A Comparative Analysis of Four Classification Algorithm for Mental Health Analysis basis on technical People

By

Afsana Sadia
(152-35-1157)

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Supervised By

Asif Khan Shakir
Lecturer
Department of Software Engineering

Daffodil International University

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

Fall-2019

APPROVAL

This thesis titled on "A Comparative Analysis of Four Classification Algorithm for Mental Health Analysis basis on technical People", submitted by **Afsana Sadia, 152-35-1157** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Prof. Dr. Touhid Bhuiyan
Professor and Head — **Chairman**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Dr. Md. Asraf Ali — **Internal Examiner 1**
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Asif Khan Shakir — **Internal Examiner 2**
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Dr. Md. Nasim Akhtar — **External Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

DECLARATION

It hereby declare that this thesis has been done by me under the supervision of **Mr. Asif Khan Shakir**, Lecturer, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Afsana Sadia

Afsana Sadia
Student ID: **152-35-1157**
Batch: 17th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by:

11.12.19

Asif Khan Shakir
Lecturer
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

ACKNOWLEDGEMENT

First of all, I am grateful to the Almighty Allah for giving me the ability to complete the final thesis.

I would like to express my gratitude to my supervisor **Mr. Asif Khan Shakir** for the consistent help of my thesis and research work, through his understanding, inspiration, energy, and knowledge sharing. His direction helped me to finding the solutions of research work and reach to my final theory.

I would like to express my extreme sincere gratitude and appreciation to all of my teachers of **Software Engineering** department for their kind help, generous advice and support during the study.

I am also express my gratitude to all of my friend's, senior, junior who, directly or indirectly, have lent their helping hand in this venture.

Afsana Sadia

TABLE OF CONTANS

LIST OF FIGURES

ABSTRACT

Mental disorder is a disorder of the mind that softens the effects of extreme discomfort in thinking and over-behaving, which fails to adapt to the normal needs and routines of life. In contrast, technology is rapidly evolving society and numerous practices now require the ability to use technology. These situations can induce problems for many people, including users with severe mental illness technology. In this study, we have tried to demonstrate a comparative analysis of four classification algorithm for mental disorder of technical people. Many researchers analyses the mental health. None of them are analysis the mental health to find out the identification reason of mental disorder among the tech people in the tech workplaces. This paper objective is Finding the predicted causes of mental illness or mental disorder in tech work places and visualizing the correlation among the causes. This paper proposed a model for identifying the reason of mental illness among the technical people. There we used Correlation to select features and showing the comparison of the result we used Decision Tree, Naïve Bayes, ANN and KNN. Elbow method is used to determine the optimal k. After the dataset analysis, the frequency of occurrence of mental disorder among these technicians is 22-24 years of age and the location of the work is system admin, back end developer, stack developer, team leader. We need to take the necessary steps to raise awareness of mental illness among tech people. If we fail to take the necessary steps, this problem will become more and more complex, and this problem will be very difficult to solve at once. Therefore, this research would yield the potential for healthier lives for countless individuals and the general advancement of the nation's well-being.

Keywords: Mental Disorder, Technology, Data mining, Analysis, Correlation, elbow method, k-means, Decision Tree, Naïve Bayes, ANN , KNN.

CHAPTER 1

INTRODUCTION

Mental health is rusted in our society. It means our emotional, psychological, and social well-being. It helps to identify how we handle stress and other things related to it. It affects our thinking, feeling and action. Mental health is important at every stage of life, from childhood to adulthood. Nowadays Mental Health is trending topic in our society. One in four people in the world will be affected by mental or neurological disorders at some point in their lives. Around 450 million people currently suffer from such conditions, mental disorders among the leading causes of ill-health and disability worldwide (WHO). A report from BIMA discovered that people working in the tech industry are five times more depressed than the general UK population by researching more than 3000 members of the UK technology community. They found that 52% had suffered from anxiety or depression at some point.

Mental illness affects our health conditions including changes in emotion, thinking or behavior. Mental illnesses are associated with extremity and problems functioning in social, work or family activities. Mental illness does not discriminate; it can affect anyone regardless of your age, gender, income, social status, religion, sexual orientation, other aspect of cultural identity. Illness can occur at any age, three-fourths of all mental illness begins by age 24.

This research targets to measure attitudes towards mental health in the tech people, and examine the condition of mental health disorders among tech people by using four data mining technique. The data will be used to measuring how mental health is viewed within the tech/IT people, and the prevalence of certain mental health disorders within the tech industry. The results can be used to raise awareness and improve conditions for those with mental health disorders in the IT people. Mental health isn't just mental illness – it is part of being human.

In the last two decades there has been a steady increase in the use of Data Mining techniques in various disciplines. Data Mining is a path to knowledge discovery and is a significant process to discover patterns in data by exploring and modeling large amounts of data.

Data Mining incorporates automatic learning algorithms to learn, extract and identify useful information and Subsequent knowledge of large databases. In the last 10 years Data Mining techniques have been used in medical research, mainly in neuroscience and biomedicine.(Alonso et al., 2018).

1.1 Background

The impacts of mental disorder are obvious and inescapable, in anguish, loss of flexibility and life opportunities, negative effects on utilization of instruction, work fulfillment and profitability, difficulties in law, organizations of human services, concentrated logical investigation into causes and fixes et cetera. Enduring, loss of working, and saw danger are among the individual and social experiences that can prompt psychological wellness administrations. Once the issues are conveyed to the consideration of emotional well-being administrations and mental issue is analyzed, a scope of conceivable results is authorized, including offer of treatment, subsidizing and maybe, contingent upon seriousness and different conditions, paid leave from work because of disease, conceivable disgrace and shame, and in outrageous cases compulsory admission to clinic, or acknowledgment of no or decreased duty in the Courts. Emotional well-being experts draw in with the issues inside institutional structures utilizing manuals for finding and giving medicines that are progressively required to be upheld by logical confirmation of viability. The social and institutional results of relegating an analysis are imperative subjects for social logical hypothesis and research. Notwithstanding, prior in the chain of occasions and outcomes are the social indications of mental issue, open for all to see, and above all the individual and relational impacts, experienced by the general population with the issues, their families and companions (Derek, 2008).

