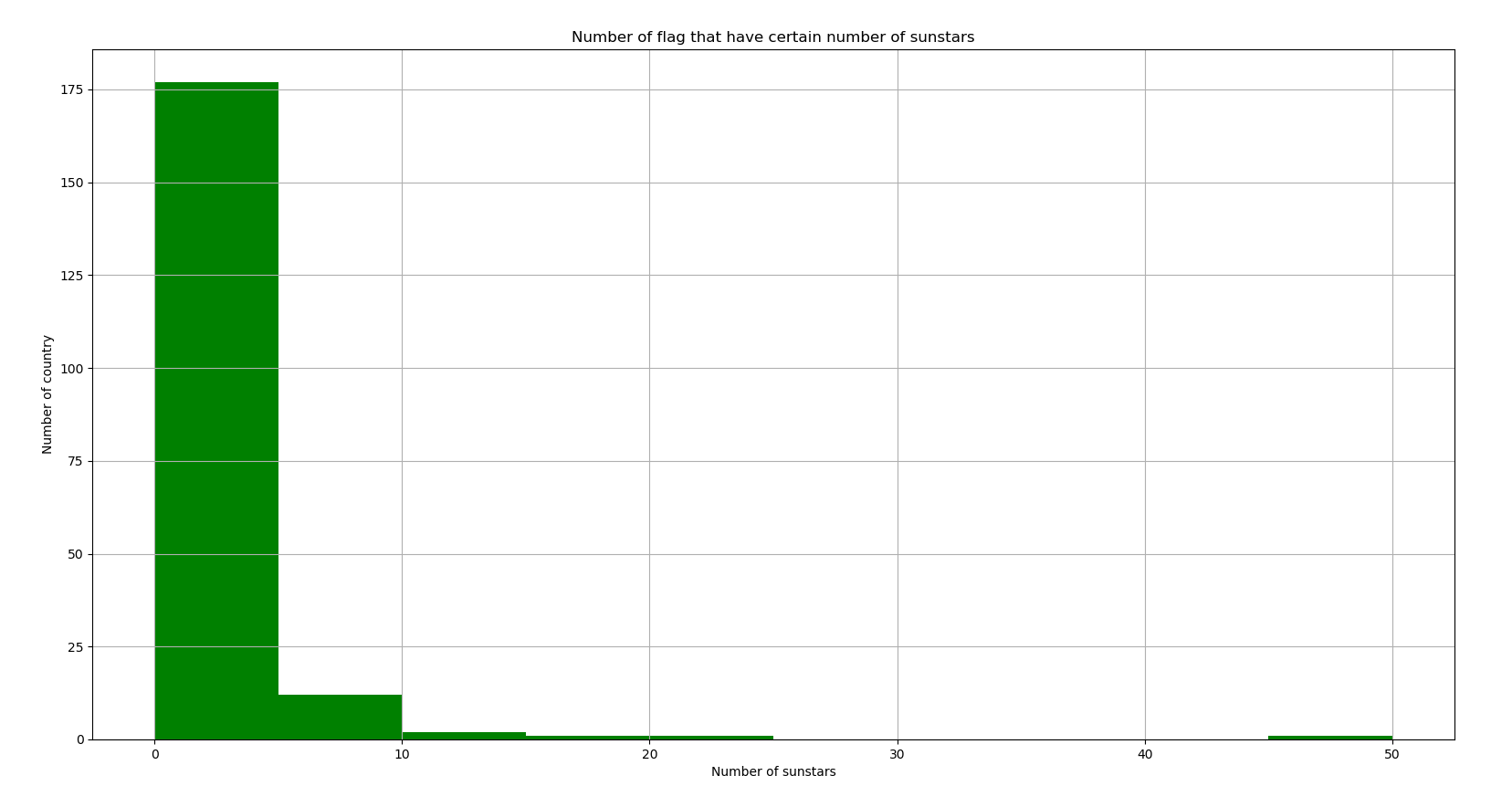






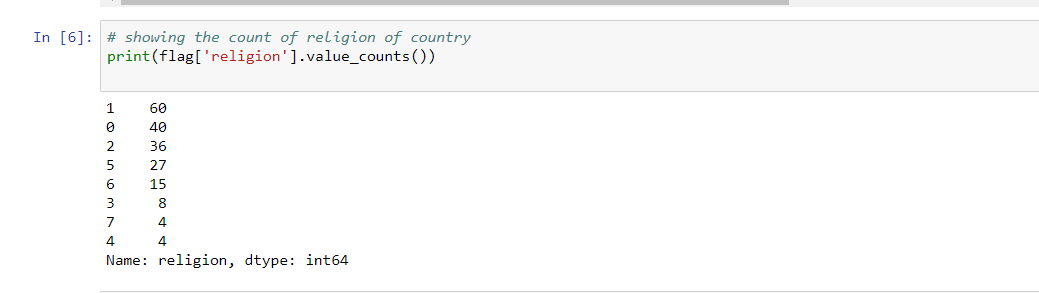






**3.2 Performance Analysis**

There are 4 algorithms had been used to train the model, which are Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbour Algorithm. All of these algorithms are used to perform multiclass classification. The performance metrics that selected to measure the performance of these model are accuracy, recall, precision and f1 score. The models are trained and evaluated using cross validation where the dataset is split into number of k set. Then it will train the model using different set of data each time and calculate the score of every dataset. The average score of it will become the final score for evaluation of the model. Due to the nature of