Analyzing & Predicting Depression Risk & Types

B.P.N Perera

2021

Analyzing & Predicting Depression Risk & Types

A dissertation submitted for the Degree of Master of Business Analytics

B.P.N Perera

University of Colombo School of Computing

2021

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name:

Registration Number:

Index Number:

____

Signature: Date: 2021/09/14

This is to certify that this thesis is based on the work of

Ms. B.P.N Perera

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr H.N.D Thilini 14/09/2021

I would like to dedicate this thesis to...

ACKNOWLEDGEMENT

I would firstly like to express my deepest gratitude to my internal supervisors, Dr.H.N.D Thilini for the guidance, dedication and helping hand throughout this study and made this research to harvest its maximum benefits through readily providing all the facilities to successfully carryout this research within the time frame.

I express my sincere gratitude to Dr.Prasadi Lokuthotahewa, Doctor of the government hospital, Hambanthota for the guidance and for giving knowledge and to help me to obtain medical data from the hospital.

And I would like to express my sincere gratitude to Prof.Gnadasa Perera, Specialized in sociology for the guidance and knowledge about depression.

Further I would like to offer my gratitude to Mrs. I.Gamage, Nurse of the government hospital, Hambanthota, for providing facilities for obtain medical data from the hospital and helping me to categorize the medical data.

I would like to express special acknowledge to my colleagues of University of Colombo School of Computing, who provided available support and contribution during my research work.

Finally I express my sincere gratitude to my parents and brothers for their support, strength and providing me with calm background to do this work properly.

ABSTRACT

This research is based on prediction of depression risk, analyzing the initial mental status of a patient and depression types by combining the knowledge of computer science, data analytics with the medical field. Depression is a one of a leading cause of disability worldwide. It's a mental illness that effects negatively on how you feel the way you think and how you act. In this research, we considered 4 types of depression (Major, Persistent, Bipolar, and Atypical) out of all main 6 types of depression. Here, we selected 1230 people for this research and collected data related to 28 attributes of depression were selected as variables. Statistical techniques were applied to predict the risk of being a patient and to analyze the dataset. Statistical analysis was done for identify the effectiveness of each risk factor on the depression prediction and the most suitable risk factors (p value <0.05) were identified and visualized based on the target variable (patient/not a patient) attribute. Statistical model has created by applying binary logistic regression model. 2 mathematical equations (to calculate Y' and probability (P)) are consisted in the created statistical model which provides the probability of having depression for any person (depression positive range >0.5, depression negative range< 0.5). Model accuracy is 92%. Hosmer-Lemenshor test shows that the model fts the data well hence the value is 0.486. A system was developed by using the result of the depression prediction with a user friendly UI to find the risk of having depression or not. Initial metal status of a person; high, medium, low was analyzed using both clustering (k-means) and classification (Naïve Bayes) to identify the best suited model for the dataset. Naïve Bayes model accuracy is 70% and the accuracy of the k-means clustering is 55% percent. To predict the depression type, decision tree (J48 tree) was used and tree was built with 31 number of leaves and 48 as the size of the tree. Accuracy of the decision tree analysis is 72% and kappa statistics takes value is 0.6397 which indicates the good performance of the model.

Table of Contents

List of Figures

List of Tables

CHAPTER1

INTRODUCTION

1.1 Project Overview

Today the most important domain is “Health Care” all over the world. The healthcare environment is more and more information enriched and there vast quantity of information to be had in hospitals and clinical associated institutions, however the quantity of understanding acquired from those data could be very less, due to the fact the shortage of data analysis tools. Every person is once a affected person who need to address healthcare environment. So, it's far greater essential to conduct studies actively with clinical statistics. Medical data is some kind of data which are belonging to the health care environment. Data set is a collection of row data which relates to information on different fields. Normally, data set can be identified as a table or a database or a data matrix. This is particularly useful for prediction of diseases. Data analysis uses for identify many trends and patterns in data to gain many benefits.

Health care records are vital for plenty activities which include analyzing disorder patterns, predicting diseases, reading remedy trends etc. So, analyzing records is extra essential because of the advantages of them withinside the clinical field. Analyzing of records in healthcare without an alertness of the concomitant peculiarities and boundaries is unstable task. A opportunity always exists for variability and discontinuities in clinical practices and in personal, demographic, favors of humans variables. This is further complex with the aid of using the passage of time, with the aid of using similarities and variations inner and in reference to demographic groups. In a great world, truthful, standardized health care records might go with the drift openly amongst patients and clinical practitioners. Administrators, researchers, planners, and reviewers of the healthcare surroundings might have quick and clean get entry to to the records vital for his or her activities. There are many advantages of data related to medical field which consider discover the data instances gained from people and to get valuable knowledge about the related field or environment. Actually, It’s all about using healthcare data to drive decisions. Because data is one of the most valuable assets that the hospital and the medical domain owns. As a response to the healthcare information, the discovery has taken advantage of medical data and analytics to make strategic forecasting decisions.

Predicting a disease is one of the main parts of handling clinical facts and the healthcare environment. Because it can reduce the fatal rate of human. A predicting is a guess what happens based on observation. In this research, it builds up the early prediction of being a depression patient and the types of depression are analyzed through the statistical techniques.

Depression is truly a severe clinical contamination that negatively influences the way you sense the manner you think and the way you act. Fortunately, it is also treatable. But, most of the time people cannot identify this disease and most people don't know whether they are suffering from depression until it becomes worse. Depression has many faces, and it directly affects to mental health of people. Mental illness is a one of a leading cause of disability worldwide. It is estimated that nearly 300 million people suffer from depression (World Health Organization, 2001).

Mental fitness consists of emotional, mental and social well-being. It influences how we think, feel, and act. It allows decide how human beings deal with pressure evaluating others and make choices. Mental fitness is critical in any respect degrees of life, beginning childhood and youth via adulthood. It consists of our emotional, mental, and social well-being. It influences how we think, sense, and act. When thinking about the sensation down from time to time, it's far a ordinary a part of existence however while feelings along with hopelessness and depression take keep and simply won't move away, you can have depression. Depression changes how you think, sense, and characteristic in everyday activities greater than simply disappointment in reaction to life's struggles and setbacks. It can covered together along with your ability to work, study, eat, sleep, and experience life. (Stringaris, 2017) The trouble of depression is that, you revel in depression; left untreated it is able to grow to be a severe health condition. If we can reduce the bad habits and the risk factors, there might be a less chance to have a depression. In this project my target is to Analyze and predict the risk of having depression and the types of the depression. So the problems of identifying the possibility of having depression and the depression types will be solved from the results and findings of the project.

Early prediction of depression is extraordinarily difficult mission for clinical practitioners due to the complex interrelationship among numerous factors. To handle such data, it is very important to connect with Information Technology. Because, it plays a vital role in healthcare.

And also the day by day the assumptions and the information regarding the depression are updating due to the improvement of the medical technology and medical knowledge. So, the attributes of this disease are little bit different from the earlier. So, the most surveys have used most unrelated, unsuitable and also unnecessary attributes. And also the attributes of depression is different from each countries; due to the traditions, cultures, the way the people eat, drink, and behave etc. So the attributes should be understood according to the life style of people well. There are some methods to calculate the risk using some of these attributes but not the all. So, this research identifies the real and most suitable attributes of the depression according to the identifications of WHO and how they affect to the patient.

There are six main types of depression which occurs in human body in different situations and levels. Most researches used only one of these types because each one has their own environment. In this research I also consider four depression types; Major, Persistence, Bipolar, Atypical. And also in medical diagnosis there are two different types of possible errors. One type is which a patient, who in reality has depression, is diagnosed as disease free. And the second type in which a patient, who in reality does not depression, is diagnosed as having that disease. So the problem is to identify these possible errors and give the solution by minimizing these errors.

The overview of this research was divided into 3 main parts. One was to identify the most effected risk factors of depression and build up a model to predict the possibility of having depression. The second one was to analyze the initial mental status of a person by using only the most significant risk factors of depression. The third one was to identify the depression types, predict the depression types based on the symptoms of depression. And the final outputs of these predictions were evaluated and accuracies were counted separately. Result of the logic of depression prediction was interpret by using Java Script, HTML and CSS as a system.

1.2 Motivation

We all once suffer from many diseases and wasting time and money to rescuer from illnesses. As a student I needed to help the heath care field by applying my computer science knowledge. Determining a disease is one of the main and important things in the healthcare environment. By doing a depression prediction research, I can contribute to the rescuing from depression.

Most of the people are suffering from depression and some people are forcing to suicide themselves and also become mentally destroyed people. The target of project was to help them. Even though, I’m studying computing and information systems I wanted give my contribution to the medical field in my country. And also I wanted to prove that not only a medical researcher can help the medical field but also the other students who are not leaning medicine can also give their contribution to the medical field. Computer science is one of the valuable learning in the world. It can apply to all the other leanings, fields, environments and also in different ways. Here, there is a relationship between a disease and the information technology; that is data. Depression is based on depression data and information technology has the ability to manage the data. The output of this relationship is a decision with understanding the knowledge. Actually the common factor that combines depression and information technology is data. By gathering all the suitable depression data, we can ensure that we can turn the data in to meaningful information and then also to valuable knowledge based on depression. In information technology, data mining is the very best task to handle such kind of situations. Data analyzing is the main purpose of data mining to achieve the research objectives. Data analyzing leads to the way of predicting depression by extracts new knowledge or patterns from gathered data. And this records are generated that helps pattern forecasting on the idea of prediction, probabilities, and visualization.

Data mining and statistics had been used intensively and notably through many organizations. In healthcare surroundings, this data mining and statistics are getting an increasing number of popular, if now no longer an increasing number of essential. Use of data mining strategies can substantially gain to expect the possibly of being a melancholy affected person or now no longer. This isn't always best treasured for humans however additionally the clinical practitioners involved withinside the healthcare industry can get extra benefit from this. The massive quantities of records and information generated through healthcare surroundings are too complicated and voluminous to be processed and analyzed through traditional methods. So, those data mining and statistics strategies offer the method and generation to convert those massive amounts of data into useful and important information for decision making. These decisions can lead to effective understanding in become a patient or not.

Motivation leads to getting high accuracy through the output result. Most of research explains many ways to depression prediction but the accuracy level is not actually high and true. People who in reality have depression, should be identified as diagnosed as having that disease. And also people who in reality do not have depression should be identified as diagnosed as disease free.

1.3 Statement of the Problem

Since so many people around the world is suffering from depression, there should be a way to identify the risk factors which are mostly significant to the depression disease and it's very important if there is a way to find the possibility of having depression. And also the initial mental state of a person should be able analyze with the most significant risk factors of depression. And the also it is very effective if we have a way to predict the depression type based on the symptoms.

1.4 Research Aims and Objectives

1.4.1 Aims

Aim of the research leads to the vision of helping the health care environment by understanding the disease culture and possibilities. Aim is primarily based totally at the value of predicting a disorder and combining the pc system and records generation with the clinical field. Applying IT and CS knowledge for healthcare domain is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Among the related technologies, data mining analysis and statistical analysis in depression include analysis of health care field for better health policy. It helps to making and prevention of hospital errors, early detection of depression and depression types and prevention of depression.

Depression is the main reason of intellectual health. By thinking about the world-extensive developing mortality of depression patients each year and the provision of large quantity of patients’ records from which to extract proper observation, the use of techniques in data mining can be used to investigate on medical data and help medical professionals for understanding more on depression is the major ambition in the research. Analyzing the clinical records in depression to assist and provide recommendation to the healthcare experts in diagnosing and presenting appropriate treatment for depression patients is important.

There were three major aims in this project. One is to predict the possibility of being a depression patient. This major aim was divided in to sub aims to reach the research in success and well important. The first sub aim is to identify all the risk factors which are directly affected to depression. And then to understand the ranges which the selected attributes affect to the predicting. Another sub aim was to develop a system to interpret the result of prediction of depression with a user friendly UI. Depression professionals store huge amounts of patient’s data. Analyzing these datasets is very important to extract useful knowledge. So the simple identification of the aim of the research is to use data mining which is an effective tool for analyzing data and to extract useful knowledge.

Second major aim was to analyze the initial metal status of a person whether he/she is a depression patient or not. When thinking about how important the early understanding the mental status of a person, it will help the people to be careful about their mental health and depression professionals also will gets more benefits by doing the treatments and checking.

The third major aim was to predict the depression type based on the depression symptom. Using the statistical analysis, the other aim declares to creating prediction model. Helping the whole world to understand their possibility to be a depression patient and also helping the depression professional to understand the depression types well are the main aims which leads to the research project. Finally the all outputs are evaluated and the accuracies are calculated separately to validate the efficiency of this research project. Hence, this project is about to statistically analyze the risk factors, initial mental status and symptoms of depression, I hope this project will be a big help to health care environment in our country.

1.4.2 Objectives

So, the objectives of this research project was categorized as part 1,part 2 and part 3.

Table 1 Objectives of the Study

Part 1	✓ Identifying the risk factors of depression ✓ Analyzing the effectiveness of each risk factor ✓ Predict the risk of having depression ✓ Validate the result output ✓ Interpret the result for a developing system
Part 2	✓ Identifying the initial mental status of the depression. ✓ Finding the best suitable method for analyzing the initial status of the depression ✓ Validate the result output ✓ Visualize the result findings
Part 3	✓ Analyzing the symptoms of each depression type ✓ Predicting the depression type based on the symptoms ✓ Validate the result output ✓ Visualize the findings of the project

1.5 Scope

Data was collected from the ‘Mana Suwa Piyasa’ department in Hambanthota District General Hospital with the help of Doctor Prasadhi Lokuthotahewa. Hence this project area covers information related to BEFORE depression and AFTER depression, data was collected from people who are identified as depression patients and non-depression patients.
Related information was gained from Professor Gnanadasa Perera who is specialized in Psychology.

The scope of this project covers with statistical analyzing, predictions and also identifications, visualizing and implementations. This research structure is divided in to three main parts hence it consists with two main predictions.

First part of the project - Predict the possibility of depression risk

When considering the first part of the project, before predicting the possibility of depression risk, the risk factors were identified as attributes. And then the significances of each risk factor were identified by using statistical analysis. Then the prediction model was created by using only the identified risk factors. Prediction logic was applied to developing a system by using Java Script. The result was validated and the suitable visualizations were applied appropriately.

Second part of the project – Analyze the initial mental status of a person

Initial mental status was identified first and the analysis was done using only the selected most significant categorial risk factors. Analysis was done using both classification and clustering with class evaluation and the most suitable method was identified by comparing the accuracies of each model. Further analysis was done using the selected best model and validation and visualizations were done appropriately.

Third part of the project - Predict the depression types from the symptoms

When considering the third part of the project, before predicting the depression types, depression symptoms were identified well. The symptoms of depression were analyzed with the class of depression type statistically. Visualizations were used to give the clear picture on each modeling findings. As for the summary of the project, statistical analysis, data mining and machine learning algorithms were used to address the problems and identify the predictions of depression risk, initial mental status and depression types.