1.1.1 Categories of Mental Disorder

Mental disorder is a condition that influences a man's reasoning, feeling, and conduct. The standard therapeutic network perceives in excess of 200 characterized kinds of mental sickness.

These conditions can adjust your capacity to identify with other individuals, work, and go to class, and can keep you from carrying on with an ordinary life. Diverse kinds of mental sickness offer distinctive encounters, and side effects may change from individual to-individual, even when they share a similar conclusion.

There are several major categories of mental disorders (Andrews, 2018):

- Organic mental disorders
- Substance abuse disorders
- Schizophrenia
- Mood disorder
- Anxiety disorders

- Somatization disorders
- Eating disorder
- Personality disorders
- Gender dysphoria
- Conduct disorders
- Neurodevelopment disorders
- Bipolar and related disorders
- Depressive disorders
- Obsessive-Compulsive and related disorders
- Sleep-walking disorders

1.1.2 Cause of Mental Disorder

There is no single reason for mental disorder; rather, they can be caused by a blend of natural, mental and ecological variables. Individuals who have a family history of psychological well-being clutters might be more inclined to creating one sooner or later. Changes in mind science from substance mishandle or changes in eating routine can likewise cause mental disorder.

Psychological factors and natural factors, for example, childhood and social introduction can frame the establishments for destructive idea designs related with mental disarranges. Just a confirmed emotional well-being proficient can give an exact finding of the reasons for a given issue.

There are several major causes of mental disorder discussed by C. Flynn (FLYNN, 2016):

- Genetics
- Prenatal damage
- Infection, disease and toxins
- Neurotransmitter systems
- Injury and brain defects
- Substance abuse

1.1.3 Mental Disorder Risk Factors

Certain factors may increase your risk of developing mental health problems, including (WHO, 2017):

1. living in a region that has few or no network assets
2. guardians who are destitute
3. dysfunctional behavior in the family
4. dependence in the family
5. practically no help from relatives
6. pompous or damaging reactions from relatives about a high scholar's understanding
7. history of injury
8. encountering a learning issue
9. history of animosity in the family or in the network
10. low IQ
11. association with medications, liquor, or tobacco
12. powerlessness to control conduct
13. shortfalls in social or psychological capacities
14. formative deferrals
15. relationship with reprobate or undesirable companions
16. social dismissal by peers
17. poor scholarly execution
18. low pledge to class
19. poor family working
20. low parental association
16. practically zero connection to guardians or parental figures

1.1.4 Mental Disorder and violence

A factual connection among violence and some mental issue is currently undoubted, however it is imperative to get this into point of view. In the first place, the dominant part of individuals with schizophrenia are not vicious. Neither the four-nor fivefold height of rate of viciousness over that in the all-inclusive community among individuals with uncomplicated schizophrenia, nor the significantly higher rate still for the individuals who too mishandle medications or liquor, ought to be misconstrued along these lines. Second, in spite of these rates, next to no of society's withdrawn brutality will be by individuals with schizophrenia – most likely around 3– 4% as it were.

Huge relationship among disorder and violence may mean one is caused by the other, yet they may happen together as a result of some normal cause. The connections might be immediate, or interceded by mediating steps. Distinctive clarifications of brutality may apply with various scatters. In the Dunedin birth companion ponder, for liquor reliance the interceding factor gave off an impression of being utilization of liquor at the season of the viciousness while for cannabis reliance it was the related way of life. A solitary clarification was weakest for the general population with schizophrenia range issue: just a single third of the savagery was clarified by a longstanding suspicious state of mind (Arseneault, Moffitt, & all, 2000).

1.2 Motivation of the Research

Over the years, technology has changed our point of view of the world. Technology has made life easier day by day. As a result, technical peoples are committed to discovering innovative things to upgrade existing technology. A technical people is busy with multiple tasks with limited timelines, manages multiple clients and so on. Therefore, every technical person is actually suffering from depression. That is why there is more chance of having a mental disorder. This is a global problem. Many researchers are working on the analysis of mental health data with different types of data mining approaches. None of them is not showing the reason why mental disorder occurs among the tech people in tech workplaces. Therefore, my interest is to find out the predicated reason for mental illness among the tech people.

1.3 Problem Statement

Mental disorder is a global problem for technical peoples in tech workplaces. As a technical people, they are suffering for depression, work pressure, handling clients, limited timeline and so on. As a result, these people are suffering for mental illness. For this reason, mental patients will increase and the normal life of human beings will be hampered. As a result, many families will become anxious and this will be one of the problems in the world.

1.4 Research Questions

1. Question 1: Make a model that find out which reason are causes of mental disorder and which algorithm is effective?
2. Question 3: Develop a technique that what are the effective wat to clustering the mental health dataset?

1.5 Research Objectives

This research objective is to propose a model for determining the reason for causes of mental illness among the tech people. Also showing which classification algorithm is better for classified the mental health analysis among the Decision Tree, Naive Bayes, ANN and K-NN. Finally, find out the relation between mental illness and working position in tech workplaces and visualizing the result.

1.6 Research Scope

Data mining techniques are used in lot of research areas such as mathematics, cybernetics, genetics and marketing and much more. This paper is analysis with four data mining algorithm over the tech people in tech workplaces.

1.7 Thesis Organization

This paper includes five sections: Introduction, Literature Review, Research Methodology, Result and Discussion, and Conclusions. Introduction section discuss about the research background, research objective, problem statement, research question and research scope.

