Public health Awareness

Phase 3 Submission Document

Name	Sharmila.E
Reg. No	410121104044
NM ID	Au410121104044
Department	CSE-III
Domain	Data Analytics with Cognos
Project Title	Public Health Awareness
Phase 3	Development Part III
College	4101-Adhi College of Engineering and Technology, Kanchipuram

Introduction:-

Mental health issues affect approximately 700 million people worldwide, accounting for about 13% of all diseases. Depression, for instance, is the second leading cause of disability, trailing only behind back pain. The primary mental health conditions are depression and anxiety, rather than schizophrenia. There is evidence to show that individuals with mental health problems may be denied employment due to their mental condition or may not seek employment because they are aware of potential discrimination.

Disclosing a mental health problem in the workplace can lead to discriminatory behaviours from supervisors and colleagues, such as social exclusion or hindering these individuals' career progression. A framework for understanding these behaviours conceptualizes stigma as comprising three issues:-

- Knowledge (ignorance or misinformation)
- Attitudes (prejudice)
- Behaviour (discrimination)

In a study conducted by Manning and Whit, the factors most commonly considered when hiring a person include the previous work record (89%), job description (87%), whether they received treatment (69%), the time they were ill the previous year (68%), and the diagnosis (64%). Fenton et al. also concluded that the employment record (78%), health record (69%), diagnosis (36%), detection under the Mental Health Act (36%), and medical opinion (7%) are important factors in hiring someone. Krupa highlighted four underlying assumptions about workplace stigma:-

1. People with mental health issues do not have the necessary skills to meet job requirements.

2. People with mental health issues are dangerous or unpredictable.

3. Working is not healthy for people with mental health problems.

4. Employing people with mental health issues is an act of charity.

These assumptions vary in intensity based on a range of organizational, individual, and social factors.

It is important to emphasize the significance of a positive work environment for enhancing the economic and social integration of individuals with mental health issues.

Dataset:-

This dataset is from a 2014 survey that measures attitudes toward mental health and the prevalence of mental health disorders in the technological workplace. The survey was conducted with 1,260 individuals from various countries, and the top 10 participating countries can be seen in Figure 1.

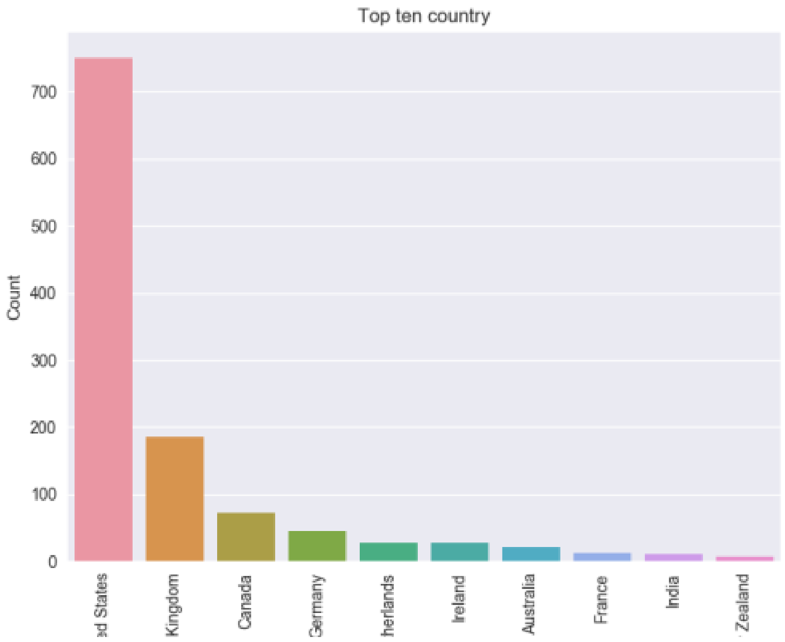


Figure 1: Top 10 Participating Countries in the Survey

The dataset contains the following information:

- Timestamp

- Age

- Gender

- Country

- State: If the person resides in the United States, in which state or territory they live.

- Self-employed: Whether the person is self-employed.
- Family history: Whether the person has a family history of mental health issues.

Necessary step to follow:

1.Import Libraries:

Program:

```
import pandas as pd
import numpy as np
from sklearn import svm, neighbors
import sklearn as sk
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
```