This project was based on medical data analysis instead of industrial based analysis. So the result of this project will be more useful for the health care sector of Sri Lanka. As from the output of this analysis, we can identify whether we have the risk of having depression or not. If we have a risk we can break bad habits which make increased the risk of depression and make sure to be in a good mental health.

And also medical related people can identify the depression types based on symptoms. This project will be more helpful to the medical related works.

1.6 Structure of the thesis

Chapter 1 is mainly focuses on the introduction of the overall thesis. The motivation of selecting the topic is defined clearly and it focuses on defining the aims and objectives of the overall project. Through the mentioned overall aims and objectives, the scope of the thesis was defined as part1 , part2 and part 3. Chapter 2 focuses on the background and the related works of the thesis. As for the background of the project, the definitions of depression and the identifications of depression types are described clearly. And as for the presentation of the statistical material, data mining process is described and the techniques of datamining which were needed to build up the overall project and aims of the project were describes separately and clearly. Hence there are many previous researches and studies on depression and also data mining and statistics, the summery contents of them were included as the literature review on the project thesis. In the methodology chapter, it goes a clear and more informative idea of the objective of the project and it describes the process of completing each and every objects by using target methodologies. It means, all the target methodologies and reason of selecting those methodologies were clearly defined for each and every objective. Chapter 4 defines the results of the analysis and the evaluation of each result was mentioned separate with each fulfilled am of the thesis. Chapter 5 defines the conclusion of the thesis. The summery overview, recommendations and also the future work were include in this chapter.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1 A Literature Review

The research combines two different fields together; Information Technology and healthcare environment. Through the improvement of the computer science knowledge the technology has come up with most scalable solutions which can impact health care environment well. People are both living and remaining active longer. Because of advances in healthcare, they can move healthy, self-sufficient lives well into old age. It's up to healthcare worker to proactively monitor the wellness of elders, including people who live in senior living communities.

"Health care is one of the main and important data-intensive and data-driven domains in the world. Numerous amounts of records arise from medical experts, public and private payers, subordinate service worker such as labs and pharmacies, and health care consumers. "The challenge is not only in repository and gets entry to, but also in making this data usable." This statement is stated by John Glaser from the Hospitals and Health Networks Magazine. In this magazine John Glaser has said that the issue of data overload inside the healthcare society perfectly. With every single new patient and unique fact added to the registry at one of these institutions, the burden put on dated practices increase and bigger. Expecting the particular infrastructures to deal with this enlarged load of names, medical histories, current treatments, and other data points is just impossible. This is one of the big data challenge and the world manage the challenge in to really useful consideration by arranging those big data in to efficient way. With the help of information technology they've turned the challenge excellent way to determine disease in the various ways. One of the special advantages they get through the big data is predicting disease and analyzing the ways of increasing the risk of the diseases. What makes things even not so good is that a weakness to correctly analyze and predict this information can lead to disastrous health implications for patients. Some issues can rise up as portion of a mismanaged or miscalculate data patterns. Such as misplaced diagnoses incorrect prediction etc. Getting a deal with on medical data is one of the most competitive things that a medical facility can do to boost treatment effectiveness and avoid unneeded healthcare problems. [John Glaser, *Hospitals afnd Health Networks Magazine*.]

When considering about the history of using medical data it has begun from the United States of America. They identified the importance of medical data at first and have got to used them properly. [S. Chandra, 2015] In 2011, the fact of Sorrell v. IMS Health, Inc., decided through the Supreme Court of the United States, dominated that pharmacies can also additionally proportion data with out of doors companies. This practice is allowed below the primary Amendment of the Constitution, defensive the "freedom of speech." However, the alternate of the [Maria Vargas Vera, 2001] Health Information Technology for Economic and Clinical Health Act (HITECH Act) helped to begin up the adoption of the digital fitness report (EHR) and helping technology withinside the United States. The HITECH Act become registered into regulation on February 17, 2009 as a part of the American Recovery and Reinvestment Act (ARRA) and assisted to sell to clinical facts mining. Prior to the signing of this regulation, estimates of most effective 20% of United States primarily based totally by and large physicians had been using automated patient records. Soren Brunak states that “the patient report flip in to as data-rich as possible”. Hence, expansion of computerized patient records have got to possible by the use of medical data mining by that designing medical line vast source of medical data analysis. The objective of the information extracting manner is to extract data from a dataset and alter it into a clear development for extra use. This is really a diagnostic purpose prepared to scrutinized the information in seek of stable patterns or coordinated associations associating variables, after which to make sure the findings by applying the detected patterns.

Background of depression

Depression is merged with mental state. It forces how humans sense, think and behave and may cause a number of emotional and bodily problems. People may also have problem when doing each day activities, and additionally every so often they will feel as though lifestyles isn't always really well worth living. Depression isn't always a weak spot and different human beings cannot simply "snap out" of it.

Depression may also require long-time period remedy. But are not getting discouraged. maximum of the human beings with depression gets a treatment through doing medication, psychotherapy or both. Depression isn't always best usual temper fluctuations and quick time period expressions responses to challenges in each day lifestyles. When long-lasting and with slight or intense depth this could come to be a heavy fitness condition. It can influences character to come to be go through substantially and overall performance poorly at work, at college and in the family. At its worst, depression can reason suicide. Nearly 800 000 human beings die because of the suicide every year. Suicide is that the second one main reason for demise for in most cases in 15-29-year-old human beings. Although there are known, powerful remedies for intellectual problems, among 76% and 85% of people in low- and middle-profits countries get hold of no remedy for their disease. Barriers to effective care include a lack of resources, loss of educated health-care providers and social stigma associated with intellectual problems. Another barrier to effective care is incorrect assessment. In countries of all profits levels, people which might be depressed are frequently now no longer effectively diagnosed, et al. who do not have the disease are too frequently misdiagnosed and prescribed antidepressants. The burden of depression and different psychological country situations is at the increase globally. A World Health Assembly decision handed in May 2013 has concerned a comprehensive, coordinated reaction to intellectual problems on the united states level.

There are main 6 varieties of depression categorized by the globe Health Organization. Among them I even have selected 4 sorts of depression types which are effecting to Sri Lanka.

• Major Depression

In major depression, the most outstanding symptom can be a intense and chronic low temper, profound sadness, or a manner of despair. The temper can occasionally seem as irritability. Or the man or woman struggling major depression won't be prepared to experience activities which might be commonly enjoyable. Major depression is quite only a passing blue temper, a "awful day" or brief sadness. a range of symptoms commonly accompany the low temper and consequently the signs can range significantly amongst specific people.

• Persistent Depression

Persistent clinical depression, additionally referred to as dysthymia, can be a continuous long-term (chronic) type of depression. you may get bored in normal each day activities, sense hopeless, lack productivity, and feature low vanity and an typical feeling of inadequacy. These emotions ultimate for years and have to considerably intervene collectively together along with your relationships, school, paintings and every day sports. If you have persistent clinical depression , you may discover it difficult to be upbeat even on happy occasions. you may be defined as having a darkish personality, continuously complaining or incapable of having fun. Though continual medical despair isn't always as intense as major depression, your current depressed temper can also be mild, moderate or intense.

• Bipolar Depression

Bipolar disorder, previously referred to as bipolar disorder , can be a mental state condition that reasons intense temper swings that encompass emotional highs (mania or hypomania) and lows (depression). when you turn out to be depressed, you may sense unhappy or hopeless and get bored or delight in most activities. When your temper shifts to mania or hypomania (much less intense than mania), you may sense euphoric, packed with strength or strangely irritable. These temper swings can have an effect on sleep, strength, activity, judgment, conduct and consequently the cap potential to assume clearly. Episodes of temper swings can also additionally arise not often or more than one instances a year. While most of the people will revel in a few emotional signs among episodes, a few won't revel in any.

• Atypical Depression

Atypical depression additionally referred to as depression with unusual capabilities means your depressed temper can brighten in reaction to effective events. Other key symptoms encompass increased appetite, sleeping an immoderate amount of , feeling that your hands or legs are heavy, and feeling rejected. Despite its name, atypical depression isn't always unusual or unusual. It can have an effect on how you're feeling , assume and behave, and it is able to reason emotional and bodily problems. you may have problem doing everyday daily sports, and sometimes you may sense as though life isn't always really well worth living.

2.2 Presentation of Statistic Material

Most of predicting approaches have built up by using data mining techniques. Because data mining is the process of analyzing data according to the different perspectives and summarizing it to get proper knowledge. Data mining has grown to be a special strategy in lots of industries to improve outputs and reduce costs. Techniques in data mining have change into wonderful capability for healthcare activities to predict health diseases by the use of systematic medical data and analysis to identify inefficiencies and best practices which improve care and reduce costs. These techniques are stable and get limited time for the forecasting structure to improve the depression with more accuracy.

Medical data as fact to be analyzed have different kind of features that are not only distinct from facts of other disciplines, but also distinct from traditional clinical epidemiology. Medical data analyzing technology has many areas in healthcare, such as predictive modeling, disease or safety surveillance, public health, and research. Data analytics frequently exploits analytic methods developed in data mining, including classification, clustering, and regression. Medical data analyzing are complicated with many technical issues, such as missing values, incomplete values, curse of dimensionality, and share the inherent limitations of observation study etc.

The knowledge can be identified as information associated with rules which allow interferences to be drawn automatically so that information can be used for purposes. In a medical field previous knowledge is used to predict and diagnose for a new condition.

Data Mining is the design of understanding in databases. Techniques of data mining assist to manner the data and transfer them into right information. Prediction outcomes popping out of data mining are beneficial in some of fields like Business Intelligence, Bioinformatics, Healthcare Management, Finance etc. Medical area has improved quantity in addition to form of data for processing and there exist several hard tasks. This area calls for best and well timed mannered analysis that can keep many patients life. Data mining strategies performs an crucial role in healthcare interpretation. Different approach might be used for numerous disorder investigations.

- **Data Mining**

Data mining is usually a process of extracting knowledge originating at enormous amount of databases. Data mining is useful most commonly in exploratory analysis due to nontrivial information in enormous amounts of data. It is used to find out observation out of data and presenting it in a form which is easily understood to humans. It is really a process to observe quite a lot of data typically collected. Data Mining techniques consist of collection, extraction, analysis and statistics of data. It is also known as Knowledge discovery process, Knowledge Mining coming out of data or data/ pattern analysis.

Because it is a logical process of discovery proper information in finding out useful data. There are two primary goals of data mining when completing prediction and description. Prediction comes to some attributes within the data set to predict unknown values or future values of other variables of interest. On any other hand description makes a specialty of discovery patterns describing the data that can be explained by humans. The disease prediction in healthcare domain plays an important role in data mining.

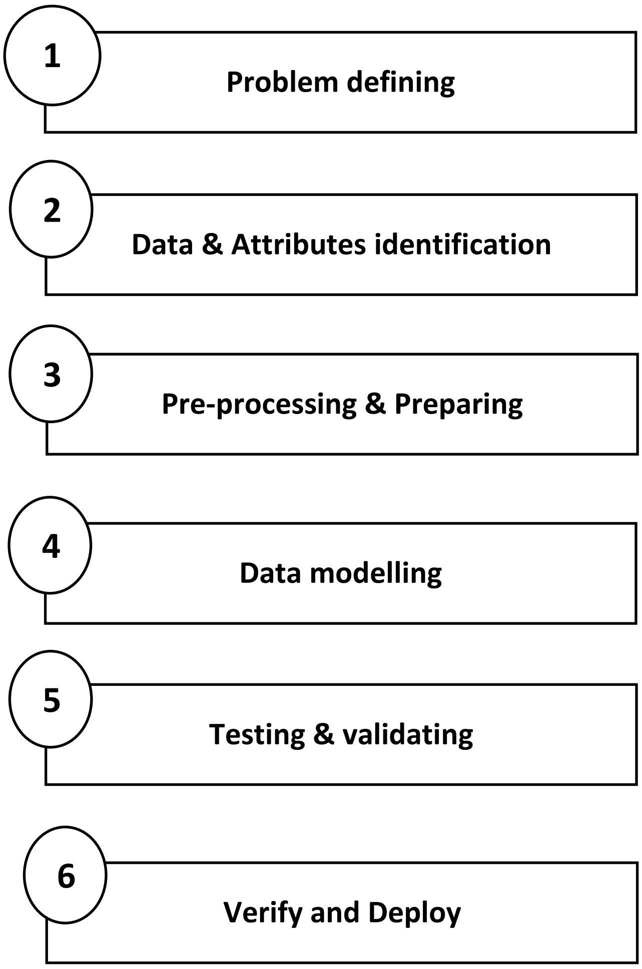


Figure 1 Steps of data mining process

1. Problem defining

First step is to discover the problem that has to solve. Which field we pick out? Which category we should identify? In which situation we proposed to analyze the data? The research is defined the problem of depression and the importance of accurate prediction model.

2. Data & attribute identification

Gathering required data and figuring out the data is the next step. Here, we have to identify the attributes. And what is the data quality of those attributes? Are the attributes very much suitable? This research identifies all the risk factors in this step.

3. Pre-processing & Preparing

In this technique cleaning and filtering of the records is finished with admire to the data and information mining set of rules hired to be able to keep away from the advent of misleading or irrelevant rules or patterns. Data mining work undergo the technique of cleaning and formatting as it should be if important. In this step, we have to recognize that we handiest want correct data sets for the research.

4. Data Modelling

Actual mining part of data mining will begin with this step. In this step, applicable algorithms for the desired task and important parameters are decided on. By this time, tools are decided on to enhance productivity. Using those tools, the model is constructed and assesses preliminary results. End purpose of data mining is set predicting. Modeling itself may also incorporate of more than one steps with admire to describing the facts.

5. Testing & validation

In this step, initial results are evaluated and the model is examined on one of a kind pattern facts units and opinions the results. Do those results throughout one of a kind samples correlate? Are there any inconsistencies? This step, version is saved iterating till the model turns into satisfied with the consistency of the results.

6. Verify and deploy

At this step, the very last model is confirmed and planed for deployment. Additionally, visualizations are used to give the story. Data mining is as lots approximately story-telling as it's miles about modeling. Findings are suggested and operationalized the technique. [Bharati M. Ramageri, 2001]

Data Mining Techniques

One of the most important tasks in data mining is to pick out the right data mining technique. Data mining technique should be picked according to the type of task and the type of problem faces. A generalized procedure should be used to enhance the accuracy and cost effectiveness of using data mining techniques. There are usually seven main Data Mining techniques.

- Statistics
- Clustering
- Visualization
- Decision Tree
- Association Rules
- Neural Networks
- Classification

In this research, five of them have been selected; statistics, clustering, classification, decision tree and visualizing.

