5-2022

Identifying Psychological Mental Disorders through Machine Learning

Mohammad Al Abdooli
mma1750@rit.edu

Identifying Psychological Mental Disorders through Machine Learning

by

Mohammad Al Abdooli

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

May 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Mohammad Saif Al Abdooli**

Graduate Capstone Title**: Identifying Psychological Mental Disorders through Machine Learning**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

Prayers and Gratitude to Allah for blessing me and for enabling me to proceed with my learning continuity to further benefit my family and country, I'd want to express my gratitude to my mentors, Dr. Ehsan Warriach for this Study and Dr. Bassem from the Dubai Police Mental Health Department for providing the required resources for this research, for their continuous advice, support, and cooperation. Also, I'd want to thank Dr. Sanjay Chair of the committee from the Rochester Institute of Technology for providing me with the needed facilitation and supervision.

Abstract

Psychology is the Study of Human Behavior and the various ways of thinking in it’s both consciousness and unconsciousness mind as this field is still growing to various fields such as Sports, Criminal and even Administrative Psychology through, Furthermore, This Study aims to shape and accommodate both fields of Sciences together, There are very few Scholars and Researches done on Medical Mental Health from a Data Analytics perspective, examining the datasets of the patients ICD-11 or DSM-5 scores and correlating it with the patients well-being, family medical history or any accident the patient have been through, so the current study approaches these strategies and enable the practitioners to discover the patients vulnerability of mental illness through data analytics, the literature review represented in this research provided an insight of the world practices and methodological examinations are being executed in approaching the patients through their care centers and illustrating the severeness of the Mental Illness the revising the risk factors that might lead the patients to even worst conditions, The Capstone Projects will analyze the influence of pandemics on the population's mental health, The Second question will reveal the most important characteristics that may turn a mentally healthy individual into a mental patient. Data Analysis and Data Mining Methods were used to translate Tweets and Internet Searches of Human Interactions into data to detect their Mental Health Condition and assess the inmate's greatest causes to encounter any sort of mental illness.

Keywords: Dubai Police Mental Health Department, The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), International Classification of Diseases, eleventh revision (ICD-11), Psychology.

LIST OF FIGURES:

Table of Contents

Chapter 1

1.1 Introduction

This Study explores the aspects of mental disorders through different point of views to work closely with both medical practitioners and the faculties managing these institutions, analyzing the impact and severeness of the nature of pandemics as the study will also add touches to the circulation and effectiveness of hardships and limitless waves the pandemic of Covid-19 had escalated on the nature of the human being, The Capstone Project will shed light on three issues by examining the effect of pandemics on the population's mental health wellness. The Second Question will shed light on the most important aspects that contribute to a mentally healthy individual developing any sort of mental disease. Data analysis and data mining techniques were used to convert descriptions of human life interactions to data in order to illustrate the inner behavior of humans, as well as to translate tweets and google searches of human interactions in order to ascertain their mental health status and to analyze the inmates' strongest motivations for facing any type of mental issue. The DSM-5 and ICD-11 are frequently used by psychiatrists to identify and interpret different patterns of behavior in order to determine whether there is a match or association. However, through in-depth research, we will determine the likelihood that a patient will be vulnerable to any mental health issue. Companies frequently make administrative decisions to increase the happiness of their employees to achieve higher rates of success at work. In this study, we will investigate the relationship between employee happiness and workplace efficiency, as well as the factors that contribute to this improvement.

1.2 Project goals

This Study aims to minimize the time taken to diagnose patients that usually takes practitioners from 12 months to a lower time period of 3 to 6 months using machine learning algorithms, while

supporting the shifting of the medical field of psychology and enabling researchers and management of mental institutions to increase and absorb the data at hand to be to predict mental ill patients through their practices and worldwide actions such as the Covid 19 pandemic, increasing the use of machine learning instead of current applying ordinary medical practices, as well as discovering the most affected type of people and levels that suffered in the of the pandemic as well as analyzing the reasons and backgrounds that would lead to the growth of mental diseases within the humans life through using machine learning to well investigate the external and internal environment of the patients, in turn psychiatrists will be able take advantage of the methods and strategies used in this research to better illustrate studies and support the mental health of human beings worldwide

1.3 Aims and Objectives

It was the goal of this project to optimize the employee's mentally to increase work efficiency, Reduce the time taken to analyze the reasons behind any decrease in the happiness indicator of any organization, while the top management of any organization or institution must strengthen continuity planning and decision-making support for the company or institution to succeed, Optimize the employee's mentally to increase work efficiency.

Predicting the number of inmates vulnerable to any mental health issue as well as investigate the reasons behind actions and diseases that cause the mental health illness, and lastly, we can identify and analyze the impact of pandemics on the general mental health of the population on a worldwide scale using artificial intelligence and machine learning, based on the data we have available.

1.4 Research Methodology

As a result of the findings of this study, the collection phase was from three sources first one, is a survey done in a private company with a total of 1260 employees while the second, is extract the data from the twitter feed and google trends to analyze the impact of the corona virus on the overall population on a worldwide standard, and the third source of data is from the Dubai Police Punitive and Correctional Institutions to analyze the reasons that caused the inmates to face mental health issues based on the Mental Disorders Diagnostic and Statistical Manual.

In this research we first applied several models and visualizations to analyze the reasons and variables behind increasing any type of mental health issue in a working place in goals to provide the tech company with actions to tackle it,

As a result of this, we applied a variety of models such as Tree Classifier, Support Vector Machine, Random Forest, and K-Nearest Neighbor (KNN) to analyze which models were the most appropriate for our study and goals. We also analyzed the accuracy of each model using the ROC curve in order to determine which model should be used on the data from the tech company.

Making use of the datasets obtained from Twitter and Google Trends, we looked at the placement and timing of each term reported by individuals seeking help for either the beginning phase of a mental health problem or even for taking medicine to treat that particular mental health issue, As a result of the analysis of past, present, and future data to assist us in identifying the gaps and severity of mental health issues during the corona virus pandemic between the years of 2019 and 2021, a heat map was created for it as well as the maximum and minimum usage of keywords illustrated was determined.

Using data from the Dubai Police Punitive and Correctional Institute, we analyzed the DSM-5 score of each patient along with their specific age and diagnoses to create a correlation matrix that will assist us in identifying key reasons for mental disease as well as shed light on the most focused age demographic for each type of mental health issue.

1.5 Limitations of the Study

To build an accurate data mining model, enough data and variables should be available to train and test as there were limited data to train the model while the more data available, the more accurate the model will perform. In this project, the mental analysis used was conducted with more than various number types of variables, resulting in different correlations and accuracy scores through the datasets. However, the number of samples for each type was not consistent, as some analysis required fixed data such as DSM-5, ICD-11 scores and external reasons such as wellness programs or workplace benefits.

Chapter 2 – Literature Review

2.1 Introduction

Psychology is a study that is related to human behavior so turning that field of science into data analytics would help serve the field much better and provide a wider understanding from a different point of view for other practitioners, the Covid-19 pandemic introduced a new understanding and environmental change in both the existing patients with severe mental illness and vulnerable patients to be heavily affected with severe mental illness, The Covid Virus 19, SARS-CoV-2 virus-induced severe respiratory disease, was found on December 31 in Wuhan, China, and afterwards spread worldwide., As infections increased in early March, a new law was passed that mandated the closure of major gathering places such as schools, colleges, theaters, and stadiums, while recommending all wise working practices such as maintaining a safe distance of 1 to 3 meters between individuals. (Bonichini, 2021)furthermore an analysis can be done as a text analysis on any social media platform using the classification models presented in the study to be classified into two kinds as simpler classification by combining the neutral and negative classes through supervised learning algorithms examples such as Naive Bayes, Decision Trees, and SVMs or even utilizing Confusion Matrix to examine the systems' ability to categorize text from the dataset in this research.(Pilgrim,2014)

Everything starts and initializes with visualizing that is where it gets its importance, digging through the roots of the data and mind for decision-making.(Pratt,2019) Data has the ability transform any unstructured, semi-structured or predictable models making relational databases harder to maintain as many platforms such as social media have billions of file interactions within seconds (Tariq,2021) Data Visualization can be used as a tool to better understand problems and implications even in engaging researchers to better understand their studies, as data visualization assists all parties of an institutions ease understanding.(AlHaddad,2018) investigating the limits of data in many research topics and the processes of it as collaboration is found on all sciences and statistics, with social media's vast amount of data considered as a psychological bridge into the public opinion and behavior. (Harlow, 2016) As the total world

population increases algorithms are fed with even more data that will cause more obstacles and problems that will require various data analysis methodologies. (Stieglitz,2016).

The study serves focuses mostly on millennials the age vary between 18 and 30 as to be the future workforce of our countries which illustrated a lot of caring and revision to put upon, as they are the people that would suffer greatly from high levels of loneliness and disturbance with low distress tolerance although people with anxiety and depression were able to build resilience as to having family members and peers around them to comfort them during the pandemic, except for the PTSD patients that suffered an in depth symptoms during the pandemic period with factors related to being alone.

Factors do affect existing patients with PTSD and ones who suffer both anxiety and depression, when shifting their studies from school to university as taking Covid-19 tests and remote studying was imposed that was somehow abnormal for most people to endure. (Liu, 2020). The study was able to analyze the factors that was associated in growing the mental illness through making use of the idea of the time periods taken to both treat the patient as well as giving the number of weeks and months the patient undergo mental illness within a timeframe that correlated reasons differently to discover without the timeframe variable as there's many similar studies that have not been taking the time variable into consideration or was able to include minor information about time that did not show a very high accurate result.(Wardenaar,2021)

The study had ineffective variables that was first introduced through using longitudinal course measures, that had itself its own psychological measurements such as IDS-SR and BAI scores, the methodology used had a quarter of the people taking and undergoing research that did not have any psychological disorders in their past medical history which will in turn illustrate uneasy results in general while the rest have shown diagnosis of dysthymia, major depressive disorder (MDD) or an anxiety disorder, although the research did acquire important variables that had a high probability of affecting the results very well out of 152 determinants.

The most remarkable limitation was that the study was based on present or absent of discrete diagnosis overtime, there was an unclear comparison between depression and anxiety when they are very strongly related, by which it gave more importance by prioritizing the smaller variables instead of the main determinants which didn't give out the right objectives and results, In putting large sets of data and having many variables while associating most of them together helped gain

stability for the algorithms although using different machine learning algorithms is proven to showcase different results, and the study has given out recommended algorithms to be used and the difference between the continuous variables and the fixed variables. (Wardenaar,2021).

Biological studies are able to connect attributes whatever size and shape it can take, so there were no genomics studies that started with a small number of attributes until it gains ability through the process of time to attain extra attributes to be added to the algorithmics studies. (Bacardit,2014). Responsibilities mentioned in the risk-based analysis gave technical grounding for technological ideas and approaches with phases including descriptive, predictive, evaluation, learning and collaborative risk analysis as using The Simpsons paradox to test the solid benefit and suspension of advertisement spending through a commercial company. (Cox Jr.,2018). Many studies are inevitably concentrating on workplace and employee characteristics within Britain and France that showcased extensive analysis that shows the fundamental demographic features of the survey groups presented by both size and precise categorical sectors with focusing on small businesses account for the bulk of jobs in any industrial economy (Amosse,2016). As some studies have presented evidence of workers taking the responsibility for their negative actions due to their Behavior, As it is one of the several elements that contribute to an accident, including environmental, engineering, managerial, cultural, and even person-states. Behavior is the result of a complex interaction of external and internal variables, and recognizing these variables helps develop and execute successful behavior modification interventions. (Clarke,2016)

The study has strong results because it was able to cover a big number of volunteers although the variables were very short in comparison to other studies having 6 strong variables only, along with the examination that was done from patients that have been already diagnosed with non-chronic depression in hopes to find strong factors that would lead these patients to chronic depression, the study discovered ways to analyze the identification of chronic depression despite its toughness and hardship as even practitioners were able to raise caution of chronic depression through increased healthcare utilization, hospitalization and higher probability of disease. (Hölzel,2010)

There was an unclear comparison between depression and anxiety when they are very strongly related by which it gave more importance by prioritizing the smaller variables instead of the main determinants which did not give out the right objectives and results. Inputting large datasets of data and having many variables while associating most of them together helped gain stability for the algorithms although through using different machine learning algorithms is proven to showcase different results, and the study has also given out recommended algorithms to be used and the difference between the continuous and fixed variables. (Hölzel,2010). Representations to learn straight from an array of data with thousands of characteristics without categorizing the data to accommodate with the universal standards that may cause a loss of inaccurate data. (Bacardit,2014).

As the variable that were analyzed through (Wardenaar,2021) had diseases severity such as worrying (PSWQ), anxiety severity (BAI), or depression severity (IDS-SR) on the other side (Hozel,2010) was able to include variables such as diagnostics scores of DSM-3 to 5 or ICD-9 to 11 that was mostly dependent on international standards examinations results that the study included both internal and external variables.

Various exams and methodological techniques were discovered to exist in extracting the data through multiple ways such as case control, cross-sectional studies or cohort studies based on the severity of the patient's disease. (Holzel,2010) care centers executed a cohort study through the ethical standards and authoritarian board approvals to analyze the populations data screening for depression and anxiety disorders from the first phase of treatment, polygraphs, interviews, and other tactical uses of the field. (wardeenar,2021) in order to identify the most essential risk variables for offending as well as empirically valid, rules by which to modify those risk factors that are possibly modifiable in a dynamic way. Hundreds of rehabilitation programs have been implemented and evaluated using RNR principles, which have been backed by several studies and meta-analyses (Bonta & Andrews, 2016). The RNR approach has encouraged jails to target their services to those who need them the most.(Bierie,2017).