multiclass classification, the score that are used to evaluate the model is “Micro Average”. It means that the metric will aggregate the contributions of all classes to compute the average metric. Micro average is preferable over macro average in this case, since there is class imbalance where the number of sample class is different in different class.

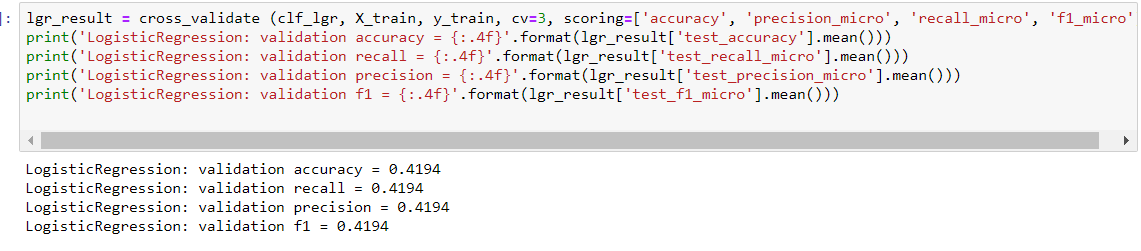


(Class imbalance exists in the dataset)

**Logistic Regression**

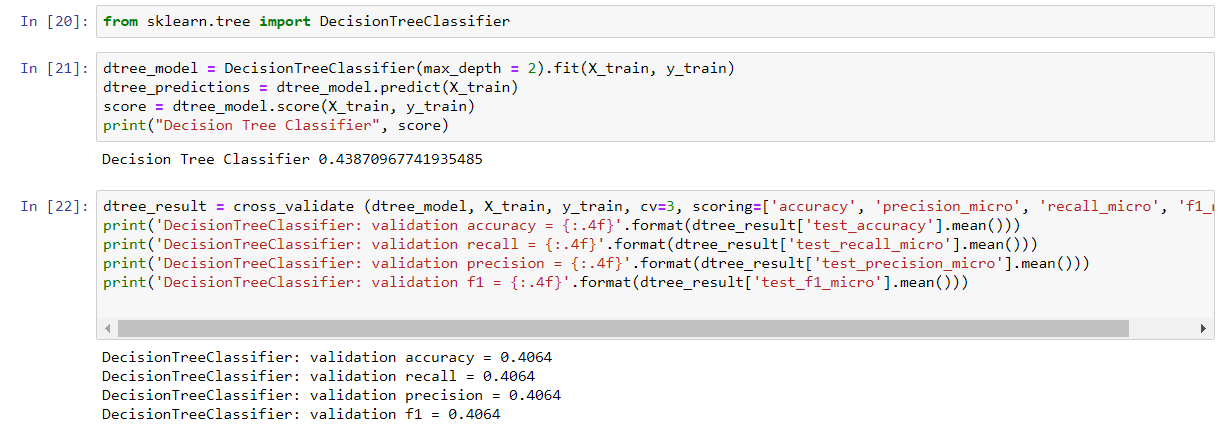
The accuracy score given by training set using logistic regression is 0.5742. When it is tested with cross validation, we can see that the average accuracy score is only 0.4194. It is the same for recall, precision and f1 score where all the metrics score is 0.4194.