2. Load the Dataset:

```
data = pd.read_csv('finalissimo.csv', sep = ',')
```

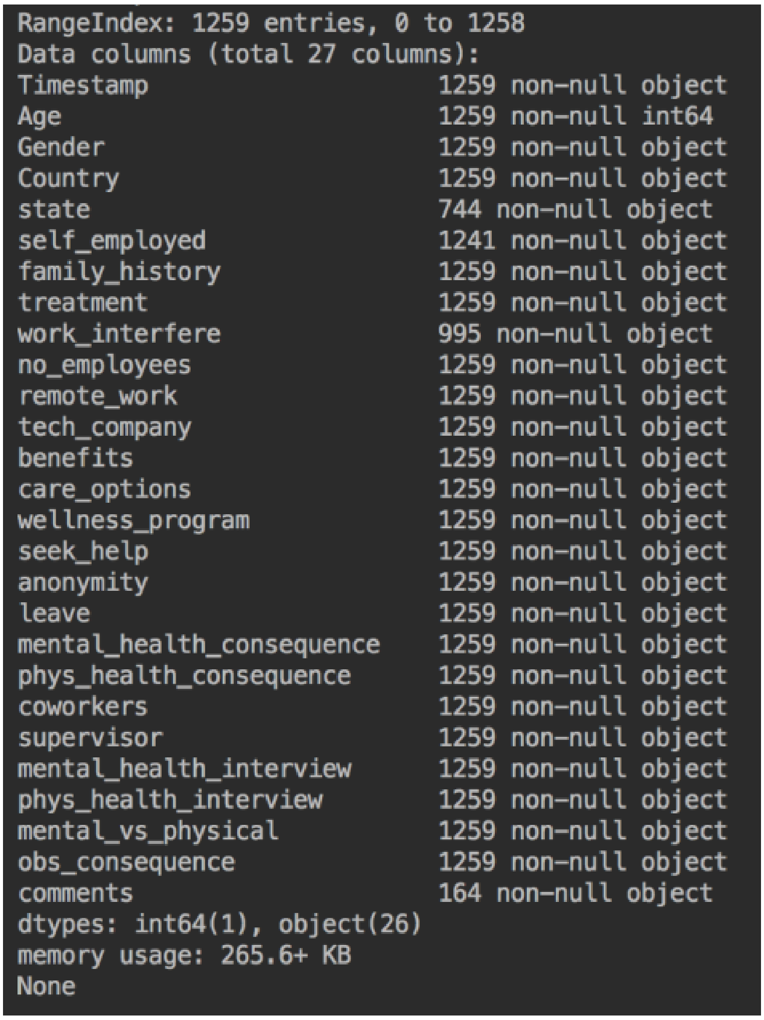
Data Preprocessing

1.Data Understanding

As mentioned earlier, the starting point is the raw data, which consists of direct responses to the questionnaire. Initially, it was necessary to understand and review all the data. It quickly became apparent that there were inconsistent, null responses, and the data was not normalized, among other issues that would prevent us from progressing with the project. Therefore, data cleaning was required.

Program:

```
data.dropna(axis=0, subset = ['work_interfere'], inplace=True)
listLabelsData = list(data)
for a in listLabelsData[1::]:
    typesOfLabels = data[a].unique()
    numericalLabels = list(range(0, len(typesOfLabels)))
    data[a].replace(typesOfLabels, numericalLabels, inplace = True)
treatment = data.loc[(data.treatment == 1)]
non_treatment = data.loc[(data.treatment == 0)]
treatment = treatment.sample(frac=1)
treatment = treatment[0:352]
newData = pd.concat([treatment, non_treatment])
newData = newData.sample(frac=1)
newData.to_csv("newData.csv")
```



```
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 27 columns):
Timestamp                    1259 non-null object
Age                          1259 non-null int64
Gender                       1259 non-null object
Country                      1259 non-null object
state                        744 non-null object
self_employed                1241 non-null object
family_history               1259 non-null object
treatment                    1259 non-null object
work_interfere               995 non-null object
no_employees                 1259 non-null object
remote_work                  1259 non-null object
tech_company                 1259 non-null object
benefits                     1259 non-null object
care_options                 1259 non-null object
wellness_program             1259 non-null object
seek_help                    1259 non-null object
anonymity                    1259 non-null object
leave                        1259 non-null object
mental_health_consequence    1259 non-null object
phys_health_consequence      1259 non-null object
coworkers                    1259 non-null object
supervisor                   1259 non-null object
mental_health_interview      1259 non-null object
phys_health_interview        1259 non-null object
mental_vs_physical           1259 non-null object
obs_consequence              1259 non-null object
comments                     164 non-null object
dtypes: int64(1), object(26)
memory usage: 265.6+ KB
None
```

2.Data Cleaning:-

To ensure that the algorithm's performance would not be compromised, certain features, namely "Timestamp," "state," and "comments," were considered irrelevant to the problem at hand and were therefore eliminated.

Next, due to the lack of normalization in responses, each feature had to be individually analyzed to check if the response ranges were as desired. "Sex," "Age," and "self-employed" required alterations.

For the "self-employed" feature, the treatment was simple, as it only required the removal of all null responses. While null responses make sense for certain attributes, for this one, the range of responses could only be "yes" or "no."

Regarding age, the "Age" feature had both negative and excessively large values. This variable was then bounded between 0 and 100, and instances that fell outside this range were removed.

Finally, the "Sex" variable was normalized to only three possible responses: "Male," "Female," and "Trans."

In this way, our dataset was reduced to 1,233 instances with 24 features. This new dataset was saved, and the remaining work focused on it.