Statistics form the core portion of data mining. The activities which we used in data mining cover the entire process of data analysis, and also statistics help in identifying patterns that further help to identify differences between random noise and significant findings. It provides a theory for estimating probability of prediction of depression. Models which having probabilities as the base of the logic, are involved in it, specifically inference, and the use of data. Here, in this research creates a statistical model with data analysis techniques.

The second data mining technique which was used in the research is clustering. Clustering could be the gathering of such set of protests in accordance with their characteristics, aggregating diehards in keeping with their similarities. Regarding to info tunneling, this technique partitions the attributes with data implementing a unique enroll set of rules, best suited for the specified instruction reasoning. This round upping opinion lets in a protest to not be part of a flock, or factually form it, province one of these gathering difficult partitioning.

As the third method Naïve Bayes was used as the classification technique. A naive Bayes classifier uses probability theory to classify data. Naive Bayes classifier algorithms make use of Bayes' theorem.

As the fourth method was decision tree. Depression types will be predicted by using this method. The fourth data mining technique which is used in the research is visualizing. Here, the visualizing models were also belonging to clustering technique. Visualization was used for representing the data expanding.

2.3 Related Works

‘Predicting Depression through Social Media’ can be a survey primarily based totally studies (M. Choudhury, et al. 2020) which became carried out to identify the social media troubles which takes place of depression. Thy have discover the ability to apply social media to discover and diagnose important medical depression in individuals. They employ crowdsourcing to assemble a collection of Twitter customers who file being identified with depressive disorder , supported a trendy psychometric instrument. Through their social media postings over a year previous the onset of depression, they degree behavioral attributes regarding social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications.

Study of Depression Analysis the use of Machine Learning Techniques (Devakunchari Ramalingam et al, 2019) via way of means of presenting an oversized dataset for identity of not unusualplace trends amongst depressed humans and perceive them the use of numerous device studying algorithms. The restriction to which they want diagnosed the depressed trends of the individual is vital to workout the volume of depression. The class performed a critical function in figuring out the kind of assist a depressed individual desires and also, the individual with suicidal mind were given to be diagnosed and helped consistent together along with his condition. This paper offers the survey approximately the usage of device studying strategies in the evaluation of depression with their studies troubles.

‘Associations among alcohol-use and depression signs and symptoms in formative years’ studies (S. Danz, 2017) tested gender variations among depression and alcohol use throughout formative years at the same time as analyzing peer and own circle of relatives pathways as feasible mediators of effects. Data became amassed longitudinally from 593 households from 3 city public center faculties in the united states.

Participants have been recruited in sixth grade and observed via ninth grade. They tested gender variations the use of a nested model comparison approach. Results indicated the association among depression and alcohol use differs by gender.

‘Depression Prediction System Using Different Methods’ research paper (Mrunal Kulkarni et al, 2019) is concentrated on the essential survey of the methods which are wont to predict depression in humans. they need studied about all the techniques which are wont to predict depression and their relative study about techniques, methods, and algorithms wont to predict depression is completed . during this project three parts have designed, i.e., question and answer part, EEG signal processing and diagnosing part and sentiment analysis part. The system can predict Depression of the user using three alternative ways . Machine learning algorithms like Naïve Bayes and Neural Network were used here for classification of data.

Detection and Diagnosis on on-line Social community Mental Disorders the use of traditional Neural Networks (S.Sridharan et al, 2018) noted that the social media systems more and more more come in the direction of emerge as a actual digitization of the human social experience. And in lots of instances humans might surely like higher to explicit themselves on-line than offline. for the duration of this paper, they want used Facebook remarks as statistics set, and supported it categorizing the customers as depressed or non-depressed

A studies became accomplished by (T. Halldorsdottir et al., ,2018) concerning the depression prediction & its medical and epidemiological results of cohorts of youths has tested the affiliation of polygenic hazard scores (PRSs) for a wide depression phenotype derived from a large-scale genome-wide association study (GWAS) in adults, and its interaction with formative years abuse, with clinically applicable despair results in medical and epidemiological adolescents cohorts.

There is some other studies paper (Melissa N Stolar, 2018) which is ready the Detection of Adolescent Depression from Speech Using Optimized Spectral Roll-Off Parameters. the intention of this paper became to examine adolescent despair detection from a medical database of sixty three adolescents (29 depressed and 34 non-depressed) interacting with a parent. loads of spectral roll-off parameters became investigated to observe an association of the frequency strength courting in connection with depression. The spectral roll-off variety progressed depression category prices in comparison to the handiest man or woman roll-off parameter. Further development became finished the use of a 2-level mRMR/SVM characteristic choice technique to optimize a roll-off parameters subset. The proposed optimized characteristic set reached an average depression detection accuracy of 82.2% for men and 70.5% for females. More acoustic spectral functions have been investigated together with flux, centroid, entropy, formants and energy spectral density to classify depression. The optimized spectral roll-off set became the most powerful of the acoustic spectral functions. All spectral functions, together with the simplest person spectral roll-off, became grouped right into a baseline feature category (S*) with an average category accuracy of 71.4% (male) and 70.6% (female). a replacement spectral category (S), with the inclusion of the proposed optimized spectral roll-off sub-set, carried out exceptional with a mean accuracy of 97.5% (males) and 92.3% (females).

Affective and Content Analysis of Online Depression Communities (Thin Nguyen et al, 2014) is some other studies that's an oversized quantity of people use on-line groups to discuss psychological state issues, therefore supplying possibilities for brand spanking new information of these groups. This paper has aimed to check the characteristics of on-line depression groups (CLINICAL) in comparison with the ones becoming a member of different on-line groups (CONTROL). they need used machine learning and statistical strategies to discriminate on-line messages among depression and manage groups the use of mood, psycholinguistic approaches and content material subjects extracted from the posts generated with the aid of using members of these groups. All components together with mood, the written content material and literary genre are determined to be appreciably exceptional among types of groups. Sentiment evaluation suggests the medical group have decrease valence than humans inside the manage group. For language patterns and subjects, statistical checks reject the speculation of equality on psycholinguistic approaches and subjects among groups. They have displayed suitable predictive validity in depression category the use of subjects and psycholinguistic clues as features. Clear discrimination among writing patterns and contents, with suitable predictive energy is a vital step in information social media and its use in psychological state.

Furthermore there may be some other studies paper (Ang Li et al, 2018) which makes a speciality of Detecting depression stigma on social media: A linguistic evaluation. Efficient detection of melancholy stigma in mass media is important for designing effective stigma discount strategies. Using linguistic evaluation strategies, this paper become aimed to create computational models for detecting stigma expressions in Chinese social media posts. First, 967 of 15,879 posts (6.09%) indicated depression stigma. 39.30%, 15.82%, and 14.99% of them advocated the stigmatizing view that “People with depression are unpredictable”, “Depression can be a sign of private weakness”, and “Depression is not a real clinical illness”, respectively. Second, the very first-rate F-Measure value for differentiating among stigma and non-stigma reached 75.2%. the very first-rate F-Measure value for differentiating amongst 3 precise types of stigma reached 86.2%.

There is some other studies paper centered Depression that's stated to tweets (Patricia A et al, 2015) examined depression-associated chatter on Twitter to glean perception into social networking approximately mental country. They want assessed subject matters of a random sample (n=2,000) of depression-associated tweets (dispatched 4-eleven to five-4-14). Tweets had been coded for expression of DSM-five signs and symptoms for Major medical depression (MDD). Supportive or useful tweets approximately depression become the most common theme (n=787, 40%), intently observed with the aid of using disclosing emotions of depression (n=625; 32%). Two-thirds of tweets discovered one or greater signs and symptoms for the analysis of MDD and/or communicated mind or thoughts that had been in keeping with struggles with depression after accounting for tweets that mentioned depression trivially. Health specialists can use our findings to tailor and goal prevention and recognition messages to the ones Twitter customers in want.

There is some other studies (Brian A et al, 2016) supported the usage of a couple of social media systems and symptoms of depression and anxiety: A nationally consultant observe amongst U.S. teens While expanded social media use (SMU) has been associated with depression and anxiety, the independent role of the use of a couple of social media systems is unclear. they want surveyed a nationally-consultant pattern of 1787 U.S. teens a long time 19-32. Dependent variables have been each depression and anxiety symptoms measured the use of the Patient-Reported Outcomes Measurement records system (PROMIS). They assessed use of a couple of social media platform with an adapted Pew Internet Research scale. They have used ordered logistic regression fashions to evaluate institutions among a couple of platform use and psychological country results even as controlling for 8 covariates, along with standard SMU. Compared to those who used 0-2 social media web sites, contributors who used 7-eleven social media web sites had notably better odds of having expanded stages of each melancholy (Adjusted Odds Ratio [AOR]=3.0, 95% CI=1.9-4.8) and anxiety symptoms (AOR=3.2, 95% CI=2.0-5.1). Associations have been linear ($p<.001$ for all) and strong to all sensitivity analyses.

There is some other studies that's simply centered on major depression (Munmun De Choudhury et al, 2017) explored the ability to apply social media to discover and diagnose major clinical depression in individuals. they want first employed crowdsourcing to collect a set of Twitter customers who file being identified with depressive disorder, supported a normal psychometric instrument. Through their social media postings over a year previous the onset of depression, they want measured behavioral attributes regarding social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. they want leveraged those behavioral cues, to create a statistical classifier that offers estimates of the threat of depression, earlier than the suggested onset. they want observed that social media includes beneficial indicators for characterizing the onset of depression in individuals, as measured via lower in institution action , raised negative affect, rather clustered ego networks, heightened relational and medicinal concerns, and more expression of spiritual involvement.

There is some other studies (Tan Tze Ern Shannon et al 2017) concerning Speech analysis and depression. throughout this paper, the correlation among the speech capabilities of the vowel /a/ and depression severity became investigated, so on derive a depression severity meter mobile software so one can as it should be discover depression quantitatively. Results confirmed a correlation among melancholy severity and speech capabilities, and an software prototype turned into created and examined to assess for predictive accuracy of BDI score

Detecting Depression in Speech (Hailiang Long et al, 2019): A Multi-classifier System with Ensemble Pruning on Kappa-Error Diagram. During this study, a totally unique multi-classifier device for depression detection in speech turned into developed and tested. They have collected speech data in numerous ways, and that they have tested the discriminative electricity of diverse speech sorts (consisting of analyzing, interview, photo description, and video description). Considering that extraordinary speech sorts might also additionally elicit extraordinary stages of cognitive attempt and deliver complementary data for the type of melancholy, they were able to make use of diverse speech statistics units to recognize a much higher end result for melancholy popularity. All man or woman learners formed a pool of classifiers, and some man or woman learners with a excessive variety and accuracy in the pool have been selected. in the process, the kappa statistics diagram helped to shape decisions. Finally, a multi-classifier device with a parallel topology turned into constructed, and each man or woman learner throughout this system used unique speech data types and speech features. In our test, a pattern of seventy four topics (37 depressed sufferers and 37 wholesome controls) turned into examined and a leave-one-out cross-validation scheme turned into used. The test end result confirmed that this new technique had a higher accuracy (89.19%) than that of unmarried classifier methods (the satisfactory is 72.97%). Additionally, they have additionally located that the overall popularity fee the use of interview speech turned into above the ones using photo description, video description, and analyzing speech.

Detecting depression (Sharifa Alghowinem et al,2013): A comparison among spontaneous and browse speech. they want used principal melancholy for this studies. throughout this paper, they want hypothesized that classifying the general traits of depressive disorder the use of spontaneous speech will deliver higher results than the use of examine speech, that there are a few acoustic capabilities which are sturdy and could deliver precise type results in each spontaneous and browse , which a `thin-slicing' technique the use of smaller components of the speech statistics will carry out in addition if now no longer higher than the use of the complete speech statistics. By inspecting and evaluating popularity consequences for acoustic capabilities on a real-international scientific dataset of 30 depressed and 30 manipulate topics the use of SVM for type and a leave-one-out cross-validation scheme, they want located that spontaneous speech has extra variability, which will increase the recognition fee of melancholy. they want additionally located that jitter, shimmer, power and loudness function organizations are sturdy in characterizing each examine and spontaneous depressive speech. Remarkably, thin-slicing the examine speech, the use of both the begin of each sentence or the number one few sentences plays higher than the use of all analyzing undertaking statistics.

There us another research (Rensik et al., 2015) based on the data gained from facebook related depression. They have create many models together as a series. Their goal was to predict postpartum type of the depression.

There is a research (Authors Orabi et al, 2018) focused on the depression based on the twiiter data. It was a depression research completely and has built up b using deep neural network techniques.

In another work (Tsugawa et al., 2015) was found and it was based on the twitter data. By suing the titter data, they have created a model to extract the features of decision list on twitter users with their depression levels.

Another research was found out (Schwartz et al,2014) and it was a model creation research based on the depression which only relates to the social media and face book. It was a survey based research and they have investigate 28749 users who are using face book daily.

CHAPTER 3

METHODOLOGY

3.1 Data collection

Data collection is the first step of doing a survey based research 1247 patient data were collected from the ‘Manasuwa Piyasa’ Department in government hospital in Hambanthota. This collection includes details of both depression patients and non-depression patients. All 1247 patient data were collected directly from the Hambanthota hospital with the help of Doctor Prasadi Lokuthotahewa. Depression details and knowledge were gathered from Professor Gnanadasa Perera who is specialized in phycology.

3.2 Attributes identification

Table 2 Attributes & Values

Attribute No.	**Attribute Name**	**Values**
1	Age	Real Numbers

2	Gender	Male = 0
		Female = 1

3	Working Hours	Real Numbers
4	Hours Spend in Social Media	Real Numbers
5	Sleeping Hours	Real Numbers
6	No of children	Real Numbers
7	Marital Status	Single Married Divorced
8	Education Level	up to O/L up to A/L Degree holder Masters/PHD
9	Fast Food Consumption	Not having much fast food consumption =0 Having much fast food consumption= 1
10	Background (Family History)	Not any relative having depression= 0 Any relative having depression=1

11	Job Related area	Gov. Privet Own Business Not doing a Job
12	Exercises	Not doing daily exercise=1 Doing daily exercises=0

13	Drugs addiction	No =0
		Yes=1

14	Alcohol consumption	No =0
		Yes=1

15	Serious Illnesses	No =0
		Yes=1

16	Smoking	No =0
		Yes=1

17	Depression patient/not	No =0
		Yes=1

18	Initial Mental Status	High/Medium/Low
19	Time range	Less than one month/ More than one month
20	Periodically	Yes/No
21	Changes in weight	High/Low/Normal
22	Changes in sleep	High/Low/Normal
23	Changes in appetite	High/Low/Normal
24	Energy	High/Low/Normal
25	Unexplained aches and pains	Yes/No
26	Thoughts of death and suicide	Yes/No
27	Trouble concentrating and memory problems	Yes/No
28	Depression Type	Major/Persistent/Bipolar/ Atypical/Not recognized as a depression patient

In this dataset some attributes (Attribute no. 1 – Attribute no.18) are identified as predictive variables for predict the depression risk since these attributes are related to the risk factors of depression. So the Attribute “Depression patient/not” was taken as the target variable. Attribute “Initial Mental Status” is related to the level of the patient whether he/she has a high risk and need quick and deep treatment or whether he/she is having not much high depression but needs to assign for treatments or whether he/she doesn’t need to assign for special treatments but just for counselling. These levels can be used to categorize the data into some clusters. Here the attributes (Attribute no.19 – Attribute no.28) is related to symptoms of depression. Here the target variable is identified as “depression type” and so that the depression type can be predict by using decision tree algorithm in classification technique. Risk factors, initial status and the symptoms of depression were identified though the World Health Organization estimates and the National Center of Biological Information.