Inmates get affected heavily by the decision made from the beginning of getting caught by police enforcement till the end of the process of the inmate getting a formal legal excuse to leave

prison, as investigation starts through self-reported mental health history as well as face to face interviews with each inmate. (Edgemon,2019).

The measurements taken through that made depression a stable dependent variable in the data algorithm. (Kimberly,2019). The study utilized risk factors connected to depression or even resulting with suicide through the inner and external environment of the prison such as if the inmate house is less than 50 km away it decreases the probability of depression and other examples such as having TV inside the cells plays a key role in decreasing the vulnerability of acquiring any type of mental disease. (Edgemon,2019).

A different analytical strategy has been developed to utilize the factors analysis and path analysis as algorithms to increase accuracy of the models, a credit score system was created for the inmates to provide them with a chance and supporting reasons to issue inmates with parole it will showcase better results as a more in-depth investigation is taken place to analyze the reasons and factors that would an inmate to get an approval of a parole, whereas a big number of inmate parole requests were analyzed and correlated for both accepted and rejected paroles (Jason,2011)although the analysis taken place had more individualistic view and the same data was extracted while using linear models.(Edgemon,2019) The addition of descriptive data allows the differentiation of races.as the white race was the majority in prison receiving a parole, although there was no effect on inmate race when associating it with depression.(Bierie,2017).

The results provided after fitting the models for prison analytics most studies have proven that inmate race and age had no direct effect on depression despite being homeless and unemployed prior to be in prison is what ignites and escalates the levels of depression this is a good indicator showing that most inmate don’t face any mental illness during prison time mostly for the violations and self-harm they have attempted to reflect on themselves despite the models showing that no TV in cells and less security in prison lowers depression with very small effects in the overall mental health of the prisoner through using a hierarchal linear modeling.(edgemon,2019)

Although the difference that was made in this study was that depression and hostility was associated with different risk factors and showcased different results, so the researchers would usually round the number to the closest and accurate answer, psychologists plan, implement and monitor prisoner treatment programs, crisis intervention and therapy under the idea as

correctional psychologists, the assessment done in the study helps test ideas and formulate policies. (Bierie,2017).

Lastly, the science of thinking and behavior can be translated to data as a data mining model must be created in order to find patterns of activity both internal and external. It is also possible that the model will identify, differentiate, and classify many kinds of variables, there are numerous DSM-5 and ICD-11 data mining principles used in the studies above to help us better understand the human life's psychological sequence and identify the identifying and prescribing pretreatment strategies to avoid psychological dysfunction are the most effective justifications in our working lives today, while, experimenting the credibility of analyzing personal information illnesses to uncover origins and trends, as well as increased population would cause an increase in social media platform users, which can lead to an increase in interactions between individuals which is exactly what is required to assist organizations and businesses in identifying obvious issues and solutions, The merge of behavioral and thinking sciences with the science of data will allow psychologists and researchers to have a deeper understanding of their area and benefit the most value from it as more variables will result in more accurate and reliable outcomes.

Chapter 3- Project Description

3.1 Project Overview

In this Study Research Project initiates the investigation on the current practices and methodologies being applied in the medical mental health industry, involving data selection, data cleaning, data characterization, data preprocessing, data mining and data classification of the world population through their interactions on social media along with an analysis and prediction on the behavior of employees in a technical company, as well as the current status of the mental health of prison inmates in the Dubai police through the punitive and correctional department, the results incorporated in this study indicates that the algorithms built in these models was a success in extracting the problems associated to the creation and birth of mental illness as well as predicting the number of mentally ill individuals, as in this study were able to differentiate and cast the light on the comparison studies to better understand the effectiveness of Diagnostic and Statistical Manual of Mental Disorders score on its fifth version(DSM-5) and the receiving mental health assistance.

A number of phases were engaged in the data analysis, such as data selection, data cleaning, characterization, data preprocessing, data mining, and categorization of data. the project's data mining model was able to correlate and separate the three kinds of data, which resulted favorably on the model's capacity to recognize and classify accurately, according to the findings of the study.

This research used numerous models and visualizations to investigate the causes and factors underlying developing mental health issues in the workplace to help the tech business take action using models like Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbor to determine which models were best for our research. We utilized the ROC curve to assess each model's accuracy and choose the best model for the tech company's data.

Using Twitter and Google Trends data, we looked at the placement and time of each phrase reported by people seeking care for a mental health condition or taking medication to address it.

The examination of historical, current, and future data let us detect the gaps and severity of mental health difficulties during the corona virus pandemic between 2019 and 2021 and data from the Dubai Police Punitive and Correctional Institute, we created a correlation matrix that will help us uncover important factors for mental illness as well as the most targeted age demographic for each sort of mental health condition.

Chapter 4- Data Analysis

4.1 Project description

The current study analyzes the views and methodologies that can be applied on the human mind to enable practitioners of the future to analyze and predict the mental health through utilizing the number of times the keywords of depression and mental health assistance were issued through two social media platforms, as people to share what they feel almost every minute in order to get feedback or support from friends and family, thus this data can prioritized for research, as well as the data obtained for the employees tech company allows for the digitization of these individuals' working environments in order to better understand their present and future mental health condition, To have a better understanding of the convicts, data was obtained from them along with their present mental health state and the factors that contribute to mental illness while in imprisonment based on age, country and DSM-5 score to improve decision-making processes for upper management and use a high degree of accuracy to handle the psychological aspects of prisoners.

4.2 Data Selection

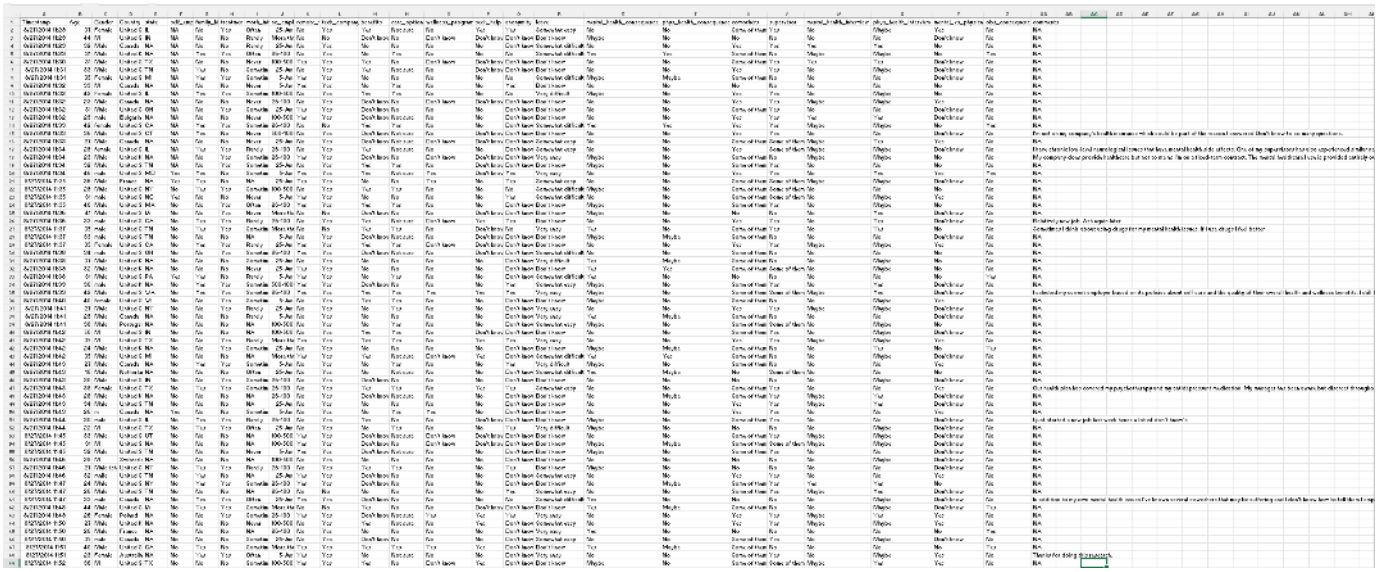


Figure 1: Display of the data matrix in the company that shows the first 27 attributes.

Data were obtained from google trends and twitter from 2019 till 2021 to extract specific keywords relating to the attempt of suicide and certain words that convey grief sadness and depression to analyze the status of the population mental health before and during the outbreak of corona virus, along data gathered from a tech company with 27 attributes relating to the current wellbeing of the employees and the state of receiving mental care or no in addition to 1260 rows, lastly, the inmates data were gathered with 6 variables and 126 inmates from the Dubai police correctional institution to analyze the current and future status of the inmates mental health through 6 attributes relating to the DSM-5 score and severeness of the mental illness faced by the patient, by first importing all of the essential python libraries into Jupyter that is the most extensively used library among data scientists, therefore it is imported it using Python's import statement. Scientific computing relies heavily on NumPy, a standard Python tool, which is also good for additional analysis identifying and replacing any null or missing values is the third step in the data cleaning process.

4.3 Data Cleaning

```
train_daf = train_daf.drop(['comments'], axis= 1)
train_daf = train_daf.drop(['state'], axis= 1)
train_daf = train_daf.drop(['Timestamp'], axis= 1)
train_daf.isnull().sum().max()
train_daf.head(5)
```

	Age	Gender	Country	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	...	anonymity	leave	mental_health_consequence	ph
0	37	Female	United States	NaN	No	Yes	Often	6-25	No	Yes	...	Yes	Somewhat easy	No	
1	44	M	United States	NaN	No	No	Rarely	More than 1000	No	No	...	Don't know	Don't know	Maybe	
2	32	Male	Canada	NaN	No	No	Rarely	6-25	No	Yes	...	Don't know	Somewhat difficult	No	
3	31	Male	United Kingdom	NaN	Yes	Yes	Often	26-100	No	Yes	...	No	Somewhat difficult	Yes	
4	31	Male	United States	NaN	No	No	Never	100-500	Yes	Yes	...	Don't know	Don't know	No	

5 rows × 24 columns

Figure 2: data cleaning and excluding NAs from the tech company dataset.

Then came data cleansing, a variety of techniques have been used to identify and eliminate/replace null values. Null values have been swapped out for the mean and median of the corresponding numerical variables. Null may be replaced with the most common category or a new category can be created for categorical variables, We checked for duplicates and missing columns and rows. There were no "na" collected through the distribution, and no similar columns or rows.different steps to clean different data columns is performed as excluding missing columns or cells, as well as removing the (Na) values, further we have classified the data to remove alike values and make it one

4.3 Data Characterization

The data were divided into 27 variables based on the standard metadata. Multiple categories depending on age distribution and density, treatment classification, family history and going through a wellness program to perform the likelihood of mental health problem. Thus, There is now a new column in the data matrix called "count" to reflect the number of individuals. Defined

training and calibration data matrices, Charts were made to see the data the rearrangement in distribution of employees by age.

Although the Twitter and Google datasets were classified and filtered according to our specific requirements, there was an increasing inability to discover any term connected with the sense of any mental health condition in the Twitter and Google datasets.

4.4 Data Preprocessing

The Analysis shown below describes the Mental Health Status of 1260 employees in a tech company during 2014 study that assesses attitudes toward mental health and the prevalence of mental health issues in the IT industry is included in the dataset below, in this task a dataset is provided and required to find the prediction of health issue with the given set, the datasets illustrated below went through a process of data cleaning prior to executing any data analytics algorithm to increase the probability of acquiring better accurate results, while datasets of both twitter and google can be shown in figure 17 and in figure 27, showcases the data for the inmates in the punitive and correctional institutions in the Dubai Police.

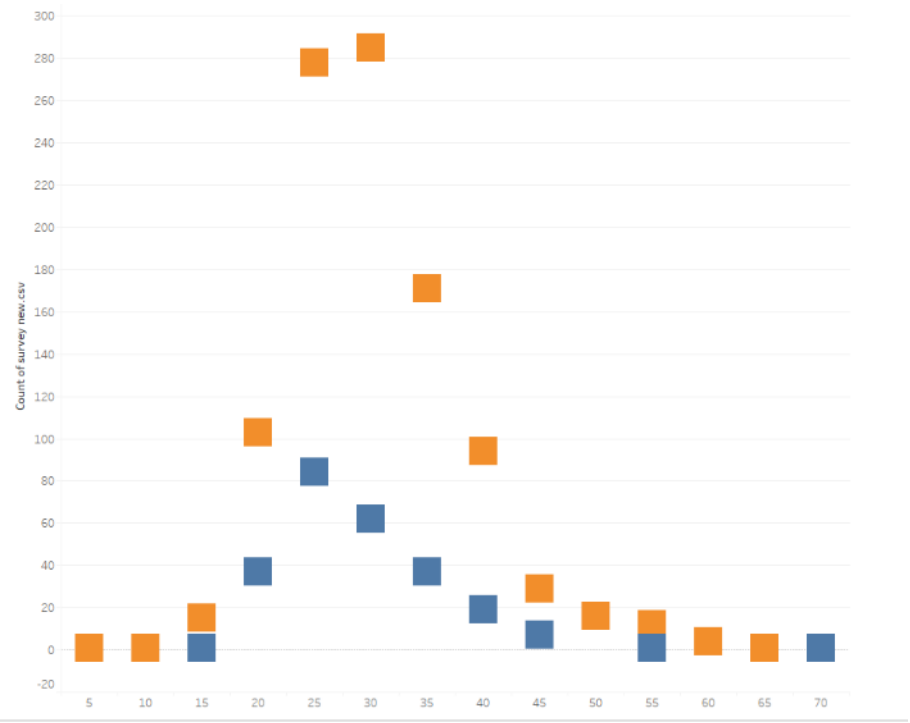


Figure 3: A bar chart that displays the distribution and density of all employees based on age and gender.

