COMP5310_Assignment Project Stage 2

Name	Yeseul Yoon	**E-mail**	yyoo0548@uni.sydney.edu.au
Unikey	yyoo0548	**Student ID**	490241543
Due Date	24/05/2019	**Submission Date**	24/05/2019

Set up

Every country has different figures in fields of land size, economy, population, penetration rate of the electric devices, industrial structure, and so on. We all may have heard about related reports or articles, and I wondered if this factor(s) could make any correlation to determine the region and if so, which figure(s) is relevant the most.

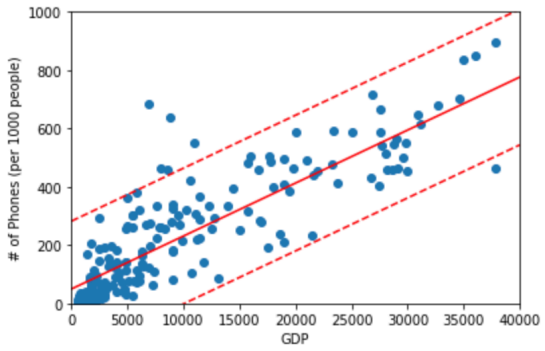
Eventually, my research question for this assignment is :

Can we predict the region with acceptable reliability using the country's economic state and the population change?

And my hypotheses for the research question are :

- Null hypothesis : Logistic Regression and Decision Tree have same error rate to predict the region.
- Alternative hypothesis : Logistic Regression and Decision Tree have different error rate to predict the region.

My chosen data set has various elements such as - country name, region, population, population density, area, net migration, infant mortality, birth/death rate, GDP(Gross Domestic Product), and penetration rate of phones. For representing the financial state, I used GDP and penetration rate of phones. It is a common sense that GDP is so different depends on regions that I chose GDP value as a 'base' in my project. In other words, for reliable prediction of the region with numerical data from various fields, I need certain elements that could be classifiers such as GDP and few others which are related to GDP. Moreover, according to the test, I figured out there is a linear relationship between them and this relationship could help to enhance the accuracy of the final test.



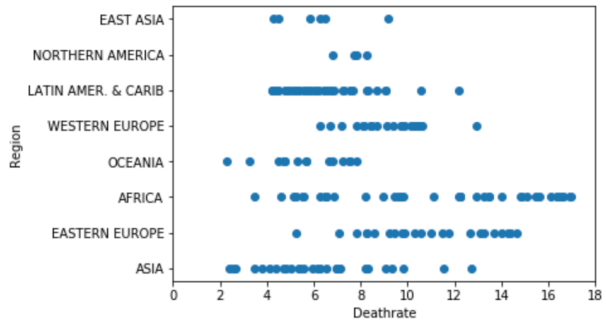

```
-------------------------------------------------
Statistic            From Scipy     From Scratch
-------------------------------------------------
alpha                     47.91            47.92
beta                       0.02             0.02
r-squared                  0.72             0.72
-------------------------------------------------

Correlation from scipy: 0.849
Correlation from scratch: 0.849
```

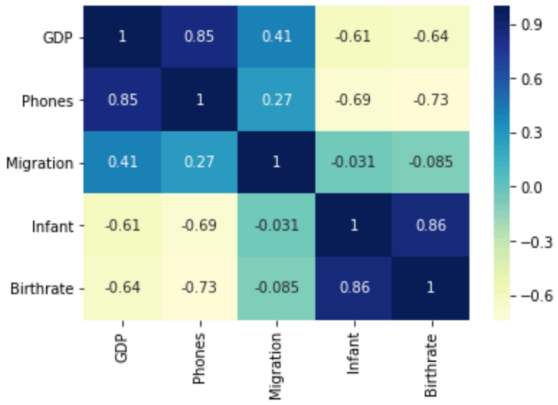
From this test and scatter plot chart, It is clear that plots seem following the linear line(—) or almost every plots are located between two other linear line(- - -) which are calculated with standard error and the 95% prediction interval. Furthermore, the correlation value (approximately 0.849) is large enough to demonstrate that there is a strong relationship between the two values.

I also did a test for other elements and found that population and area have an extremely weak correlation with GDP as -0.034 and 0.068, respectively. On the other hand, the correlation value between GDP and Infant mortality (per 1000 births), Birth rate, Literacy, and Net migration was -0.614, -0.644, 0.517, and 0.413, respectively, and it is much higher than previous two results. In addition, I decided to add Death rate

data as one of the parameters to determine regions because I found that they have different distributions by groups, and it helped to classify the regions with higher accuracy. This scatter graph below enabled me to understand how values are distributed by regions.