3.3 Data Preprocessing

Data preprocessing can be diagnosed as the first step in data mining. In Data preprocessing the misclassified data is removed. In this technique cleansing and filtering of the data is performed with appreciate to the statistics and data mining set of rules hired on the way to keep away from the introduction of misleading or irrelevant policies or patterns. In preprocessing first off an attribute became decided on for selecting a subset of attributes with accurate predicting capability. It handles all lacking values and investigates every possibility. If an characteristic has extra than 5% lacking values then the data must now no longer be deleted and it's miles beneficial to impute values in which data is lacking, the use of a appropriate method. Data withinside the actual international can be dirty, incomplete and noisy. Incomplete in missing variable values or containing most effective aggregate values are referred to as noisy and it containing mistakes or outliers and inconsistent containing discrepancies in names or codes. But why is the data turns into dirty? Because incomplete data may also come from now no longer applicable‖ statistics value whilst statistics needs to be collected and the main difficulty is a unique attention among the instances whilst the data became analyzed and human hardware and software program troubles are common. Noisy and incomplete data may also arise whilst a human enters the incorrect value on the time of data entry as no one is perfect. Inconsistent data may also come from the unique statistics sources. Duplicates information additionally want data cleansing. Why data preprocessing is critical in analytics? Data isn't always clean, duplicity data, no best data leads no best end result so data preprocessing is critical. Quality selections need to be primarily based totally at the quality data. Data warehouse needs regular integration of quality data. By the processing of data, data quality may be measured in term of accuracy, completeness, consistency, timeliness, believability, interpretability.

3.4 Statistical Analysis

Statistics is a collection of processes for analyzing, organizing and presenting quantitative data. Data is the term for records which have been received and therefore recorded. Data generally refers to quantitative data which means numbers. Essentially then, statistics is called a systematic method that's used to reading numerical records as a manner to allow us to widen our figuring out, translation and use. This means that statistics enables us to turn data into information. Information is data that have been interpreted, understood and are beneficial to us. Actually, statistics is actually the methodical series and assessment of numerical records, in order to test out or discover relationships in reference to phenomena so one can supply an explanation, predict and manage their occurrence. The opportunity of confusion comes from the reality that now no longer simply is statistics the strategies used on quantitative data, however the identical phrase is likewise used to consult the numerical results from statistical analysis. Statistics help in summarizing the data and calculating it. It additionally enables in presenting information approximately the data very easily.

3.4.1 Predict the depression risk

According to the type of the data and attributes, the most suitable way to analyze these data is creating a binary logistic regression model.

3.4.1.1 Binary Logistic Regression Model

Logistic regression is identified as an extension of simple linear regression. Since the dependent variable is binary or diploid in naturally, the appropriate model is binary logistic regression. Binary logistic regression is one of the logistic regressions and it is the statistical technique which is used to predict the relationship between them. In Binary logistic regression, there should be dependent variable and independent variables for apply this method. Independent variables also called as predictors and dependent variable also called as predicted variable. Here, the dependent variable should be binary. There should be at least two independent variables. And predicted variables can be continuous (interval/ratio) or categorical (ordinal/nominal). All predictor variables are tested through this model to evaluate their predictive ability while supervising for the effects of other predictors in the model.

As mentioned earlier, for the depression prediction we use has 16 predictive attributes. The dependent variable (response variable) is whether the person is a depression patient or not which has binary values as 1 and 0.

- If the person is a depression patient = 1
- If the person is not a depression patient= 0

Here, the ‘depression patient/not’ variable is set as the response of the model and set the response event as the positive of depression result (value 1). By creating the model it shows the p value for the all the attributes.

From this analysis, we have identified the significance level of each attribute. For that P- value was considered hence it is which is a measurement included in each attribute.

- **P – Value**

Deviation table shows the p- value of the attributes of the depression data. It means it performs the hypothesis test in statistics and here, the p-value is a number which helps to understand and figure out the importance of our result. Normally, hypothesis tests are used to check the effectiveness of a claim which is made about the data population. This claim which is under examination of analysis is also referred to as the null hypothesis. But here, the all hypothesis are true attributes of depression.

Generally, the alternative hypothesis may be the one the researchers would believe if the null hypothesis is concluded to be untrue. Here, the data is the evidence in under the examination of analysis and the statistics that go together with them. The p-value is used to weight the strength of the evidence by all hypothesis tests. Actually, the p-value is in between 0-1 and it provides us a clear idea on what the data convincing on the data population.

- P-value which are typically less than 0.05 (< 0.05), provides the strong evidence against the null hypothesis. It means they are mostly affected to the logistic model.
- P-value which are greater than 0.05 (> 0.05) provides the weak evidence against the null hypothesis. It means they are not mostly affected to the logistic model.

To create a model with only significant attributes, the unnecessary attributes were removed from the model. As mentioned earlier, the attributes which are most effect to the statistical model or not can be identified.

- **stepwise regression**

Stepwise regression is used for the exploratory stages of model building. It is used to perceive a useful subset of predictors. It runs the method to systematically add the most important, effected variable or removes the least important variable for the duration of each step. There are three common stepwise procedures in stepwise regression. But here, I used the standard stepwise regression which is doing both adding and removing the predictors as required for every single step. Model iterations stops running in two situation occurs-

- When all variables not inside the model have p-values which are bigger than the specified Alpha-to-Enter value. Here, α to enter = 0.15.
- When all variables inside the model have p-values which are equal or less than the specified Alpha-to-Remove value. Here, α to remove = 0.15.

The last step is validating the created statistical model. In statistics, validation of the regression results is the process of deciding whether the numerical results quantifying hypothesized relationships between attributes which were obtained from regression analysis, are acceptable as descriptions of the data. The created statistical model is validated in this step. The validation checks whether a patient, who in reality has depression, is diagnosed as disease free or not and a patient, who in reality does not have depression, is diagnosed as having depression or not.

3.4.2 Analyzing the Initial Mental Status of a person

When a person comes to get treatments from the hospital, their initial mental statement at the moment is identified by the doctors while taking the information; demographical and social aspect. This variable was taken as the response variable and the analysis was done using only the most significant categorical variables which were identified though the previous analysis ; predicting the depression risk. Since there were categories in the response variable both classification and clustering technique were used and best suitable technique was identified through the accuracies of each model.

3.4.2.1 Clustering VS Classification

- **Clustering**

Clustering algorithm separates a data set into a few groups in accordance with the principle of maximizing the intra-class sameness and minimizing the inter-class sameness. Regarding to data mining, clustering technique segregations the data implementing a specialized join algorithm, most fitted for the desired information analysis. Partitioning a set of objects in databases into homogeneous groups or aggregates is a fundamental operation in data mining.

Here, the data set was divided in to four clusters. A cluster is usually a subgroup of objects that are “similar”. A subgroup of objects such which the space between any two objects within the cluster is lower than the space between any object within the cluster and any object not placed inside it. A connected region of a multidimensional distance containing almost strong density of object.

- **Classification**

Classification may be a most familiar and most popular data processing technique which is known as process of finding a model (or function) that describes and distinguishes data classes or concepts, for the aim of having the ability to use the model to predict the category of objects whose class label is unknown. The derived model is predicated on the analysis of a group of training data (i.e., data objects whose class label is known). Training set is employed to develop specific parameters required by the technique.

The goal of classification is to create a concise model which will be wont to predict the category of records whose class label isn't known. Classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions. That is, it's wont to predict missing or unavailable numerical data values instead of class labels.

3.4.2.2 K-means Clustering

K-means clustering is really a method of vector quantization. It is a kind of unsupervised learning, which is used when you have unlabeled data. Rather than defining groups before looking at the data, clustering enables us to find and analyze the groups which have formed naturally. The "Choosing K" section describes how the number of groups can be decided. The goal of this algorithm is to find groups within the data, with the number of groups described by the variable K. The algorithm normally runs iteratively to assign each data point to one of K groups by considering the features which are provided. Data points are clustered according to feature sameness. But here, the test ‘classes to cluster evaluation’ was used as the clustering option. It first remove the class variable and then start clustering the data set. Here, 1230 data points (instances) randomly possess the clusters leading to clusters that have roughly an identical number of data points by considering the class variable. Then, it calculates the distance from the data point to every single cluster. If the data point is closest to its own cluster, this algorithm leaves that data point where it is. If the data point is not closest to its own cluster, the data point is moved it into the closest cluster. After that it repeats the above step until a whole pass through all the data points ends up in no data point are moving from one cluster to another. At this situation the clusters are stable. So that the clustering process will end. The selection of basic partition can very much affect the final clusters that result, in terms of inter-cluster and intra cluster distances and cohesion.

The results of the K-means clustering algorithm are:

- The centroids of the 3 clusters, which can be used to label new data
- Each data point is assigned to a single cluster

3.4.2.3 Naïve Bayes Classification

A naive Bayes classifier is an set of rules that makes use of Bayes' theorem to categorize objects. Naive Bayes classifiers assume strong, or naive, independence among attributes of knowledge points. Popular makes use of naive Bayes classifiers consist of unsolicited mail filters, textual content evaluation and diagnosing. These classifiers are broadly used for machine learning due to the fact they're easy to implement. A naive Bayes classifier makes use of probability theory to classify data. Naive Bayes classifier algorithms make use of Bayes' theorem. The key perception of Bayes' theorem is that the probability of an incidence might be adjusted as new data is introduced. 'Initial status of the person' attribute is taken because the class variable of this model.

3.4.2.4 Precision and Recall

Precision is a ratio which shows the correctly identified items (also called as true positive values) and all items which returned (true positives + false positives). In another words, precision is a ratio which is calculated by dividing the correctly predicted positive values from the total predicted positive values. This value highlights the correct positive predictions out of all the positive predictions. High precision indicates low false positive rate.

Recall can be defined as a quotient of correctly identified items (also called as true positives values) and all relevant items (true positives + false negatives). In another words, recall can be calculated by dividing the correctly predicted positive values from the actual positive values. Recall is defined the sensitivity of the algorithm either by considering all the actual positives how many were captured by the program.

3.4.3 Predict the depression type based on the symptoms

• Decision tree

Decision tree is defined as a classification technology. It creates a tree architecture with the set of rules. Here the cross validation techniques was used as the test option of the data model. The tree structure is like a flow chart. It’s internal node denotes a test on an attribute. It divides the data in to smaller subsets while at an equivalent time an associated decision tree is incrementally developed.

Outcome of the dataset is represents from the branches of the tree structure. Branch represents an outcome of the test and therefore the leaf nodes in the represent class labels or class distribution or rule. Here the utilization of the choice tree is to classifying an unknown sample and also to check the attribute values of the sample against the choice tree. There are many types of decision trees in classification. But here the J48 decision tree was used to predict the depression types based on the symptoms. This j48 decision tree is based on the entropy values and it is identified as the principle of entropy reduction and information gain. Most of the time it leads to overfitting and so that it improves the rate of classification and overcome over-fitting. Pruning was used when creating the tree architecture. Here, the last 10 attributes (variable no.19 –variable no.28) was selected from the dataset. “Depression type “variable was selected as the target variable and it is a categorical nominal variable which is having 5 values (major, bipolar, persistent, atypical, none).

3.5 Solution Design

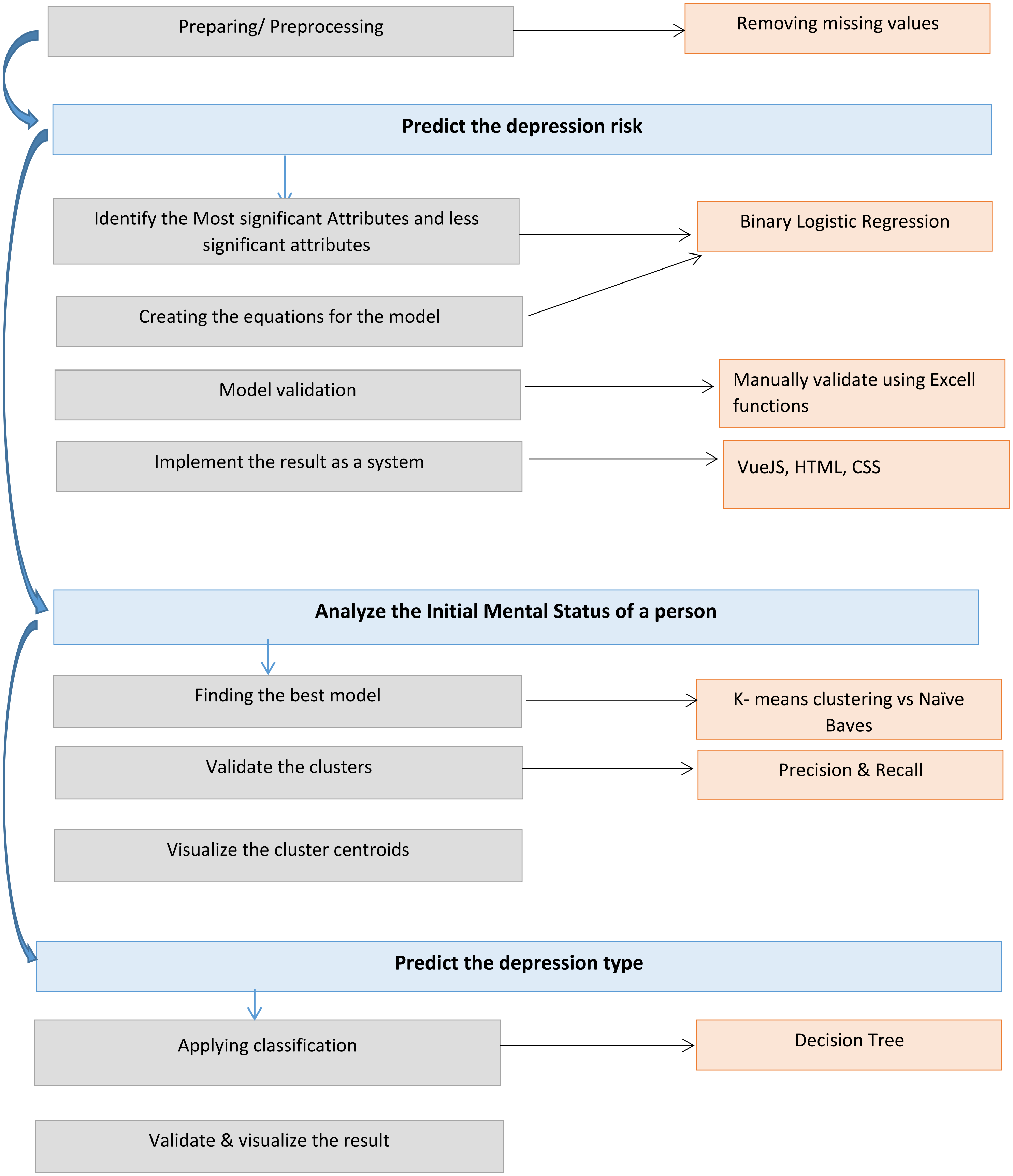

Preparing/ Preprocessing
Removing missing values
Predict the depression risk
Identify the Most significant Attributes and less significant attributes
Binary Logistic Regression
Creating the equations for the model
Model validation
Manually validate using Excell functions
Implement the result as a system
VueJS, HTML, CSS
Analyze the Initial Mental Status of a person
Finding the best model
K- means clustering vs Naïve Bayes
Validate the clusters
Precision & Recall
Visualize the cluster centroids
Predict the depression type
Applying classification
Decision Tree
Validate & visualize the result

Figure 2 Solution design

CHAPTER 4

EVALUATION & RESULT

4.1 Predict the depression risk

Before predicting the risk of having depression, the most significant risk factors were identified. Then by using only the significant risk factors the prediction model was created.