As the chart above showcases the age and gender differences within the datasets as the highest point in orange is the categorized male employees that make an average age between 30 and 35 while being count for between 280 to 300 of the employees of the company.

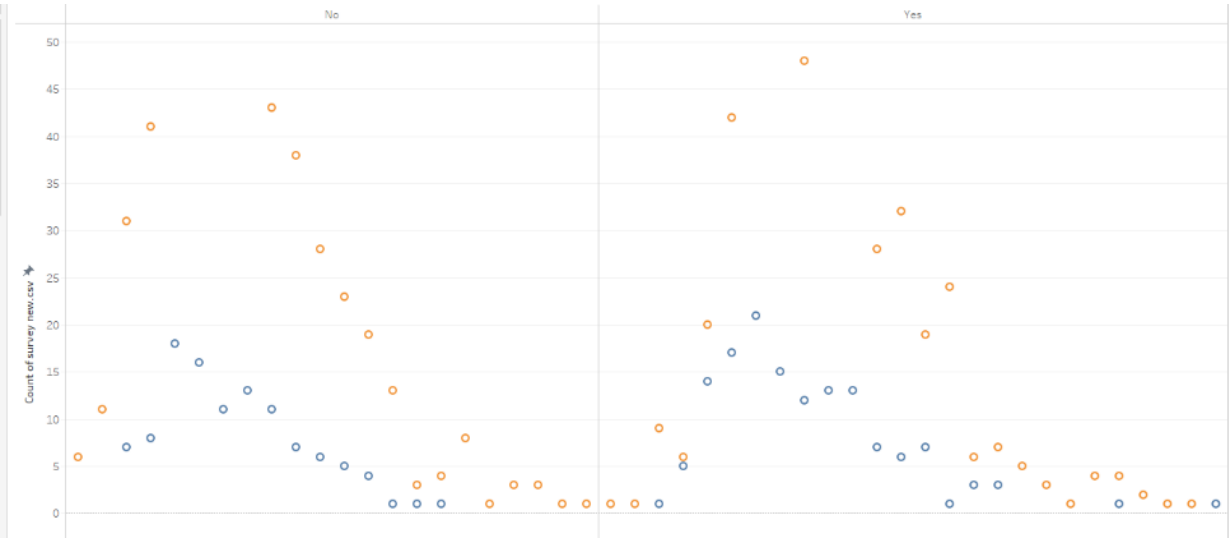


Figure 4: A bar chart that displays the average ages of employees receiving treatment

We looked further into the typical ages of workers undergoing treatment in the chart above to continue the process of gathering explanations for mental health difficulties as it appears that males mostly receive treatment for their mental issues between the age of 24 and 30 while female account for mostly between 24 and 26.

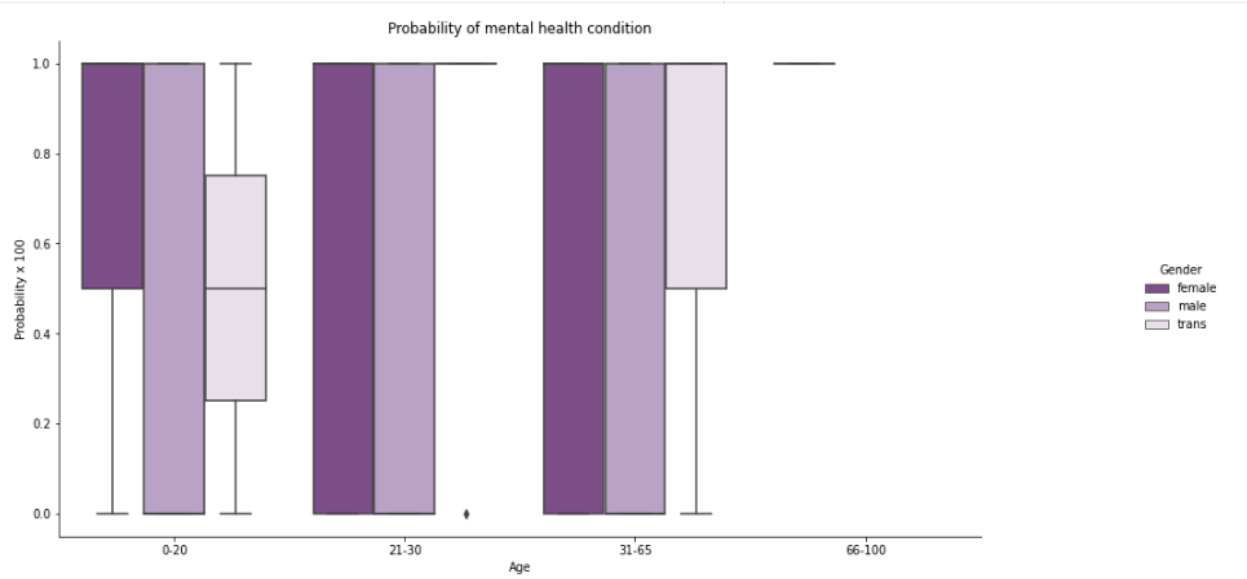


Figure 5: The bar chart illustrates the likelihood of people developing any mental health disorder as a function of their gender and age.

With both figure 4 and 5 we are able to analyze that employees between the ages 25 and 30 are less possible to acquire any mental issue and are more intending to receive treatments while the employees between the ages 30 and 40 are more likely to receive mental health issues and even more likely not to visit a practitioner for treatment.

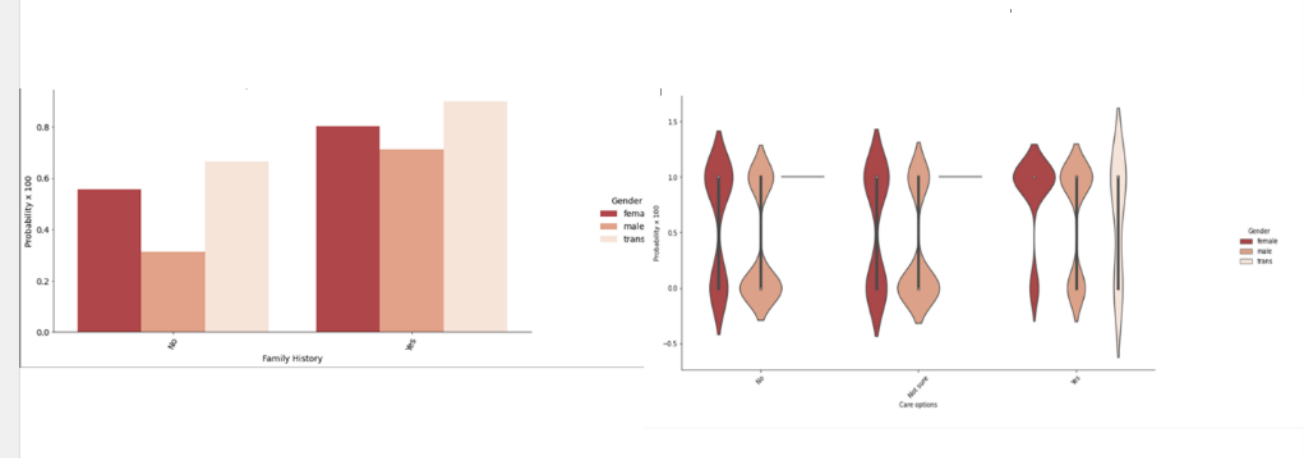


Figure 6: the chart on the left is the occurrence of a person's family history of mental illness is considered, whereas the chart on the right is the occurrence of when an employee's company offers mental health services.

The bar chart above depicts the likelihood of a mental health issue occurring because of the previously medical mental health issue condition in the family, whereas the chart on the right depicts the likelihood of employees developing a mental health condition as a result of the care options available to them within the organization. And after doing the probability test on the data set, we get some charts that illustrate the chance of mental health vulnerability in different genders and ages, Additionally, we are reviewing documents to see whether the patient is dealing with familial concerns that impact their mental health and their ability to get effective therapy.

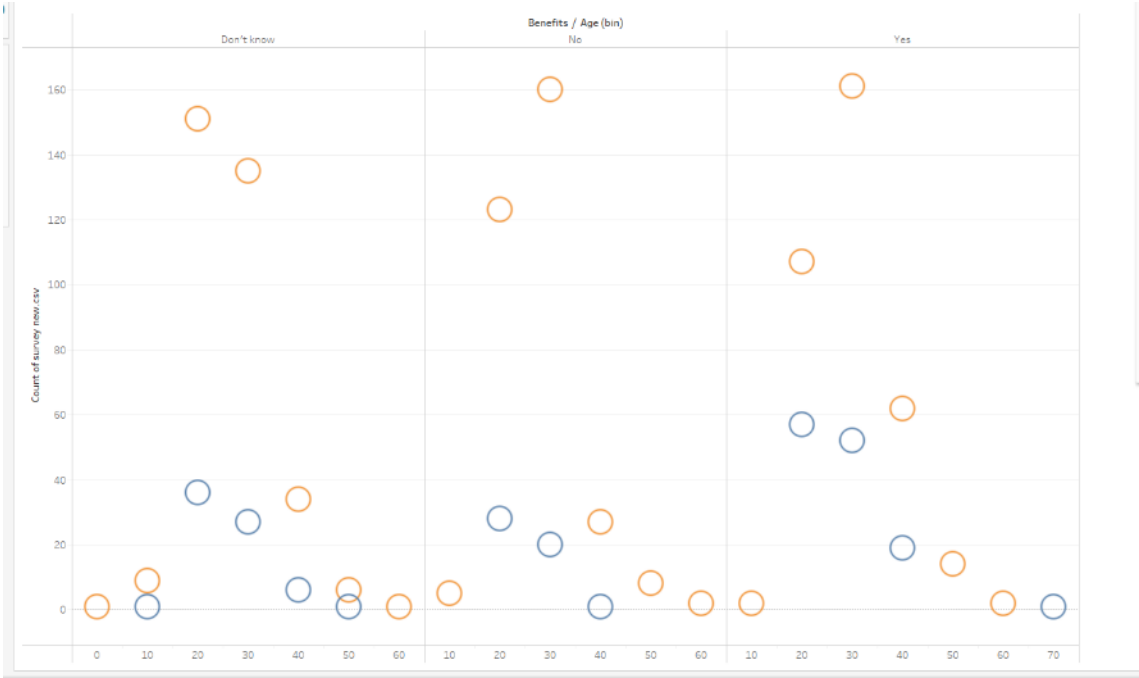


Figure 7: The figure above illustrates the likelihood of experiencing a mental health issue according on gender and benefits.

Following that, more study is conducted, and the dataset is re-examined to determine if they are aware of the advantages offered by the workplace that may have an influence on those suffering from a mental health condition or not, after doing the data analysis, it is apparent that very few of them are aware that the work interface is producing the mental health problem.

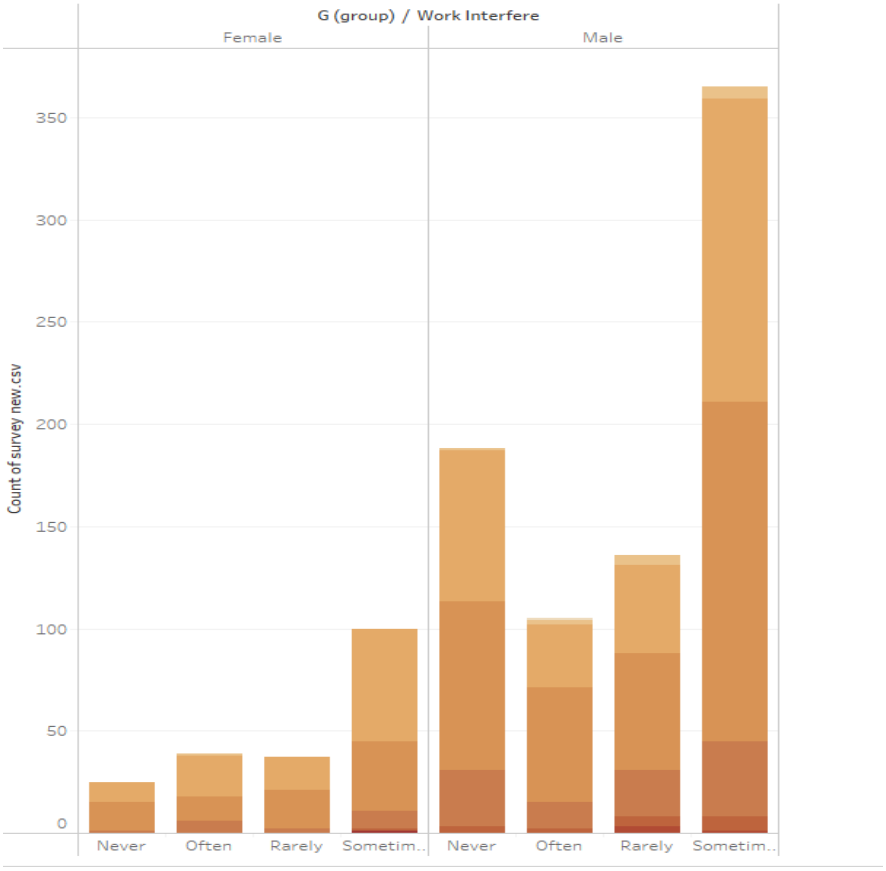


Figure 8: The bar chart above illustrates the likelihood of experiencing a mental health disorder as a function of job interference and gender.

