SISA Final Project – Team 7:

Team Members:

Shahar Oded, Nir Rahav, Guy Zagorsky

Section 1- Introduction:

Research Question:

Are there trends differentiating drug addicts' admissions in the US for addicts who are submitted to rehab multiple times compared to addicts who are submitted to rehab for the first time?

Our objective is to create clusters (differentiated by trends in the different groups) for individuals who have participated in rehabilitation programs multiple times and those who are participating for the first time. Our research aims to provide a better understanding of these groups and their personal status prior to their admission, so that matching them with an appropriate treatment will perhaps be an easier task (based on their group tendencies).

Our vision is that if we can create characteristics (trends) for each group, that might affect the treatment given to the individual, and funds will be used more accurately, which can have a significant contribution to the mental health and drug addiction field. For Example, this could lead to the creation of hospitals that are proficient with first-time patients and others that are proficient with returning drug addicts, based on personal background.

This problem is significant as many recovered drug addicts struggle with relapses, and a better, more group focused treatment could help them recover and improve their lives.

Creating this profiling is difficult for many reasons, the first being the credibility of the data, as many addicts might withhold information on their admission to the rehab facility (for personal reasons, skepticism in the process and in their anonymity etc.) Moreover, the data is not serial data, meaning we can't use prior admissions to rehab for returning addicts, and a single addict may have a few rows in our dataset (from different admissions). This issue is handled by using primarily "day of admission" data, in oppose to patient background data.

Previous research has emphasized the role of social factors in addiction and recovery, and our study extends this research by mapping trends in the different groups (Appendixes).

Section 2- Data overview:

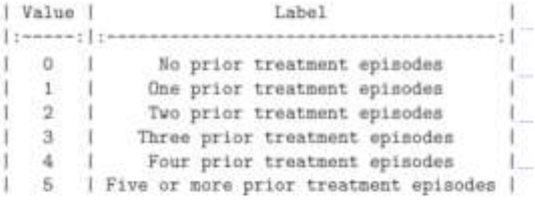
Our data set holds admission records to rehab facilities (US) during the year 2020. The data has features from before the admission, day of admission and of the treatment offered to the patient, while each row (entity) represents a patient's rehab admission.

In our analysis we chose to focus on data from **before and from the day of admission** as personal features that might show differences in the groups of our research question. The relevant feature families are composed from data such as state characteristics, first age of legal and illegal drug use (based on state), worse drug type in use, frequency and so on, that can paint a profile for the admitted (for example: "**recurring users** are often those who used their first illegal drug at **an older age**"). We created 8 calculated features that provide more suitable information for our purpose by incorporating the original data. The meaning and calculation method of these features, as well as other used features, are explained in the attached README file (see Appendixes).

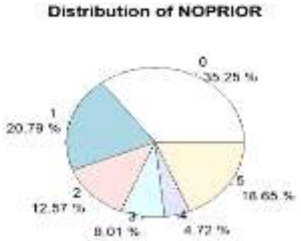
As described in the data processing chapter in the proposal, our final data set is composed of 22 features and 579254 rows (admission records). Due to computational power limitations, given the complex models we had to analyze, we sampled 5% (~29,000) of the data randomly, while keeping the ratio of the target value in our sampled data.

Section 3- Methods and results:

The objective of this study was to identify the essential features and trends distinguishing individuals' admissions to rehab. These records were categorized based on the number of prior admissions to rehabilitation ('NOPRIOR'), as described in the research proposal:



Value	Label
0	No prior treatment episodes
1	One prior treatment episodes
2	Two prior treatment episodes
3	Three prior treatment episodes
4	Four prior treatment episodes
5	Five or more prior treatment episodes



To identify the most prevalent features and trends within our groups, a regression tree model was developed. This model was constructed with the intention of classifying each admission record based on the corresponding number of prior admissions, enabling the extraction of essential and shared features from the model's decisions.

It is noteworthy that the model's accuracy on test cases is of secondary importance, as its primary purpose lies in serving to identify trends in key features. Our model's purpose is **not** to predict each record's 'NOPRIOR' value, but to create a low biased- high variance model (on the train), to later use its many branches to identify trends in the different classifications the model made.