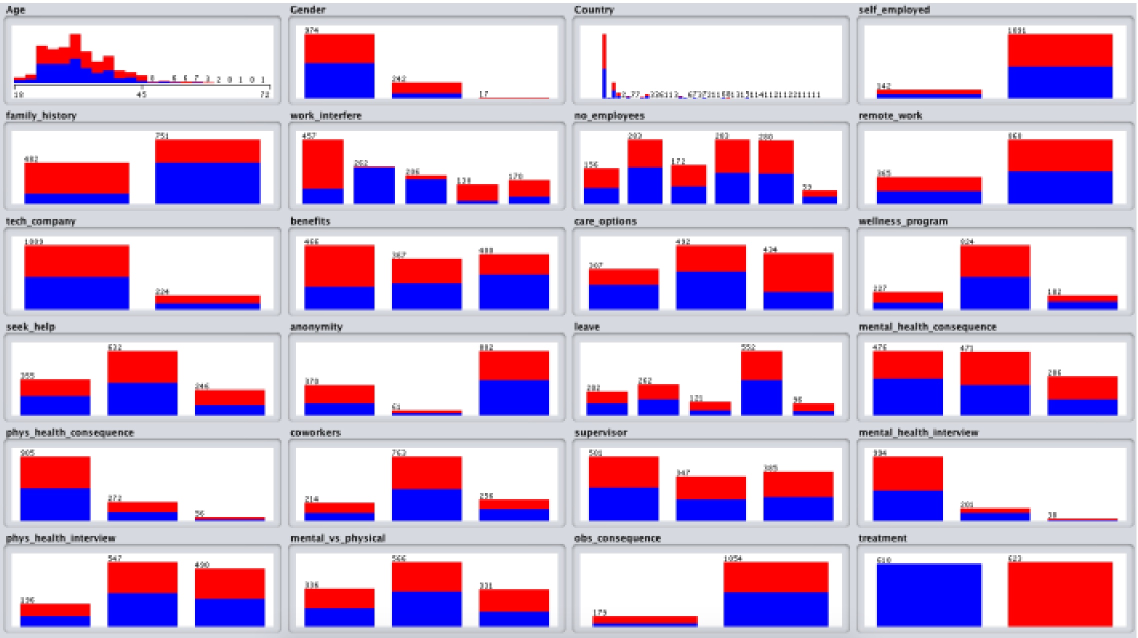
3.Data Visualization:

With the dataset now normalized and preprocessed, we proceeded to visualize our data. To do this, we used the WEKA tool introduced in the course lectures. It was necessary to adapt the data format to fit the tool's accepted format.

Using WEKA and defining our class variable, we were able to examine how the other attributes behave concerning the class.

Program:

```
fig,ax = plt.subplots(figsize=(8,6))
sns.countplot(data=df,x = 'Age_Group',hue= 'remote_work',ax=ax)
plt.title('Remote Work vs Age Group')
plt.show()
country_count = Counter(df['Country'].dropna().tolist()).most_ common(10)
country_idx = [country[0] for country in country_count]
country_val = [country[1] for country in country_count]
fig,ax = plt.subplots(figsize=(8,6))
sns.barplot(x = country_idx,y=country_val ,ax =ax)
plt.title('Top ten country')
plt.xlabel('Country') plt.ylabel('Count') ticks =
plt.setp(ax.get_xticklabels(),rotation=90)
plt.show()
fig,ax = plt.subplots(figsize=(8,6))
sns.countplot(data=df,x = 'mental_vs_physical',hue=
'mental_vs_physical',ax=ax)
plt.title('Mental Health vs Physical Health')
plt.show()
fig,ax = plt.subplots(figsize=(8,6))
sns.countplot(data=df,x = 'phys_health_interview',hue=
'phys_health_interview',ax=ax) plt.title('Physical Health Interview')
plt.show()
```



With this initial step, it is possible to quickly identify which features may play an important role in the classification. This is the case for the variables "family_ history," "work_ interfere," and "care_ options." Therefore, it makes sense to focus the visualization on these attributes.

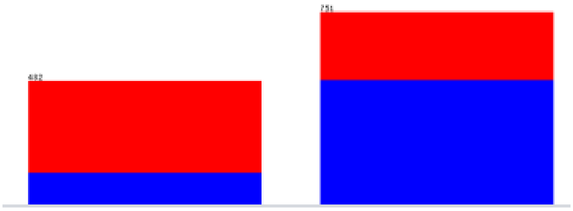


Figura 4: Visualização do atributo family_history

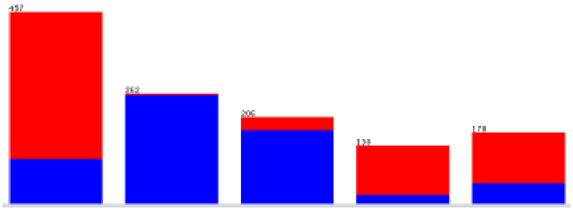
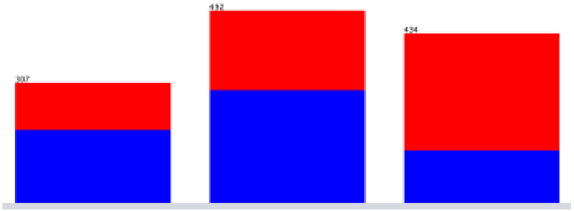