As we can see, age is a significant element in our data set, and we conducted various tests to demonstrate its significance, by which the darker the color of orange the older the age is and the lighter the color orange gets illustrates the younger percentages of employees in the company by count and work interfere.

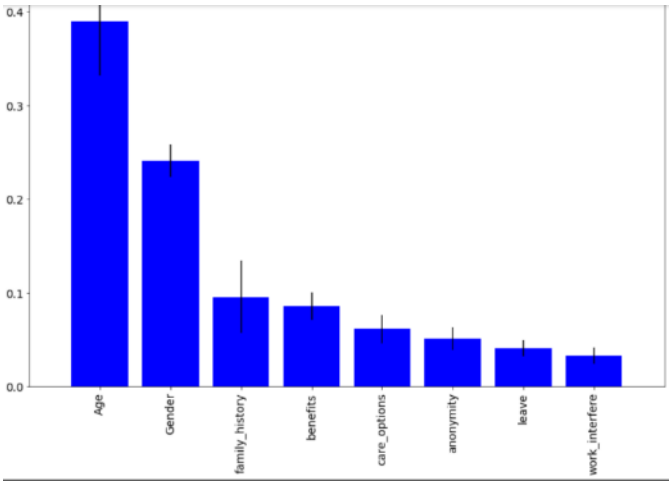
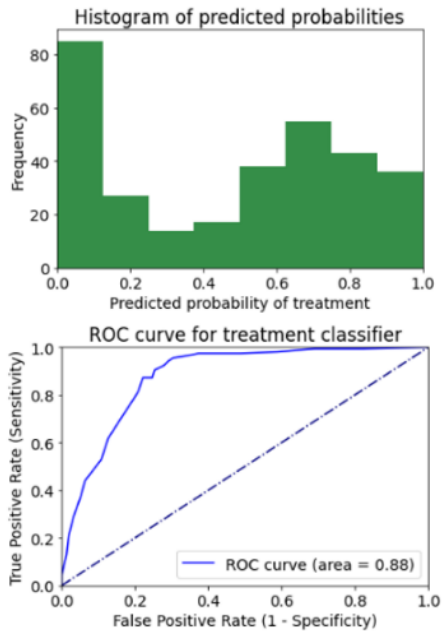


Figure 9: The figure above illustrates the single most critical data variable in the dataset that influences mental health.

The work is being done to analyze the risk factors that contribute to mental disorders in employees as well as to track their receipt of treatment and attention in the workplace, through specific attributes to understand the reasons correlating over the employees' mental issue, we have tested and run five different algorithms in order to determine which model is the best at predicting mental health issues.

K-nearest neighbor



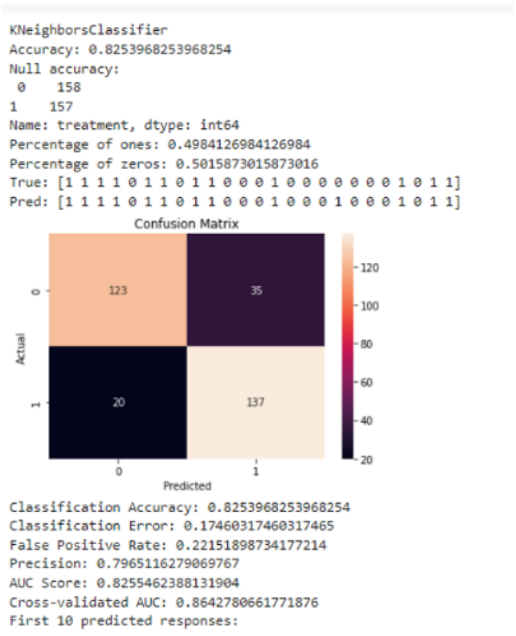


Figure 10: the observations above showcase the accuracy of the k-nearest neighbor.

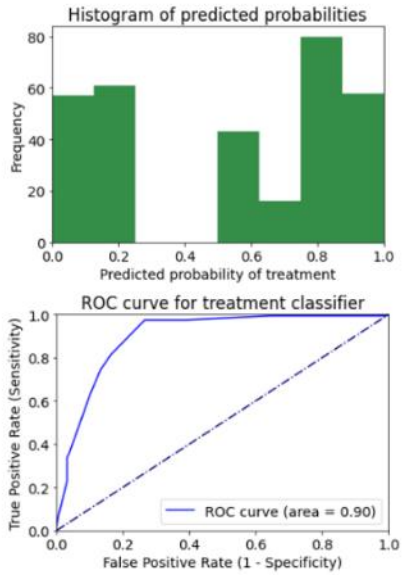
Confusion matrix has 4 different colored boxes that were created these with heat map the boxes represents values such as true positive and negative, false positive and negative, these boxes are supposed to showcase us what values are correct and what are not, so for example a true positive mean value is correct and the model would succeed in illustrating a true positive and true negative mean although by which the actual value is to be not true, are typically used to illustrate the relationship/trade-off between clinical sensitivity and specificity for each feasible cut-off value for a test or a group of tests graphically. Additionally, the area under the receiver operating characteristic curve provides insight into the utility of the test(s) in issue, that attained above through KNN model an accuracy of 0.8253968253968254 and an ROC curve of 0.88 which quite the equivalent of the accuracy of the logistic regression in figure (13)

The lines after the bar chart is ROC curve emphasizes on how much the accuracy of the model is achieved thus the higher the lines reach a mean number the more accurate it becomes.

These below algorithms are used to determine whether or not a person has a mental health problem, a total of five different algorithms or models were tested in order to determine which model was the most accurate in predicting mental health issues. Following the training of the models, we discover that the random forest approach is the most accurate prediction algorithm. Then we run the random forest algorithm and generate a CSV file as an output file.

The box with four boxes that you see is referred to be a confusion matrix, and we built them using a heat map, in a confusion matrix, there are four boxes being referred to as true positive and negative, false positive and negative, This box informs us of which values are right and which are incorrect. For example, true positive values indicate that the value was correct and that our model also indicated true, while true negative values indicate that the actual value was correct but that our model indicated false and finally, the lines following the bar chart are the ORC curve, which tells us how accurate the model is; the higher the lines, the more accurate the model.

Tree Classifier:



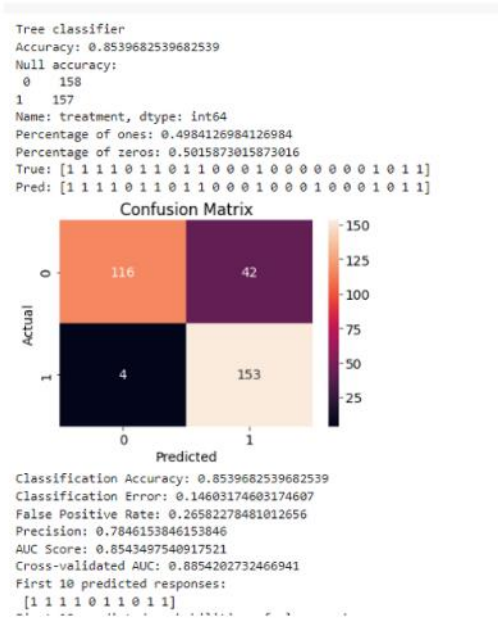
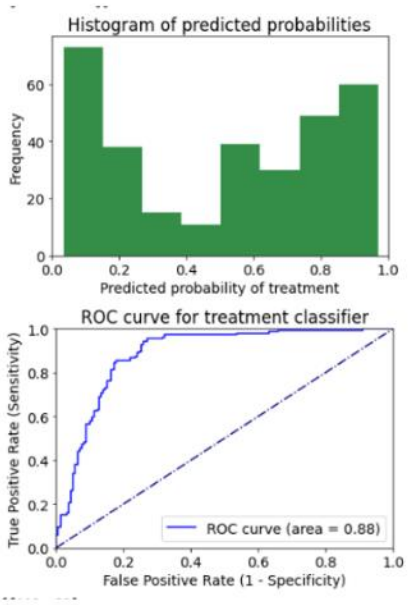


Figure 11: the accuracy model done for the Tree Classifier Forest algorithm.

in the form of a tree structure, generating classifications and regression models as the information is broken down into smaller and smaller pieces, a decision tree is being constructed in the background. Finally, the output is a tree of decision nodes, and has a achieved a classification accuracy of 0.853968253 which has got the highest accuracy between the models as well as the highest cross validation AUC of 0.88542027 and an ROC curve of 0.90.

Logistic Regression:



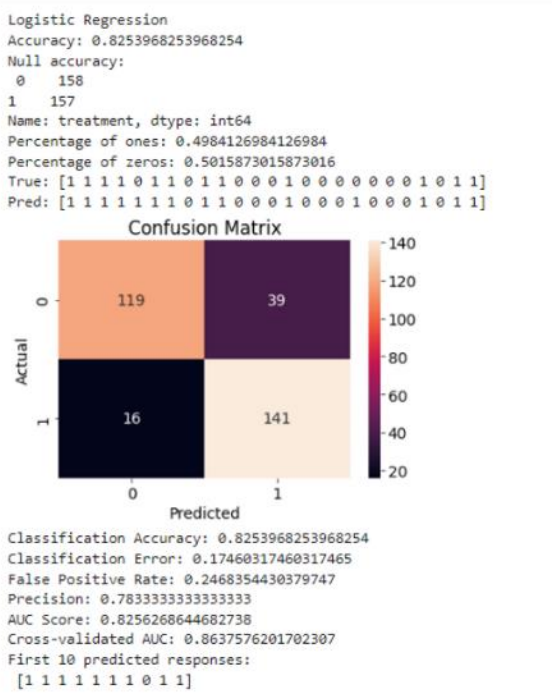


Figure 12: the accuracy model done for the logistic regression algorithm.

The likelihood of a discrete result given an input variable is modelled using logistic regression. Logistic regression is most often used to model a binary result, such as true or false, yes or no, the lines following the bar chart represent the ORC curve, which indicates the model's accuracy, the higher the lines showcasing in the predicted probability using logistic regression will have more accurate results that achieved the following 0.825396825 and an ROC curve of 0.88

Random Forest:

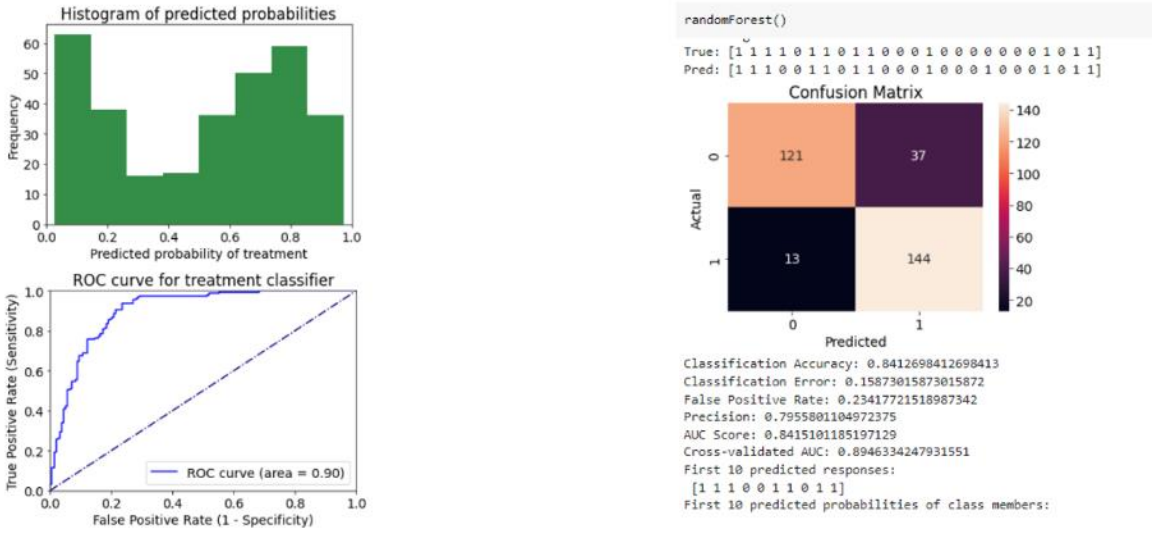
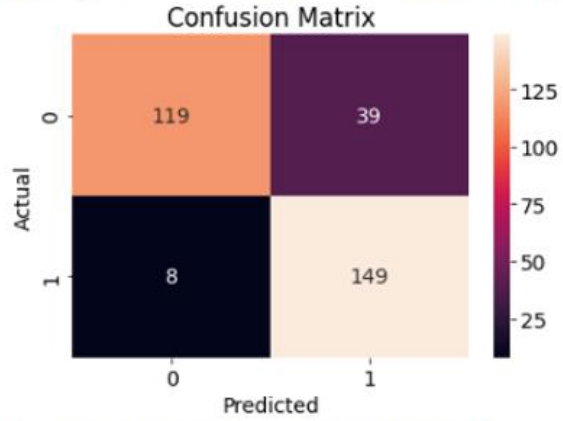


Figure 13: the accuracy model done for the Random Forest algorithm.