The heat map below shows the correlation value between GDP and other components. I could figure out the component that has the highest correlation with GDP is Phones, Birthrate as the second highest and Net migration has the lowest correlation.



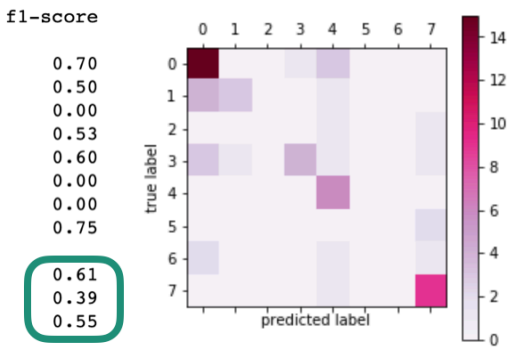


Approach

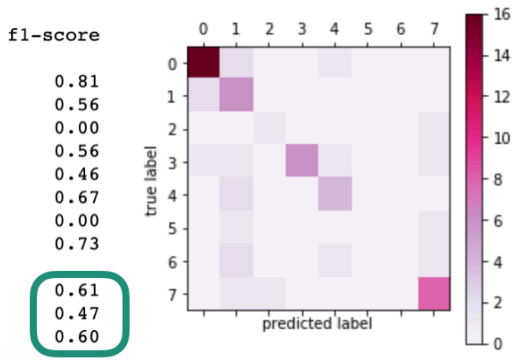
In this project, I compare two models, logistic regression and decision tree.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary and used for prediction. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. (https://www.statisticssolutions.com/what-is-logistic-regression/) Indeed, since the prediction from my data set is expected more than two outcomes, multinomial logistic regression could work better to classification. On the other hand, the decision tree identifies, classifies, and predicts the relationship between the data and the targets for continuous data. It divides the data into features and labels to form conditions that can be created using those features. This model is relatively easy to read and more accurate compared to other machine learning models. Also, it is usually applied to the classification problem.

First of all, I randomly divided my data into the training set and test set using [train_test_split] function as three parts of the test set to seven of the training set. Eventually, I have got 142 value as test data and 61 value as training data. Next, I fit the models for both logistic regression and decision tree and calculated the accuracy of them without any tuning. Two models have a quite similar accuracy (nearly 61%) and also f-1 score as we can see from the following classification reports below.



Logistic Regression



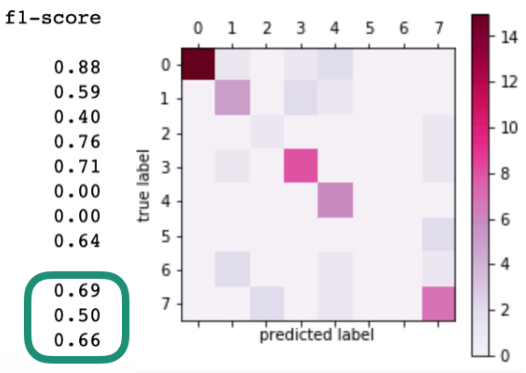
Decision Tree

The classification report includes the precision, recall, F1, and support scores for the model and three averages (micro, macro, weighted). The report shows a representation of the main classification metrics on a per-class basis. Precision is the ability of a classifier not to label an instance positive that is actually negative, recall is the ability of a classifier to find all positive instances, and the f1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0 (https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html). Micro average which represents averaging the total true positives, false negatives and false positives is only shown for multi-label or multi-class with a subset of classes.

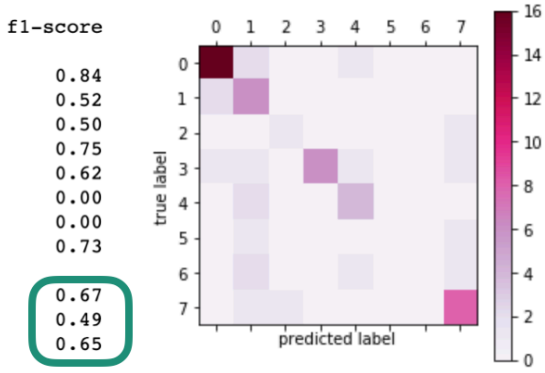
The confusion matrix is used to compare the quality of the output or result. The diagonal elements represent the number of points for which the predicted label is equal to the true label (https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html). From this matrix, we can clearly catch the differences between regions (0=AFRICA, 1=ASIA, 2=EAST ASIA, 3=EASTERN EUROPE, 4=LATIN AMER. & CARIB, 5=NORTHERN AMERICA, 6=OCEANIA, 7=WESTERN EUROPE) and also two different models.