4.1.1 Finding most significant factors

When following preprocessing, 37 missing values were identified and removed from the dataset. Full data set was taken as the training data (1230) and randomly picked amount of data was used as the test data (50). Statistical analysis was done to the training dataset.

Following table shows the P- values of each attributes. Here, the P-value of attribute ‘Gender’ ‘Working Hours’, ‘No of children’, ‘Background’, ‘Alcohol consumption’, ‘Marital status’, ‘Education Level’ is 0.000. So, these are the lowest P- value among all the attributes. The highest P- value is 0.783 which is belongs to the attribute ‘Age’. As mentioned in the methodology of the statistical analysis, the attributes can be identified as most significant or less significant according to the P-values.

Table 3.P- values of the attributes

Source	P-Value
Age	0.783
Gender	0.00
Working Hours	0.00
Social Media	0.635
Sleeping Hours	0.015
No of children	0.00
Fast Food Consumption	0.223
Background (Family History)	0.00
Exercise	0.013
Drugs addiction	0.002
Alcohol consumption	0.00

Serious Illnesses	0.027
Smoking	0.039
Marital Status	0.00
Education Level	0.00
Job Related to	0.355

- A small *p*-value (typically < 0.05) shows the strong evidence against the null hypothesis. It means they are mostly affected to the logistic model.

Table 4.significant attributes

Gender	0.00
Working Hours	0.00
Sleeping Hours	0.015
No of children	0.00
Background (Family History)	0.00
Exercise	0.013
Drugs addiction	0.002
Alcohol consumption	0.00
Serious Illnesses	0.027
Smoking	0.039
Marital Status	0.00
Education Level	0.00

- A large *p*-value (> 0.05) shows the weak evidence against the null hypothesis. It means they are not mostly affected to the logistic model.

Table 5.Less significant attributes

Age	0.783
Social Media	0.635
Fast Food Consumption	0.223
Job Related to	0.355

4.1.2 Binary Logistic Regression

According to the significant of the attributes, as per the stepwise technique less significant attributes were removed from the Binary Logistic Regression model. It built up a statistical model which consists of mathematical equation. Since there are two categorical significant variables as ‘Marital Status’ and ‘Education Level’, the model includes equations with relevant combinations. They generate a way to predict the possibility of having depression.

- **Equation1**

$$P(1) = \exp(Y')/(1 + \exp(Y'))$$

- **Equation2:**

Table 6.Calculating Y’

Marital Status	Education Level	Y'
Divorced	up to A/L	Y' = -1.820 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Divorced	up to degree	Y' = -2.943 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Divorced	up to Msc/Phd	Y' = -1.868 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Divorced	up to O/L	Y' = -3.822 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking

Married	up to A/L	Y' = -3.322 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Married	up to degree	Y' = -4.445 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Married	up to Msc/Phd	Y' = -3.369 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Married	up to O/L	Y' = -5.324 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Single	up to A/L	Y' = -2.473 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Single	up to degree	Y' = -3.596 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Single	up to Msc/Phd	Y' = -2.521 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking
Single	up to O/L	Y' = -4.475 - 1.430 Gender + 0.3158 Working Hours + 0.1368 Sleeping Hours + 0.7142 No of children + 1.788 Background (Family History) - 0.3690 Exercise + 1.821 Drugs addiction + 3.067 Alcohol consumption + 1.151 Serious Illnesses - 0.5560 Smoking

Equation 2 consists of most relevant attributes which are most affected to the statistical model as per the combination of categorical variables. The attributes values should be replace to this equation and then the value of Y' should be calculated well.

Then the calculated Y' value should be replace to the equation 2 and should calculate the value of P(1). Prediction of having depression on this P(1) value.

- If the P(1) value is equal or greater than 0.5, The result is close to 1 (one). It means the prediction risk is positive. In other words, the person has the risk of having depression

- If the P(1) value is less than 0.5, The result is close to 0 (zero). It means the prediction risk is negative. In other words, the person has not the risk of having depression.

So the positive range is identified as the value which is equal or greater than 0.5 and the negative range is identified as the value which is less than 0.5.

4.1.3 Validating the statistical result

Validations were applied to find out the accuracy of the created model. Randomly selected 50 instances were used and checked manually. Randomly selected data was got through the real data set and checked weather the created algorithm give the correct possibility of being positive or negative in depression It was checked by the response column (depression patient/not) manually.

No of correct predicting - 46
No of wrong predicting– 4
So the manually calculated accuracy result is 92%

Table 7 Goodness of fit test

Test	DF	Chi-Square	P-Value
Deviance	1214	988.83	1
Pearson	1214	1102.58	0.990
Hosmer-Lemeshow	8	7.48	0.486

Hosmer-Lmeshow value is 0.466. This is typically a good value. It proves that the data fits the model well.

4.1.4 Implement the analysis result

By using VueJS the development logic was implemented. For developing the UI of the system HTML and CSS was used.

UI of the system was developed to get the data from the user only for the most significant attributes which were identified though the regression analysis.. Text boxes were created for ‘working horse’. ‘sleeping hours’ and the ‘no of children’ attributes.

For the other attributes; Gender, Background (Family History), Exercise, Drugs addiction, Alcohol consumption, Serious Illnesses, Smoking, Marital Status, Education Level, radio buttons were created.

```
<div class="item">
              <label for="fname">Working Hours<span>*</span></label>
              <input id="fname" type="number" name="fname" required v-
model="working_hours" />
            </div>
```

After finishing the form filling, the user can submit the filled form.

```
<div class="btn-block" style="text-align: left;">

            <button @click="formSubmit" type="button" href="/">Submit</button>

        </div>
```

Below displays the way of calling the development logic of finding ‘Y’ value.

```
processData() {
if (this.marital_status == 'divorced' && this.education == 'a_l') {
const y = this.findY(1.820, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'divorced' && this.education == 'degree') {
const y = this.findY(2.943, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'divorced' && this.education == 'master') {
const y = this.findY(1.868, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'divorced' && this.education == 'o_l') {
const y = this.findY(3.822, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'married' && this.education == 'a_l') {
const y = this.findY(3.322, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'married' && this.education == 'degree') {
const y = this.findY(4.445, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'married' && this.education == 'master') {
const y = this.findY(3.369, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'married' && this.education == 'o_l') {
const y = this.findY(5.324, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'single' && this.education == 'a_l') {
const y = this.findY(2.473, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'single' && this.education == 'degree') {
const y = this.findY(3.596, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'single' && this.education == 'master') {
const y = this.findY(2.521, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);

} else if (this.marital_status == 'single' && this.education == 'o_l') {
const y = this.findY(4.475, 1.430, 0.3158, 0.1368, 0.7142, 1.788, 0.3690, 1.821, 3.067, 1.151, 0.5560);
this.depre_level = this.findP(y);
}
},

findY(c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11) {
return -c1 - (c2 * this.gender) + (c3 * this.working_hours) + (c4 * this.sleeping_hours) + (c5 * this.no_of_cildren)
+ (c6 * this.family_depression) - (c7 * this.do_excercise) + (c8 * this.do_drugs) + (c9 * this.alcohol) + (c10 *
this.illness) - (c11 * this.smoking);
}
```

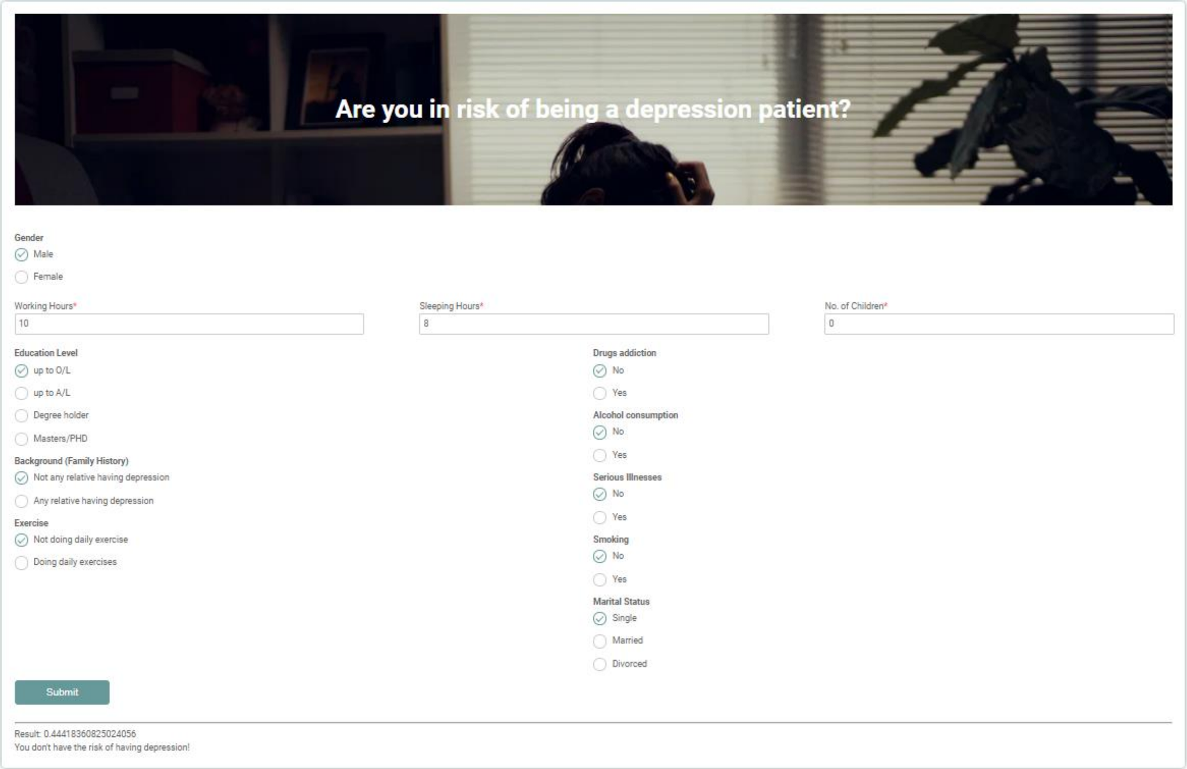
Below displays the way of calling the development logic of finding ‘P’.

```
findP(y) {
            return Math.exp(y) / (1 + Math.exp(y));
          }
```

If P is equal or greater than 0.5, the prediction is positive. If the P is less than 0.5, the prediction is negative.

```
watch:
{
        depre_level: function (val) {
        this.message = val >= 0.5 ? "You have the risk of having depression!" :
        "You don't have the risk of having depression!"
        },
```

Figure 3 shows the UI of the system.



Gender
✓ Male
○ Female

Working Hours* 10

Sleeping Hours* 8

No. of Children* 0

Education Level
✓ up to O/L
○ up to A/L
○ Degree holder
○ Masters/PHD

Background (Family History)
✓ Not any relative having depression
○ Any relative having depression

Exercise
✓ Not doing daily exercise
○ Doing daily exercises

Drugs addiction
✓ No
○ Yes

Alcohol consumption
✓ No
○ Yes

Serious Illnesses
✓ No
○ Yes

Smoking
✓ No
○ Yes

Marital Status
✓ Single
○ Married
○ Divorced

Result: 0.44418360825024056
You don't have the risk of having depression!

Figure 3 UI of the system

4.2 Analyzing the initial mental status of a person

At the initial stage of the diagnosing, the status of the mental health is identified by the doctor. This is simply introduced as high, medium or low categories. Whether a person is a depression patient or not, he/she is definitely belongs to a one category either low, medium or high based on his/her current condition of mental health.

Here, the target was to identify this status of a person by using the most significant attributes which were identified through the results of the first objective of the project. So, the 'status' variable was analyzed with the background (family history), alcohol consumption, gender, marital status, education level hence they are the most significant categorical variables.

Both k-means clustering and naive Bayes classification were used separately to identify which model is most suitable for categorize the initial mental status of a person with the risk factors.

4.2.1 Applying Clustering

Hence there are three statuses, Analysis was built up for K=3.

And cluster mode was selected as 'Classes to cluster evaluation', since there is a class variable as 'status'.

Below table displays the cluster analysis and values of each variable was divided according to the identified clusters.

Cluster 0 is for Medium
Cluster 1 is for High
Cluster 2 is for Low

Table 8 Clustering data

Attribute	**Cluster 0**	**Cluster 1**	**Cluster 2**
Background (Family History)	Y	N	N
Alcohol consumption	N	N	N
Gender	F	M	M
Marital Status	Married	Divorced	Married
Education Level	up to A/L	up to A/L	up to A/L

But here, the accuracy is 55%.

Incorrectly clustered instances- 556.0 (Percentage - 45.2033 %)

4.2.2 Applying Classification

Data set was classified by using the 'status' variable as the class variable. Here, the 'Cross-validation' technique was used with 10 folds to get test the classified result to get its best. Because in the cross validation it divides the data set in to the given folds and test the data set iteratively based on the given folds.

Here, the accuracy is 70%

Incorrectly Classified Instances- 363 (Percentage- 29.5122 %)

Correctly Classified Instances – 867 (Percentage-70.4878 %)

4.2.3 Select the best model

To analyze the categories of the initial mental status of a person, Both Naïve Bayes classification and k-means clustering can be used. But here, the Naïve Bayes classification has classified the data well than clustering the data set. Marjory of the correct instances are classified in the Naive Bayes classification model rather than k-means clustering. And the accuracy of the Naïve Bayes classifier is 70% and the accuracy of the k-means clustering is 55%. So, as the best model to analyze the initial status of a person is Naïve Bayes.

