**Exploring and Modeling**

**on Mental Health Data**

**: Insights from Kaggle Data Analysis**

Prepared for: Prof. Divya Pandove

(TA: Bhumika Mallikarjunhorapet)

Prepared by: Team 7

Haeyeon Jeong, Sai Rachana Kandikattu

Abirham Getie, Yonathan Shimelis

Dec 17, 2024

Introduction1

Research Methods and Results2

*Task 1. Data Acquisition and Data Preprocessing 2*

*Task 2. Exploratory Data Analysis and Statistical Testing* 5

*Task 3. Modeling* 7

*Task 4. Model Evaluation* 8

Results12

Conclusions12

Recommendations12

References 1

Appendix 1

**Introduction**

According to Johns Hopkins, 1 in 4 American adults suffers from a diagnosable mental disorder in a given year. A statistic that proves mental health challenges are becoming more visible across different sectors, affecting individuals in diverse ways. The first step to developing effective support systems is understanding the factors contributing to mental stress. Our project will explore mental health data using various classification models to identify key factors influencing mental health outcomes. Specifically, we will examine how variables such as mood swings, family history, treatment status, occupations, and time spent indoors affect mental health. We chose this topic because we hope to uncover patterns that can help improve mental health management and prevention strategies by analyzing these factors through data models.

This study seeks to address the following key question and other additional questions:

***Initial Research Question:***

How can we develop models to assess mental stress, and which factors are most influential in predicting mental health outcomes?

***Additional Research Questions:***

Question 1: What are the top five factors contributing to the escalation of stress levels in individuals?

Question 2: Does a family history of mental health issues influence whether individuals seek treatment for mental stress?

Question 3: What specific factors contribute the most to the growing stress among students?

**Research Methods**

To achieve these objectives, we will perform the following tasks:

***Task 1. Data Acquisition and Preparation***

A screenshot of a computer

Description automatically generated

The dataset used in this study is available on Kaggle at this ([link](https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset)). It contains 292,364 observations with 17 variables. The original source of the data is Ourworldindata.org. The method of collection is via survey. We preprocessed and changed the raw dataset we were working with before doing any EDA or model building. The process for this preprocessing is listed below:

1. Missing Values Handling (drop rows with NA values)
2. Drop unnecessary or irrelevant columns like Timestamp, Country, and Self\_Employed
3. Feature transformation for our categorical variables to numerical
4. One-hot encoding

It’s important to highlight key differences in our changed dataset versus the raw dataset we downloaded. Most notably, the changes we made to the target variable of our models, “Growing\_Stress”. The variable has three different values: no, yes, and maybe. We first created a map to define the unique values with 0 representing no, 1 representing yes, and 2 representing maybe. We then removed the rows containing the response “maybe” for our “growing stress” variable to make outcomes for this variable binary.

After making these changes, we began EDA to gain a basic level understanding of our data.

***Task 2. Exploratory Data Analysis and Data Visualization***

We will conduct exploratory data analysis to identify key trends, relationships, and distributions in the dataset.

***Task 3. Modeling and Evaluation***

Various classification models will be employed, including Logistic Regression, SVM, KNN, and Random Forest, to assess the impact of identified factors on mental health outcomes. By doing so, we aim to answer the SMART questions we developed at the start of this project.

**Results**

**Question 3: What contributes to the growing stress among students?**

In our dataset, students had the second-highest proportion of survey responses that indicated experiencing feelings of growing stress. To add to this, we found this would be an interesting question to gain insight into since we’re students. By first doing EDA and statistical testing, we found growing stress was not statistically independent from our growing stress target variable. After discovering so, we moved forward to try and observe any correlation between growing stress and other variables for the students in our dataset. To our surprise, our variables had a weak correlation between growing stress and our other variables. The strongest correlation is -0.14 between “changes\_habits” and our target variable. So moving forward in trying to answer this question, we used logistic regression modeling.

The model didn’t perform very well. With an accuracy of 61% and a AUC of 0.62, it left much to be desired. The confusion matrix revealed the model was eager to predict a student reporting growing stress when students didn’t report growing stress. The model’s shortcomings can be proof of the lack of predictive power our current variables give us just for students. It could be possible that if the data included variables more tailored for students, we’d have a logistic regression model with stronger predictive power. Variables like academic performance or financial strife could provide key insight in predicting whether or not a student is experiencing growing stress. The model did, however, leave us with more revealing coefficients.

However, we were able to use this logistic regression model to derive feature importance for our target variable in this subset of data focused on just students. We conclude days indoors, gender, mood swings, and social weakness being the top five factors that impact growing stress for students. Of these variables, only changes habits and gender had an inverse relationship with growing stress.

**Conclusions**

In conclusion, the top five factors contributing to growing stress across all occupations include Mental Health History, Days Indoors, Work Interest, Mood Swings, and Changes in Habits. For students, the primary stressors are Changes in Habits, Days Indoors, Gender, Mood Swings, and Social Weakness. Individuals with a family history of mental health issues may be more likely to recognize their symptoms or seek treatment. To mitigate stress, it is important across all sectors to spend more time outdoors, acknowledge stress when experiencing changes in habits or frequent mood swings, and seek treatment when symptoms arise. The negative relationship between growing stress and treatment underscores the critical need to address stress proactively.

**Recommendations**

**References**

Johns Hopkins Medicine. (2024). *Mental health disorder statistics*. John Hopkins Medicine. https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics

‌

Our World in Data. (2024). *Our World in Data*. Our World in Data. https://ourworldindata.org/

Git Repository: [**github**](https://github.com/sairachanak/6103-Team7.git)