Random forest has been utilized extensively for classification and regression issues in supervised machine learning algorithms that attained an accuracy of 0.8412698 For classification and regression, it creates decision trees on a variety of samples and gets the majority vote from each one, the confusion matrix that helps us identify the sensitivity and specificity of the algorithm that clearly shows a 13 that is false negative, To conduct an analysis of the techniques described above in order to prevent erroneous data placement in growing anti-error data.

Support vector model:

```
SVM
Accuracy: 0.8507936507936508
Null accuracy:
 0    158
1    157
Name: treatment, dtype: int64
Percentage of ones: 0.4984126984126984
Percentage of zeros: 0.5015873015873016
True: [1 1 1 1 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1]
Pred: [1 1 1 1 0 1 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 1]
```




```
Classification Accuracy: 0.8507936507936508
Classification Error: 0.14920634920634923
False Positive Rate: 0.2468354430379747
Precision: 0.7925531914893617
AUC Score: 0.8511045714746432
Cross-validated AUC: 0.8461187452504897
First 10 predicted responses:
 [1 1 1 1 0 1 1 0 1 1]
```

Figure 14: the above is the confusion matrix for the support victor machine that displays the accuracy of the modeling.

support vector machines (SVMs) use classification techniques, to solve two-classification issues, the goal is to find the optimum line or decision boundary that can divide n-dimensional space into distinct classes so that additional data points may be readily categorized in the future.

Analysis of risk factors for mental diseases is now being studied forth to contribute towards mental disorders in employees as well as to track their receipt of treatment and attention in the workplace.

The algorithms illustrated above are being used to detect mental health issues in employees of various ages and nationalities. We have tested and run five different algorithms in order to determine which model is the best at predicting mental health issues. After training the models, it was determined that the random forest model is the most predictive algorithm; Because of this, a random forest is utilized, which generates an output CSV file.

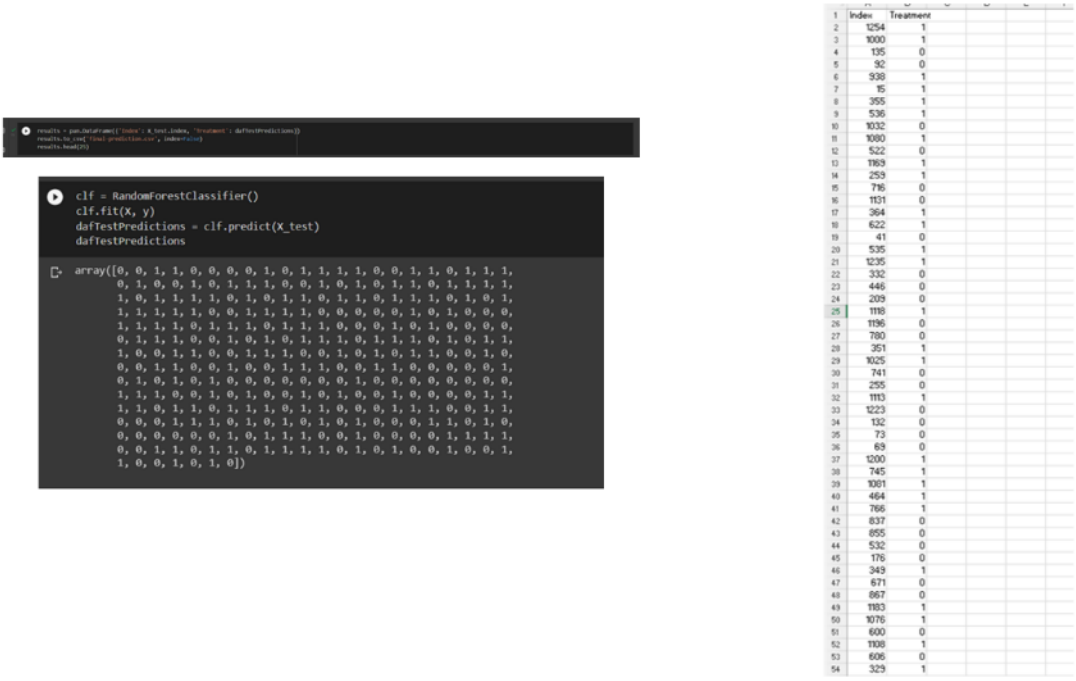


Figure 15: the output data on the left side and the number of treatments resulting from the random forest on the right.

The output data set in the array on the left side was used to calculate the number of treatments shown on the right side. Printing the resulting data set reveals our predictive models. The

outcome of creating an excel file that allowed us to compute the number of treatments received by the organization is included in the dataset below.

Twitter feed

7/28/2019	77	77	76	69	20	32	34	64
8/4/2019	74	75	78	64	22	0	35	70
8/11/2019	74	75	81	70	25	0	35	67
8/18/2019	82	79	79	67	19	31	37	67
8/25/2019	73	76	77	66	19	30	38	73
9/1/2019	72	76	79	70	22	31	36	73
9/8/2019	78	81	81	73	29	0	39	76
9/15/2019	74	82	87	84	23	30	39	76
9/22/2019	73	78	86	80	21	29	38	82
9/29/2019	76	79	88	91	23	0	38	80
10/6/2019	75	82	85	83	25	59	42	99
10/13/2019	73	81	86	85	25	30	40	88
10/20/2019	75	79	92	85	23	29	41	88
10/27/2019	72	78	86	81	22	0	39	83
11/3/2019	76	79	84	79	20	29	42	84
11/10/2019	77	80	86	78	19	0	44	78
11/17/2019	74	82	83	100	20	30	43	84
11/24/2019	82	76	88	69	18	0	36	76
12/1/2019	74	79	82	87	19	58	43	82
12/8/2019	76	76	84	85	18	29	40	77
12/15/2019	74	73	80	73	16	60	38	67
12/22/2019	72	63	78	57	17	0	31	54
12/29/2019	76	72	71	65	18	31	38	57
1/5/2020	81	82	70	78	17	29	44	71
1/12/2020	79	79	79	69	20	27	42	74
1/19/2020	75	78	80	74	19	0	42	76
1/26/2020	86	81	78	78	20	28	41	78
2/2/2020	80	79	77	83	21	28	41	79
2/9/2020	77	79	80	80	22	0	43	80
2/16/2020	84	83	81	86	25	0	43	79
2/23/2020	83	82	85	86	20	27	57	79
3/1/2020	86	79	84	77	19	53	53	80
3/8/2020	81	74	83	81	17	0	45	72
3/15/2020	94	74	76	58	15	0	35	65
3/22/2020	100	81	65	57	17	25	35	69
3/29/2020	97	82	65	67	21	0	38	72
4/5/2020	94	78	70	76	18	0	43	71

Google Trends

Week	depression	anxiety	ptsd	suicide	sleep deprivation	panic attack	mental Disorder	sertraline
6/16/2019	81	87	61	58	75	72	45	86
6/23/2019	76	85	75	60	76	73	44	83
6/30/2019	73	83	63	57	70	73	38	76
7/7/2019	80	90	74	62	77	76	45	85
7/14/2019	80	90	72	65	76	79	44	91
7/21/2019	79	92	74	59	77	78	43	88
7/28/2019	80	93	67	60	79	78	42	86
8/4/2019	78	92	80	59	78	74	45	86
8/11/2019	78	92	52	60	84	79	44	90
8/18/2019	79	95	63	59	86	80	44	93
8/25/2019	81	92	72	63	82	74	44	90
9/1/2019	80	92	67	62	81	76	45	87
9/8/2019	90	96	74	64	79	76	53	96
9/15/2019	93	96	67	61	83	76	53	100
9/22/2019	89	97	76	67	82	76	55	94
9/29/2019	87	96	75	62	80	80	56	92
10/6/2019	92	98	81	67	80	77	100	91
10/13/2019	97	97	73	67	81	73	55	95
10/20/2019	92	97	77	70	81	75	56	94
10/27/2019	91	94	77	65	79	73	52	91
11/3/2019	95	95	89	67	77	73	55	96
11/10/2019	97	95	91	74	75	80	53	88
11/17/2019	98	99	94	74	78	77	55	90
11/24/2019	84	90	78	68	72	79	45	73
12/1/2019	94	93	84	74	80	76	51	89
12/8/2019	93	89	84	67	76	76	46	83
12/15/2019	83	84	67	67	74	74	42	77
12/22/2019	68	77	59	62	73	67	28	66
12/29/2019	78	87	67	65	86	77	35	77
1/5/2020	88	96	69	69	89	78	51	93
1/12/2020	89	94	72	71	85	78	53	94
1/19/2020	89	97	70	70	83	83	53	94
1/26/2020	93	99	75	72	81	86	55	94
2/2/2020	92	95	82	69	81	82	54	95
2/9/2020	95	96	75	67	84	76	55	96
2/16/2020	97	100	83	69	80	82	57	97

Figure 16: the above datasets displaying the occurrences of the specific keywords that convey sadness and depression.

In this task specific keywords were retrieved that were found mostly used prior and during Covid-19 pandemic outbreak as figure showcases the data gathered from twitter and figure showcases the data gathered from google searches on a worldwide standard, that also went through the process of data cleansing and preprocessing, creating a more understood dataset as data quality should be evaluated prior to applying machine learning or data mining algorithms to a given set of information in data mining, we are unable to interact with it until we have this step completed with raw data in this manner.

Separating the keywords from the twitter feed of each tweet while safekeeping the time and place the tweet was executed, as well as extract each google search with the required keywords

that connects the user with other links related to the required treatment and challenges of the mental health issue.

Google Trends Searchs Flow chart

```
# Plot the search interest worldwide.
plt.figure(figsize=(20, 10))
plt.plot(worldwide_data_google.index, worldwide_data_google['anxiety'], color='gray')
plt.plot(worldwide_data_google.index, worldwide_data_google['Depression'], color='y')
plt.plot(worldwide_data_google.index, worldwide_data_google['suicide'], color='orange')
plt.plot(worldwide_data_google.index, worldwide_data_google['anxiety'], color='m')
plt.plot(worldwide_data_google.index, worldwide_data_google['sertraline'], color='black')
plt.plot(worldwide_data_google.index, worldwide_data_google['panic attack'], color='r')
plt.plot(worldwide_data_google.index, worldwide_data_google['ptsd'], color='pink')
plt.plot(worldwide_data_google.index, worldwide_data_google['sleep deprivation'], color='blue')
plt.plot(worldwide_data_google.index, worldwide_data_google['Mental disorder'], color='green')
plt.axvline(x='2020-02', color='brown')
#axvline(x=1)

plt.legend(['depression', 'anxiety','suicide', 'panic attack','sertraline', 'ptsd','sleep deprivation','Mental disorder'])
plt.xlabel('Week')
plt.ylabel('Search Interest')
plt.title('Search Interest Worldwide');
```

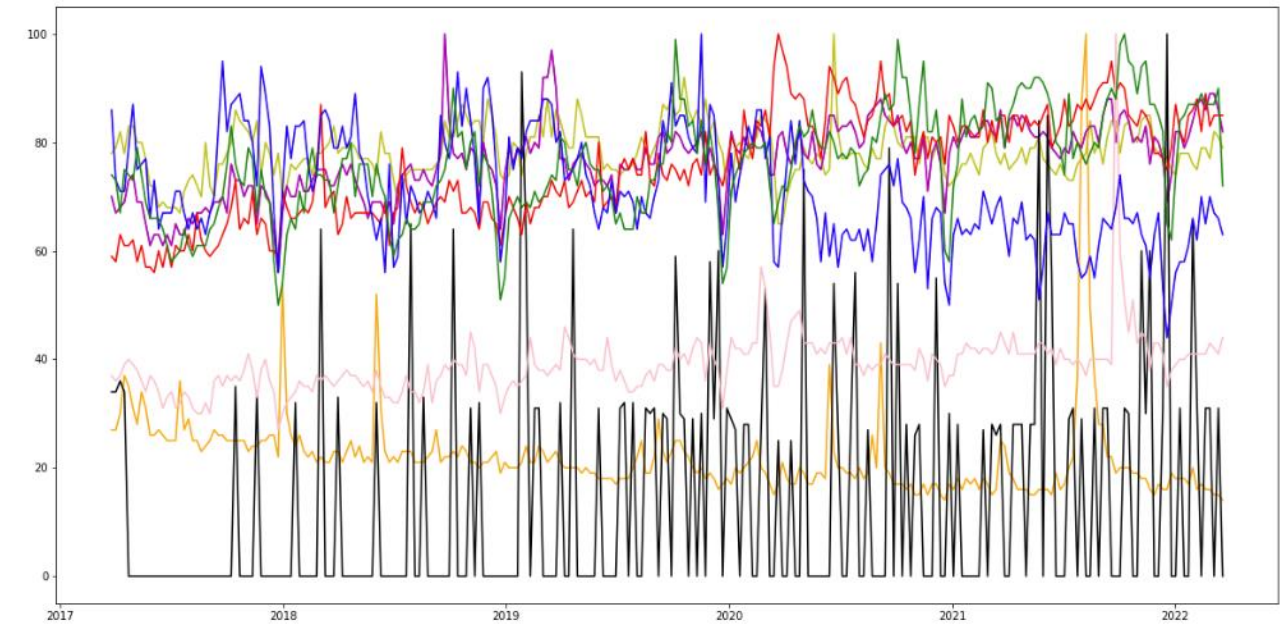


Figure 17: : A line plot that displays the keywords utilization after assigning each a distinct line color

Then, a Sample Grouping feature was used, which allocates a different color to each group in order to facilitate comparison. Figure (3) shows a fresh line plot that was developed and shown.

