Who Hurts Their Liver The Most?
Çağatay Doruk Balcı – Uras Varolgüneş

Dataset:

Survey data about Slovakian students aged between 15-30. Consists of 150 variables, that reflect habits, personality traits and demography.

Objective:

To predict drinking habits of students based on other variables in the dataset and see the relationship between alcohol consumption and other variables.

Methodology:

-To distinguish between heavy drinkers and the others, we label students who drink a lot as 1 and the rest as 0.

-The data set consists of 150 variables, to come up with a parsimonious model, picking out the most distinguishing variables is the key.

-First we conduct explaratory data analysis to discover interesting connections between alcohol consumption and the other variables.

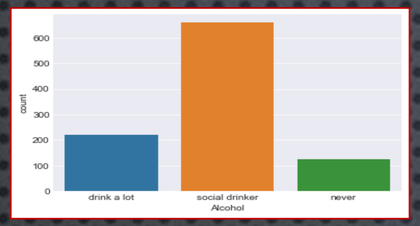
-Then to construct our model, we apply feature selection to the variables which correlate with alcohol consumption the most.

-To solve the classification problem, we apply logistic regression, decision tree and KNN algorithms.

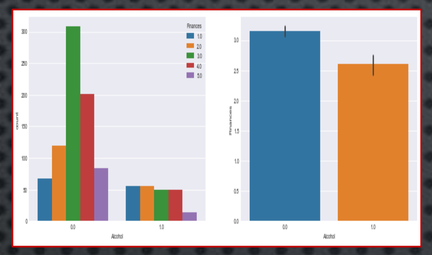
-After the feature selection process, we proceed with hyperparameter tuning to optimize the algorithms.
  -Regularization parameter in Logistic Regression
  -K in K nearest neighbours.

EDA and Important Figures

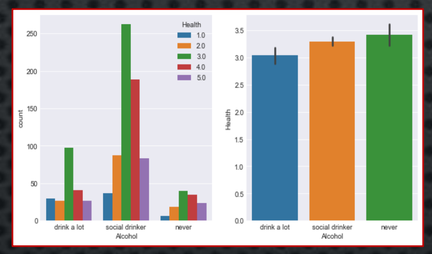
General Distribution of Students



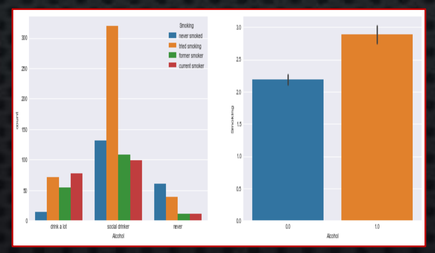
Finances



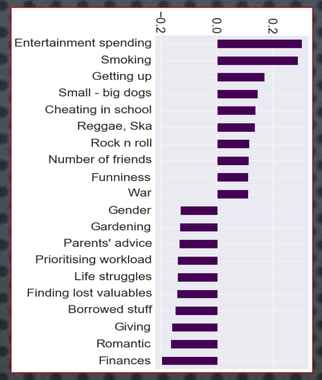
Health



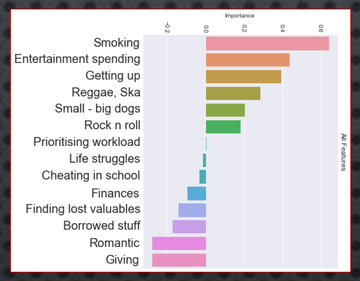
Smoking



Important Features Correlation



Logistic Regression Impacts

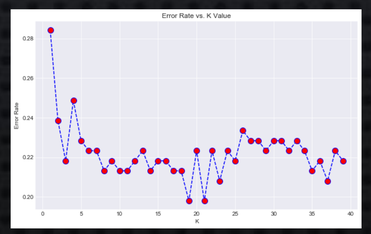


Decision Tree Variables

-Entertainment Spending
-Finances
-Smoking

KNN Variables

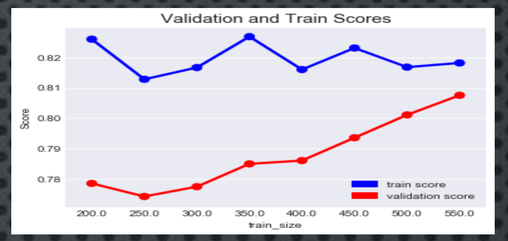
-Entertainment Spending
-Finances
-Smoking
-Romantic
-Difficulty Getting Up
-Giving
-Preferring Big Dogs over Small Dogs



Results

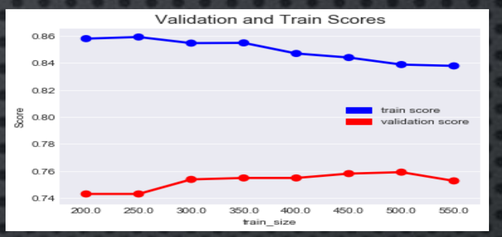
Logistic Regression

	Precision	Recall	F-1 Score
0	0.80	0.95	0.87
1	0.74	0.37	0.49
Avg / Total	0.78	0.78	0.74



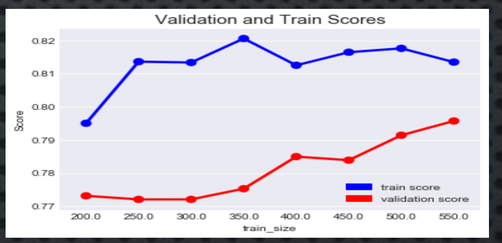
Decision Tree

	Precision	Recall	F-1 Score
0	0.79	0.91	0.85
1	0.50	0.28	0.36
Avg / Total	0.72	0.75	0.73



KNN

	Precision	Recall	F-1 Score
0	0.81	0.95	0.87
1	0.60	0.23	0.34
Avg / Total	0.76	0.79	0.75



	Average Accuracy Score on 5-Fold Cross Validation Set
Logistic Regression	0.833
Decision Tree	0.799
KNN	0.808

Conclusion

-We should keep in mind that this is a survey data set, so there is a possibility of response bias, which might have caused some distortion in the results.
-Observing the learning curves for Logistic Regression and KNN algorithms, we conclude that if we had more data we could have achieved better results.
-Average Accuracy Scores imply that Logistic Regression provides the best fit overall.
-None of the models have a high success rate predicting students that identify themselves as drinking a lot. Recall scores for positive labels are considerably low, but compared to the other two methods, Logistic Regression provides an almost satisfactory recall score for positive labels and a decent precision score.