Literature review section discuss about the related work of this research and research gap. Research methodology section, shown a proposed model for the research and discuss about the research methodology. Result and Discussion section, shown the result of the methodology with discussion. Finally, conclusion section, discuss the final output of the result and future recommendations.

CHAPTER 2

LITERATURE REVIEW

Intelligent data mining and Machine Learning for Mental Health Diagnosis using genetic Algorithm (Azar et al., 2015), a novel study introduces a semi-automated system that helps preliminary diagnosis of the psychological disorder patient. They ensure their classifier is aware of all possible mental health illnesses could match patient's symptoms. They follow "Diagnosis and statistical manual of mental disorders" is text format used a softcopy. Each criteria of mental disorder have been identified and formed into a question for asking to the user and loaded into the database. The test urns did not find as good a solution with k-means as the lowest solution from the genetic algorithm. This algorithm solved increasing the success of information retrieval and relevancy between keywords-matching and relevant user's symptoms as shown in future, they work with target data set generalization and investigation the possibility of integrating multiple sources of data for improving the data extraction quality.

(Yuan, 2014), In order to determine the students' mental health level, the collected on mental health of college students and use the data mining tools. IDA to, based on supervised learning. Their aim to explore data mining technique to detect three types of SNMDs which are Cyber-Relationship Addiction net compulsion and Information overload, They categorize two types of features are social interaction feature and personal feature. In social interaction feature they consider the para-social relationship, Online and offline interaction ratio (ONOFF), social searching vs browsing (SSD). And personal features are self-disclosure based feature (SD), Temporal behavior features (TEMP). They find 30% of college students have problems in their opposite-sex contacts and 17% of college students have the sense of self contempt, envy and even resentment. 14% of student believe their college lives are interesting and 61% of students believe their college lives are flat and ordinary.

Another interesting example is (Shuai et al., 2016), here they propose a Machine Learning framework that exploits features namely social Network Mental Disorder Detection. They

also exploit based Tensor Model (STM) to improve the performance In result each community in the dataset is represented by three different types of points, CR, NC and IO. They saw the IO/NC point are similar and each SNMD type, when the average SNM users in the community.

(Sumathi & B., 2016) This paper is about the early diagnosis of mental health problem. In this paper describe machine learning algorithm and compared to measure accuracy in diagnosing five basic mental health problems. They words with the attention problem, Academic problem, Anxiety problem, Attention Deficit Hyperactivity Disorder (ADHD), pervasive Developmental Disorder (PDD). They used The Averaged One-dependence Estimator; Multilayer perceptron, Radial Network, IB1, KStar, Multiclass classifier and LAD Tree produce more accurate than others. In future, they interested to work with a large dataset.

(Deziel, Olawo, Truchon, & Golab, 2013)This is a survey based on guidelines from the Canadian Mental health Association and applied classification and regression algorithm. They do a survey on first and final year students. The first step they collected answer about potential academic influences on mental health. Second part they categorized five aspects Ability to Enjoy Life, Resilience, Balance, Emotional Flexibility, Self-Actualization. They used WEKA tool to data mining discretized the numeric attributes. In the result they found homework is the greatest effects, second-year students had the highest scores and final-year students are the lowest scores. Electrical Engineering students had lower mental health, Systems design students had higher scores and women in all Engineering programs have lower overall mental health.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Each issue solver takes after some preprocessing way to deal with take care of their issues. This research additionally takes after some preprocessing system on logical methodologies. The research methodology separated into a few sections for getting the outcome as much as literal such as data collection, data preprocessing, data analysis and Visualizing the outcome. In Figure:3.1 showing the thesis proposed model:

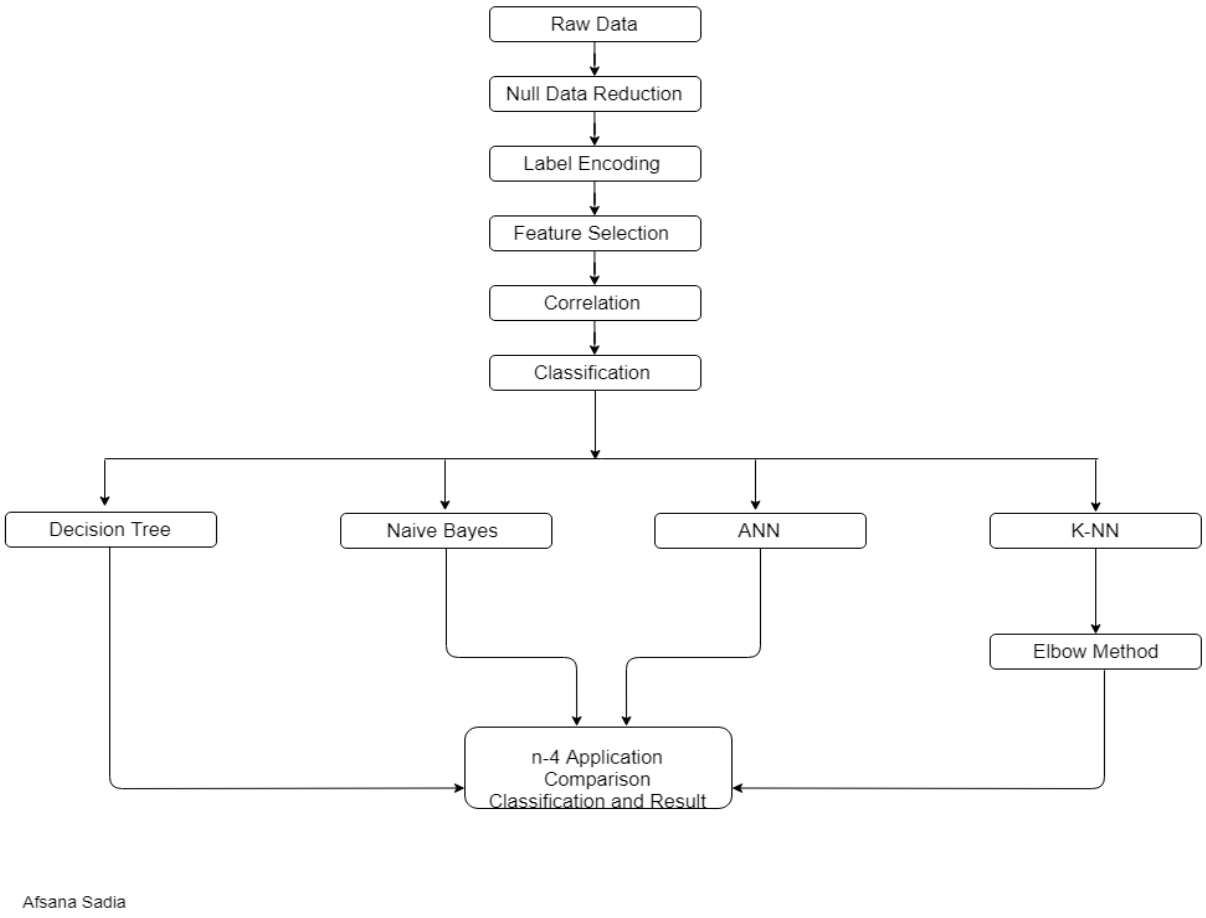

Figure 3.1: Proposed model for mental health data analysis

3.2 Data Collection

Data Collection Data collection empowers a person or organization to answer relevant inquiries, evaluate results, and make predictions about future prospects and patterns in order to obtain a complete and precise picture of a conspiracy's territory. Accurate data collection is fundamental to maintaining the credibility of the research, setting educated business choices and guaranteeing quality assurance. (OSMI Mental Health in Tech Survey 2016) Currently over 1400 responses, the ongoing 2016 survey aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers. Form "Kaggle" downloaded the dataset.

3.3 Data Preprocessing

It is a data mining technique that transform raw data into an understandable format. Raw data (real world data) is always incomplete and that data can’t be sent through a model. That would cause certain errors. That is why we need to preprocess data before sending through a model. Steps of Data Preprocessing are the steps I have followed: Import libraries, Read data, checking for missing values, checking for categorical data, Standardize the data. For the research, used different process for processing the data such as Null value reduction, Label Encoding Demonstrated the all preprocessing steps bellow Import libraries, Read data, checking for missing values, checking for categorical data, Standardize the data.

3.3.1 Null data reduction

In real world database and data warehouse, there are some commonplace properties such as incomplete, noisy and inconsistent data. There are several fields that have no recorded data for several column, then the null values are removed by ignore the row. This is actually used when the class label is not found. This method is poor when the percentage of missing values are increase. The algorithm for remove null data from dataset is shown in Figure 3.2.

```
Load_DataSet()
For each row of rows (Number of total records) {
        IF row contain null value {
                Then remove entire row
        } Else { continue
        }
}
Save_DataSet()
```

Figure 3.2: Algorithm for remove null data from dataset

3.3.2 Label Encoding

Label encoding means, notice which text is converted to a number. These are as evidence of the conversion of text values. Further, make the decision easily evidence file. Although numerical values are more efficient for machine learning but think out or understand the result need these label coding evidence. When we perform classification, we often don't handle too many names. These marks can be as words, numbers, or something unique. The ability to learn machine in sklearn Guess what their number will be. Therefore, in the event that they are as of now numbers, at that point we can utilize them specifically to begin preparing. In any case, this is not generally the case. In reality, names are as words, since words are human readable. We name our preparation information with words so the mapping can be followed.

To change word names to numbers, we need to use a name encoder. Name encoding Refer to the way toward changing the word names into numerical frame. This rifle the calculations to work on our data. The classified value encoding label contains a function of the data frame in the Panda library to encode the data and reverse the encoded values. The algorithm for label encoding is show in Figure 3.3.

```
Load_DataSet()
For each column of columns (Number of total columns/attribute)
        { unique values ← Finding unique values of the column
        For I = 0 to M – 1 (M number of unique values) {
                Encoding ← Encoded index of I unique values
        }
}
Save_DataSet()
```

Figure 3.3: Algorithm for label encoding of the dataset

3.4 Feature Selection

Feature selection has served as a field of design recognition, insight and effective exploration in data mining networks. The basic idea of feature selection is to select a subset of the information that is deleted without any natural information. Feature selection can fundamentally enhance the concepts of the next classification models, and often create a model that adds a better amount of latent focus. Furthermore, it is generally seen that finding the right subset of natural features is an important issue in its own particular right. For example, a doctor may decide whether a hazardous treatment modality is necessary to treat, considering treatment highlights. Feature selection is fundamental for building a decent model for a few reasons. One is that some material preferences suggest some level of cardinality reduction to force a cut-off on the amount of properties that can be considered when building a model. Often the data contains more data or the wrong type of data than expected to make the model. For example, someone has a dataset with a thousand sections that illustrates the client's characteristics; In any case, if the information in any of the sections is extremely inadequate which will add almost no benefit to them from being added to the model, and if one section of the sections duplicates each other, using two categories can affect the model. For feature selection in the research we apply Correlation Coefficient algorithm to find out the features which are more important to the target attribute.