Figura 5: Visualização do atributo work_interfere



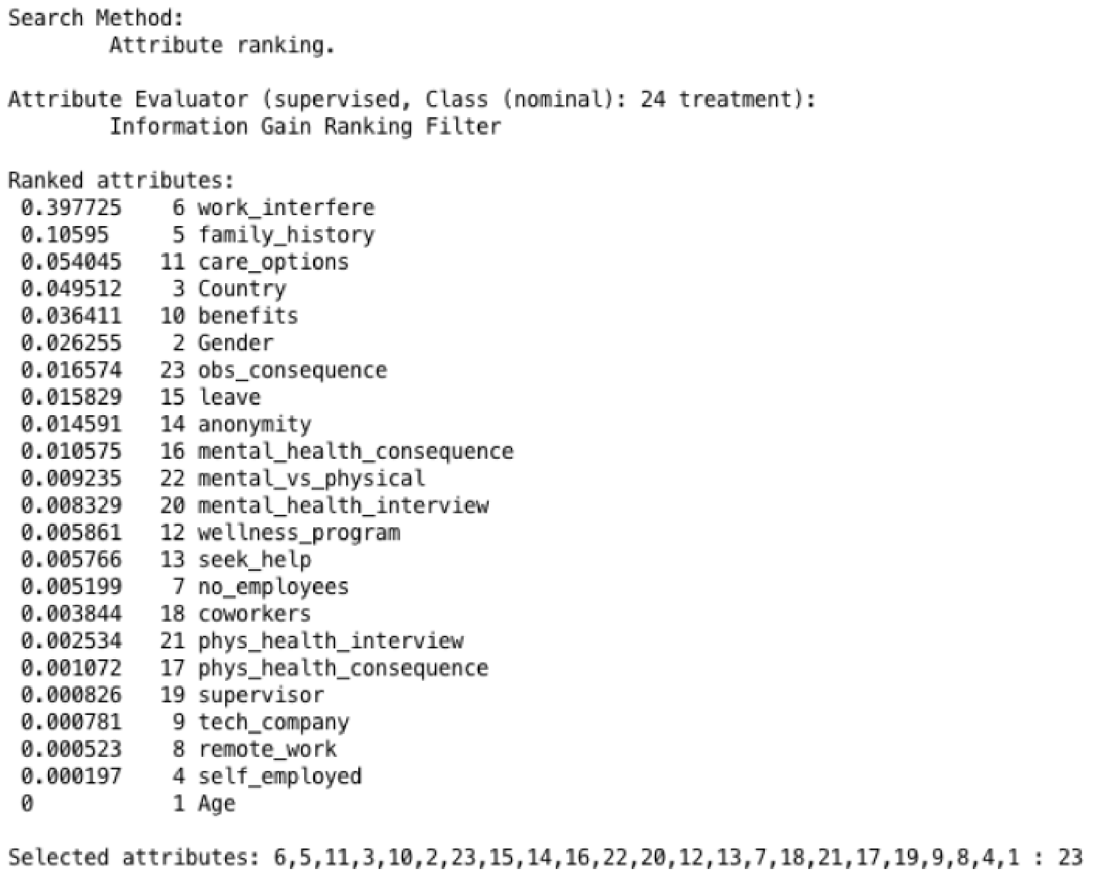
As can be observed, these attributes are good discriminants, as the distribution of the Class across their value spectrum is heterogeneous. For example, in the "family _ history" attribute, for the negative response, the predominance of the negative class is obviously higher, and vice versa. For the other two attributes, the behaviour is similar.

4.Feature Selection:

Although it is possible to determine the most relevant features through attribute visualization, it is necessary to confirm this using algorithms designed for that purpose. Therefore, feature selection/extraction methods need to be applied. However, since this is a highly nominal problem, the loss of information and meaning of the attributes is not desired.

As a result, we limited this section to feature selection. In other words, out of the 24 features included in the dataset, we will select only the most relevant ones for the classification process.

Using the "Select Attributes" option in Weka, we were able to carry out this process. The methods used were InfoGainAttributeEval and CorrelationAttributeEval. Both methods perform feature selection functions, as discussed earlier. The results obtained for each method were as follows:

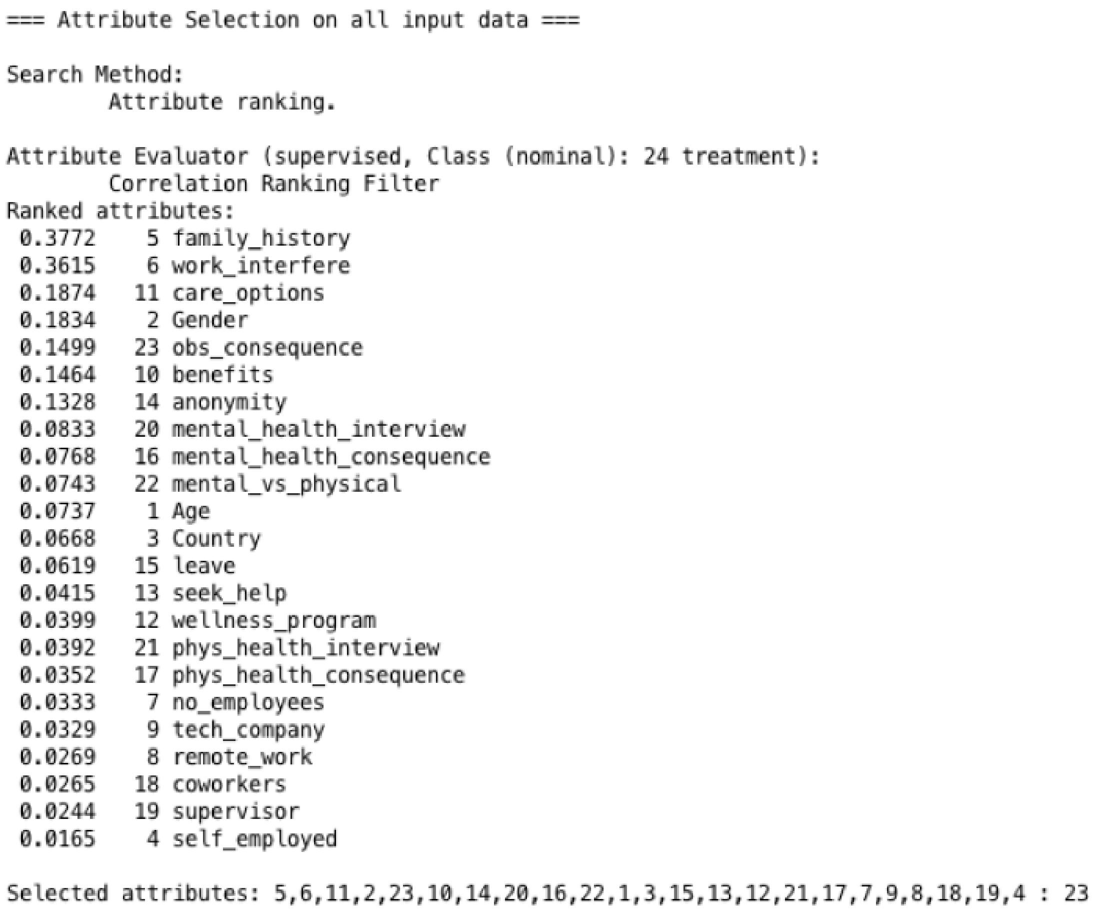


```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 24 treatment):
        Information Gain Ranking Filter

Ranked attributes:
 0.397725    6 work_interfere
 0.10595     5 family_history
 0.054045   11 care_options
 0.049512    3 Country
 0.036411   10 benefits
 0.026255    2 Gender
 0.016574   23 obs_consequence
 0.015829   15 leave
 0.014591   14 anonymity
 0.010575   16 mental_health_consequence
 0.009235   22 mental_vs_physical
 0.008329   20 mental_health_interview
 0.005861   12 wellness_program
 0.005766   13 seek_help
 0.005199    7 no_employees
 0.003844   18 coworkers
 0.002534   21 phys_health_interview
 0.001072   17 phys_health_consequence
 0.000826   19 supervisor
 0.000781    9 tech_company
 0.000523    8 remote_work
 0.000197    4 self_employed
 0           1 Age

Selected attributes: 6,5,11,3,10,2,23,15,14,16,22,20,12,13,7,18,21,17,19,9,8,4,1 : 23
```



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 24 treatment):
        Correlation Ranking Filter
Ranked attributes:
 0.3772    5 family_history
 0.3615    6 work_interfere
 0.1874   11 care_options
 0.1834    2 Gender
 0.1499   23 obs_consequence
 0.1464   10 benefits
 0.1328   14 anonymity
 0.0833   20 mental_health_interview
 0.0768   16 mental_health_consequence
 0.0743   22 mental_vs_physical
 0.0737    1 Age
 0.0668    3 Country
 0.0619   15 leave
 0.0415   13 seek_help
 0.0399   12 wellness_program
 0.0392   21 phys_health_interview
 0.0352   17 phys_health_consequence
 0.0333    7 no_employees
 0.0329    9 tech_company
 0.0269    8 remote_work
 0.0265   18 coworkers
 0.0244   19 supervisor
 0.0165    4 self_employed

