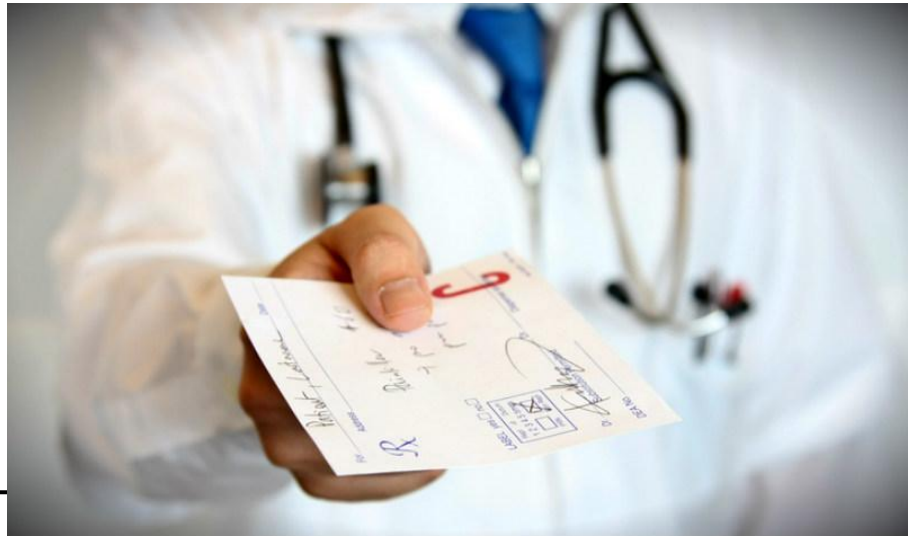
Predicting Drug Use Using Personality Metrics

Amir Edris

Goal of my project

Helping doctors make a more informed decision when assessing the risk vs reward for prescribing drugs that have potential for abuse.



About the dataset

This Data Set is from the uci machine learning repository

- 1885 observations
- 32 columns
- 16 columns of drug consumption data
- 7 columns relating to different measures of personality

Quantifying “High Risk”

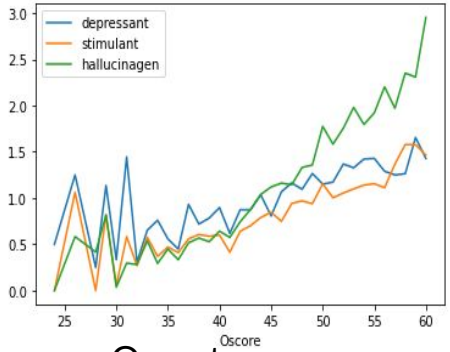
Instead of risk our model is actually going to predict likelihood that a patient has consumed hard drugs in the past few months.

For this to be accurate High risk assumes that you have taken at least 2 of the following drugs

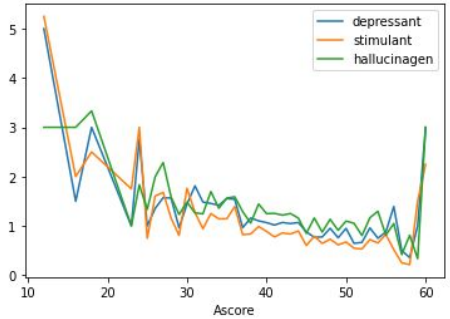
- Cocaine, Crack, Amphetamine, Methamphetamine

Verifying that our tagert captures what we are looking for

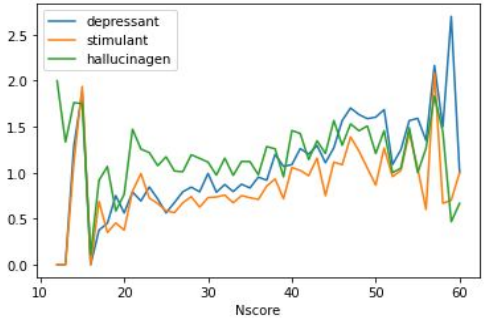
The relationships here make sense for instance negative correlation for drug use and how cautious an individual is.