3.4.3 Correlation

A statistical method of correlation that measures and breaks the level of connection between two variables. Correlation investigation manages the relationship between at least two variables. Correlation signifies the interdependency among the factors for corresponding two wonder, it should be two wonder cause-and-effect relationships that are fundamental and if there is no such relation, then two wonder relations cannot be involved. There are two factors that vary, so that development involves the development of the other, these factors are called the relationship between the situation and the end result. The correlation coefficient, r, is an outline measure that portrays the degree of the factual connection between two interim or proportion level variables. The correlation coefficient is scaled with the goal that it is dependably between - 1 and +1. At the point when r is near 0 this implies there is little correlation between the factors and the more distant far from 0 r is, in either the positive. The degree of relationship between the variables under consideration is measure through the correlation analysis. If X is a attribute whereas Y is a target attribute, then the correlation of X & Y is

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Where N is the total number of records. The correlation esteem is should between - 1 to +1. On the off chance that the connection is lower than 0 and close to - 1 then X and Y are negative related. In the event that connection esteem is close to positive 1 at that point there is a positive connection between X and Y. In the event that the correlation is 0, at that point there is no connection between X and Y.

3.5 Classification

Classification is the task of data analysis, that is, the process of finding a model that describes and differentiates data classes and concepts. The classification involves the problem of defining a set of a category (subpopulation), a new set of observations based on the training set associated with the observation, and whose membership of the categories is known. Categorization is a data mining technique that drills data collection sections to help more accurately predict and test. Additionally, in some cases, known as decision trees, a classification is expected to be one of the few techniques that make extensive dataset analysis compelling. Construction of classification models different algorithms are used to create a hierarchy by teaching models using the training set available. The model needs to be trained to predict accurate results. The model is used to predict class labels and test the model built into the test data, and therefore estimates the accuracy of the classification rules.

3.5.1 Decision Tree

Decision trees are an exceptionally compelling strategy for supervised learning. Its points are the partition of a dataset into bunches as homogeneous as conceivable as far as the variable to be predicted. It takes as info an arrangement of characterized information, and yields a tree that looks like to an introduction outline where each node (leaf) is a decision (a class) and each non-last node (inward) represents to a test. Each leaf present to the decision of having a place with a class of information checking all tests way from the root to the leaf. Here fig3.3 (Lernverfahren et al., n.d.)

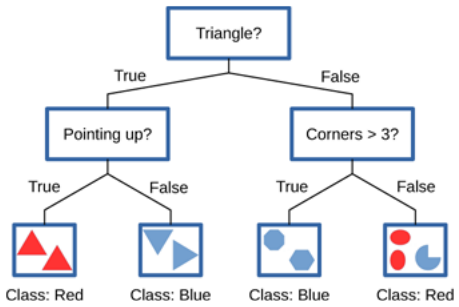

Figure 3.4: Decision tree example

The two most important things to consider in the decision tree are: How to make the best split node decision? 2 when Stop splitting? Because the original information cannot be pure. Missing data attribute, data must be present wrong, the noise in this situation, which will result. Over fit may reduce accuracy Classification and decision making predict the tree and increase the complexity of the tree Structure. So after creating a tree, we also need to prune.

3.5.2 Naïve Bayes:

It has good classification efficiency and stable classification effect. The main thought of Naive Bayes is when a classification given, calculate the happening possibility of each condition .(Russell & Domm, 1995)

$$\mathbf{P(A|B)} = \frac{\boldsymbol{P(B|A)P(A)}}{\boldsymbol{P(B)}}$$

Figure 3.5: Naïve Bayes example

It tells us how often A happens *given that B happens*, written **P(A|B)**, when we know how often B happens given that A happens, written **P(B|A)** , and how likely A and B are on their own.

- P(A|B) is “Probability of A given B”, the probability of A given that B happens
- P(A) is Probability of A
- P(B|A) is “Probability of B given A”, the probability of B given that A happens
- P(B) is Probability of B

This classifier assumes the features are independent. Hence the word naive. Even with this it is powerful algorithm used for

- Real time Prediction
- Text classification/ Spam Filtering
- Recommendation System

3.5.3 ANN:

Artificial neural networks are composed of elementary computational units called neurons (McCulloch & Pitts, 1943) combined according to different architectures. For example, they can be arranged in layers (multi-layer network), or they may have a connection topology. Layered networks consist of:

• Input layer, made of n neurons (one for each network input);

• Hidden layer, composed of one or more hidden (or intermediate) layers consisting of m neurons;

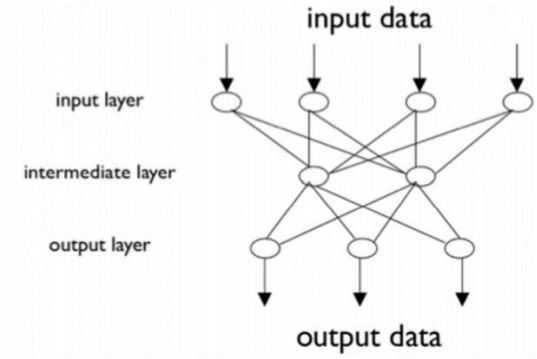

Figure 3.6: ANN example

Output layer, consisting of p neurons (one for each network output). The connection mode allows distinguishing between two types of architectures:

• The feedback architecture, with connections between neurons of the same or previous layer;

• The feedforward architecture (Hornik, Stinchcombe, & White, 1989), without feedback connections (signals go only to the next layer's neurons).

3.5.4 KNN:

K-nearest neighbors uses the local neighborhood to obtain a prediction The K memorized examples more similar to the one that is being classified are retrieved A distance function is needed to compare the examples similarity Euclidean distance (d(xj , xk) = $\sqrt{\sum_i (x_{i,j} - x_{k,i})^2}$) 2) Manhattan distance (d(x_j , x_k) = $\sum_i |x_{j,i} - x_{k,i}|$)This means that if we change the distance function, we change how examples are classified.

We can use locality sensitive hashing (approximate k-nn) Examples are inserted in multiple hash tables that use hash functions that with high probability put together examples that are close We retrieve from all the hash tables the examples that are in the bin of the query example We compute the k-nn only with these examples.

There are different possibilities for computing the class from the k nearest neighbors Majority vote Distance weighted vote Inverse of the distance Inverse of the square of the distance Kernel functions (gaussian kernel, tricube kernel, ...) Once we use weights for the prediction we can relax the constraint of using only k neighbors 1 We can use k examples (local model) 2 We can use all examples (global model).

