Extra Credit: Demographic Composition of Liberia Report

# Introduction

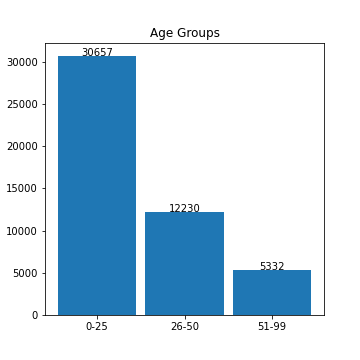
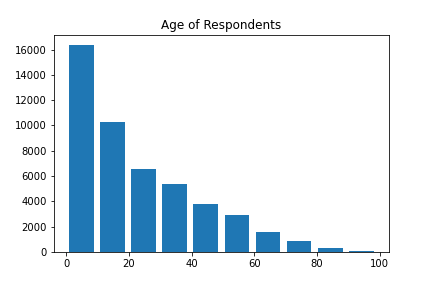
This dataset comprises of 48219 data points that hold demographic of the survey respondents. The names of the variables are location, size, wealth, gender, age and education. The target variable (y) in this project was the education variable which has 6 unique values namely [0,1,2,3,8,9]. Upon doing summary statistics for the target variable it was revealed that the average value for education is 0.66 which suggests that a large amount of the respondents are uneducated.



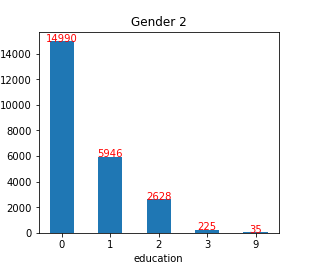
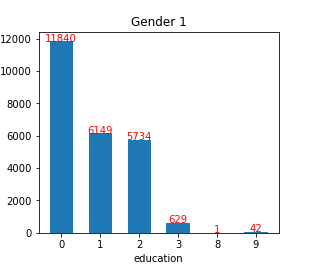
Data points with education label ‘0’ account for more than 55% of the entire dataset, while data points with education label ‘8’ and ‘9’ account for less than 1% of the dataset. This suggests class imbalance which makes it very likely that no classification model will ever reach a high level of classification with the given data. Also, label ‘8’ may have been a transcription error since there was only one data point with that label. It should be noted that the highest level of education in this dataset is 3. Values 8 & 9 refer to someone not knowing their education level/grade.



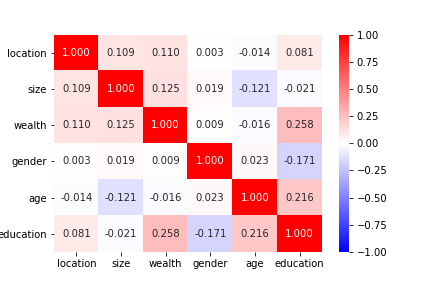
Creating a histogram of the age data revealed that most of the respondents in the survey were young people between the ages of 0-25. This and the fact that more than half the dataset has education label ‘0’ suggests that most of the young people are not very educated which may be a problem in the future.



Analysis on the gender portion of the data revealed that data points with the gender label ‘1’ were much more likely on average to have a higher education than gender ‘2’. The mean education value of gender 1 is 0.81 and gender 2 is 0.51. This variation in education levels across genders is because of gender 1 having larger amounts of people with education levels of 2 (68% of total) and 3 (74%) of total. This suggests that people in gender 1 are the ones that would typically have a higher interest in attaining higher levels of education since those people are more spread across the education levels than gender 2.



The heatmap below didn’t reveal any interesting relationships between the variables as they were all marginally correlated with each other with no variable having a higher correlation value than 0.26. Education correlated with wealth with a value of 0.26 and this was the highest correlation. This was not particularly interesting because this is a typical outcome in most societies.



# Logistic Regression with Scaled and Unscaled Data



Various classification models were trained using unscaled and differently scaled data. The worst performing model was a tie between the Min Max Scaled data and the Robust Scaled data. They each managed to classify 10998 data points correctly with the same test score of 57.02%. The best performing Logistic Regression Model used data that was Normalized. It achieved a test score of 64.65%, training score of 65.02%, and it correctly classified 12470 of the total 19288 test points

# KNN Model



This next classifier was trained using the KNN method. A loop to iterate over the values between 50-61 inclusive was used to find the value of k. This range of values was chosen after trial and error with larger spaced values over a much wider range. Interestingly, all the KNN models with the 5 types of data had the same testing score and the same optimal value of k so to preserve space I decided to only list the results of the unscaled version. This model had a test score of 69.88%, training score of 71.63%, optimal k-value of 52, and 13478 correctly classified points. The training score suggests that the model is slightly overfit but it still outperforms the best Logistic Regression model

# Decision Tree Classifier



This next classifier was trained using the Decision Tree Classifier. For both min\_samples\_split and max\_depth hyperparameters, a loop to iterate over values between 2-21 was used to find the optimal values. The best max\_depth was 7 and the best min\_samples\_split was 13. The test score for this model is 71.81% and it correctly classified 13851 data points.

# Random Forest Classifier



The last model was a Random Forest Classifier. For the n\_estimators hyperparameter, a loop to iterate over the range [50,100,500,1000,2000,3000,4500] was used to find the optimal value. For the max\_depth and min\_samples\_split hyperparameters a loop over the values (2-9) were used to find the optimal value. I only decided to do something that would be so computationally demanding because I wanted to try out the GPU feature in google colab.

For the three hyperparameters manipulated, the optimal values are 8 for max\_depth, 7 for min\_samples\_split , and 2000 for n\_estimators. The test score for this model is 72.23% and it correctly classified 13932 data points. The Random Forest Classifier achieved the highest testing score so this was the best model.

# Conclusion

The best model out of all the trials was the Random Forest. Given the model’s fair test & train scores above 70%, I think that it is likely that this model would be fairly accurate in predicting education levels in Liberia if used on another dataset.