Selected attributes: 5,6,11,2,23,10,14,20,16,22,1,3,15,13,12,21,17,7,9,8,18,19,4 : 23
```

Both methods return a value for each attribute. The higher this value, the greater the impact that this feature has on the classification process. As expected, the attributes that perform best are those previously identified solely through the visualization of their distribution.

It is now possible to exclude attributes with insignificant importance in classification. Therefore, we chose to eliminate "self_employed," "supervisor," "tech_company," and "remote_work." We are now proceeding with only 19 features.

5.Clustering:

Clustering is a data mining technique used to automatically group data based on their degree of similarity. In this work, the k-means algorithm was used to group data, attempting to separate samples into n groups of equal variance, minimizing a criterion known as inertia or within-cluster sum of squares. In this algorithm, the number of clusters must be defined in advance.

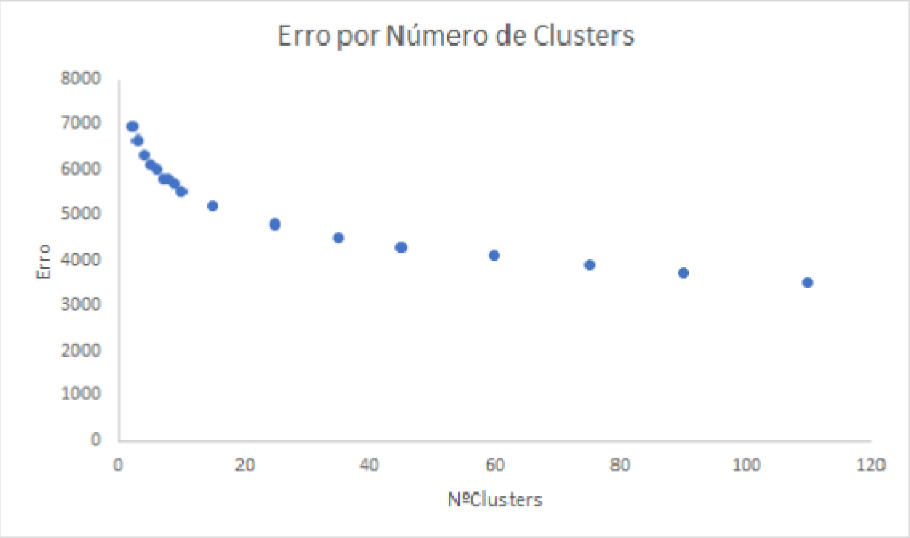
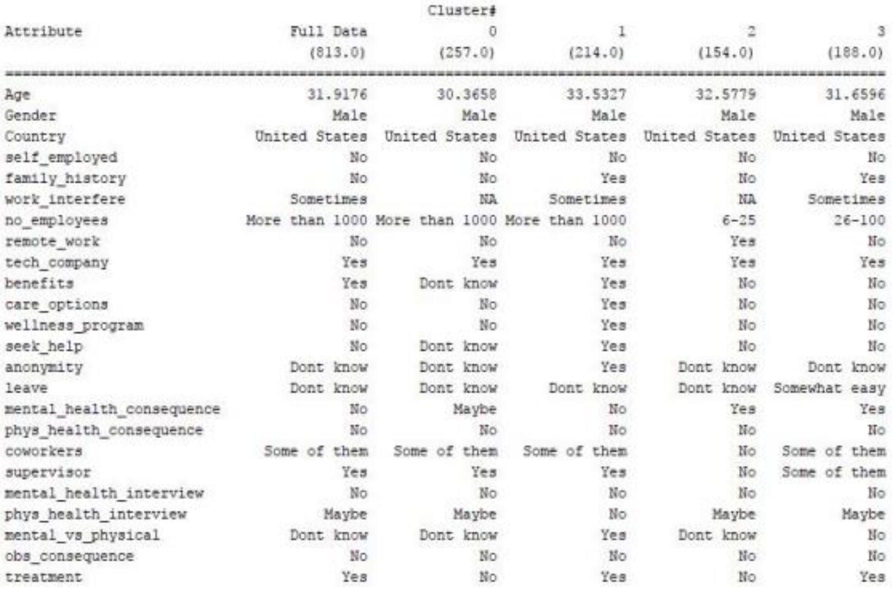


Figure 13 shows the error per number of clusters for the test set, representing 34% of the entire dataset. As can be observed, the error begins to stabilize after 20 clusters, meaning that it decreases less. However, we decided to choose only 4 clusters to describe our problem. This is because describing 20 clusters for a binary classification problem with 23 features is very complicated. By choosing 4 clusters, it is possible to qualitatively describe each cluster. The graph of error per number of clusters was created using Excel.