For making much more efficient and reliable results from models, I did the parameter selection tests. For logistic regression model, I used [GridSearchCV] with parameter grid dictionary, and on the other hand, for the decision tree, I used [DecisionTreeClassifier] with 'gini' and 'entropy' to measure impurity. Gini is a criterion to minimize the probability of misclassification, and it is maximal when the classes are entirely mixed (https://www.bogotobogo.com/python/scikit-learn/scikt_machine_learning_Decision_Tree_Learning_Informatioin_Gain_IG_Impurity_Entropy_Gini_Classification_Error.php). Entropy is a way to measure impurity, and we can more successfully classify the data with lower entropy of the classification condition.

According to the Figure 1, I chose 5 as a minimum leaf size parameter because it would surely increase the accuracy of my further new model. After changing the parameter for decision tree, I got new accuracy of test set as 67% and result proves that my new model is more reliable now compared to the previous one. Furthermore, the accuracy of my logistic regression model fitted with the best parameters has also increased to 69%. The partial part of conditional reports and the confusion matrix graphs below demonstrate how the figures have changed with best parameters.



Logistic Regression



Decision Tree

Result

- Summary of the result

From all the test I've done until now, I realised that it is possible to predict regions using both linear regression and decision tree with around 69% and 67% of reliability, respectively. In the beginning, accuracy rates were both around 61%, and they became higher after applying the best parameters. Next, I did McNemar test to check if these two models have the same error rate. What I found from the result is only a little difference exist between them. Therefore, I cannot reject my null hypothesis because the p-value is too, and there is no strong evidence that error rates are significantly different.

```
[1 1 1 1 0 0 1 1 1 1 0 1 1 1 0 1 1 1 0 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0
 1 1 1 1 0 1 1 1 1 0 0 1 1 1 0 1 0 0 0 1 1 1 1 0]
[1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 0 1 0 1 0 1 0 0 1 0 0 1 1
 1 1 1 1 0 1 1 1 1 0 0 0 1 1 0 1 1 0 0 1 1 1 0 1]
('Can we reject H0?', 'No')
```

The arrays above show the result of calculation whether each test prediction is correct. The first array is for logistic regression, and another one is for decision tree. The arrays were quite clear to identify the McNemar test's result.

Then, since decision tree model is generally more efficient and appropriate to classify the categorical data set like mine and handy to interpret the result, I made the decision tree with my data set. According to parameter selection test for my decision tree, I found the best parameter which could help enhancing the reliability. The reliability has increased to 67% from 61% and I successfully illustrated the final decision tree.

- Analysis of the result

The accuracy rates of my prediction of regions were 69% and 67%, which are lower than other highly precise works such as examples from online or textbook. I suppose one of the reasons is due to several countries that have markedly different figures in several features. For example, the average GDP of Africa region is only $2,603, but the GDP of South Africa is $10,700, which is remarkably higher than the average. This kind of trap element has the possibility to decrease the accuracy of the prediction result. Moreover, another reason is the insufficient number of data in groups. East Asia and Northern America have only 6 and 4 countries in their groups, and it also results in the poor accuracy in prediction compared to other regions.

Besides, the Oceania region has even both limitations I mentioned above. Oceania group consists of only 15 countries and countries tend to have a too wide range of values. Specifically, American Samoa and N. Mariana Islands have net migration figures in -20.7 and 9.6, respectively. Also, Oceania's GDP is distributed from $1,600 to $29,000. Eventually, as we can see from the classification report, the accuracy of Oceania remains at 0 in both models. What I would like to do is adding the unique character that could help to distinguish Oceania easily and improve the accuracy expectation.

Conclusion

From this data analysis project, I have learned various things. Firstly, finding out the proper data set and the way to read and clean the data. I think this is not only the basic technique but also one of the most essential steps followed by further next steps. Because if the data set was inappropriate and dirty, I could face the imprecise result in some reasons. Second, the differences logistic regression and decision tree. It is always important to choose the model for machine learning depends on the value in the data set, and I could learn the way to determine the model to fit. Furthermore, I learned the parameter selection is also critical for higher accuracy, and I saw the reliability of the result has grown up when I applied the best parameters. Lastly, one more thing I learned is choosing the appropriate graph for visualisation and demonstrating the results in various types of figure. I have used the scatter plot to illustrate the distribution and used the heat map for values contained in a matrix. At last, I could draw the decision tree which classifies the regions using six characters.