Plotting the dataset and drawing a red line in the plot to divide it into two halves, one before the pandemic and one after pandemic, is completed when the dataset has been loaded, as the red lone shows a high usage from 2017 till 2022 with a gradual increase and the word “depression” which is the yellow line reaching its peak between 2021 and 2022 after the pandemic that can be shown from reasons relating to enforcing people to stay home, wear masks in some countries, restricting movements and after a very long time of high unemployment etc.

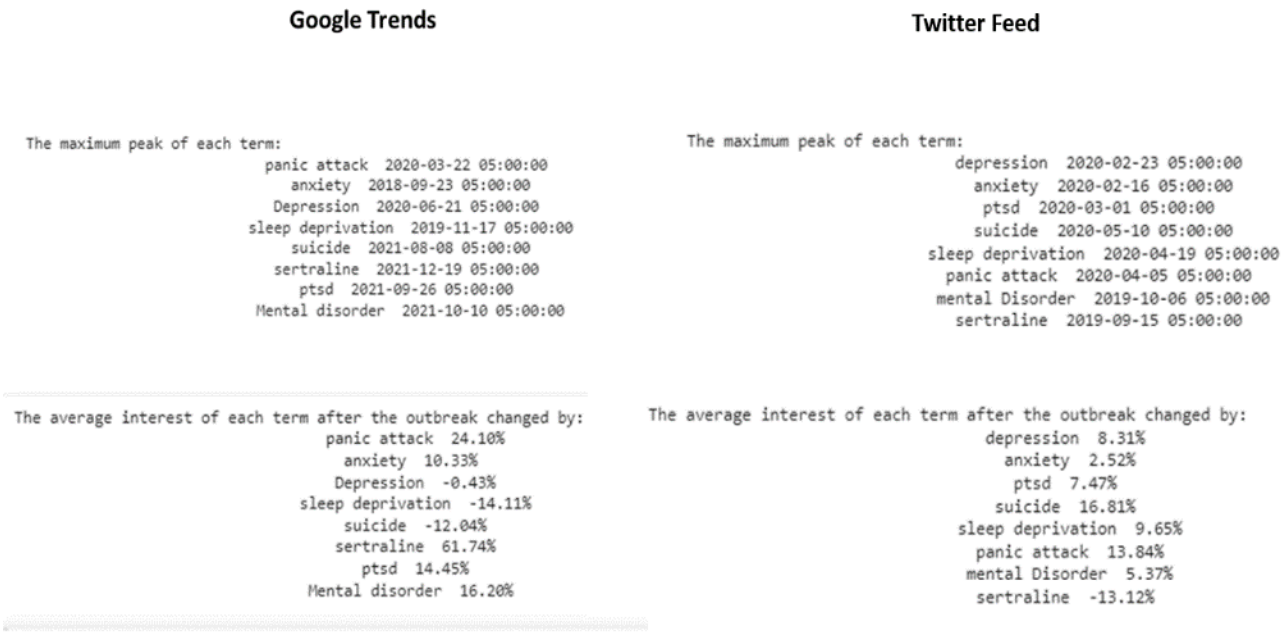


Figure 18: The number of occurrences during a time period was studied to verify for each dataset's maxima on date.

In this phase the number of occurrences within a specific period was examined to check for the peaks of each dataset on date, then finding out how much in percentage difference is occurred over the time that will assist us in demonstrating the impact of the catching the corona virus from those who were seeking help and those who were putting those preparations into action in the worst-case situation.

Google Trends

```
The average interest of each term over the 12 months:
panic attack    75.34
anxiety    76.80
Depression    78.36
sleep deprivation    71.20
suicide    22.09
sertraline    13.77
ptsd    39.02
Mental disorder    76.08

The average interest of each term before Jan, 31, 2020:
panic attack    68.28
anxiety    73.54
Depression    78.51
sleep deprivation    75.79
suicide    23.30
sertraline    10.89
ptsd    36.74
Mental disorder    71.13

The average interest of each term after Jan, 31, 2020:
panic attack    84.73
anxiety    81.14
Depression    78.17
sleep deprivation    65.10
suicide    20.49
sertraline    17.61
ptsd    42.05
Mental disorder    82.66
```

Twitter Feed

```
The average interest of each term over the 12 months:
depression    75.34
anxiety    76.80
ptsd    78.36
suicide    71.20
sleep deprivation    22.09
panic attack    13.77
mental Disorder    39.02
sertraline    76.08

The average interest of each term before Jan, 31, 2020:
depression    68.28
anxiety    73.54
ptsd    78.51
suicide    75.79
sleep deprivation    23.30
panic attack    10.89
mental Disorder    36.74
sertraline    71.13

The average interest of each term after Jan, 31, 2020:
depression    84.73
anxiety    81.14
ptsd    78.17
suicide    65.10
sleep deprivation    20.49
panic attack    17.61
mental Disorder    42.05
sertraline    82.66
```

Figure 19: The graph above depicts the average amount of interest in popular keywords during a certain time period through the pandemic's past and present.

The graph above depicts the average level of interest in important keywords over a period of time spanning the pandemic's past and current state, analyzing the interest of users searching the terms above like "Depression" and "Anxiety" to either tackle and emotions or seeking assistance for it, with the word "Panic" in both Google and Twitter before the pandemic as them being at the low points of 68.28 and 10.89 then increasing to a shocking 84.73 and 17.61 to showcase the use of both platforms for these specific keywords during the pandemic, as well as using Medications for Depression such "Sertraline" being both on 10 and 71 then increasing to 17 and 82.6 to showcase the need for this specific medication and treatment for depression.

Google Trends

```
The variance of each term over the 12 months:
panic attack    96.86
anxiety         45.63
Depression      22.15
sleep deprivation    99.33
suicide         71.59
sertraline     451.67
ptsd           32.93
Mental disorder    93.94

The variance of each term before Jan, 31:
panic attack    36.79
anxiety         43.67
Depression      25.55
sleep deprivation    80.90
suicide         25.79
sertraline     364.21
ptsd           11.92
Mental disorder    61.18

The variance of each term after Jan, 31:
panic attack    22.16
anxiety         15.27
Depression      17.57
sleep deprivation    58.57
suicide        128.02
sertraline     542.24
ptsd           44.78
Mental disorder    61.67
```

Twitter Feed

```
The variance of each term over the 12 months:
depression      56.89
anxiety         20.91
ptsd            90.49
suicide         97.76
sleep deprivation    55.04
panic attack    49.39
mental Disorder    90.03
sertraline     125.65

The variance of each term before Jan, 31:
depression      57.81
anxiety         24.84
ptsd            80.09
suicide         23.63
sleep deprivation    17.88
panic attack    11.88
mental Disorder    123.64
sertraline      53.74

The variance of each term after Jan, 31:
depression      22.42
anxiety         10.20
ptsd            90.17
suicide        156.22
sleep deprivation    85.32
panic attack    45.78
mental Disorder    23.88
sertraline     170.91
```

Figure 20:The illustration above showcases variance of each primarily keywords over a period of time through the pandemic's past, present and future.

Using particular keywords like "Panic Attack," which had a pre-pandemic use of 68.2% and a post-pandemic usage rate of 84.73 percent, the graph above depicts the average interest in major keywords through time, the proportion of variance for the word suicide ranged from 25.7 to 120, with an average of 71.59, during the course of the pandemic's past and present average.

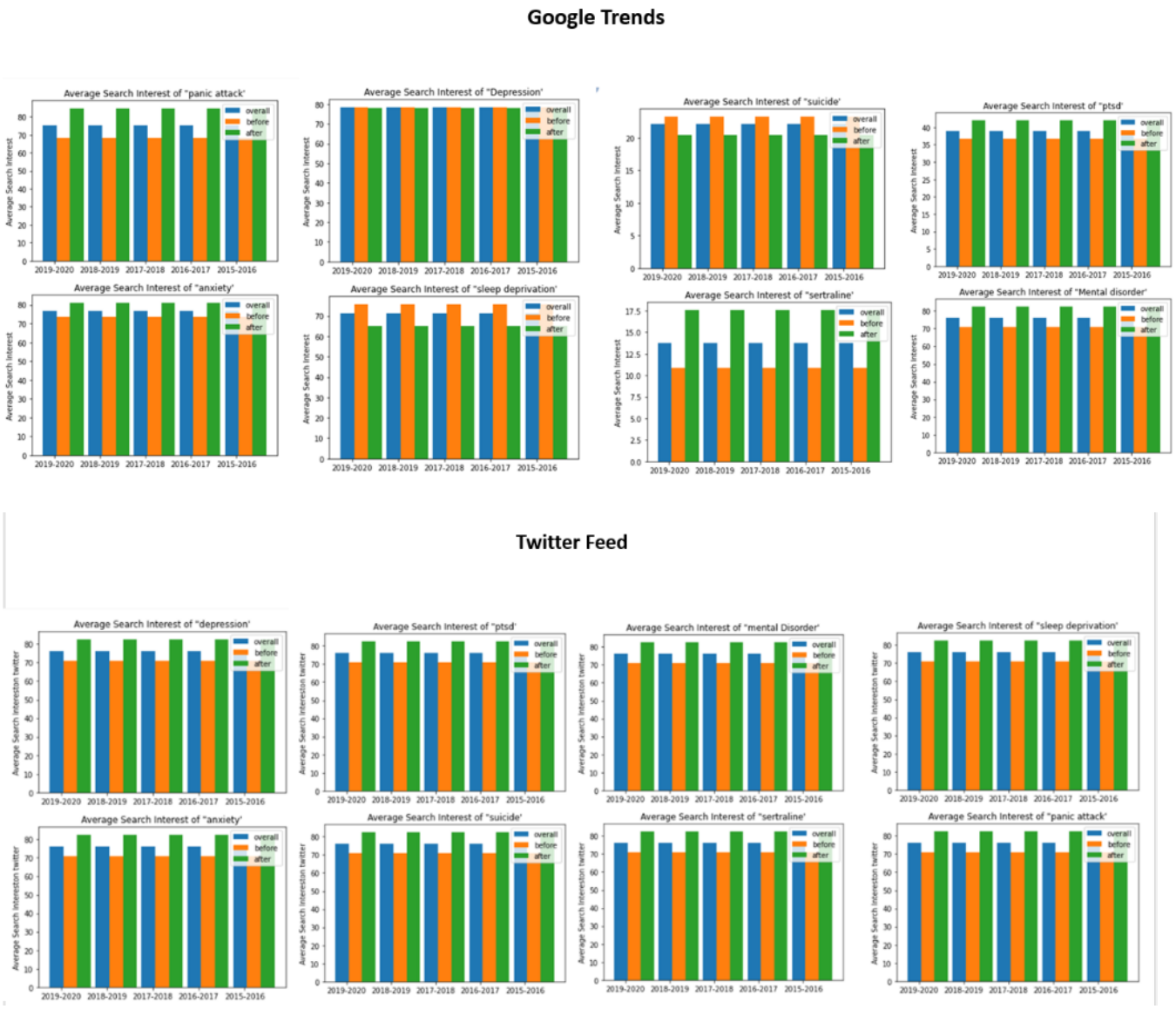


Figure 21: the groups of bar charts above showcase the keywords occurrences on both google and twitter platform.

the groups of bar charts above showcase the keywords illustrating an increase in panic within the worldwide population, the charts above clarify using these keywords in a straightforward manner with a nice visual representation of the before and after of the pandemic in addition to the overall usage of the keyword and comparing them on both the past and present time of the spread of the disease.