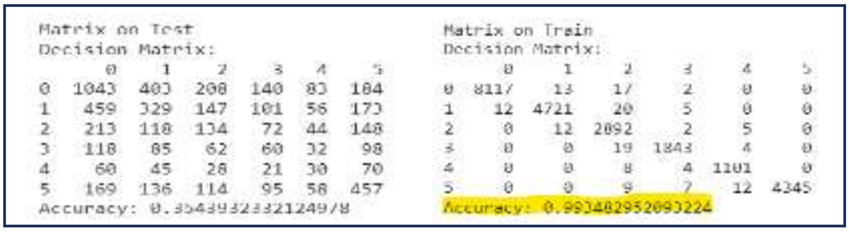
Our Model:

We utilized a regression tree as our main model. Since our model is continuous, we rounded and constrained its predictions to align with the original target values. Despite a classification model appearing more suitable, we opted for a continuous model to preserve the hierarchical order of our target values. Using a classification model would have treated each category differently and disregarded the order (see the model in the Appendixes).

To reinforce our findings, we generated 18 K-Means clusters (3 clusters for each 'NOPRIOR' category). By comparing the behavior of the feature analyzed in the tree model with the **correlation** and **slope** of the regression line in the cluster visualization, we validated the accuracy of our conclusions.

Model's Results on the train / test cases (Accuracy):

Given that this model is a multi-class classifier model, we used a truth table:

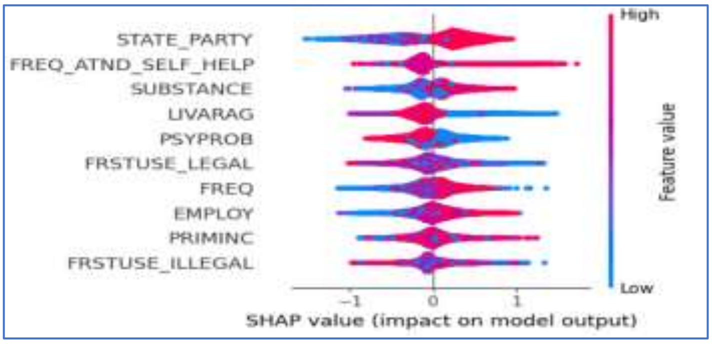


```
Matrix on Test                                  Matrix on Train
Decision Matrix:                                Decision Matrix:
      0     1     2     3    4     5                0     1     2     3     4     5
0  1043   403   208   140   83   184            0  8117    13    17     2     0     0
1   459   329   147   101   56   173            1    12  4721    20     5     0     0
2   213   118   134    72   44   148            2     0    12  2092     2     5     0
3   118    85    62    60   32    98            3     0     0    19  1843     4     0
4    60    45    28    21   30    70            4     0     0     8     4  1101     0
5   169   136   114    95   58   457            5     0     0     9     7    12  4345
Accuracy: 0.3543932332124978                    Accuracy: 0.993482952093224
```

As an exploratory model, our primary focus was on the accuracy of the model on the training cases. This allowed us to analyze the features that were frequently selected by the model to identify the groups with maximum precision.

Our Analysis and ID's creation:

We used a continuous SHAP visualization (a "bee swarm" viz), that shows the most meaningful features of the model and their trends (viz explanation, README in Appendixes):

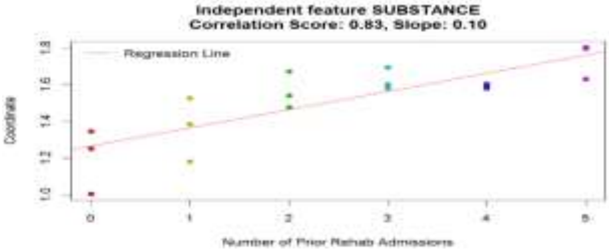


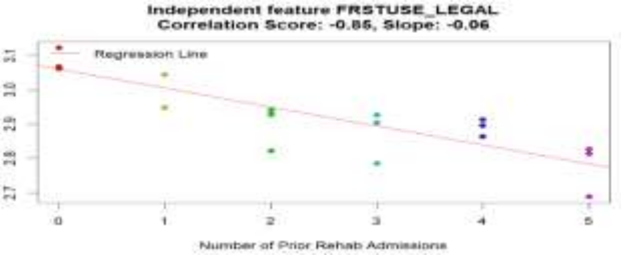
Key Results:

- Users with multiple rehab admissions tend to use harder narcotics, based on our scale.
- Users with multiple rehab admissions started using their first legal drug at a younger age, but this trend was not observed with illegal drugs.
- Users with multiple rehab admissions often experience co-occurring mental health issues alongside drug use.
- Users with multiple rehab admissions generally have poorer employment situations and dependent living arrangements.
- Users with multiple rehab admissions report higher frequencies of narcotic use.

Attached are the correlations of the variable centroids, supporting our findings. A positive slope/correlation indicates a positive relationship between the feature and the number of admissions for a patient. Analysis for all the influential features in the Appendixes).

To prevent any bias in the K-Means clustering, we applied scaling to our variables within the range of [0,1]. However, despite this adjustment, we did not observe any substantial changes in the results. Application of this scaling can be found in the codebook.





Section 4- Limitations and Future Work:

Limited data credibility and the collected features provided a restricted understanding of each group. With more time and resources, in the longer run, we would delve into mental evaluations and childhood information, if could be collected, as they hold valuable insights for characterizing each group and can guide experts in providing specialized care. On a closer future, creating a serialized dataset would enable us to compare admissions and identify changes over time for individuals returning to rehab.

In addition, conducting this research on the entire dataset, along with increased computational resources, would be desirable, given the proper resources.

Appendixes:

path to proposal:

Project Folder -> proposal (folder) -> proposal.pdf

Project Folder -> proposal (folder) -> proposal.RMD

Path to README of variables and code replication

Project Folder -> README (Markdown Source File)

** Check the README file before proceeding to the code files

Path to Final Results / model's code

Project Folder -> Final Results -> Models.pynb (tree model + SHAP analysis)

Project Folder -> Final Results -> model.RMD (basic analysis + K-means model)

Path to data

Project Folder -> data -> TEDSA_PUF_2020.csv (original data source)

Project Folder -> data -> admission_data_cooked.csv (processed data after proposal)

Project Folder -> data -> data_sample.csv (data sample of 5% as stated in the final results)

Link to prior research as stated:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4663247/

How to read a Bees warm SHAP visualization?

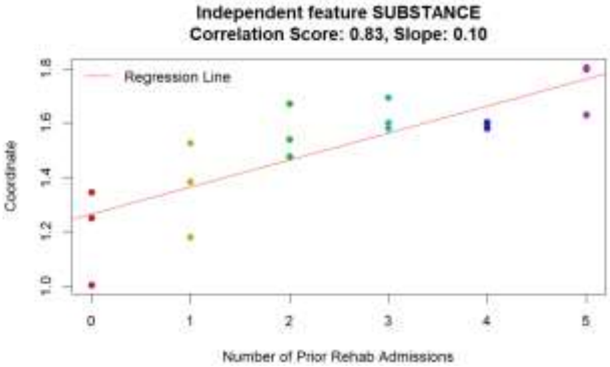
There are a few rules in understanding visualizations such as this one:

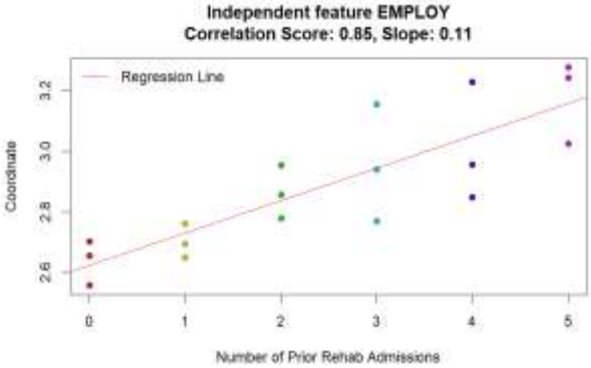
- The features are sorted top to bottom by their order of importance to the model.
- The colors as shown in the legend are for increase / decrease in the feature value.
- The horizonal scale indicates if the influence is positive or negative on the target value of the model.
- For example, a dark dot placed on the right side of the horizonal bar means that a feature sample with **low value in this feature**, was affecting in classifying target values with relatively **higher value**.