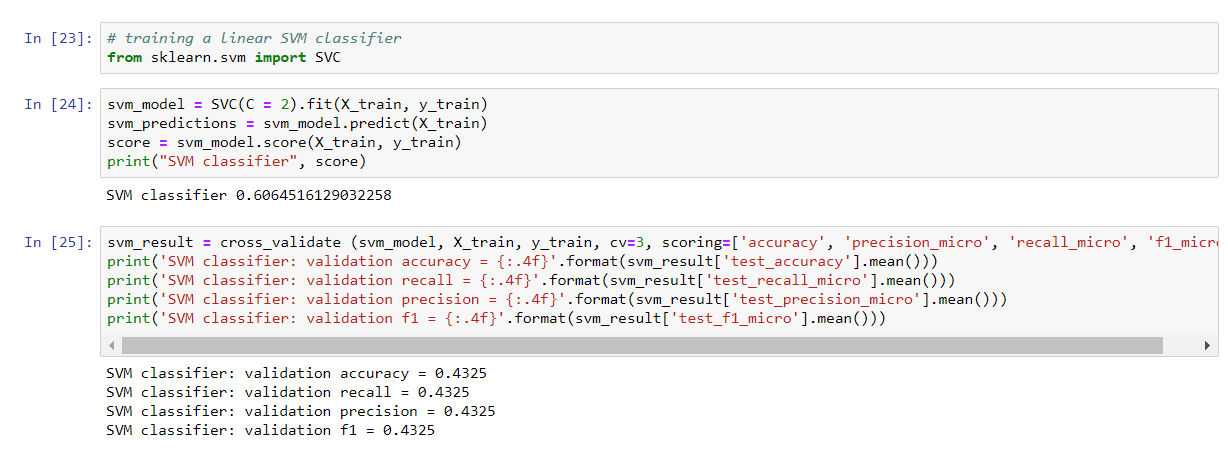
**Decision Tree Classifier**

In the case of Decision Tree Classifier, the score return from predicting training set is 0.4387, which is the lowest score among all the algorithm. In the cross validation, the decision tree model has achieved 0.4064 accuracy. The recall, precision and f1 score is also 0.4064. It is the worst in term of performance among all the algorithm model.



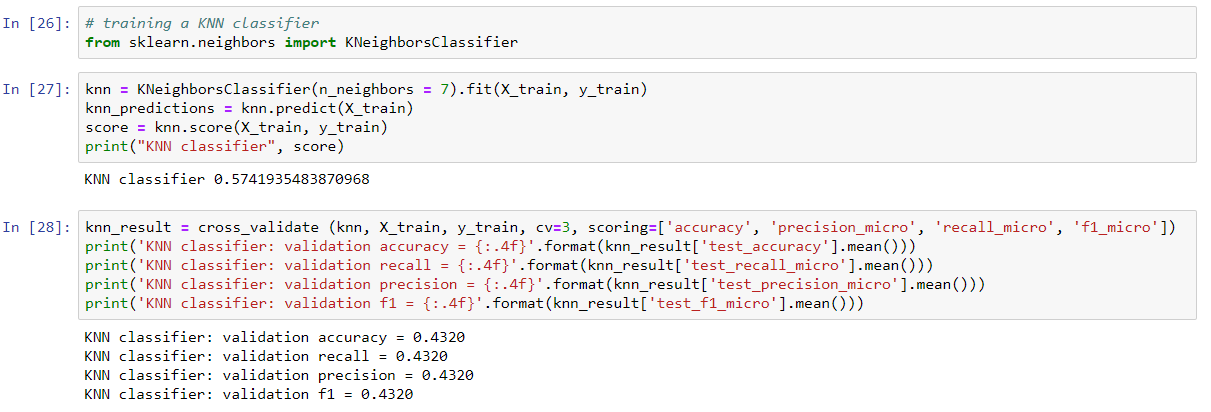
**Support Vector Machine Classifier (SVC)**