3.6 Clustering

Clustering is the collection of a specific group of items based of their attributes, collecting them as indicated by their likenesses. Regarding to data mining, this procedure parcels the data actualizing a particular join calculation, most reasonable for the desire data analysis.

3.6.1 Finding Number of Cluster using elbow method

Determining the ideal number of clusters in an informational index is a central issue in apportioning clustering, for example, k-means clustering, which requires the client to indicate the quantity of cluster k to be produced. Unfortunately, there is no definitive answer to this question. There are several methods to identify the optimal k for k-means clustering such as elbow method, Average silhouette method, Gap statistic method and etc. To determine the optimal k for the research, used elbow method.

The Elbow technique is a strategy for understanding and approval of consistency inside group investigation intended to help finding the proper number of clusters in a dataset. Determining the ideal number of clusters in a data set is a principal issue in parceling grouping, for example, k-means cluster, which requires the number of clusters that defined by user.

Unfortunately, there is no rules or formula to define the exact clusters number. The optimal number of clusters is by one means or another subjective and relies upon the strategy. Utilized for estimating likenesses and the parameters utilized for partitioning. A familiar and well-known solution consists of examining the dendrogram delivered utilizing various leveled cluster to check whether it recommends a specific number of groups.

This technique exists upon the possibility that one should to pick a number of clusters with the goal that including another cluster doesn't give much better demonstrating of the information. The level of difference clarified by the groups is plotted against the quantity of clusters. The primary groups will include much data yet at some point the minimal pick up will drop drastically and gives an edge in the diagram. The right k i.e. number of groups is picked now, consequently the "elbow standard". The thought is that Start with K=2, and continue expanding it in each progression by 1, computing your groups and the cost that accompanies the preparing. At some incentive for K the cost drops dramatically, and after that it achieves a level when you increment it further. This is the K esteem you need. The method of reasoning is that after this, you increment the quantity of groups yet the new cluster is extremely close to a portion of the current. The average internal sum of squares is the average distance between points inside of a cluster. Mathematically.

Where the number of clusters, is the number of points in cluster and is the sum of distances between all points in a cluster.

Algorithm of elbow method to determine the number of k of k-means:

1. Initial k = 1;
2. Begin
3. Increment the value of k by 1
4. Measure the distance for optimal quality solution
5. If any point the distance of the solution is drops dramatically
6. That's point is the optimal k
7. End

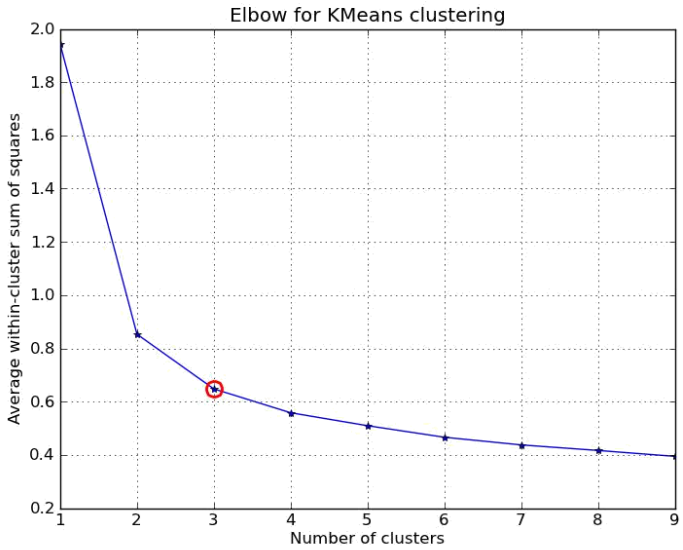

Figure 3.7: Elbow Method Example

In the above graph, see that at point 3, the distance drops dramatically. So, this point is the optimal k = 3.

3.7 Visualization

For visualizing the result requires a presentation method that is comfortable, meaningful and spectacular for present the result. The Matplotlib package is an awesome package that are able to visualizing one dimensional, two dimensional and especially three-dimensional graph and can serve a GUI. For 3D dimensional plotting, apply the axes3d library which is more powerful for plot 3D objects on 2D matplotlib figure. The library supported poly collection, line collection and patch collection. The library modifying the collection of 2D objects and converted to a 3D object and adding z coordinate information.

CHAPTER 4

RESUTL AND DECISION

In this paper, there are several parts in result section such as Feature Selection (Correlation) result and combined result of these four features selection algorithms, Comparison Decision Tree, Naïve Bayes, ANN and KNN, elbow method result, and visualization the result and finally discussion with the result.

4.1 Feature Selection Result

4.1.1 Correlation Result

In Chapter 3, Section 3.1.1, discussed about the Correlation algorithm and show how to calculate correlation between attribute and target attribute. After applying correlation algorithm on the dataset, then found a result, which indicates the correlation value for each attribute and show the result state. Correlation result is shown in Table: 4.1

	0	1	2	3	4	5	6	7	8	9
0	1	0.35	0.4	0.46	0.073	-0.23	-0.73	0.48	-0.44	0.015
1	0.35	1	-0.28	0.57	-0.29	0.38	-0.36	0.64	0.25	0.19
2	0.4	-0.28	1	-0.52	0.15	-0.14	-0.093	0.016	-0.43	-0.38
3	0.46	0.57	-0.52	1	-0.23	-0.23	-0.48	0.47	0.28	0.45
4	0.073	-0.29	0.15	-0.23	1	-0.1	-0.15	-0.52	-0.61	-0.19
5	-0.23	0.38	-0.14	-0.23	-0.1	1	-0.03	0.42	0.21	0.095
6	-0.73	-0.36	-0.093	-0.48	-0.15	-0.03	1	-0.49	0.38	-0.35
7	0.48	0.64	0.016	0.47	-0.52	0.42	-0.49	1	0.38	0.42
8	-0.44	0.25	-0.43	0.28	-0.61	0.21	0.38	0.38	1	0.15
9	0.015	0.19	-0.38	0.45	-0.19	0.095	-0.35	0.42	0.15	1

Figure 4.1: Correlation Table

Here, correlation value is always between -1 to +1. If the correlation value is 0, then the attribute has no relation with the target attribute. On the other hand, if the value is greater than 0, then there is a positive relation between attribute and target attribute and the relation level is depending on nearest value of +1. Similarly, if the value is lower than 0, then there is a negative relation between attribute and target attribute and the relation level is depending on nearest value of -1.

