Final Project 4300

Jake Hasenfratz

April 2025

1 Introduction

For this project, I am trying to predict whether a person is likely to consume alcohol frequently or not. This is done using the Drug Consumption (Quantified) dataset from UCI Machine Learning Repository (1). I chose this because I have a number of friends who are on both sides of the spectrum. The frequency is separated by drinking alcohol within the last week or not. Originally the dataset had 6 levels for alcohol consumption; CL0 - Never used, CL1 - Used over a Decade ago, CL2 - Used in the last Decade, CL3 - Used in the last Year, CL4 - Used in the last month, CL5 - Used in the last week, CL6 - Used in the last day. However, I condensed it to just having level 5-6 combined and 0-4 combined. The metrics that I will be focused on for accepting my model will be Accuracy and F1 score.

2 Phase 1

My dataset looks at various levels of usage for drugs/alcohol based on 12 different features. The main features include:

1. Age
2. Gender
3. Education
4. Ethnicity
5. Country
6. Nscore where this denotes Neuroticism
7. Escore where this denotes Extraversion
8. Oscore where this denotes Openness
9. Ascore where this denotes Agreeableness
10. Cscore where this denotes Conscientiousness
11. Impulsive
12. Sensation Seeing

There were not any missing values for the dataset that needed to be fixed.

During the initial phase of exploring my dataset, I plotted the various distributions of all my variables as shown in Figure 1. After that, I took a look at my output variable, which was alcohol consumption to ensure that it was not too skewed one way or the other, which would make the model have a hard time learning and predicting the outputs as shown in Figure 2. Even though the data is slightly skewed towards drinking alcohol at 67% opposed to not drinking alcohol frequently at 33% it wasn't a steep enough skewness to be worried about the model being unable to differentiate between the two. For normalizing my data, I used the MinMaxScaler function from sklearn.