I would like to recommend my decision tree model for my research problem with several reasons. First of all, it has acceptable accuracy, and I believe it could be much higher if I added any other supporting character. Second, even though not all the characters in my data set have strong linear relationships with each other, it does not affect the performance and this is one of the advantages of the decision tree. Lastly, the best advantage is that the decision tree I made is easy to follow and interpret.

Appendix

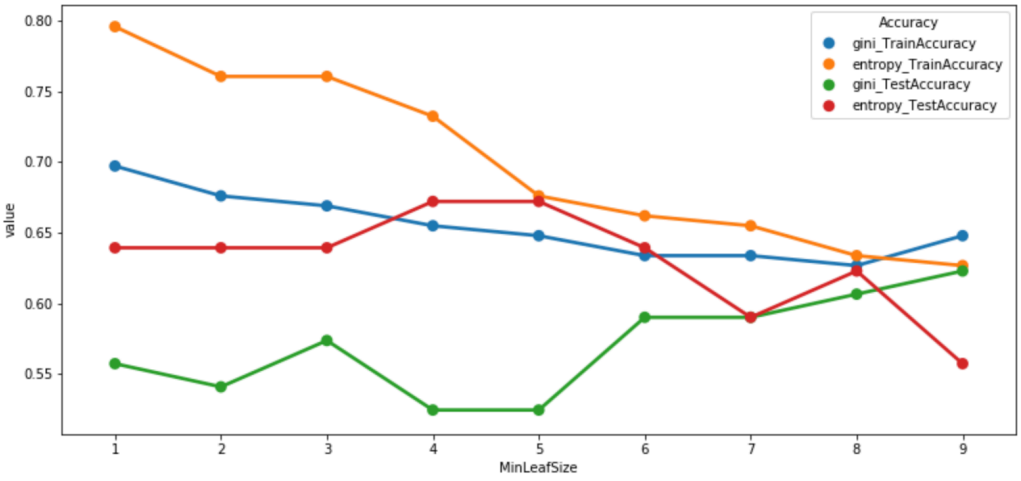


Figure 1. The graph of evaluating the parameter for the decision tree.

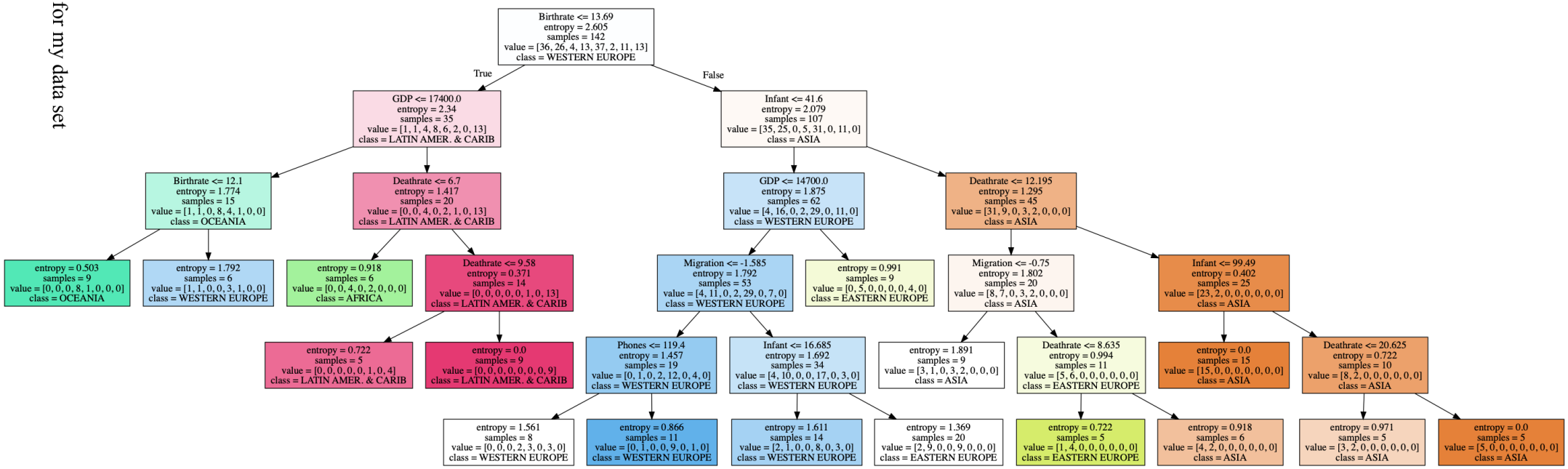


Figure 2. The final decision tree for my data set