Support Vector Classifier is getting 0.6065 accuracy score when predicting training set, which is the highest accuracy score among all the model in predicting training set. However, the average accuracy in cross validation testing is only 0.4325 for SVC. It is still the best average accuracy score among all the model. For recall, precision and f1 score, it remains the same which is 0.4325 for SVC model.

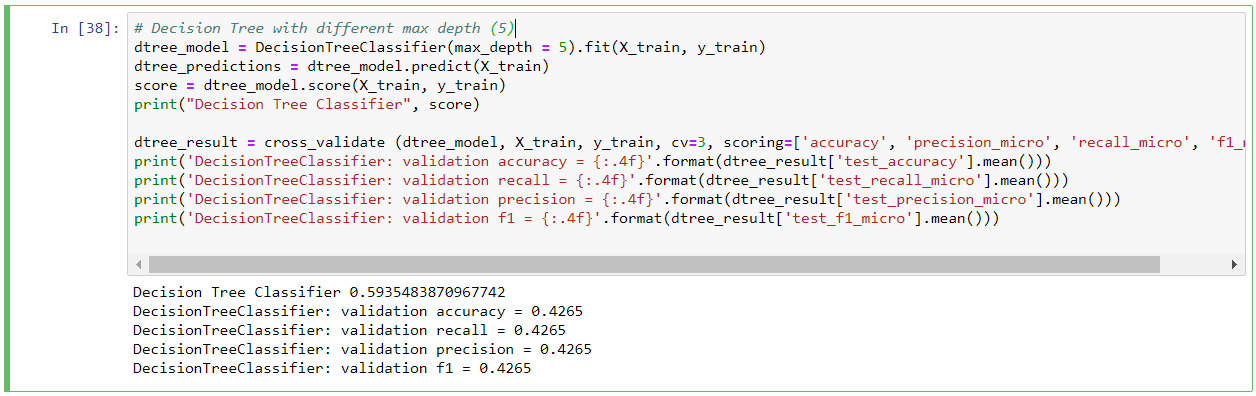


**K-Nearest Neighbors Classifier (KNN)**

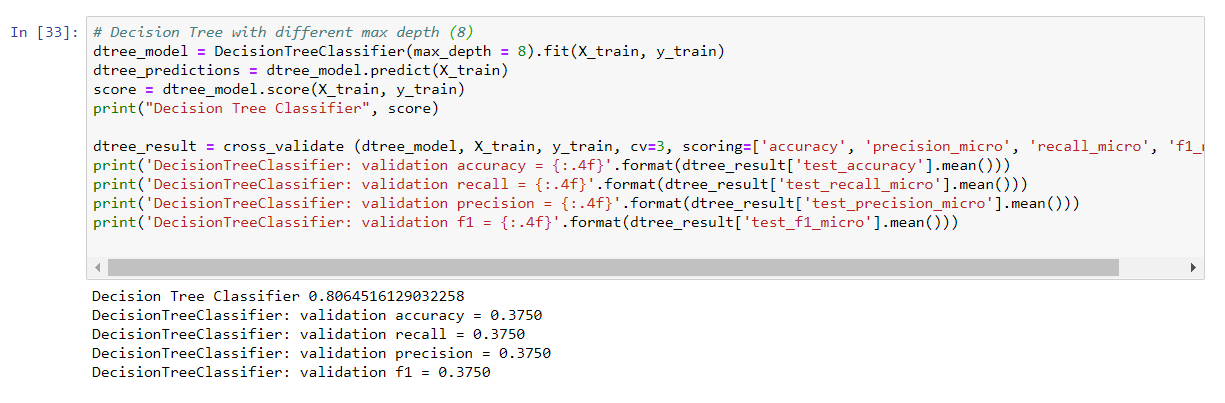
With only default value of KNN setting, the accuracy of training set is 0.5742. In the cross-validation testing, the average accuracy is 0.432, slightly worse than SVC, but still the second best among the other model. The score 0.432 also has been attained in recall, precision and f1 score.



