Matteo Shafer, Sucheen Sundaram, Kyle Lew, and Cameron Stivers

Alex Dekhtyar

CSC 466

14 December 2023

Final Project Report

Contents

Authors and Affiliations

Matteo Shafer ~ mshafe01@calpoly.edu

Sucheen Sundaram ~ sssundar@calpoly.edu

Cameron Stivers ~ ctstiver@calpoly.edu

Kyle Lew ~ klew06@calpoly.edu

Abstract

This analytical report dives into mental and physical health in contemporary society, which are important to us. Our objective was to explore predictors influencing health metrics and identify their significant impacts. We hoped to discover new trends in health-related to music, lifestyle, and in the tech workforce. We selected three diverse datasets from Kaggle, a prominent data science platform, and conducted an extensive analysis of the data implementing different methods for each dataset.

Introduction

The approach our team took for this project, we decided we wanted to cover mental and physical health in modern society. Health and well-being are things that we as college students care about so we wanted to focus our analytical project on these topics. Specifically, we answered the question, are there predictors that affect health metrics? And if so, which predictors have the biggest impact? To cover this topic, we explored 3 datasets that all relate to health, each with their own specific analytical questions.

Dataset Descriptions

We selected all 3 of our datasets from Kaggle, a well-known data science competition platform and online community of data scientists.

Sleep and Health Lifestyle

The Sleep Health and Lifestyle Dataset contains around 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

Mental Health in Tech

This dataset is from a 2014 survey that measures attitudes towards mental health and the frequency of mental health disorders in the tech workplace. This was a survey conducted in tech companies all over the world with 60% of the data coming from U.S tech workers, 15% from the United Kingdom, and the last 25% coming from a variety of different countries such as Bulgaria, Netherlands, and many other countries.

Music and Mental Health

This dataset compares the mental health and music preferences/listening habits of respondents. For the music variables, this dataset tracks things such as hours listened, favorite genre, do you listen while working, and many more. For the mental health variables, it has Anxiety, Depression, Insomnia, and OCD all on a scale of 1-10.

Analytical Questions

For dataset 1: Sleep and Health Lifestyle

- Can we classify a person's sleeping disorder status based on sleep quality and lifestyle predictors? Which of these predictors is best at classifying?

For dataset 2: Mental Health in Tech

- How can different tech workers be grouped together based on their responses to the mental health survey? What do these two groups represent?

For dataset 3: Music Mental Health

- What are frequent combinations of music preferences?
- Are certain music genres more associated with mental health disorders?
- Can we predict whether music improves mental health based on listening habits?

Methods Deployed

a. We implemented a KNN classifier on our Sleep and Health Lifestyle dataset to classify whether somebody had a sleeping disorder and which one they had. We also decided to use a k of 10 for our neighbors as we thought this would be a good number of neighbors as our dataset was about 700 rows. We computed the accuracy using cross-validation with 5 folds.

b. For our Mental Health and Tech dataset we implemented a clustering algorithm, specifically k-means. We decided to set k=2 as we wanted two clusters of the respondents who were strong opinioned on mental health and respondents who did not have strong opinions. We tried to implement other clustering algorithms such as DBScan and Hierarchical, but they produced either one cluster or way too many.

c. In the Music and Mental Health dataset, we chose to perform frequent item set and association rules mining of music preferences and mental health statuses. For music genres, we consider the subject a listener if they answered one of ‘Sometimes’ and ‘Very Frequently’ to how often they listen to the genre. For mental health conditions, if the subject rated the prevalence of the condition as at least a 7 out of 10, then we considered it significant enough to consider them as having a strong case of the condition. The main objective of finding frequent item sets is to see if the users in the data have any common combinations of music preferences, while the intent of association rules mining is to discover any associations between music preferences and significant reported mental health conditions. After some parameter tuning, we decided to use a minimum support of 0.15, and a minimum confidence of 0.4.

d. Using the Music and Mental Health dataset, we fit a Logistic Regression model to predict whether a respondent believes music improves their mental health using several predictive variables about music habits. Our main motivation behind creating this model was to output the coefficient values for these predictive variables, giving us a quantitative interpretation of the how music habits affect mental health. We still used a test/train split on the model to check the accuracy of the model and our coefficients.

Results

a. *Model 1: Sleep Level*

Predictors	**Accuracy**
Sleep Duration, Quality of Sleep	0.630

Model 2: Physical Activity

Predictors	**Accuracy**
Physical Activity Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps	0.851

Model 3: Stress Level

Predictors	**Accuracy**

Stress Level	0.668

b.

