Tools Used and Why

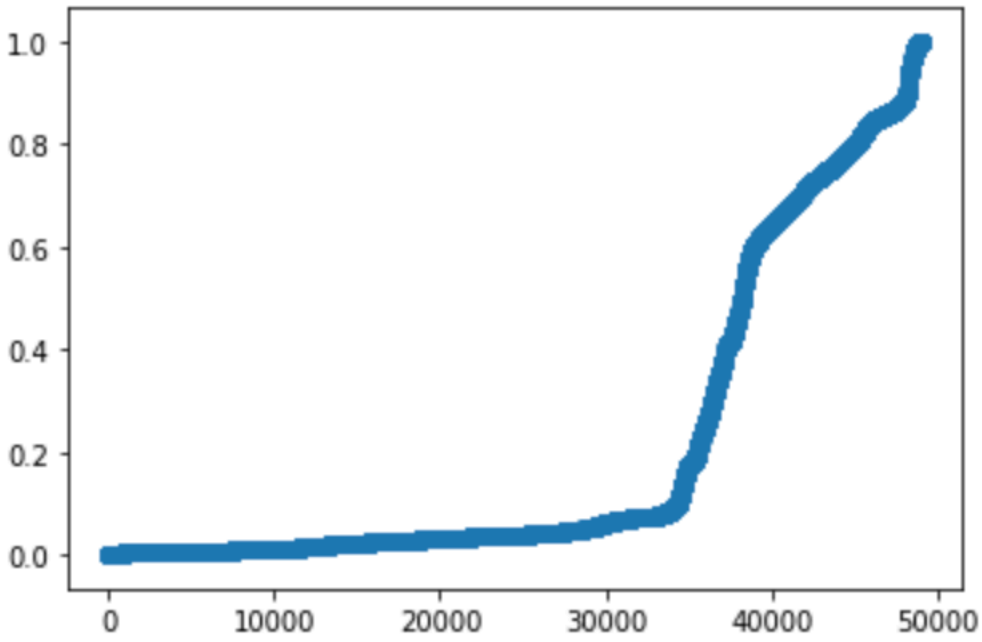
For this assignment, I used the "LogisticRegression" classifier from the class "sklearn.linear_model". I used this because it allows you to implement logistic regression on a dataset. I then used the "fit" method to the regressor object that I created to the different folds of the data set. I used the "predict" and the "predict_proba" methods on the classifier in order to predict the output. The "predict" method was followed by the "accuracy_score" method, applied on both the test and validation sets for each fold, which outputs the accuracies I used in my table below. The "predict_proba" method was used to visualize the probability distribution with a graph, for which I used "pyplot. scatter" from "matplotlib". To break the data into the appropriate training, test, and validation sets I simply followed the example provided to us in the notebook.

Metrics

	ACCURACY	
FOLD	VAL	TEST
1	0.9188961717972914	0.9189178160450502
2	0.9177229717666863	0.9188974128784787
3	0.9197123109490168	0.9187341875459071
4	0.9184370935244459	0.9193462825430507
5	0.9196357979035426	0.9189178160450502
AVG	0.9189627030115073	0.9188808691881967

These metrics were the result of applying the "accuracy_score" method from sklearn to both the test and validation data on each of the five folds of data.

Visualization



I was having trouble with the visualizations. Here is the probability curve for the first fold, I only included it because the other four look very similar. I left some code in the notebook to show some attempts at visualization, but I could not get what I wanted. I knew I could plot the relationship between TWO of the RBG values from the dataset and create a scatter plot with colorized dots to show which class they belonged to, but I could not figure out how to include all THREE of the independent variables without having a 3d graph.

Conclusion

In conclusion, I think I understood the assignment well, other than not being able to get the visualizations that I wanted. I noticed that a good performance on test data does not mean that you will have an equally good performance on the validation data. After averaging the

accuracies together, the validation accuracy was slightly higher than the test data accuracy. This was so close however, that it could just be by chance due to how the data was partitioned

https://github.com/MitchellWilsonTXST/Assignment2_ML