

**HIGHER NATIONAL DIPLOMA IN SOFTWARE
ENGINEERING
KAHDSE24.1F**

Datawarehouse and Business Intelligence

Course Work Report

Mental Health Awareness Data Analysis

SUBMITTED BY

B.T Bogahalanda	KAHDSE241F-030
P.D.C Perera	KAHDSE241F-032
A.G.G.B. Premasiri	KAHDSE241F-033
M.A.Y.R Perera	KAHDSE241F-036
C.B Herath	KAHDSE241F-054

Table of content

01. Introduction.....	1
02. Data Analysis and Visualization.....	1
2.1. Data Visualization.....	2
2.2. Discussion and Interpretation.....	3
2.3. Key Trends Observed.....	3
03. Selection of Data Mining Algorithm and Data Pre-processing.....	4
3.1. Data Mining Algorithm Selection.....	4
3.2. Data Pre-processing Steps.....	4
04. Data Mining.....	5
4.1. Implementation of Decision Tree.....	5
4.2. Interpretation of Results.....	5
05. Data Ethics.....	7
06. Conclusion.....	8
6.1. Overall Visualization Results.....	8
6.2. Overall Data Mining Results.....	8
6.3. Business Intelligence Insights.....	9

01. Introduction

Considering millions of people worldwide suffering from afflictions including anxiety, depression, and stress-related disorders, mental health has become a major public health concern. Despite its significance, mental health frequently gets less attention than physical health, which results in inequalities in support and care. By examining a large dataset on mental health, this aims to address this problem by determining the trends, patterns, and variables that affect mental health outcomes. This dataset offers plenty of information for comprehending the complexity of mental health. It includes characteristics such as gender, country, stress levels, treatment history, family history of mental health, occupations, and self-reported mental health status. This project intends in using data analytical tools to find insights that can help healthcare professionals to create focused interventions and enhance the mental health services.

To make sure the dataset was suitable for the analysis, the data was cleaned and prepared, such as removing inconsistent data, fixing missing values which might affect results, dealing with outliers, normalizing numerical data to guarantee consistency, and categorical values transformed into numerical formats. For the analysis to be accurate pre-processing stage is essential.

02. Data Analysis and Visualization

The mental health dataset selected for this investigation is an extensive compilation of information about mental health conditions, therapies, and results. This presents an in-depth of mental wellbeing using power BI. The dataset contains data approximately individuals' conditions that was conducted by an interview on work life, country, mental health history, treatment, care options, habits and so on.

The objective of this investigation is to reveal patterns, designs, and the relationships that affect mental wellbeing, using information visualization strategies. There are ten visualization techniques that was used in to analyze patterns and insights.

2.1. Data Visualizations



- I. Gender Distribution using a Pie Chart – Shows the number of respondents by gender.
- II. Count of Treatment using a Clustered Column Chart – Comparing the individuals with and without a mental health history of health conditions and the likelihood of seeking treatment.
- III. Social Weakness using a Donut Chart – Showing the social weaknesses of individuals that they might face during outings.
- IV. Count of Interview using a Gauge Chart – Represents the count of individuals cooperation for the mental health interview conducted.
- V. Growing Stress Count using a Line Chart – Comparing the individual with a certain occupation and the likelihood of growing stress counts of them.
- VI. Mood Swings likelihood using an Area Chart – Compares the mood swings of individuals and the count of treatment likely to get.

- VII. Work of interest Count using a Tree Map – Represents the occupations of the interviewed individuals and the work of interest count.
- VIII. Country by using a Map – Represents the countries of individuals and the genders of interviewed individuals' dataset.
- IX. Coping Struggles by using a Clustered Bar Chart – Comparing the in-door days or out door days of individuals with the likelihood of coping struggles they get to affected.
- X. Family History Count using a Stacked Column Chart – Comparing the individuals with and without a family history of health conditions and the likelihood of seeking treatment.

2.2. Discussion and Interpretations

- More female respondents indicating a higher willingness of females to participate in this survey and the male underrepresentation shows the gap of mental health awareness of males.
- Individuals without mental health history, mostly females, seeks for treatment at a high rate, this may indicate the rising mental health awareness.
- Individuals with a mental health history also seek for treatment often this suggests the genetic or the environmental factors may play a role.
- A high percentage (35.47%) of respondents reports for social weakness that indicates mental health issues affecting confidence.
- Certain jobs show higher stress levels, and these industries require better mental health support.
- The higher number of people with frequent mood swings seek treatment, suggesting self-awareness and proactive health behavior.
- People in stressful jobs have low work interest indicating burnout or job satisfaction.
- Individuals with family history of mental conditions seek treatment often, showing the awareness and genetic influence, and those without family history seeking treatment suggest that there is a shift towards mental health acceptance.