4.2.4 Further on Naïve Bayes

Below table shows the categorization of class variables with the values of each attributes.

Table 9 Classification model

Attribute	Class		
	Low (0.24)	**Medium (0.41)**	**High (0.36)**
Background (Family History)			
N	228	378	191
Y	63	127	249
Alcohol consumption			
Y	8	29	150
N	283	476	290
Gender			

F	28	432	228
M	263	73	212
Marital status			
Single	56	115	70
Married	196	308	266
Divorced	40	83	105
Education Level			
up to O/L	61	37	33
up to A/L	103	312	273
up to degree	103	136	102
up to Msc/Phd	26	22	34

Kappa statistics is a range from −1 to +1. It is one of the most commonly used statistics to test interrater reliability. In this classification model, value of the kappa statistic is and it is normally a good value. So, it proves that this model is reliable to the dataset.

4.2.5 Validate the Result

Table 10 Class variable - Confusion matrix

	Actual Class		
Predicted Class	**Low**	**Medium**	**High**
Low	248	21	20
Medium	37	391	75
High	139	71	228

Here, this model shows that 248 instances were correctly classified as class low. And 391 instances were correctly classified as medium and also 228 instances were correctly classified as class high. It proves that the majority of the instances are correctly classified for each class, hence the model is a effective model to explore the initial mental status of a person.

Table 11 Detailed accuracy by class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Low	0.858	0.187	0.585	0.858	0.696	0.599	0.895	0.719
Medium	0.777	0.127	0.81	0.777	0.793	0.655	0.878	0.893
High	0.521	0.12	0.706	0.521	0.599	0.436	0.823	0.729
Weighted Avg	0.705	0.138	0.72	0.705	0.701	0.564	0.862	0.794

TP rate is indicating the True Positive rate of the each class which defines the number of true positives classified by the model. Here, Class low has the highest TP rate (TP Rate = TP/(TP+FN)). FN rate is indicating the number of false negatives classified by the model. For all three classes, the TP rate is a higher value and the FN rate is a lower value. This is a good performance of the data set (TN Rate = TN/(TN+FP).

Precision is the fraction of true positive examples among the examples that the model classified as positive. In other words, the number of true positives divided by the number of false positives plus true positives. And the Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples. In other words, the number of true positives divided by the number of true positives plus false negatives. So, the good sensitivity also gained from the class low. However, the study proves that the dataset is very sensitive to the model hence it gives the good performance of the model.

F-score is a factor indicating how much more important recall is than precision. For example, if we consider recall to be twice as important as precision, we can set β to 2. The standard F-score is equivalent to setting β to one. So, this model shows that the recall of the class low is 0.696 important as precision, class medium is 0.793 important as precision and the class high is 0.599 important as precision.

4.2.6 Visualize the Result

Figure 5 shows the visualization of how the dataset was categorized with each symptoms according to the ‘depression type’ variable; Major, Bipolar, Persistent, Atypical and None.

- Major -Red
- Light Blue – Persistent
- Bipolar – Ash
- Atypical – Pink
- None – Dark Blue

X axis of each chart was displayed as follows.

- Background (Family History) – No, Yes
- Alcohol consumption - No, Yes
- Gender – Female, Male
- Marital Status -Single, Married, Divorce
- Education Level - up to O/L , up to A/L, Degree holder, Masters/PHD

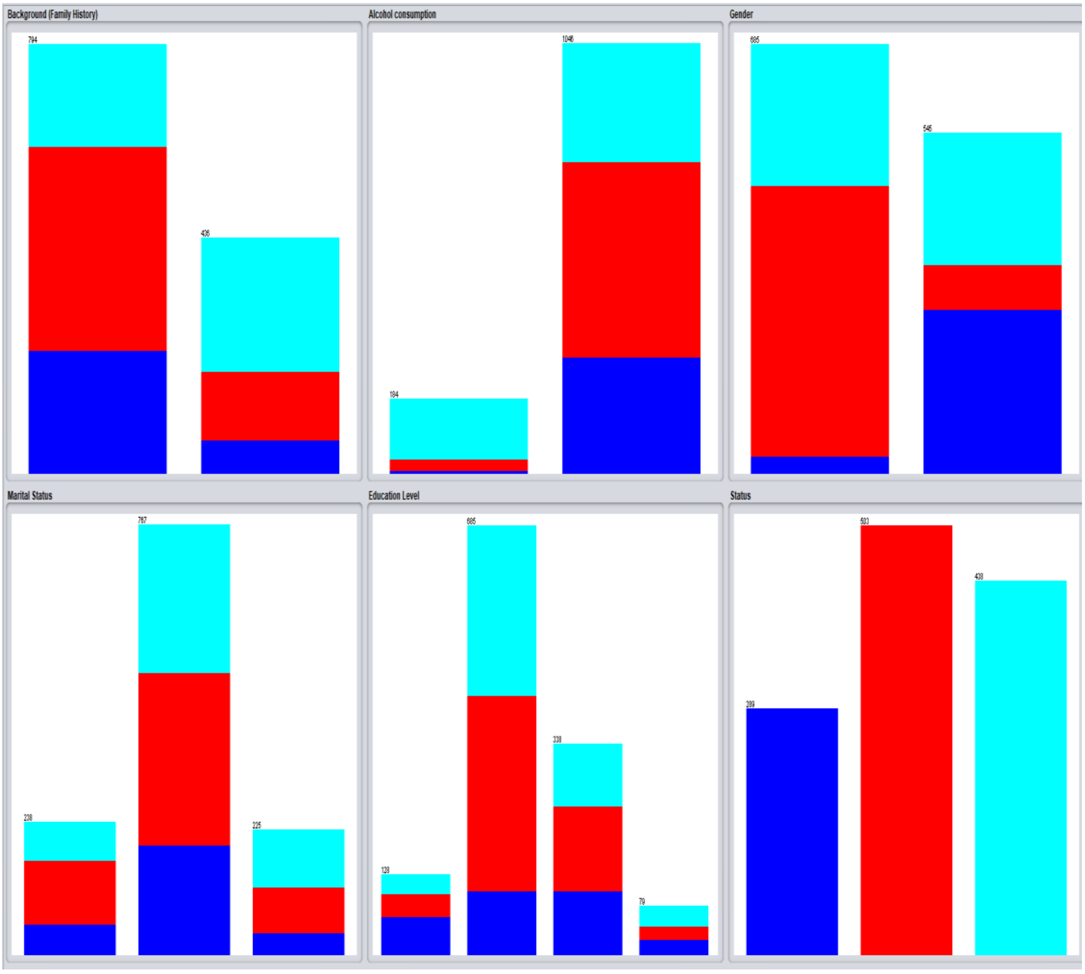


Figure 4 Depression types with exploratory variables

4.3 Predict the depression type

To predict the depression type based on the symptoms, J48 decision tree was used by selecting the ‘depression type’ attribute as the class variable. Attribute no – Attribute no were used to the model as the exploratory variables.

4.3.1 J48 decision tree

```
Time range = less than more than 1 month
|   Periodically = No
|   |   Trouble concentrating and memory problems = No
|   |   |   Thoughts of death and suicide = No: None (211.0/35.0)
|   |   |   Thoughts of death and suicide = Yes: Artypical (5.0)
|   |   |   Thoughts of death and suicide = yes: None (0.0)
|   |   Trouble concentrating and memory problems = Yes
|   |   |   Thoughts of death and suicide = No
|   |   |   |   Energy = High
|   |   |   |   |   Unexplained aches and pains = Yes: None (3.0)
|   |   |   |   |   Unexplained aches and pains = No: Artypical (42.0/13.0)
|   |   |   |   Energy = Medium: Artypical (11.0/3.0)
|   |   |   |   Energy = Low: None (17.0/4.0)
|   |   |   Thoughts of death and suicide = Yes: Major (4.0)
|   |   |   Thoughts of death and suicide = yes: Artypical (0.0)
|   |   Trouble concentrating and memory problems = yes: Major (16.0/6.0)
|   Periodically = Yes
|   |   Trouble concentrating and memory problems = No
|   |   |   Thoughts of death and suicide = No: None (231.0/82.0)
|   |   |   Thoughts of death and suicide = Yes: Bipolar (4.0)
|   |   |   Thoughts of death and suicide = yes: None (0.0)
|   |   Trouble concentrating and memory problems = Yes: Bipolar (116.0/33.0)
|   |   Trouble concentrating and memory problems = yes: None (0.0)
Time range = more than 1 month
|   Periodically = No
|   |   Thoughts of death and suicide = No
|   |   |   Trouble concentrating and memory problems = No: Artypical (18.0)
|   |   |   Trouble concentrating and memory problems = Yes: Persistence (115.0/44.0)
|   |   |   Trouble concentrating and memory problems = yes: Persistence (104.0/43.0)
|   |   Thoughts of death and suicide = Yes: Major (102.0/6.0)
|   |   Thoughts of death and suicide = yes: Major (10.0)
|   Periodically = Yes
|   |   Trouble concentrating and memory problems = No: Bipolar (74.0/4.0)
|   |   Trouble concentrating and memory problems = Yes
|   |   |   Energy = High
|   |   |   |   Changes in weight = Low: Persistence (14.0/5.0)
|   |   |   |   Changes in weight = Medium: Bipolar (5.0/3.0)
|   |   |   |   Changes in weight = High: Bipolar (8.0/4.0)
|   |   |   Energy = Medium: Bipolar (20.0/9.0)
|   |   |   Energy = Low
|   |   |   |   Changes in appetite = Low: Persistence (6.0/2.0)
|   |   |   |   Changes in appetite = High: Bipolar (21.0/8.0)
|   |   |   |   Changes in appetite = Medium
|   |   |   |   |   Changes in weight = Low: Major (2.0/1.0)
|   |   |   |   |   Changes in weight = Medium: Persistence (0.0)
|   |   |   |   |   Changes in weight = High: Persistence (9.0/3.0)
|   |   Trouble concentrating and memory problems = yes: Persistence (62.0/8.0)
```

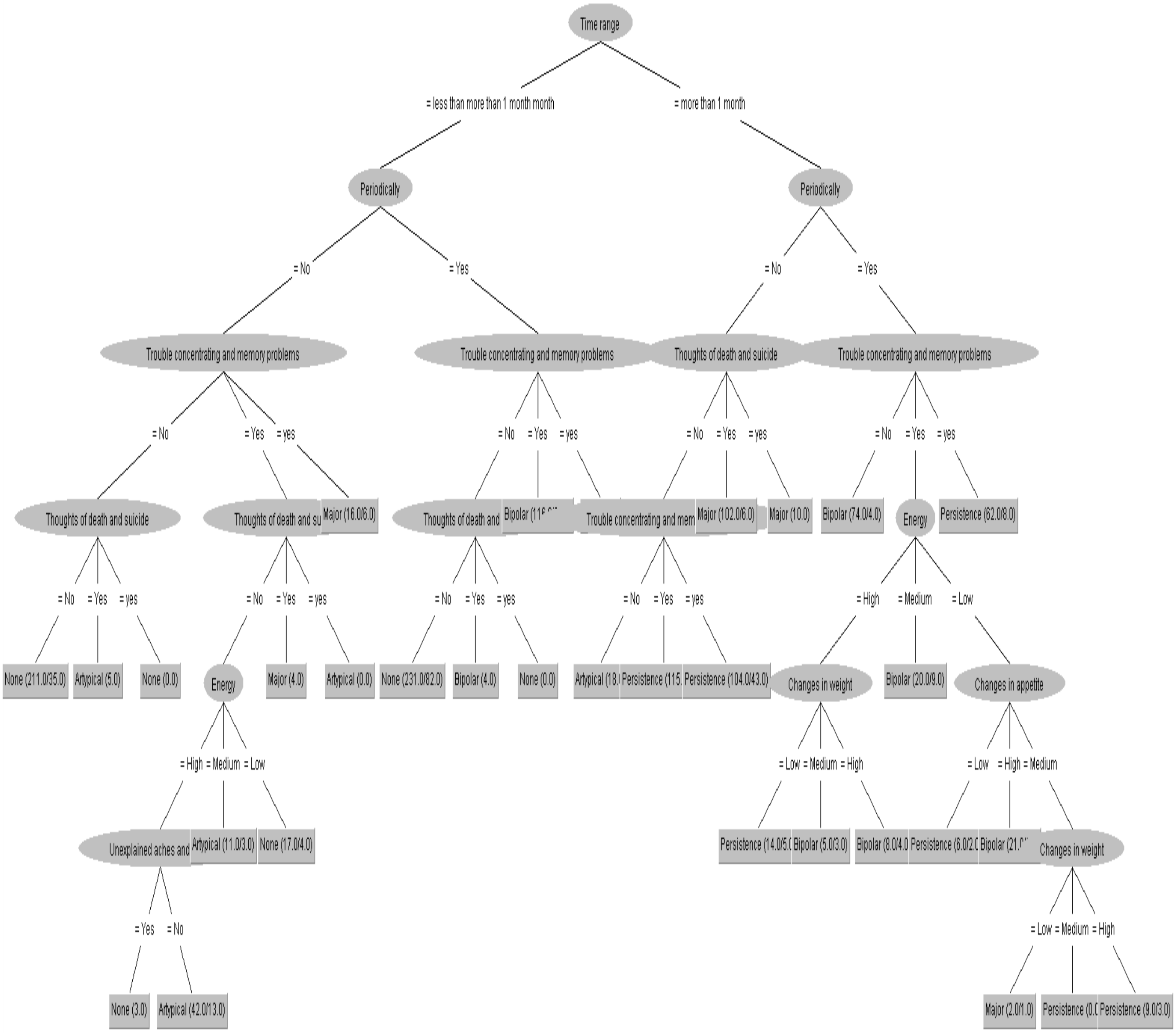


Figure 5 Decision Tree

Decision tree is consists with 31 leave and the size of the tree is 48. From this decision tree 31 predictions regarding the depression types can be identified appropriately.

4.3.2 Accuracy of the decision tree

Here, the accuracy is 72%

Incorrectly Classified Instances- 340 (Percentage- 26.6423%)

Correctly Classified Instances – 890 (Percentage- 72.3577%)

In this classification model, value of the kappa statistic is 0.6379 and it is normally a good value. So, it proves that this model is reliable to the dataset.