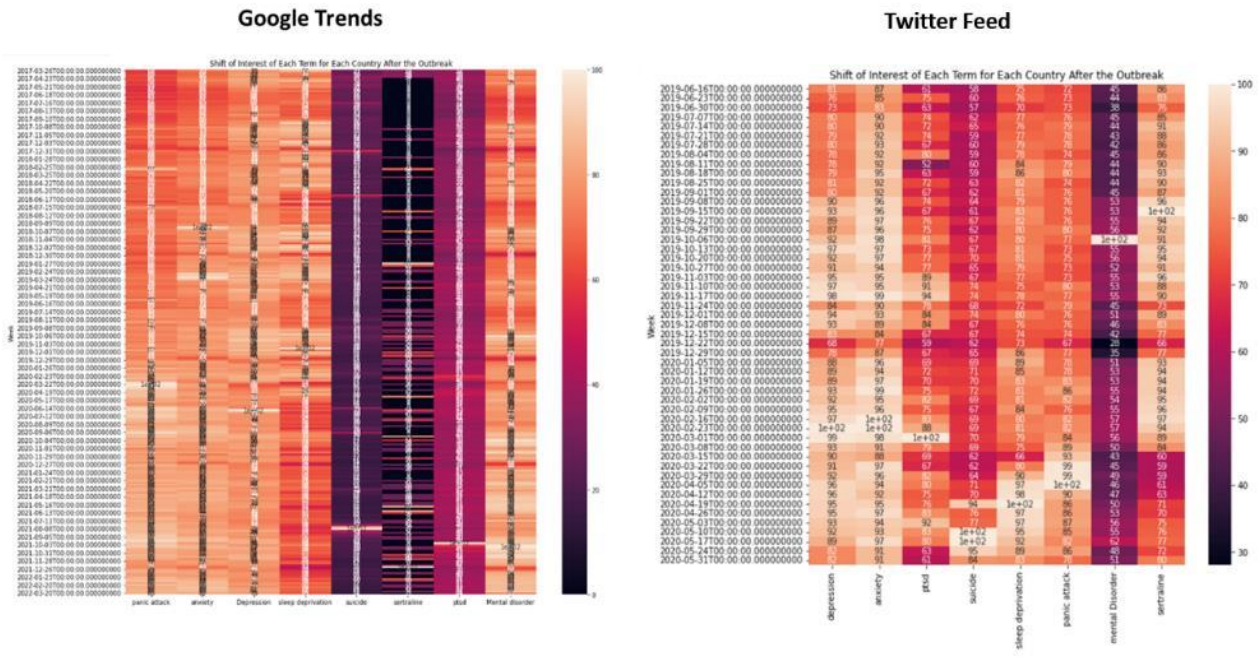


Figure 22: heat map of the dataset indicating peak weeks

Then showing the histogram of each dataset value before after and difference, showing dataset values for better visualization, with the heat map showcasing an increase in the usage of words such as “Suicide”, “Sertraline” and “PTSD” while the twitter heat map showcasing a usage of “Mental Disorder” on a worldwide standard.

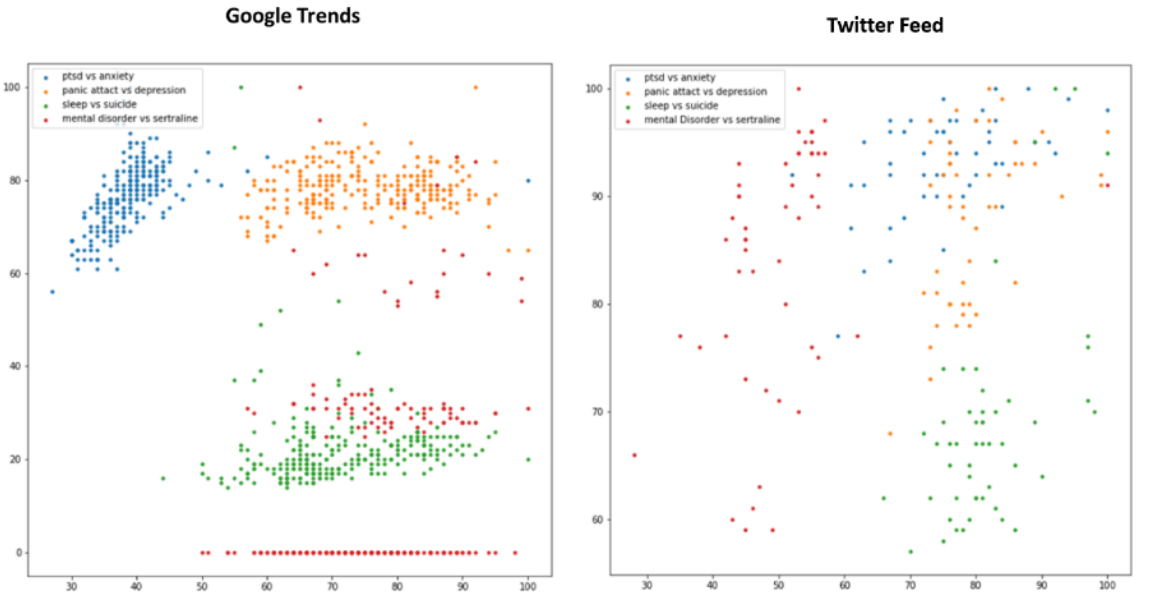


Figure 23: the scatter plot above showcases the variations between google trends and twitter.

This is the scatter plot, used to show 4 pairs of data variances of each other and showing the focus and concentration of these values with the orange being the “Panic Attack” and “Depression” in Google searches being higher than the usage of sleep and suicide as it is against each other, along with sleep and suicide being the lowest in the Twitter feed than “PTSD” and “Anxiety” as per a lot of users using twitter might be mentioning their diagnosed syndromes and sharing with other people while people shoeing empathy for that kind of disease.

The use of “Mental Disorders” and “Sertraline” is demonstrated to be steady over the majority of the count and time in the Twitter feed, and there are no significant differences in peak use.

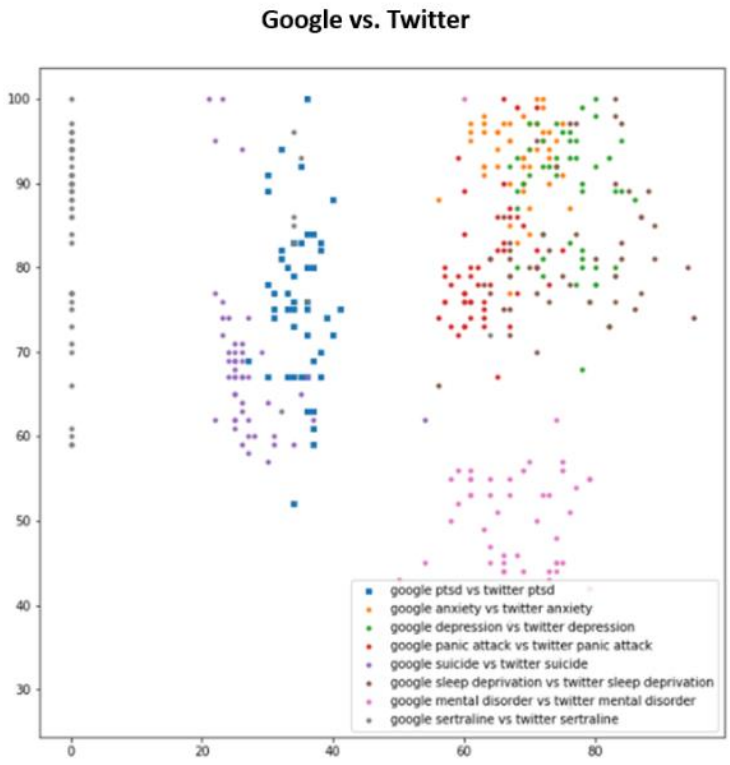


Figure 24: the scatter plot above showcases the variations between google trends and twitter together.

This Scatter Plot displays the graphs that depict the connection between two variables in Google Trends and Twitter in a dataset, and it is available for download. It is a two-dimensional plane on which data points are represented. A high correlation exists between the word's "Anxiety" and "Depression" in both Twitter and Google searches, whereas there is a low correlation between the word's "anxiety" and "Depression" in both Twitter and Google searches, whereas there is a low correlation between the words "Anxiety" and "Depression" in both Twitter and Google searches,

Twitter Line Chart

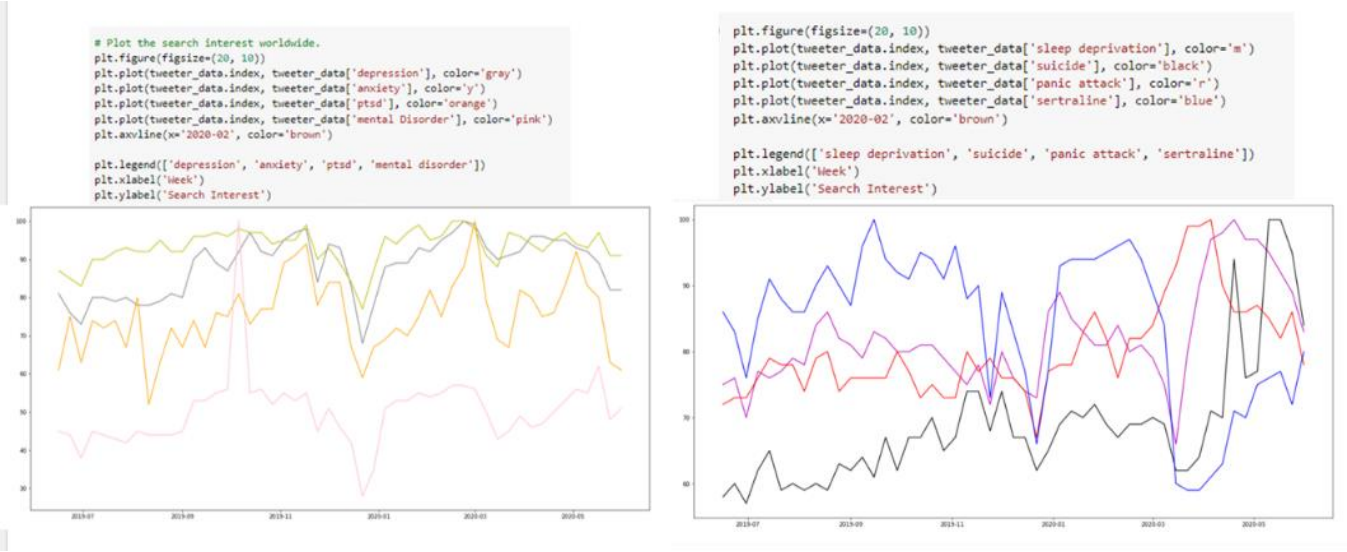


Figure 25: the Line Chart above showcases the line graphs of keywords before. During and after the pandemic.

The Line Chart is a visual representation of the sequence in which the various phases of a process are completed. As a general tool that may be used to describe a number of processes, including industrial processes, administrative or service processes, or project plans, it can be customized for a broad range of applications, plotting the dataset to view before and after values, the red line show the Covid-19 occurrence

	A	B	C	D	E	F	G	H
1	S.NO	Gender	Country	family_history	DSM	Age	5 score	
2	1	M	UAE	YES	F	38	13	1
3	2	M	UAE	YES	F	24	13	1
4	3	M	UAE	YES	F	35	13	1
5	4	M	JORDAN	NO	F	32	13	1
6	5	M	PAK	YES	F	45	13	1
7	5	M	RAWANDAN	NO	G	24	40	909
8	6	M	UAE	NO	F	27	13	1
9	7	M	PAK	YES	F	70	2	8
10	8	M	PAK	NO	F	38	41	9
11	9	M	UAE	NO	F	36	41	9
12	10	M	UAE	NO	F	36	13	1
13	11	M	UAE	NO	F	39	31	75
14	12	M	UAE	NO	F	26	13	1
15	13	M	PAK	NO	F	38	25	1
16	14	M	UAE	NO	F	32	13	1
17	15	M	JORDAN	NO	F	38	32	1
18	16	M	UAE	YES	F	34	13	1
19	17	M	JORDAN	YES	F	38	13	1
20	18	M	EGYPT	YES	F	28	13	1
21	19	M	EGYPT	YES	F	33	25	1
22	20	M	UAE	NO	F	54	32	1
23	21	M	UK	YES	F	59	41	3
24	22	M	SUDAN	NO	F	38	23	
25	23	M	YEMEN	NO	F	55	31	75
26	24	M	PAK	NO	F	40	32	1
27	25	M	UAE	NO	F	37	23	
28	26	M	INDIAN	NO	F	38	41	3
29	27	M	NIGERIA	NO	F	53	32	1
30	28	M	RUSSIAN	NO	F	31	51	2
31	29	M	ARMENIA	NO	F	32	23	
32	30	M	UAE	NO	F	25	23	
33	31	M	UAE	YES	F	35	13	1
34	32	M	IRAN	NO	F	25	51	2
35	33	M	NIGERIA	NO	F	67	41	3
36	34	M	UAE	NO	R	51	46	81
37	35	M	MOROCCO	NO	F	22	41	
38	36	M	CAMERON	NO	F	37	43	2
39	37	M	EGYPT	NO	F	24	41	9
40	38	M	CAMERON	NO	F	33	23	
41	39	M	BELGUM	NO	G	62	40	909
42	40	M	CHINA	NO	F	57	41	3

Figure 26: the above datasets displaying the inmate's data in the punitive and correctional institutions in the Dubai police.

In the dataset above we went through data cleaning and data preprocessing separating the DSM-5 scores letters from the numbers to show us better results and correlation, as this research was done using six variables and 126 inmates from the Dubai Police Correctional Institution to investigate the inmates current and prospective mental health condition using six factors connected to their DSM-5 scores as well as the degree of mental illness they suffered.