4.2 Classification Results:

In this experiment, we consider different analysis to examine the four data mining classification techniques for the classification of mental health problem dataset. Figure 4.2 shows the confusion matrix of prediction results for Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN) and artificial neural network (ANN) algorithms.

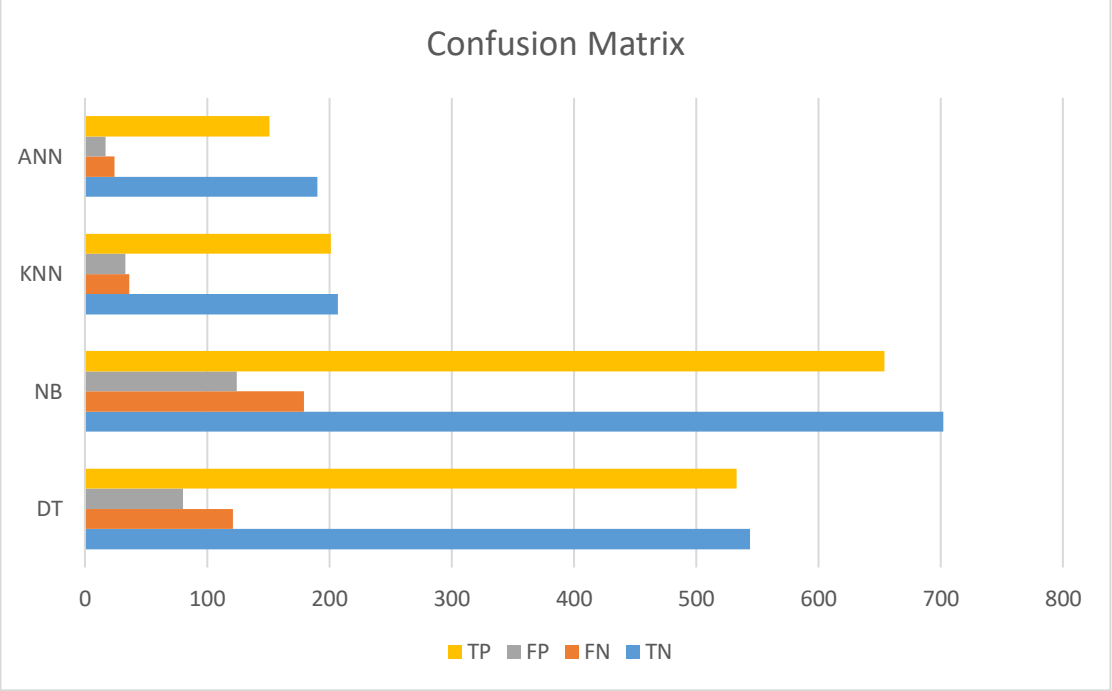

Figure 4.2: Classification Result using Confusion matrix

Figure 4.3 shows the prediction accuracy of these datamining algorithm for mental health problem detection. Here, ANN achieved better performance than the other classification techniques by obtaining 90% accuracy, whereas NB shows the worst performance by attaining 81% accuracy.

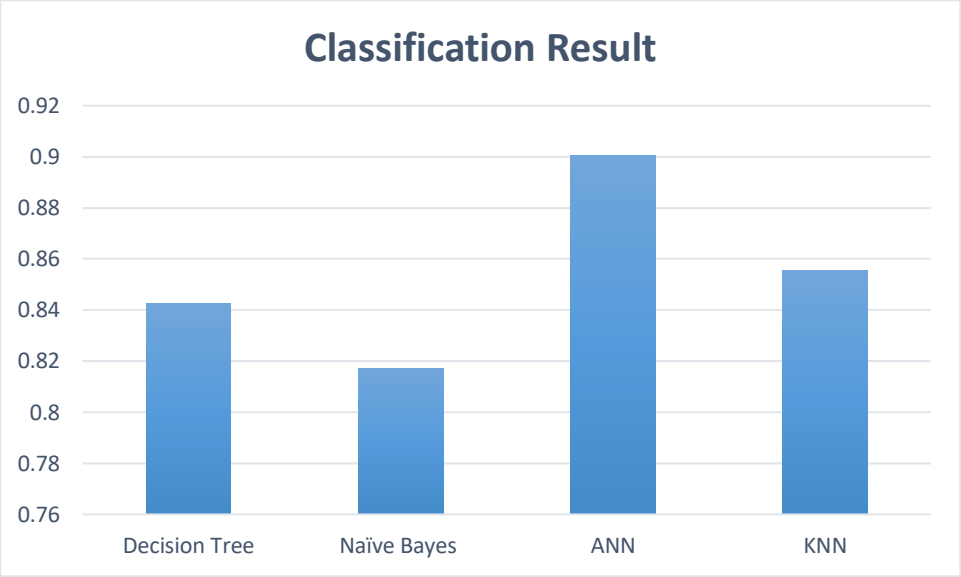

Figure 4.3: Prediction accuracy of classification techniques

Algorithm	Result
Decision Tree	0.8427
Naïve Bayes	0.8173
ANN	0.9005
KNN	0.8553

Figure 4.4: Classification Result Table

4.3 Elbow Result

In this paper Chapter 3, Section 3.6.1, already discussed about the elbow method and how to calculate the optimal K. After executing the elbow method against the dataset, then get a graph is called elbow graph and shown in Figure 4.3.