```
                                                Cluster#
Attribute                       Full Data              0              1              2              3
                                  (813.0)        (257.0)        (214.0)        (154.0)        (188.0)
=====================================================================================================
Age                               31.9176        30.3658        33.5327        32.5779        31.6596
Gender                               Male           Male           Male           Male           Male
Country                     United States  United States  United States  United States  United States
self_employed                          No             No             No             No             No
family_history                         No             No            Yes             No            Yes
work_interfere                  Sometimes             NA      Sometimes             NA      Sometimes
no_employees               More than 1000 More than 1000 More than 1000           6-25         26-100
remote_work                            No             No             No            Yes             No
tech_company                          Yes            Yes            Yes            Yes            Yes
benefits                              Yes      Dont know            Yes             No             No
care_options                           No             No            Yes             No             No
wellness_program                       No             No            Yes             No             No
seek_help                              No      Dont know            Yes             No             No
anonymity                       Dont know      Dont know            Yes      Dont know      Dont know
leave                           Dont know      Dont know      Dont know      Dont know  Somewhat easy
mental_health_consequence              No          Maybe             No            Yes            Yes
phys_health_consequence                No             No             No             No             No
coworkers                    Some of them   Some of them   Some of them             No   Some of them
supervisor                            Yes            Yes            Yes             No   Some of them
mental_health_interview                No             No             No             No             No
phys_health_interview               Maybe          Maybe             No          Maybe          Maybe
mental_vs_physical              Dont know      Dont know            Yes      Dont know             No
obs_consequence                        No             No             No             No             No
treatment                             Yes             No            Yes             No            Yes
```

Figure 14 represents the distribution of features using 4 clusters.

Cluster 1 and 3 relate to "yes" for treatment, our target class. The differences between these clusters are related to the following attributes:

- "no_ employees," where in cluster 1, the number of employees in the company/organization is >1000, and in cluster 3, it is between 26-100.

- "benefits," where in cluster 1, the employer provides benefits for mental health, and in cluster 3, there are no benefits.

- "care_ options," where in cluster 1, the person is aware of the options for mental health care, and in cluster 3, they are not aware.

- "wellness_ program," where in cluster 1, the employer has discussed the well-being program with the employee, and in cluster 3, they have not.

- "seek_ help," where in cluster 1, the employer provides resources to learn more about mental health issues and how to seek help, and in cluster 3, this does not happen.

- "anonymity," where in cluster 1, the worker's anonymity is protected, and in cluster 3, they are unsure if it is protected.

- "leave," in cluster 1, the worker is unsure if taking a medical leave is easy, and in cluster 3, it is easy.

- "mental_ health_ consequence," in cluster 1, the response to the question "Do you think discussing mental health with the employer will have negative consequences?" is no, and in cluster 3, the response is yes.

- "supervisor," in cluster 1, workers are willing to discuss their mental state with their immediate supervisors, and in cluster 3, only some of them are willing.

- "physics_ health_ interview," in cluster 1, workers do not discuss a physical problem during an interview, and in cluster 3, they might.

- "mental_ vs_ physical," in cluster 1, the employer takes mental health as seriously as physical health, and in cluster 3, this is not the case.

Cluster 0 and 2 relate to "no" for treatment. The differences between these clusters are related to the following attributes:

- "no_ employees," where in cluster 0, the number of employees in the company/organization is >1000, and in cluster 2, it is between 6-25.

- "remote_ work," in cluster 0, people do not usually work outside the office for at least 50% of their time, and in cluster 2, they do.

- "benefits," in cluster 0, people are unsure if the employer provides benefits for mental health, and in cluster 2, there are no benefits.

- "seek_ help," in cluster 0, the employer provides resources to learn more about mental health issues and how to seek help, and in cluster 2, this does not happen.

- "mental_ health_ consequence," in cluster 0, the response to the question "Do you think discussing mental health with the employer will have negative consequences?" is maybe, and in cluster 2, the response is yes.

- "coworkers," in cluster 0, workers are willing to discuss their mental state with some coworkers, and in cluster 2, they do not discuss it with any coworkers.

- "supervisor," in cluster 0, workers are willing to discuss their mental state with their immediate supervisors, and in cluster 2, they do not discuss it with any supervisors.

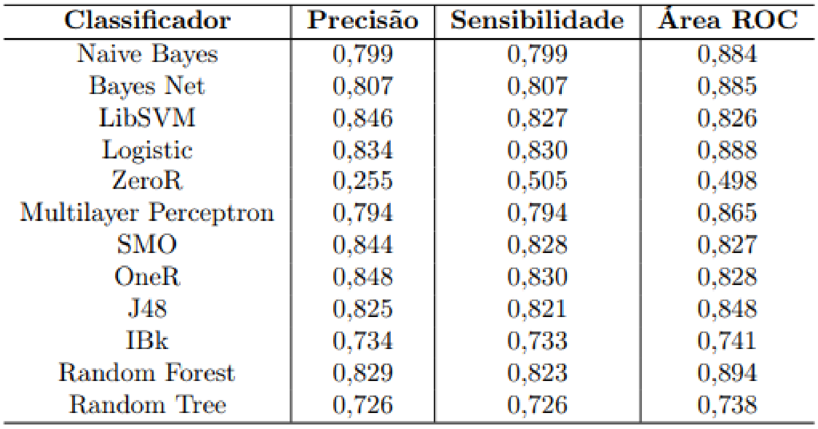
6.Classification:

In the first part of the work, the WEKA software was used to test the performance of various classifiers to determine which one performs best. Cross-validation was used for data classification. To evaluate the performance of each classifier, the precision and sensitivity of each classifier were compared. The results can be seen in Table 1.