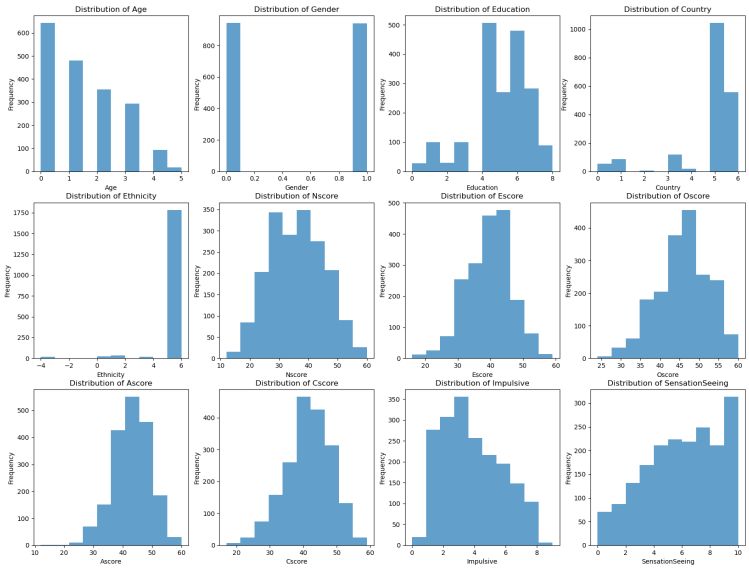

Figure 1: Distribution of objects in Dataset

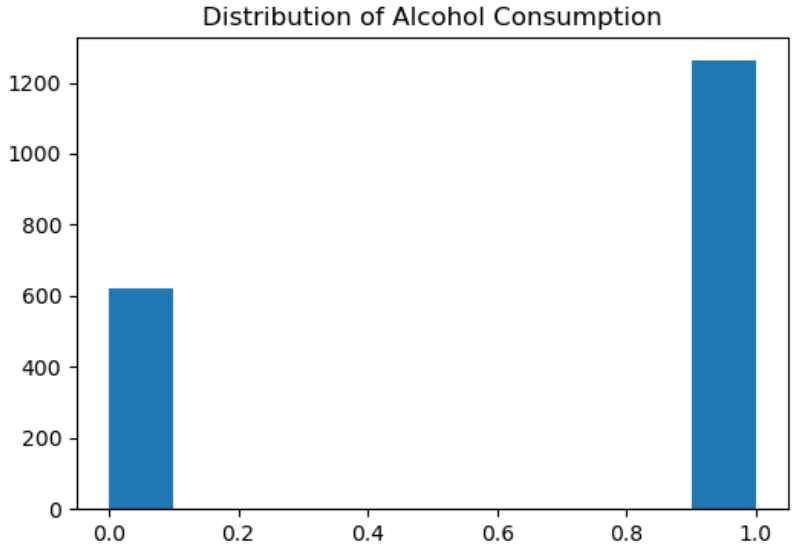

Figure 2: Distribution of drinking Alcohol frequency

Model	Training Set	Validation Set	F1 Score
Baseline	66.7%	69.4%	1
Logistic Regression	67.2%	64.7%	0.784
Neural network model (2-1)	67.9%	65.8%	0.785
Neural network model (4-1)	67.9%	64.7%	0.793
Neural network model (8-1)	68.9%	65.8%	0.793
Neural network model (16-8-1)	71.8%	68.2%	0.771
Neural network model (32-16-8-1)	74.1%	63.5%	0.762
Neural network model (64-32-16-8-1)	85.6%	54.1%	0.712

Table 1: Training and Validation Set performance

3 Phase 3

For Phase 3 I started working towards evaluating my model with all of the variables included in training. The results are shown in Table 1. With that I found that for the validation set the best overall accuracy% was with the 16-8-1 model using 256 epochs. However with the check-pointing the best model was the 32-16-8-1 with 256 epochs which had a loss of .558 and an accuracy of 70.8%. Using the best model that I wanted to check the precision, recall and F1 scores using the sklearn methods and those ended up as the following. Precision: 0.72, Recall: 0.88 and F1 Score: 0.793.

One of the interesting things that happened, as seen in Table 1, was as the training set accuracy got higher it had a big problem with overfitting and underperforming on my validation set. This was due to the model essentially memorizing the data that it was training on and not being able to make reasonable decisions for the correct output.

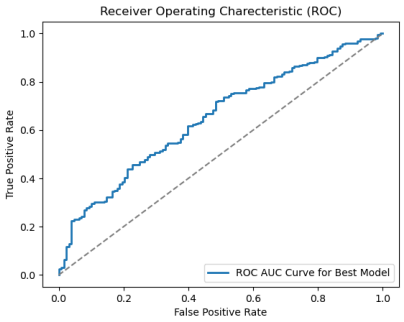

Figure 3: ROC AUC Curve

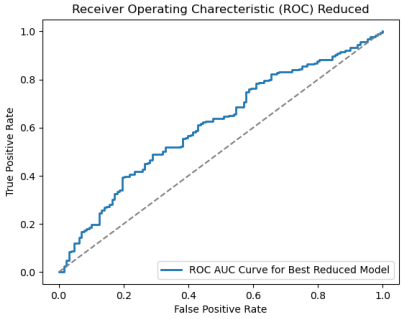

Figure 4: ROC AUC Curve (Reduced Parameters)

4 Phase 4

Ending off with Phase 4 I separated all the variables to find out what had the most impact on training my model and allowing for the most accurate predictions. After running the tests on each individual column, I found there to be almost the exact same Accuracy at 70.6%, Precision at 0.694, Recall at 1.0 and F1 score at .819 which showed me that all of my columns ended up training into predicting 1 for all values which wasn't very helpful. Therefore, for the final model I decided to include just the psychological scores in order to reduce the amount of features that my data needed to learn from.

The final validation I used was to compare the ROC AUC curve for my best model and the best model after reducing the features to just the psychology scores. You can see the original model with all of the features included in Figure 3. After that I removed Age, Gender, Ethnicity and Country. From removing those features I found the ROC AUC Curve to be as shown in Figure 4.

5 Conclusion

To cap off my model I ended up using my best model with 32-16-8-1 size and 256 epochs with only my 4 score features as well as the Impulsiveness and Sensation Seeing. The reason for choosing those features is the model didn't have any stronger of a performance when I included the remaining features and I was able to reach the same conclusion at a quicker speed by omitting them. With my goal of trying to maximize Accuracy and F1 score I ended up with a decent testing accuracy around 71% with my F1 score being .793.

6 Sources

(1) Fehrman, E., Egan, V., & Mirkes, E. (2015). Drug Consumption (Quantified) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TC7S.