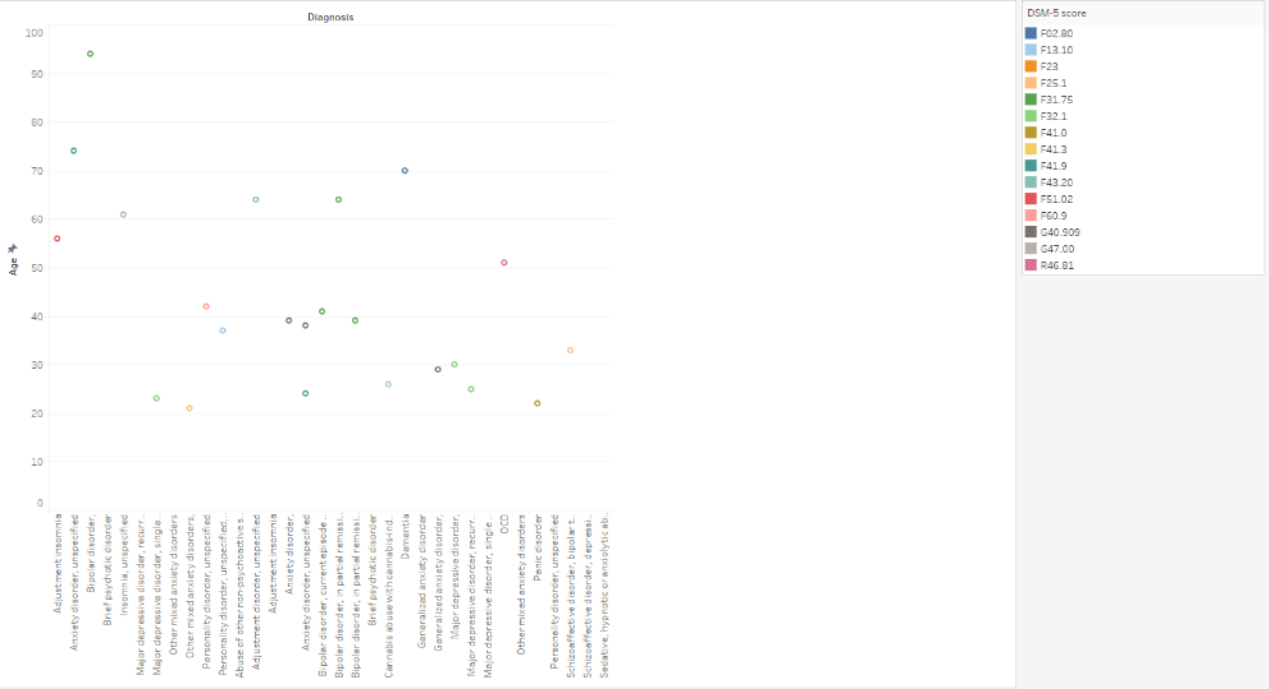


Figure 27: The graph above depicts the average age of prisoners with a given DSM-5 score in their mental wellness program.

Using the data from figure (27), we first used tableau to represent the average age of inmates that suffer from a specific DSM-5 score while being put in their suitable and accommodated mental wellness program, with the highest point being the DSM-5 score is F31.75 94 at the ages around 94 and the lowest points being with the DSM-5 score F41.3 and F41.0 between the ages 20 and 21 in the chart above. Diseases were classified based on their DSM-5 score, which is shown on the right side of the image, in order to aid in the identification of mental illnesses by a single focus group of inmates.

```
[ ]  sn.set(font_scale=1)
     sn.heatmap(Corr_Matrix, annot=True, cmap="YlGnBu", cbar=True)
     plt.show()
```

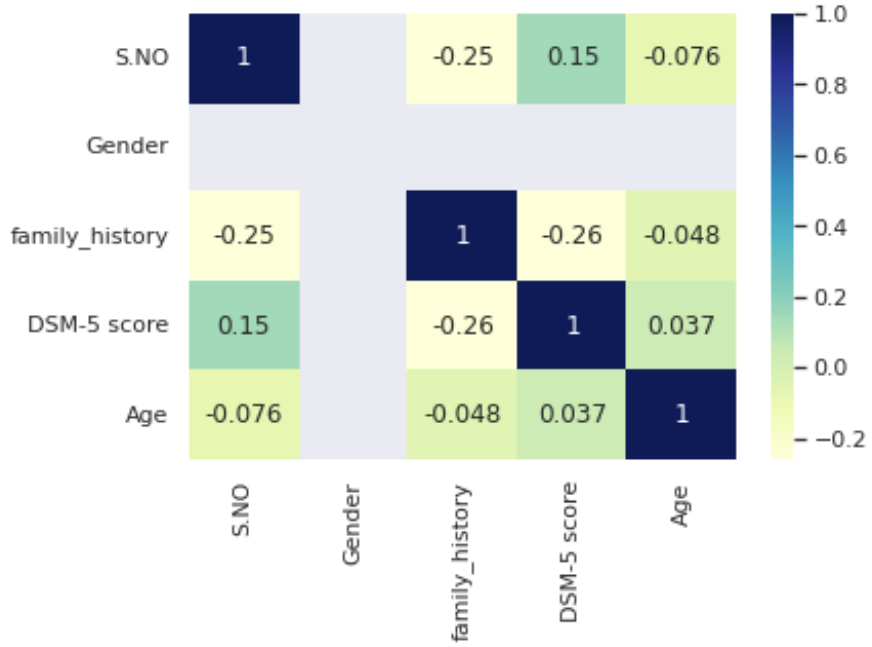


Figure 28: the above is a correlation matrix between the 5 variables of the inmate's dataset.

A Correlation Matrix is a table of variable correlations. The table above illustrates the correlation between two variables of each row and column, this correlation matrix is used to summarize the data, as well as to diagnose problems as the highest correlation is between the DSM-5 score and age which indicates that age is the highest associated variable to cause specific syndromes, as a result, it will most likely occur depending on the inmate's age, as after finalizing the model we received a percentage of 0.657 accuracy for the correlation matrix by which it resulted with a successful accuracy rate but a very low percentage that is needed to be either mixed with another models or re-running the codes again.

#	Model	Accuracy	Percentage
1	Correlation matrix	0.657	65%

#	Model	Accuracy	Percentage
1	K-nearest neighbor	0.82	82%
2	Tree Classifier	0.8539	85%
3	Logistic Regression	0.82	82%
4	Random forest	0.8412	84%
5	Support Vector Model	0.8507	85%

Figure 30: a table containing the different accuracy model results.

After making the comparisons needed from models above the current the accuracy models implemented it showed that the accuracy of the Tree Classifier is the highest accuracy with 0.8539 more than the SVM model by 0.0032 while the KNN model is 0.8253968253968254 and the ROC curve is 0.88, which is equal to the accuracy of the logistic regression model.

Since age is the most likely associated variable to cause specific syndromes, it will most likely occur depending on the inmate's age. After finalizing the model, we received a percentage of 0.657 accuracy for the correlation matrix, which resulted in a successful accuracy rate but a very low percentage that is needed to be combined with other models or re-run with other configurations.

Chapter 5

5.1 Conclusion

In conclusion, to sum up all organizations and medical institutions face different kind of challenges when aiming to succeed in their scope of not recognizing that a better comprehension of the data may lead them to better opportunities and solutions within their line of work, issue within their administrational hierarchy a psychological support department to be always on spot when needed for assistance as well as coordinate and monitor the current mental health situation for its subordinates

In this project most of the data mining processes were unsupervised and text analysis methodological practices were made. this study was rich in data from an employee's point of view to their lives interaction of twitter and ending with the inmate's data that were diagnosed with a mental health issue, in order to get the best results, large amounts of data created by groups of analysis must be put to many models and data mining methods in order to experiment with and train the dataset for improved outcomes. In this study, 1260 individuals were divided into 27 distinct factors, which were then evaluated using Random Forest, Decision Tree, SVM and KNN to extract the best possible findings from the data. The random forest provided a large amount of data that was used to create a probability equation. Because it was utilized to handle the data analysis portion of this project, the bar chart contributed to the extraction of essential points to be presented in this research. First, the data was exposed to a series of data cleaning and preparation procedures. It was critical to preprocess the data by splitting the DSM-5 scores into two parts in order to smooth the data while also eliminating algorithm errors and assuring the best outcomes when running the data analysis models, the correlation matrix was then used to highlight the most essential and effective causes behind the occurrence of mental health difficulties among prisoners, based on the facts included within the dataset that had been collected. When it came to projecting original factors onto newly defined axes and computing the components representing those factors

that may lead to newly formed variables, this method was critical. Furthermore, the models that were used were critical in identifying the most important traits or factors that are associated with significant effects.

5.2 Recommendations

In this project, we faced multiple data scarcity as only 5 types of variables were used in the data analysis part of the inmates analyses model due to lack of data available from the correctional institutions concerning in depth survey with the inmates themselves as per our literature review, the Distance is one of many additional forms of data gathered via an employee survey in a company from home to work in addition to number of siblings and children would've had a direct association with the mental issue of the person as well as correlate for us different results. Thus, a valuable recommendation would be analyzing more samples and variables of different types of individuals that are desired to be included in the classification models. Moreover, it is preferable to have a consistent number of datasets in to identify psychological disorders using machine learning as well as using aa set of models instead of one.

1. Analyzing more types of variables concerning both internal and external environment of the human being thus include more categorization methods to the classification model.

2. Developing other data mining and classification models for either employee and inmates' psychology that will relatively work as a supporting reason to achieve organizational and institutional goals, such as supervising and monitoring its workers' mental health to increase happiness levels from the 4th level skilled workers to the CEO of a corporation.

3. Applying data mining and classification models to other different individuals based on the company's genre of work as well as we may collect employee data in the future. We can track how long their employees spend on various tasks. We may then utilize this data to increase our

workers' efficiency. We may also make a list of their actions, put comparable ones together, and eliminate the ones that don't provide value. Using this exercise, we may better allocate and manage staff's work and better serve their clients or customers.

References

1.Alhadad, S. S. J. (2018). Visualizing data to support judgement, inference, and decision making in learning analytics: Insights from cognitive psychology and visualization science. Journal of Learning Analytics, 5(2)https://doi.org/10.18608/jla.2018.52.5

2.Amossé, T., Bryson, A., Forth, J., Petit, H., & SpringerLink (Online service). (2016). Comparative workplace employment relations: An analysis of practice in britain and france (1st 2016. ed.). Palgrave Macmillan UK.

3. Bacardit, J., Widera, P., Lazzarini, N., & Krasnogor, N. (2014). HARD DATA ANALYTICS PROBLEMS MAKE FOR BETTER DATA ANALYSIS ALGORITHMS: Bioinformatics as an example. Big Data, 2(3), 164-176. https://doi.org/10.1089/big.2014.0023

4.Bierie, D. M., & Mann, R. E. (2017). The history and future of prison psychology. *Psychology, Public Policy, and Law*, *23*(4), 478–489. https://doi.org/10.1037/law0000143

5.Bonichini, S., & Tremolada, M. (2021). Quality of life and symptoms of PTSD during the COVID-19 lockdown in italy. International Journal of Environmental Research and Public Health, 18(8), 4385.

6.Clarke, S. (2016). The wiley-blackwell handbook of the psychology of occupational safety and workplace health. Wiley Blackwell.

7.Cox Jr., L. A., Popken, D. A., Sun, R. X., & SpringerLink (Online service). (2018). Causal analytics for applied risk analysis (1st 2018. ed.). Springer International Publishing.

8.Edgemon, T. G., & Clay-Warner, J. (2019). Inmate mental health and the pains of imprisonment. Society and Mental Health, 9(1), 33-50.

9.Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. Psychological Methods, 21(4), 447-457. https://doi.org/10.1037/met0000120

10.Hölzel, L., Härter, M., Reese, C., & Kriston, L. (2010;2011;). Risk factors for chronic depression — A systematic review. Journal of Affective Disorders, 129(1), 1-13.

11.Houser, K. A., Vilcica, E. R., Saum, C. A., & Hiller, M. L. (2019). Mental health risk factors and parole decisions: Does inmate mental health status affect who gets released. International Journal of Environmental Research and Public Health, 16(16), 2950.

12. Liu, C. H., Zhang, E., Wong, G. T. F., Hyun, S., & Hahm, H. ". (2020). Factors associated with depression, anxiety, and PTSD symptomatology during the COVID-19 pandemic: Clinical implications for U.S. young adult mental health. Psychiatry Research, 290, 113172-113172.

13.Sykes, G. M., Western, B., & JSTOR (Organization). (2007). The society of captives: A study of a maximum security prison. Princeton University Press.

14.Matejkowski, J., Draine, J., Solomon, P., & Salzer, M. S. (2011). Mental illness, criminal risk factors and parole release decisions. Behavioral Sciences & the Law, 29(4), 528-553. https://doi.org/10.1002/bsl.991

15.Passos, I. C., Mwangi, B., Kapczinski, F., & SpringerLink (Online service). (2019). Personalized psychiatry: Big data analytics in mental health (1st 2019. ed.). Springer International Publishing.

16. Pilgrim, D. (2014). Influencing mental health policy and planning: DSM-5 as a disciplinary challenge for psychology. Review of General Psychology, 18(4), 293-301. https://doi.org/10.1037/gpr0000021

17.Pratt, L. (2019). Link: How decision intelligence connects data, actions, and outcomes for a better world (First ed.). Emerald Publishing Limited.

18.Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, 39, 156-168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

19.Tariq, M. U., Babar, M., Poulin, M., Khattak, A. S., Alshehri, M. D., & Kaleem, S. (2021). Human behavior analysis using intelligent big data analytics. Frontiers in Psychology, 12, 686610-686610. https://doi.org/10.3389/fpsyg.2021.686610

20.Wardenaar, K. J., Riese, H., Giltay, E. J., Eikelenboom, M., van Hemert, A. J., Beekman, A. F., Penninx, B. W. J. H., & Schoevers, R. A. (2021). Common and specific determinants of 9-year depression and anxiety course-trajectories: A machine-learning investigation in the netherlands study of depression and anxiety (NESDA). Journal of Affective Disorders, 293, 295-304.