2.3. Key Trends Observed

- Individuals with family history are more likely to seek treatment
- Females are more likely to seek treatment suggesting they are more proactive about mental health.
- There is an awareness among women than men about mental health awareness.
- People spending more time indoors struggle with coping and social weaknesses.

03. Selection of Data Mining Algorithm and Data Pre-processing

3.1. Data Mining Algorithm Selection

To further analysis the mental health awareness dataset, the best algorithm for this is Decision Tree Algorithm for classification. This algorithm is suited for handling both categorical and numerical data which aligns with data visualizations and metrics in dataset, such as gender, occupation, treatment likelihood, coping struggles. The categorical relationships such as the comparison of treatment likelihood with mental health history from clustered column chart or the tree map that displays the work interest by their occupation, captures the decision tree relationships. This also can help in identifying factors that influence decisions such as seeking treatment or coping with stress.

The Classification – Decision Tree Algorithm can help to predict the treatment seeking behavior and determine the factors such as mood swings, social weakness that impact treatment decision. This will help in clear understanding of the mental health decision factors.

3.2. Data Pre-processing Steps

First, starting to examine the dataset for any missing values and addressing them through techniques, based on the extent of issue. Next analyzing the dataset for outliers in continuous variables such as stress levels or probability of seeking treatment. Outliers can distort the results, therefore, eliminating them to adjust the extreme values into a manageable range. Furthermore, normalizing or standardize all numerical to make each variable hold equal weight in the model.

- I. Handling missing values – When the dataset contains missing records, filling the values with appropriate mode.
Ex.- self_employed column
- II. Converting Boolean values to Binary – Replacing TRUE/FALSE with 1/0 for a better analysis.
Ex. Family_History. Treatment, Coping_Struggles. Columns
- III. Standardize Categorical Variables – Converting text-based responses into numeric categories and ensuring consistency.
Ex. - Days_Indoors, Growing_Stress, Changing_Habits, Work_Interest, Social_Weakness, Mental_health_interview, Care_options columns
- IV. Removing duplicates in the dataset.

04. Data Mining

After the process of pre-processing, the Decision Tree algorithm can be applied to analyze the relationships of the dataset. The chosen algorithm decision tree splits the dataset into subsets on the features that best separate data. This creates nodes that represents the outcomes on tests. This is useful classification tasks and handling both the categorial and continuous data.

The goal of this analysis is to predict the target variable treatment likelihood based on the input features.

4.1. Implementation of Decision Tree

Model Training – Once pre-procession is completed, training the decision tree model is done using features. As we want to predict whether the individual is likely to seek treatment based on the factors stress count, coping struggles, occupation, and family history. We select feathers Treatment likelihood, Growing Stress count, Occupation, and Coping Struggles. The target variable is selected to be treatment likelihood.

Model Evaluation – When the model is trained, the performance is evaluated using various metric such as accuracy, precision, recall, and F1-score. These metrics can predict the likelihood of the individuals that are seeking treatment based on the provided features.

4.2. Interpretation of Results

Accuracy – Low accuracy and the model correctly classifies 42.94% of instances but it does capture the patterns of occupation or stress level influence the likelihood of seeking treatment.

Precision – The measures of the predicted positive outcomes are actually correct

- Class 1 Precision 39.0% of predicted class 1 instances are actually correct
- Class 2 Precision 43.1% of predicted class 2 instances are actually correct.
- Class 3 Precision 46.9% of predicted class 3 instances are actually correct

Recall – The positive measures are correctively identified by the model

- Class 1 Recall 26.6% of actual Class 1 instances
- Class 2 Recall 64.3% of actual Class 2 instances
- Class 11 Recall 31.6% of actual Class 11 instances

F1-Score – Providing a balanced measure between the recall and precision

- Class 1 F1-Score 31.6% (poor)
- Class 2 F1-Score 51.6% (moderate)
- Class 11 F1-Score 37.8% (poor)
- Weighted Average F1-Score 41.3%

From the given dataset that consist of 9000 instances, the model was trained to classify them into three classes as Class 1, Class 2, and Class 11. However, the model's performance suggests that there are some possible issues as imbalanced for some classes and model may not be generalizing well some of the unseen data. But this model is reliable at predicting treatment likelihood, predicting positive those who are likely to seek treatments. This model has balanced precision and recall which provides a strong prediction performance.

