DSCC265 Final Predictive Project - Prospectus

Erjia Meng, Jiayi Hao, Shiyu Liu, Xinyi Liu, and Yulin Feng

I. INTRODUCTION

The United States is experiencing an increase death resulting from drug overdoses. It is a serious cause of deaths, and finding solutions is an ongoing effort. In this project, our goal is to train models to complete a two-class classification that predicts whether a prescriber is overprescribing drugs. We wish to do models comparison by using different methods to analyze which models are most suitable for prediction in this dataset.

II. DATA

The dataset we used from [1]Kaggle contains 25,000 licensed healthcare professionals in the United States wrote 250 prescriptions for opioids and non-opioids in 2014 for citizens. The dataset also includes some metadata about the doctors.

There are 3 csv files in this dataset. The prescriber-info data has 262 columns, including some basic information of a prescriber a long list of drugs with the total number of prescriptions per year, and Opioid.Prescriber - a boolean variable indicating whether one individual prescribed opiate drugs over 10 times in the year. We will mainly work on this dataset to predict the target variable Opioid.Prescriber, which suggest whether that individual is potentially overdosing opiate drugs. The other two files may also provide additional information about drug generics and overdose population.

III. LITERATURE REVIEW

According to [2], "Millions of Americans suffer from pain and are often prescribed opioids to treat their conditions. However, the dangers of prescription misuse, use disorder, and overdose have been a growing problem throughout the United States. Specifically, While prescription opioids were involved over 28% of all opioid overdose deaths in 2019, there was a nearly 7% decrease in prescription opioid-involved death rates from 2018 to 2019. From 1999 to 2019, nearly 247,000 people died in the United States from overdoses involving prescription opioids." As a result, there's a crucial significance to distinguish doctors that might be such drug dealers.

IV. GOALS

In this project, we are conducting a predicative analysis to identify potential 'pill whales' - prescribers who are over-prescribing opiates drugs. Our main goal is to develop a model that gives the highest possible prediction accuracy in predicting whether a prescriber is prescribing given the information of this prescriber. To achieve this goal, we will mainly focus on training and comparing different predictive models that use (1) different sets of input attributes, and (2) different ML models. In the prior direction, we plan to perform attribute selection by using different dimension reduction techniques, including PCA and L1-regression, and compare the results produced by different attribute spaces. In the later direction, we plan to train various ML models, such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest, KNN, SVM, Neural Network, and Clustering, and contrast the performance of each model using confusion matrix, F-1 score, as well as ROC curve.

V. EXPECTATION AND RESULTS

At this stage, we cannot approximate the expected accuracy for each model due to the complexity of the dataset and the large number of combinations of variables. However, we can still hypothesize the set of attributes and models that may give the best results according to the characteristics of the dataset. First of all, we speculate the significant attributes are: the death-to-population ratio of each state, the medical and the type of medicinal practice degree of each prescriber. Some of the opioids might also be significant predictors but we cannot identify them yet. Secondly, we deduce that the most effective models for prediction are Logistic Regression, Naïve Bayes, and Random Forrest and Neural Network. The reason is that we are predicting a Boolean variable and these models are suitable for it. At last, we will compare the performances of all the models and provide a visualization for the best one.

VI. TIMELINE

After group discussions, we made the following arrangement. (1) Xinyi Liu will be handling with the data and EDA. Since lacking or unreasonable data will greatly impact the model prediction, it is significant to have cleaned and useful data. (2) Erjia Meng will be working on selecting the important attributes since if more important attributes are selected as prediction data, it will improve the calculation results to a great extent. (3) Yulin Feng and Shiyu Liu will be working coding and testing different models. Build different models to prepare the ground for the selection of Attributes. se behavior. (4) Jiayi Hao, as the team leader, will work with other members in all the aspects, helps in handling workload or technical difficulties, and assures all the members are in the same pace. In the week of April 18-23, we will mainly work on building and test the model, after that we will mainly focus on writing the final report.

REFERENCES

[1] "U.S. Opiate Prescriptions/Overdoses". Accessed on: Apr. 15, 2022. [Online]. Available: https://www.kaggle.com/datasets/apryor6/us-opiate-prescriptions?select=prescriber-info.csv

[2] "Drug Overdose Deaths in the U.S. Top 100,000 Annually," Nov. 17, 2021. Accessed on: Apr. 15, 2022. [Online]. Available: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2021/20211117.htm