Table 12 Confusion matrix

	Predicted Class				
Actual Class	**None**	**Major**	**Persistence**	**Bipolar**	**Artypical**
None	335	0	0	38	14
Major	4	115	88	2	1
Persistence	3	10	198	23	2
Bipolar	82	0	13	182	2
Artypical	35	2	10	11	60

Table 13 Detailed accuracy by class

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
None	0.866	0.147	0.73	0.866	0.792	0.69	0.914	0.755
Major	0.548	0.012	0.906	0.548	0.682	0.663	0.94	0.787
Persistence	0.839	0.112	0.641	0.839	0.727	0.66	0.929	0.656
Bipolar	0.652	0.078	0.711	0.652	0.68	0.593	0.921	0.746
Artypical	0.508	0.017	0.759	0.508	0.609	0.59	0.905	0.611

TP rate is indicating the True Positive rate of the each class which defines the number of true positives classified by the model. Here, Class Persistent and None have the highest TP rates. FN rate is indicating the number of false negatives classified by the model. For all five classes, the TP rate is a higher value and the FN rate is a lower value. This is a good performance of the data set. Precision is the fraction of true positive examples among the examples that the model classified as positive. And the Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples. So, the good sensitivity also gained from the class persistent and class None. However, the study proves that the dataset is very sensitive to the model hence it gives the good performance of the model. F-score is a factor indicating how much more important recall is than precision. So, this model shows that the recall of the class persistent is 0.727 important as precision, class major is 0.682 important as precision, class bipolar is 0.68important as precision, class atypical is 0.609 important as precision and class none is 0.792important as precision.

4.3.3 Visualization of each instances with depression symptoms

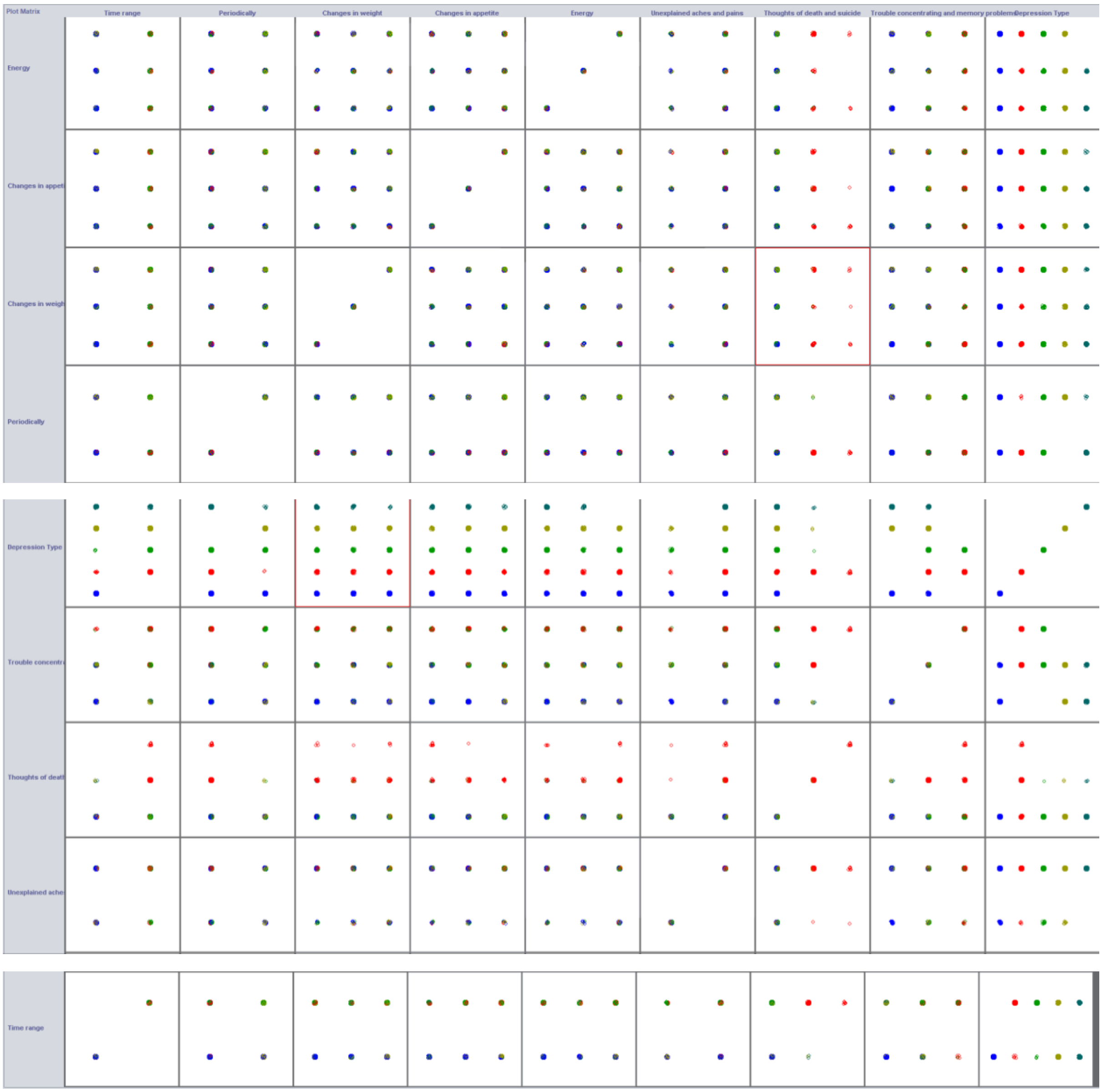
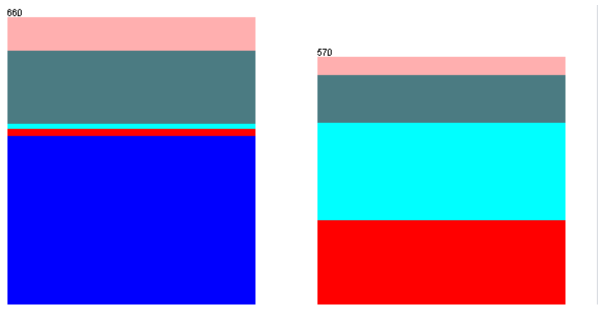


Figure 6 Instances with variables

Figure 7 shows how the data were spread on each symptoms of depression according to the clusters of depression types.

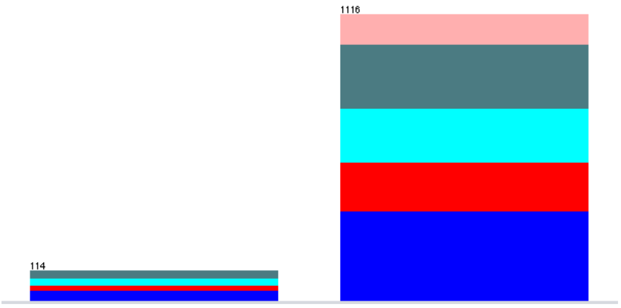
Below figures 8 to Figure 15 displays on how the depression types categorize with each symptoms.

- Pink- Atypical
- Ash – Bipolar
- Red- Major
- Light blue- Persistent
- Dark blue- None



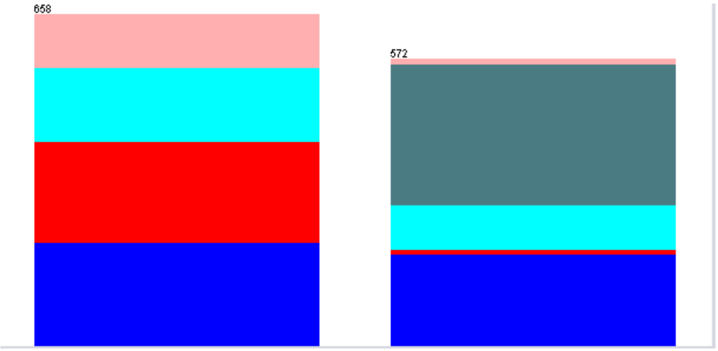
Note - X axis – less than one month, more than one month respectively

Figure 7 Time range



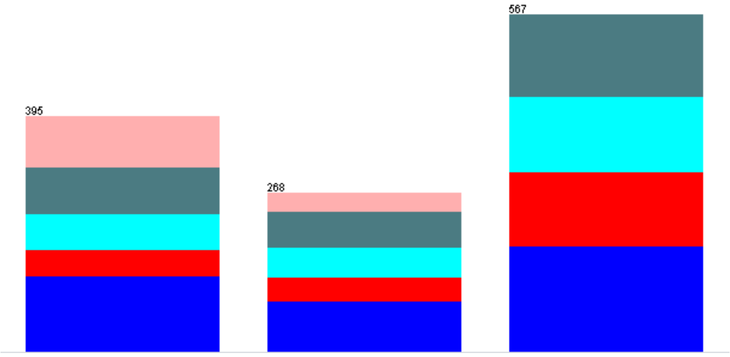
Note- X axis – No, Yes respectively

Figure 8 Unexplained aches and pains



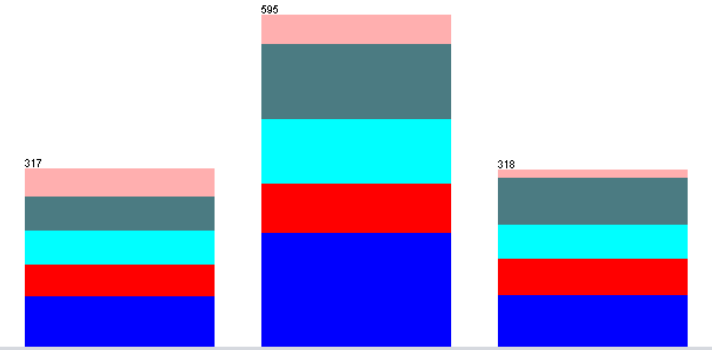
Note- X axis – No, Yes respectively

Figure 9 Periodically



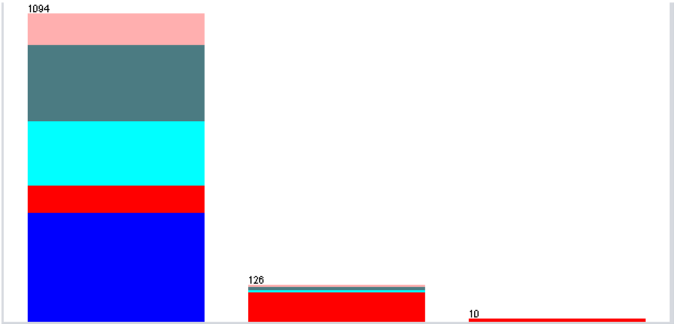
Note - X axis – high, medium, low respectively

Figure 10 Energy



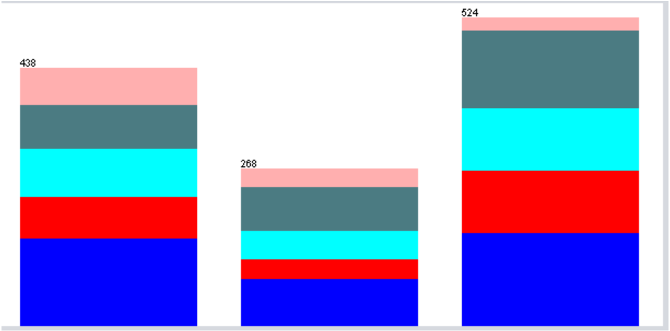
Note- X axis – high, medium, low respectively

Figure 11 Changes in appetite



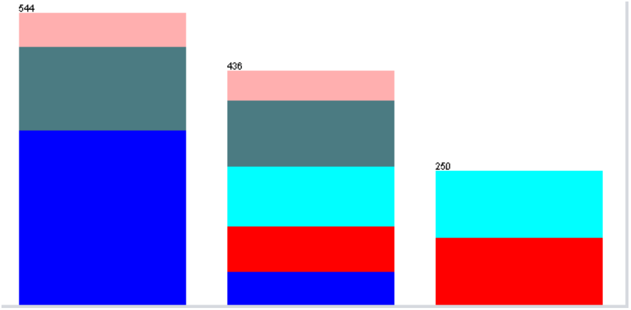
Note- X axis – high, medium, low respectively

Figure 12 Thoughts of death and suicide



Note- X axis – high, medium, low respectively

Figure 13   Changes in weight



Note- X axis – high, medium, low respectively

Figure 14 Trouble concentrating and memory problems

CHAPTER 5

CONCLUSION AND FUTURE WORK

This project is based on medical data analysis instead of industrial based analysis. So the result of this project will be more useful for the health care sector of Sri Lanka. As from the output of this analysis, we can identify whether we have the risk of having depression or not. If we have a risk we can break bad habits which is effecting to increase the risk of depression and make sure to be in a good mental health. And the initial mental status was analyzed though the most significant categorical risk factors. And also we can predict the depression type based on the symptoms of depression. This project will be more helpful to the medical related works.

In this project, data of people who are suffering from depression and not suffering with depression. Data which belongs to the risk factors, symptoms of depression were included in the dataset. By applying binary logistic regression technique in statistic, the most significant risk factors which effects to the depression disease were identified. Since the data set is having the target attribute as "depression patient/not" the values of the target attribute consist of binary "0/1" values. So, binary logistic regression model was the most suitable one which should be applied for this kind of situation. By considering the p-values of each risk factors the most significant risk factors and less significant risk factors were identified. The most significant risk factors are Gender, Working Hours, Sleeping Hours, No of children, Background (Family History), Exercise, Drugs addiction, Alcohol consumption, Serious Illnesses, Smoking, Marital Status, Education Level. The less significant risk factors which are not effects to the depression risk are Age, Social Media, Fast Food Consumption, Job Related to. By using only the identified risk factors, 2 equations were created as the prediction model. One equation is consists of combination of 12 sub equations (To calculate Y value). Another equation is for calculate the probability (P). Hosmer-Lemenshow test value of the model is 0.486 and it proves that the model fits the data well. And the model accuracy is 86%. By using the logic of the model, the result was interpreted by developing a system using Java Scripts, HTML and CS. By using the system, we can enter data and identify whether we are having the risk of being depression patient or not.

All the people were recognized as per their initial mental status on this depression as High, Medium, and Low. So, both Naïve Bayes classified and k-means clustering (classes to cluster evaluation) were applied to analyze the dataset. Accuracy of the Naïve Bayes is 70% and accuracy of the k-means clustering was 55%. According to the accuracy which was identified true the correct classified instances, Naïve Bayes was selected as the best suitable model for analyzing the initial mental status of a person. To predict the depression type, the symptoms of the depression were used. Decision tree was used as the technique and there were 31 leaves of the tree. So, 31 basic predictions can be identified through this model with the probabilities of occurrence. Model accuracy was 72%.

As for the future work of the project, I recommend to enhanced the created system to interpret the outcome of the last objective of the project; Predicting the depression type based on the symptoms. Created model displays the outcome of the 1st objective of the project. But as a future work we can enhance the system by applying the outcome of our last object too.

REFERENCES

Arabi, A.H., Buddhitha, P., Orabi, M.H. &Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic(pp. 88-97).

Ang Li, Dongdong Jiao, Tingshao Zhu, “Detecting depression stigma on social media: A linguistic analysis”; Journal of Affective Disorders; Volume 232; pp. 358-362; 2018.

Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," Modern Institute of Information Technology and Research, Department of Computer Application, Yamunanagar, Nigdi Pune, Maharashtra, 2001, India-411044

Brian A. Primack, ArielShensa, César G.Escobar-Viera, Erica L. Barrett, Jaime E.Sidani, Jason B. Colditz, A. EveretteJames, “Use of multiple social media platforms and symptoms of depression and anxiety: A nationallyrepresentative study among U.S. young adults”; Depression and Anxiety; Volume 33; pp. 323-331; 2016

Coppersmith, G., Dredze, M. & Harman, C. (2014b). Quantifying mental health signals in Twitter. In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality (pp. 51-60).

Coppersmith, G., Harman, C. &Dredze, M.(2014a). Measuring post traumatic stress disorder in Twitter. In Eighth international AAAI conference on weblogs and social media

Choudhury, M. Gamon, S. Counts and E. Horvitz, "Predicting Depression via Social Media", *Microsoft Research*, 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/,

De Choudhury, M., Counts, S., Horvitz, E.J. and Hoff, A., 2014, February. Characterizing and predicting postpartum depression from shared facebook data. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 626-638). ACM

De Choudhury, M., Counts, S. and Horvitz, E., 2013, May. Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47-56). ACM.

Drugs.com. 2021. *Major Depression Guide: Causes, Symptoms and Treatment Options*. [online] Availableat:<https://www.drugs.com/health-guide/major depression. [Accessed 1 March 2021].

International Journal of Innovative Technology and Exploring Engineering, 2019. Study of Depression Analysis using Machine Learning Techniques. 9(2S), pp.540-543.--

Hailiang Long, Xia Wu, Zhenghao Guo, J. Liu, B. Hu less,2017 'Journal of Health and Medical Informatic', *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* (), pp. .

Halldorsdottir et al., "F63. Polygenic Risk: Predicting Depressive Symptoms in Clinical and Epidemiological Cohorts of Adolescents", *Biological Psychiatry*, vol. 83, no. 9, p. S262, 2018. Available: 10.1016/j.biopsych.2018.02.676.

hen, J.H. &Rudzicz, F., 2017. Detecting anxiety through reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology–- From Linguistic Signal to Clinical Reality (pp. 58-65).

John Glaser, *Hospitals and Health Networks Magazine*.

Maria Vargas Vera, "Knowledge Extraction by using an Ontology based Annotation Tool," Knowledge Media Institute (KMi), The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom, 2001.

Melissa N Stolar, Margaret Lech, Shannon J Stolar, Nicholas B Allen; "Detection of Adolescent Depression from Speech Using Optimised Spectral Roll-Off Parameters"; Biomedical Journal of Scientific & Technical Research; 2018.Margaret Lech. Biomed J Sci & Tech research

Mowery, D., Bryan, C. & Conway, M. (2017). Feature studies to inform the classification of depressive symptoms from Twitter data for population health. arXiv preprint arXiv:1701.08229

Mrunal Kulkarni1, Prof. Arti R.Wadhekar, 'Depression Prediction System Using Different Methods ', 2019, 01(), pp.

Munmun De Choudhury Michael Gamon Scot (n.d.) 'Predicting Depression via Social Media ', *Seventh International AAAI Conference on Weblogs and Social Media,2017,* WA 98052 (), pp. .

Nadeem, M. Identifying depression on Twitter.,2016 arXiv preprint arXiv:1607.07384

Park, M., McDonald, D.W. & Cha, M. (2013). Perception differences between the depressed and non-depressed users in twitter. In Seventh International AAAI Conference on Weblogs and Social Media.

Patricia A. Cavazos-Rehg, Ph.D.Melissa J. Krauss, M.P.H,Shaina Sowles, M.P.H.,Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut, M.D. (2015) 'A content analysis of depression-related Tweets', *ournal of Engineering Science and Computing,* (), pp. 351–357..

Pedersen, T. (2015). Screening Twitter users for depression and PTSD with lexical decision lists. In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality (pp. 46-53).

Reece, A.G., Reagan, A.J., Lix, K.L., Dodds, P.S., Danforth, C.M. & Langer, E.J. (2017). Forecasting the onset and course of mental illness with Twitter data. Scientific reports, 7(1), p.13006.

S. Chandra, "Creation of an Adaptive Classifier to enhance the classification accuracy of existing classification algorithms in the field of Medical Data Mining," in International Conf. of Computing for Sustainable Global Development (INDIA Com), 2015, pp. pp 376 – 381.

S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 7547-7551, doi: 10.1109/ICASSP.2013.6639130.

Schwartz, H.A., Eichstaedt, J., Kern, M.L., Park, G., Sap, M., Stillwell, D., Kosinski, M. and Ungar, L., 2014. Towards assessing changes in degree of depression through facebook. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 118- 125)

S. Danzo, A. Connell and E. Stormshak, "Associations between alcohol-use and depression symptoms in adolescence: Examining gender differences and pathways over time", *Journal of Adolescence*, vol. 56, pp. 64-74, 2017. Available: 10.1016/j.adolescence.2017.01.007.

S.Sridharan, AkilaBanu, M. Bakkiyalakshmi, A. Buvana P (2018) 'Detection and Diagnosis on online Social network Mental Disorders using conventional Neural Networks', *ournal of Engineering Science and Computing,* (), pp. International J.

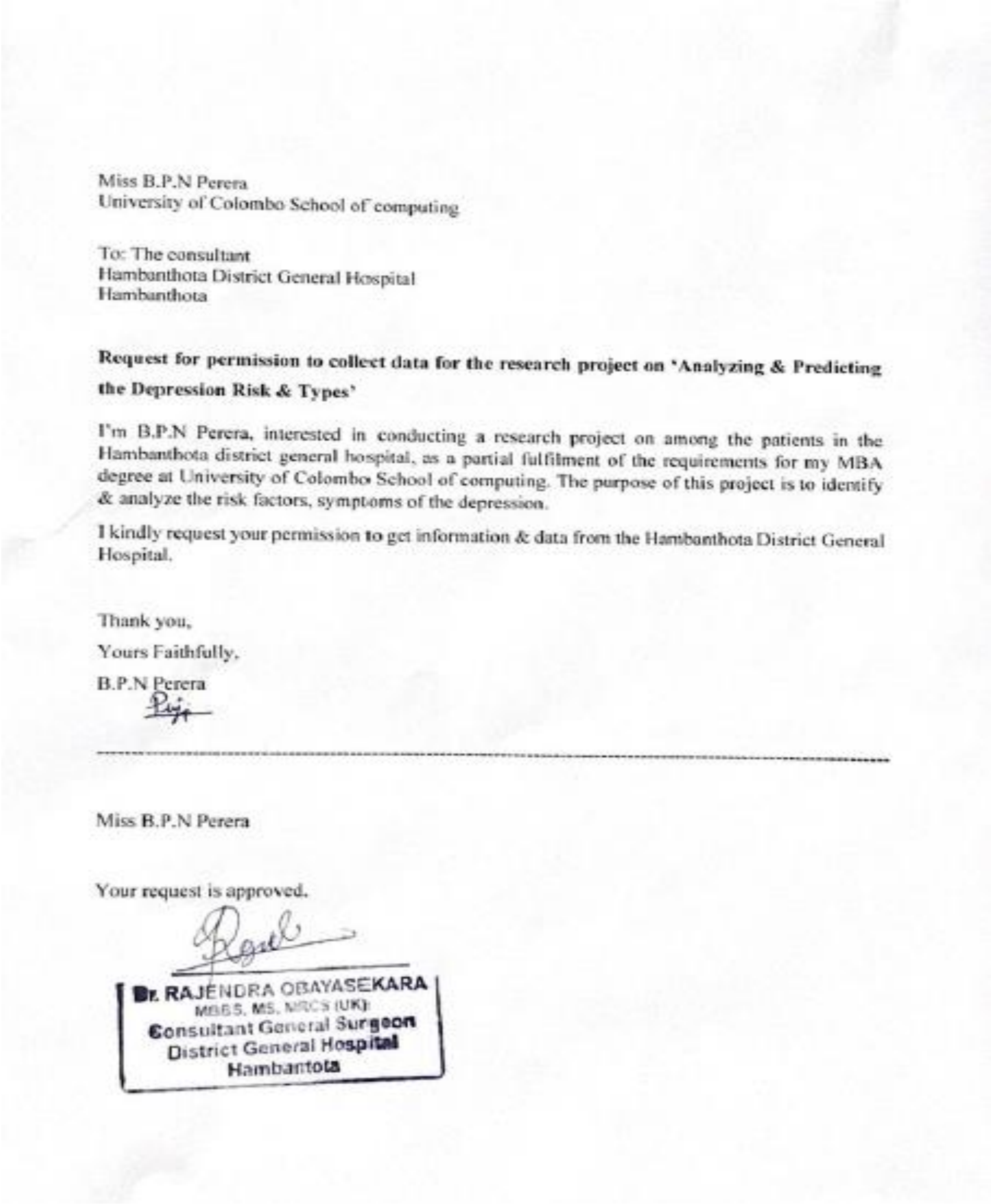
Thin Nguyen, Dinh Phung, Bo Dao, SvethaVenkatesh, MichaelBerk; “Affective and Content Analysis of Online Depression Communities”; IEEE Transactions on Affective Computing; Volume 5; pp. 217- 226; 2014.

Trotzek, M., Koitka, S. & Friedrich, C.M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L. & Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. In Pacific Asia Conference on Knowledge Discovery and Data Mining (pp. 201-213). Springer, Berlin, Heidelberg

APPENDICES

Appendix A – Permission letter



Miss B.P.N Perera
University of Colombo School of computing

To: The consultant
Hambanthota District General Hospital
Hambanthota

Request for permission to collect data for the research project on 'Analyzing & Predicting the Depression Risk & Types'

I'm B.P.N Perera, interested in conducting a research project on among the patients in the Hambanthota district general hospital, as a partial fulfilment of the requirements for my MBA degree at University of Colombo School of computing. The purpose of this project is to identify & analyze the risk factors, symptoms of the depression.

I kindly request your permission to get information & data from the Hambanthota District General Hospital.

Thank you,

Yours Faithfully,

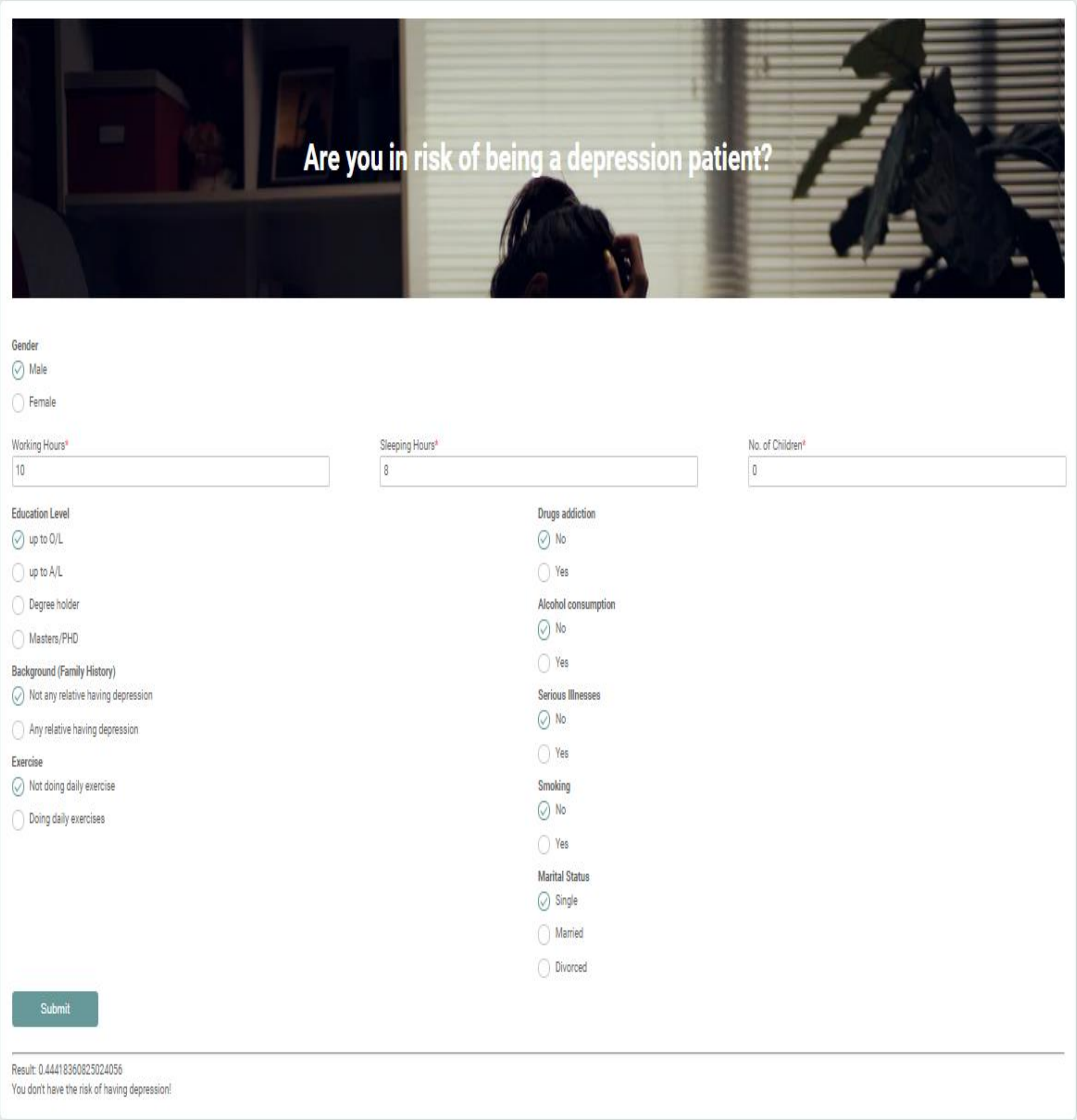
B.P.N Perera

Miss B.P.N Perera

Your request is approved.

Dr. RAJENDRA OBAYASEKARA
MBBS, MS, MRCS (UK)
Consultant General Surgeon
District General Hospital
Hambantota

Appendix B- System UI



Are you in risk of being a depression patient?

Gender

● Male

○ Female

Working Hours*: 10

Sleeping Hours*: 8

No. of Children*: 0

Education Level

● up to O/L

○ up to A/L

○ Degree holder

○ Masters/PHD

Background (Family History)

● Not any relative having depression

○ Any relative having depression

Exercise

● Not doing daily exercise

○ Doing daily exercises

Drugs addiction

● No

○ Yes

Alcohol consumption

● No

○ Yes

Serious Illnesses

● No

○ Yes

Smoking

● No

○ Yes

Marital Status

● Single

○ Married

○ Divorced

Result: 0.44418360825024056

You don't have the risk of having depression!