05. Data Ethics

Analyzing and using business data when related to mental health, treatments, social issues, and other data the ethical concerns must be looked to ensure that the data is handled responsibly and fairly. When dealing with sensitive information it is critical to ensure the rights of individuals are respected and the insights and prediction results are used responsibly.

By addressing the concerns, the businesses can ensure that the data they use promotes fairly by taking accountability while minimizing harm to individuals and communities.

- Mental health data and personal data are highly sensitive therefore, privacy concerns arise when the information is collected and analyzed. It is crucial to ensure the informed consent is obtained before collection of the data. The data must be anonymized to protect the individuals' identities also taking measures to restrict access to sensitive data limiting access.
- Data that is used for analysis should be accurate, complete, and reliable. Inaccurate data might lead into misleading results that can harm individuals leading into incorrect predictions. Therefore, must ensure the data is cleaned and pre-processed correctly, avoiding errors.
- Decision making should be transparent and when making certain decisions it should be made clear how and why those decisions were made. Lack of transparency in algorithms make wrong predictions therefore, this decision making in data analysis should be clear and concise.
- Unauthorized access to sensitive data, especially mental health information can lead to harm or exploitation of individuals. It is essential to protect the personal and sensitive data from being exposed or misused in the contexts where these individuals may face discrimination based on the information they provide.
- When analyzing data in mental health related areas this may affect a vulnerable population, such as marginalized communities, individuals with pre-existing mental health conditions, and other groups. There is a high risk in that the predictions based on incomplete and biased data can negatively affect the vulnerable individuals. Therefore, by using inclusive datasets and fair algorithms to evaluate the impact of the analysis on different groups ensures that they are not unfairly excluded.

06. Conclusion

6.1. Overall Visualization Results

The visualization of the dataset about mental health awareness provides a comprehensive understanding of the key factors that influence the mental health treatment, growing stress levels, occupations, work interest and the coping struggles among the individuals that participated in this survey. The data visualization dashboard contains ten data visualization cards;

- Gender distribution – Providing a gender distribution to understand the gender role plays in mental health awareness.
- Treatment Count – Proportion of individuals likely to seek treatment and also showing a trend in some individuals with the history of mental health conditions still do hesitate to seek treatment.
- Social weakness – High number of individuals reporting experiencing social weaknesses that might be linked to social anxiety, stress or workplace burnout.
- Growing stress count – Shows increase in stress levels with specific occupations.
- Mood Swings Likelihood – Individuals with frequent mood swings showing a higher probability of seeking treatment.
- Family History Count – Individuals with a family history of mental health conditions are more likely to seek treatment and there are some who still hesitate to seek treatment.
- Coping Struggles – The ones spending more time in doors experience great struggles in coping with stress.
- Country representation – Dataset has covered a wide range of individuals across multiple countries, showing trends in mental health awareness.
- Work interest – The occupational distribution reveals that certain jobs can be correlated with increasing stress levels.
- Interview Cooperation – This dataset contains 9000 data obtained by a survey conducted revealing a moderate to high participation in mental health interviews.

6.2. Overall Data Mining Results

By using the decision tree Algorithm, we were able to analyze the dataset and predict the likelihood of treatment based on certain factors such as stress count, occupation, coping struggles, family history, work interest, mental health history.

Model fit and Accuracy – The decision tree classifier achieved an accuracy of approximately 42,94% indicating a low prediction capability. The performance Metrics are Precision (46.9%), Recall (64.3%), F1-Score (51.6%)

6.3. Business Intelligence Insights

The data visualizations and the data mining analysis provides a valuable business intelligence for organizations such as healthcare providers.

- Mental Health Awareness Campaigns – The findings indicating individuals with mood swings and stress that don't seek treatments, this suggests mental health awareness programs can be organized for individuals.
- Treatment Plans – By analyzing stress patterns and coping struggles health care professionals can offer personalized plans.
- Workplace Mental health Initiatives – The correlation between the occupation and stress levels can suggest workplaces to introduce employee wellness programs, stress management programs.

In conclusion, Important trends and patterns referring to stress levels, coping struggles, and mental health treatment were found through the dataset analysis. Based on the stress levels, occupations, work interest, family history, care options, mental health history, the Decision Tree Model accurately predicted the likelihood of treatment granting the mental health organizations, Business organizations, healthcare providers useful business intelligence. By using these insights, companies and healthcare organizations can improve their methods for helping people with mental health issues, raises awareness, and designing specialized counselling programs to enhance the general wellbeing of individuals.