SUMMARY PAPER

Team Odyssey

Shiyan Chen, Pratik Rath, Siddhartha Vanam, Wenshu Wu

INITIAL RESEARCH & DATA EXPLORATION

The United States is in the midst of an unprecedented opioid epidemic. The CDC estimates overdose deaths increased by 5% from 2018-2019, quadrupling since 1999. Between 1999-2019 there were nearly 500,000 opioids involved-overdose deaths (prescription & illicit). Even though opioid-related fatalities were similar across racial/ethnic groups in the '90s, they have risen disproportionately among non-Hispanic Whites over the years. A study conducted by NSDUH using data from 2003-2014, found prescription opioid misuse was consistently higher among White adolescents relative to their Hispanic & Black peers. Historically, prescription opioids misuse has been overlooked among Asians & Native Americans. However, recent data suggests opioid-related fatalities among Native Americans are second only to Whites & a 30% rise in treatment admissions for misuse of prescription opioids among Asian individuals between 2000-12.

We examined two main demographic factors, the income & education level, & their correlations to the misuse of opioids. According to the US Year 19 dataset, a relatively balanced distribution of people who have ever used opioids across all income levels was observed, indicating a weak, positive correlation between the income level & lifetime use. But there may be a possible negative correlation between the income level & non-medical use of opioids. Since income level depends on education, we continued exploring the dataset & found a possible inverted U relationship between education level & opioid lifetime use and a negative relationship between education level & non-medical use. After identifying these potential relationships, we investigated the data further by building models.

GOALS, HYPOTHESES, DATA HANDLING, & MODELS

Our goal is to model the data for the US in the Year 2019 & explore the effects of demographic profiles (e.g. income level, education level, race, etc.,) on opioid misuse as well as use our model to help construct a shorter survey & predict opioid misuse tendencies. Hypotheses are (1) Education & Income will have a significant negative effect on opioid abuse; (2) Demographic Info will provide high predictive capability in identifying respondents that have abused opioids.

Firstly, we dropped features with more than 30% missing data to ensure the data ingested by the model was as credible as possible. This left six variables with 20% missing information on average. We had imputed these features with 0 since we assumed the missingness mechanism was 'Missing at Random' (MAR). After wrangling the data it was ready to be fed into a Classification Model pending some feature transformations. Our models have a two-fold purpose: to be explanatory and predictive.

We created 3 Logit prediction models to predict whether a respondent had *ever* misused Opioids - which is denoted as [OP_NMU] using the accuracy metric. We established the first as a baseline model (A) using five main demographic features: age, gender, region, education, & income. This baseline model yielded an R^2 of 6.63% & an accuracy of 88.47% Models B and C had 22 & 55 features and yielded R^2s of 17.5% and 36.7%. Additionally, this jump in feature selection increased the accuracy by 1.5% from 89.4% to 90.9%.

FINDINGS & CONCLUSIONS

Our final model picked up 55 variables, employing a combination of statistically significant & insignificant variables. We kept the race variable because logic states that it plays an important role in opioid misuse in society. The new model has an accuracy of 90.9%. But we chose Model A as our final model because doubling the variables only increased the accuracy by 1.5%. Our model could be used to reduce the survey completion time from 15 to 2 minutes. It would also increase respondents' willingness to complete the surveys and decrease high data storage. Our final models are limited in that statistically insignificant variables and that they did not account for the weighted distributions of the population.