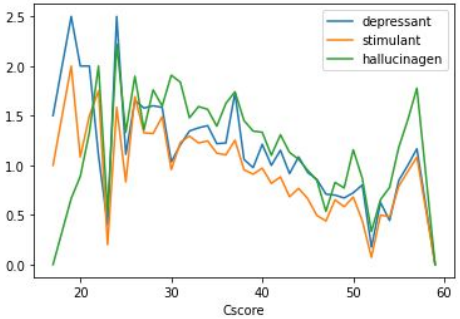
Open to new Experiences



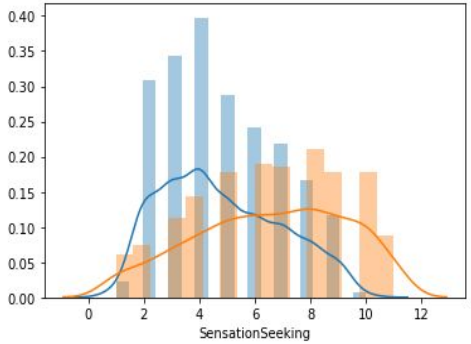
Willingness to Cooperate with others



Level of Anxiety



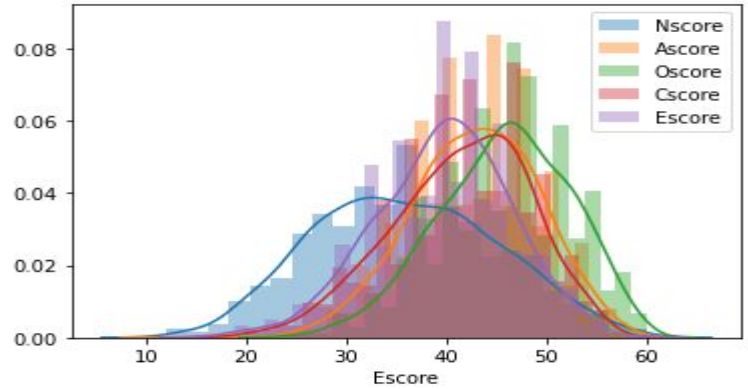
Cautiousness



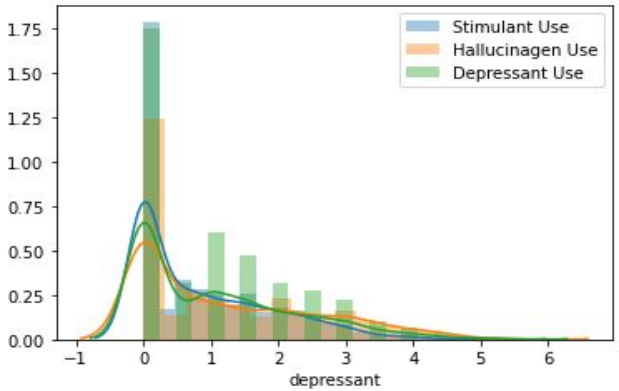
Now that we are ok with our features whats next?

Now we have to address the imbalance in our data

~400 high:1400 low



Distribution by personality scores on the left vs distribution of drug use on the right.



Oversampling

Using SMOTE and Tomec links on my training data to make synthetic data points similar to my positive class.

Final voting classifier was made up of

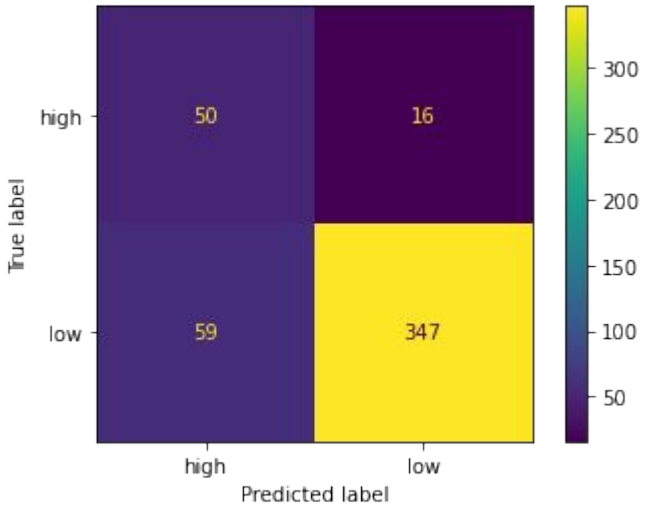
Voting Classifier 1: Gridsearchcv on the range of models we've covered as a group

Voting Classifier 2: Using pipeline degree 3 interaction and selected 43 best features on the same set of models from before

Voting Classifier 3: Using Adaboost on selected models from

The previous 2 classifiers

Is this actually a good model?



It predicted 109 people are high risk

There were 66 people who were actually high risk

And of those 66, 50 were actually high risk

Good progress but work would still need to be done

What are the next steps to improve upon this could improve on this

-More Feature engineering

-Model Parameters aren't the best due to gridsearch and time constraints

-A dataset with more observations and drug use columns would be very beneficial to a project like this

Conclusion

In my opinion this model would be helpful to practitioners as context for the practitioner but instead of patients self evaluating their personality scores there would have to be a more consistent method for measurement. There is a lot of work to be done if you want to use something like this in the medical field but overall.

Thanks for Listening

Any Questions?