Cluster Sizes

Cluster	**Cluster Size**
Group 1	585
Group 2	673

Cluster Means

Cluster	Age	Sought Treatment
Group 1	25.9	47.2%
Group 2	37.18	52.8%

c. A total of 70 frequent item sets were found, and below is a small sample of skyline frequent combinations of music preferences:

Frequent Item Sets

Item Set	**Support**
Rock, Rap, Hip Hop, Pop	0.224
Rock, Rap, Hip Hop, R&B	0.171
Rock, R&B, Pop	0.225
Rock, Video Game Music, Pop	0.25
Hip Hop, R&B, Pop	0.257

Using the parameters described previously, we found 3 association rules with a mental health condition on the left side and a music preference on the right:

Association Rules

Left Side	**Right Side**	**Confidence**
Anxiety	Jazz	0.500
Anxiety	EDM	0.489
Depression	Metal	0.414

d. Logistic Regression

Logistic Regression Model

Variables	**Coefficient**
Hours per day	0.21
While working	-0.03
Instrumental	0.29
Composer	0.21
Exploratory	0.58
Foreign language	-0.14
BPM	-0.17

This model achieved an accuracy of 0.77 when fit on a test/train split

Explanatory variable descriptions:

Hours per day – Number of hours the respondent listens to music per day
While working - Does the respondent listen to music while studying/working?
Instrumental - Does the respondent play an instrument regularly?
Composer - Does the respondent compose music?
Exploratory - Does the respondent actively explore new artists/genres?
Foreign Language - Does the respondent regularly listen to music with lyrics in a language they are not fluent in?

BPM - Beats per minute of favorite genre

Conclusion

Through our comprehensive analysis of these distinct datasets, our investigation revealed intriguing insights. In our first dataset looking at KNN to classify whether someone had a sleep disorder (Sleep Apnea or Insomnia) or none we looked at three different models. The first model we looked at was sleep level where we looked at sleep duration and quality of sleep to see how our predictions did. Surprisingly, we only got an accuracy of 63.0%, which was lower than expected. The second model we looked at was physical activity where we looked at predictors such as physical activity level, BMI category, blood pressure, heart rate, and daily steps. This model did very well with an accuracy of 85.1% which is fairly high for classifying someone correctly with their sleeping disorder or not having one. Lastly, we looked at a stress level model where our only predictor was stress level. This model did okay with an accuracy of 66.8% however, this was more expected as we only had stress level as a predictor. Overall, these results were surprising as we thought that the sleep model would perform best, but it ended up performing the worst. However, based on this data, physical activity level seems to be the strongest predictor of whether someone has a sleep disorder.

We used cluster analysis in our second dataset focusing on the survey responses of technology workers and their opinions on mental health. The analysis, which focused on the survey responses of technology workers regarding mental health, revealed two distinct groups with differing characteristics. Group 1, with an average age of 25.9 years, had a lower percentage (47.2%) of individuals who wanted treatment for mental health issues. On the other hand, Group 2, which is older with an average age of 37.8 years, showed a higher likelihood (52.8%) to seek treatment. This indicates a potential association between age and the likelihood of seeking treatment among technology workers. It suggests that older individuals in this sector may be more inclined to seek help for these types of issues. However, the small difference in percentage shows that the association isn't very significant, and clustering did not separate groups as effectively as we would have desired.

A key takeaway from the table of frequent item sets of the Music and Mental Hearth data is that rock and pop seems to be very popular genres of music among the listeners in the data, appearing in many frequent item sets with high support. From the association rules, we see that 50% of subjects with self-reported anxiety are jazz listeners and 48.9% of them are EDM listeners. 41.4% of those with self-reported depression listen to the metal genre. Overall, these do not expose causation, but they do indicate associations between music preferences and mental health conditions, which can be indicative of how users with a mental health condition might treat themselves, as we explored using Logistic Regression.

From our Logistic Regression model on the Music and Mental Health dataset, we were able to find coefficients for certain predictor variables, quantifying the effects they have on whether a respondent uses music to improve their own mental health. Looking specifically at some of these, the variable with the largest coefficient of 0.58 is 'Exploratory', meaning that if a respondent actively explores new music, they are more likely to use music to improve mental health. Conversely, the BPM variable has the lowest coefficient at –0.17, meaning respondents whose favorite genre has faster BPMs are less likely to use music to improve mental health. By analyzing the coefficients like this, we can both get a sense of the direction in which the predictors affect our target as well as the magnitude of these effects.