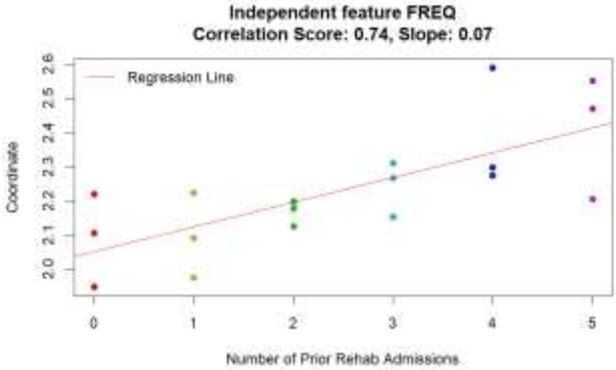
Validation of SHAP conclusion using K-means plot:

** Remember to check the variables encoding when checking these slopes.

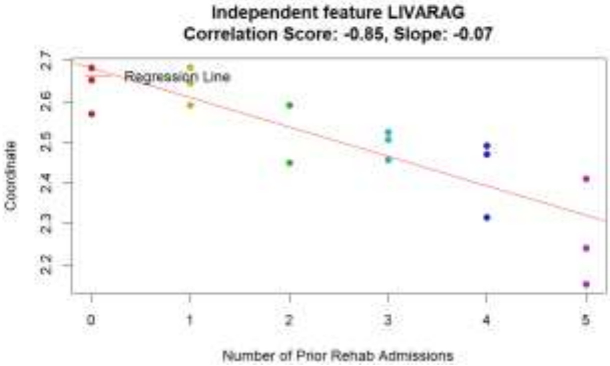
Higher independent feature values -> increase in target feature values:

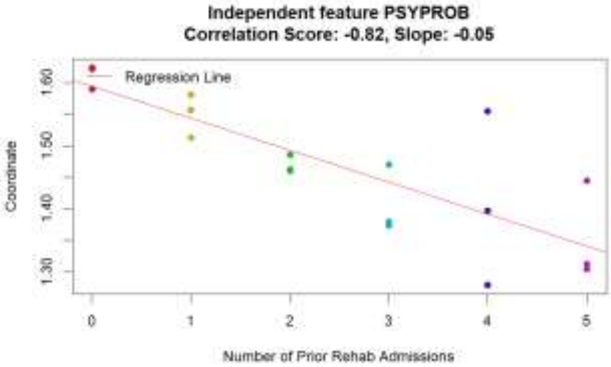


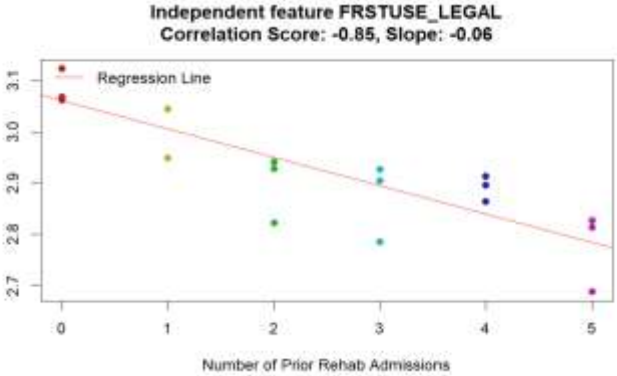




Higher independent feature values -> decrease in target feature values:


Independent feature LIVARAG
Correlation Score: -0.85, Slope: -0.07
Regression Line
Coordinate
Number of Prior Rehab Admissions


Independent feature PSYPROB
Correlation Score: -0.82, Slope: -0.05
Regression Line
Coordinate
Number of Prior Rehab Admissions


Independent feature FRSTUSE_LEGAL
Correlation Score: -0.85, Slope: -0.06
Regression Line
Number of Prior Rehab Admissions