From the result obtained, we know that Decision Tree Classifier has the worst result. As the Decision Tree above only have max\_depth = 2, this might be caused by underfitting, thus having the worst result among the models. However, even when we found the best fit of max\_depth, which is max\_depth = 5 in this case. The result is still not quite statisfying as other model. Its score is 0.5935, while cross validation accuracy is 0.4265, same for recall, precision and f1 score. It seems like it is quite high, but the cross validation score is still lower than SVM classifier and KNN classifier. It may seem to have a overfitting issue with higher max\_depth. For example, max\_depth = 8, even the score is 0.8065. However, the cross validation accuracy, recall, percision and f1 score is only 0.375. It is clearly an overfitting issue.



(Decision Tree with max\_depth = 5)



(Decision Tree with max\_depth = 8)

Hence, the best 2 model will only be selected to proceed to the next step, which is tuning and testing on testing set. The best 2 model from above result is SVM Classifier and KNN classifier. Both models have cross-validation score of 0.4325 and 0.432 respectively.

**3.3 Tuning with Grid Search**

Grid search pair every combination of the parameter to find out the best parameter for the model that gives the best result.

SVM Classifier

The parameter selected to test with SVM Classifier is C (Regularization parameter) and kernel type is used in the algorithm. From the result, we know that the best hyperparameter setting is C = 2, kernel = ‘rbf’. It gives the best result with score 0.4325. It is not much different before and after grid search tuning.



(Grid Search Result with SVM Classifier)

**KNN Classifier**

The parameters selected to use grid search tuning for KNN Classifier are n\_neighbor (number of neighbour used in KNN), algorithm (algorithm used to compute nearest neighbour) and metric (metrics that used to calculate the distance). From the result, it can be clearly seen that algorithm = ‘ball\_tree’, metric = ‘manhattan’ and n\_neighbor = 4 gives the best combination. The score for this model is 0.471 which is a better model than SVM Classifier. Hence, the best model is KNN Classifier. It is time to proceed to test with the testing set.



(Grid Search Result with KNN Classifier)

**3.4 Testing on testing set with best models**

Since KNN Classifier gives the best result among all other model, so the model is now selected to test with testing set. The metric used to measure performance is still the same to measure training set, which are the micro average of accuracy, precision, recall and f1 score. This time, the score we get for the test is 0.6154, which are slightly better than the training result.

****

**Conclusion**

In today’s world, machine learning and data classification is indispensable. It benefits plenty of sectors and eases our daily life. As beneficial as it may sound, there are some risks to it too. If the accuracy is not good enough, it will undermine the decision making process and produce a wrong result which will cause a lot of errors and disasters. Therefore, accuracy is crucial for machine learning process. Choosing the right algorithm and method is very important.

A national flag is not only a representation of a country but its characteristics and features conceive the cultures, ideals, histories and also religion. This project investigates the correlation between national flag characteristics with religion of countries. We used different algorithm such as k-nearest neighbours(KNN), Support Vector Machine(SVM), Logistic Regression and also decision tree. Based on our result of classification, we found out that KNN have better accuracy than the rest. In this classification, a total of 25 attribute from 194 countries after removing insignificant data. The model with the best accuracy is KNN which is 53.85%.

As a result, we have learned that there are many factors that affect the outcome and there are close correlations between these factors when we take certain characteristics of countries. The model can be more effective if there is a greater sample size and better algorithm. One approach we can take is to reduce the categories and features to a point where there are more obvious identifiers.