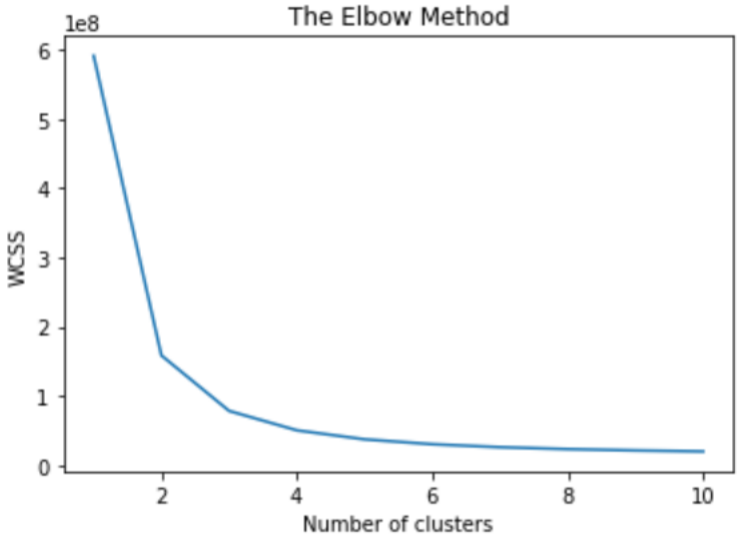

Figure 4.5: Elbow graph for determine the optimal k

In Figure 4.3, the distance is drop dramatically at cluster number 2. Hence, in this point the sum of squared errors is small than the other clusters. Therefore, we can pick the optimal cluster number 2 for k-means clustering. As a result, number of cluster k equal 2's values are more variance and effective than the other clusters value.

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

5.1 Findings and Contributions

Prediction is not always defined the exact result but shown the assumption. In Figure 4.3, shown the predicted result between age, work position and finally the medical professional decision. The prediction result describe that, tech people are affected by mental illness whose age is between 26-46 years and whose work position is System Admin, Front End Developer, Back End Developer, Executive Leadership, Team Leader, Developer according to the encoding label. This result is not claimed 100% accurate, but based on statistics and data analysis that can be happened. Data analysis is not a simple task. The real-world data is not organized. In this paper, we apply four classification algorithm Decision Tree, Naive Bayes, ANN and KNN. In figure 4.2, we show Naïve Bayes give us best accuracy.

5.2 Recommendations for Future Works

This analysis research is very long process. In this paper, within short time we tries to cover the data analysis process. Hence, the research topic is very importance, because the mental disorder problem is global problem. In this paper, shown the reason for mental illness frequency among the tech people. In fact mental illness not only occurred among the tech people. Civilians are also affected by mental illness. In future, want to analysis the frequency of mental illness among the civilians and comparison the result between the tech people and civilians.

REFERENCES

1. Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. *Journal of Medical Systems*, *42*(9). https://doi.org/10.1007/s10916-018-1018-2

2. Azar, G., Gloster, C., El-Bathy, N., Yu, S., Neela, R. H., & Alothman, I. (2015). Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm. *IEEE International Conference on Electro Information Technology*, *2015-June*, 201–206. https://doi.org/10.1109/EIT.2015.7293425

3. Deziel, M., Olawo, D., Truchon, L., & Golab, L. (2013). Analyzing the Mental Health of Engineering Students using Classification and Regression. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*.

4. Gallo, C. (2018). *Artificial Neural Networks : tutorial*. (July).

5. Lernverfahren, M., Master-thesis, K. R., Arnold, T., Date, L., Weihe, K., & Frank, A. (n.d.). *A Machine Learning Approach for Coreference Resolution*.
originally published in: Bulletin of Mathematical Biophysics, Vol. 5, 1943, p. 115-133. (2008). *5*, 115–133.

6. Russell, C. J., & Domm, D. R. (1995). *Two field tests of an explanation of assessment centre validity*. 25–47.

7. Shuai, H. H., Shen, C. Y., Yang, D. N., Lan, Y. F., Lee, W. C., Yu, P. S., & Chen, M. S. (2016). Mining online social data for detecting social network mental disorders. *25th International World Wide Web Conference, WWW 2016*, 275–285. https://doi.org/10.1145/2872427.2882996

8. Sumathi, M., & B., D. (2016). Prediction of Mental Health Problems Among Children Using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, *7*(1), 552–557. https://doi.org/10.14569/ijacsa.2016.070176

9. Yuan, C. (2014). Data mining techniques with its application to the dataset of mental health of college students. *Proceedings - 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications, WARTIA 2014*, 391–393. https://doi.org/10.1109/WARTIA.2014.6976277

10. Bhukya, D. P., & Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. *International Journal of Computer and Electrical Engineering*, 660–665. doi: 10.7763/ijcee.2010.v2.208

11. Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre

validity. *Journal of Occupational and Organizational Psychology*, *68*(1), 25–47. doi: 10.1111/j.2044-8325.1995.tb00686.x

12. Yurtoğlu, N. (2018). http://www.historystudies.net/dergi//birinci-dunya-savasinda-bir-asayis-sorunu-sebinkarahisar-ermeni-isyani20181092a4a8f.pdf. *History Studies International Journal of History*, *10*(7), 241–264. doi: 10.9737/hist.2018.658

13. Lustig, M., & Hill, W. E. (1967). The .mu.-oxo-difluorophosphines (CF3)2C(OPF2)I, (CF3)-2C(OPF2)Br, and (CF3)2C(OPF2)H. *Inorganic Chemistry*, *6*(8), 1448–1450. doi: 10.1021/ic50054a003