Precision is given by Vp / (Vp + Fp), meaning of all the elements classified as positive, how many are truly positive. It represents the proportion of relevant items among the selected items.

Sensitivity is given by Vp / (Vp + Fn), representing the percentage of true positives among all examples whose expected class is the positive class.

The area under the ROC curve is obtained by representing the true positive rate (Vp / Pt) versus the false positive rate (Fp / Nt). An area with a value of 1 represents a perfect test, while an area with a value of 0.5 represents a test with no value.



Classificador	Precisão	Sensibilidade	Área ROC
Naive Bayes	0,799	0,799	0,884
Bayes Net	0,807	0,807	0,885
LibSVM	0,846	0,827	0,826
Logistic	0,834	0,830	0,888
ZeroR	0,255	0,505	0,498
Multilayer Perceptron	0,794	0,794	0,865
SMO	0,844	0,828	0,827
OneR	0,848	0,830	0,828
J48	0,825	0,821	0,848
IBk	0,734	0,733	0,741
Random Forest	0,829	0,823	0,894
Random Tree	0,726	0,726	0,738

The classifier that performed the best is OneR. This algorithm creates a rule for each attribute in the training data and selects the rule with the lowest error percentage as the single rule . To create a rule for an attribute, it is necessary to determine the class that appears most frequently for that attribute. A "rule" is simply a set of attribute values bounded by their majority class. The error percentage of a rule is the number of training instances in which the class of an attribute value is not concordant with the classification of that attribute in the rule.

The classifier that performed the worst is ZeroR. This algorithm is target-based and ignores all predictors. It simply predicts the majority class .

In the second part of the work, Python programming language was used. This part began with 19 features, with 18 being nominal and 1 being numeric. Therefore, it was necessary to transform the nominal variables into numeric variables. This step is crucial for classification because many classifiers only accept numeric variables, and some only accept binary variables, such as in the Random Forest algorithm, where the target vector must be binary.

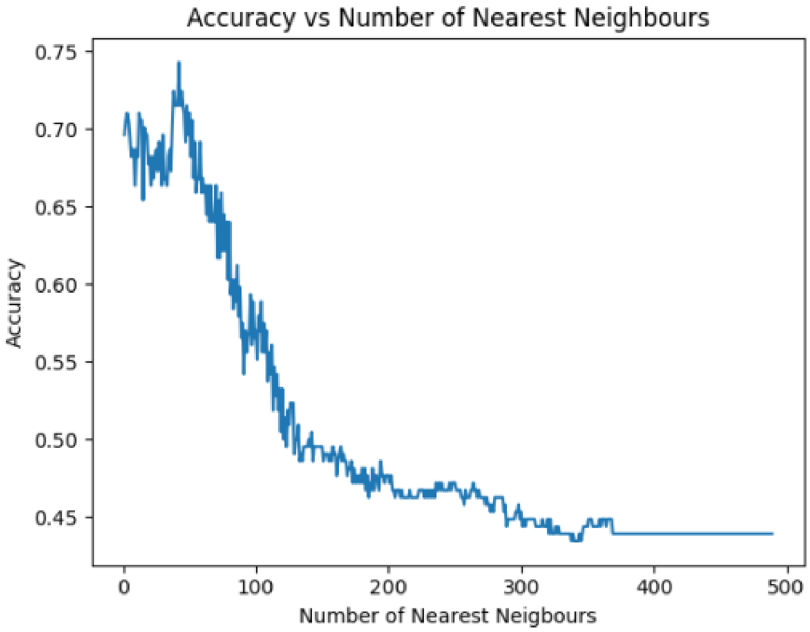
Before this transformation, samples with NaN values in the "work_interfere" variable were removed. Removing them is important because if this variable were included, the classifier might consider a NaN value as important. Once the values of the variable were changed, it was studied whether the data set was balanced or not to ensure fair classification. By removing samples where the "work_interfere" variable had NaN values, the data set became imbalanced because most samples with NaN values were from subjects who had never sought treatment. Therefore, 260 samples of subjects who did not seek treatment were removed, resulting in 352 samples for each class. Finally, the data set was shuffled.

With the data set ready for classification, it was divided into training and testing data. It was decided to divide it into 70% training and 30% testing, making the data set ready for classification.

In this section, the following classifiers were used: Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbours (KNN). For Random Forest and KNN, a parameter study was conducted to maximize performance.

In the Random Forest classifier, the number of branches in the classifier and the minimum number of samples required to be in a node were varied. Initially, the classifier presented accuracy, sensitivity, and specificity of 1.0. In order to avoid overfitting, the number of branches in the classifier and the minimum number of samples in a node were increased, but it was in vain.

In the KNN classifier, the number of neighbors is the most important factor to optimize. Through a loop, classification started with 1 neighbor up to the same number of samples in the training group. From Figure 15, it can be seen that the number of neighbors that maximizes the classifier's performance is 42, achieving an accuracy of 75.3%.



Discussion and Conclusion:

The work was based on a recent Kaggle challenge. We chose a challenge for which there were not many previous attempts, making our work more practical and exploratory rather than purely theoretical or focused on improving existing work.

The work spanned various components of Data Mining. Starting with raw data that had to be properly processed, we went through all the steps to build a classifier with satisfactory performance. The best classifier obtained, SVM, has shown very good results, allowing for correct classification of its class when presented with new data. It's important to note that a high sensitivity value means that the classifier correctly identifies individuals who have genuinely needed or sought treatment. The Random Forest classifier showed signs of overfitting, which can mislead decision-makers who rely on insights from such models.

We also attempted to use a feedforward neural network for classification, but configuration issues prevented us from observing the results, likely due to a bug in the Keras package or a conflict between Python and Anaconda. A neural network could have been an excellent option since it has proven to be highly adaptable to the type of variables and typically shows high performance.

In addition to the machine learning components, throughout the process, we were able to draw some more theoretical conclusions related to the chosen topic of study. One of the questions we aimed to answer was how the frequency of mental health disorders and attitudes varies geographically. Although the majority of instances originated in the United States, it is possible to create a global distribution of the sample, as shown in Figure 16.

It is also possible to conclude what the best predictive factors of a mental illness are. The answer to this question arises during the Feature Selection phase, where, effectively, family history is undoubtedly the factor with the greatest predictive power.

In summary, we consider that the work fulfills the imposed requirements. The learning component was very present, as was self-directed learning. We had the opportunity to study and apply knowledge in an area that is currently in a phase of global expansion.

*****