PREDICTING POTENTIAL PATIENT OF DEPRESSION USING DATA MINING

MCS (Final) – Morning

Abstract

Depression has been outlined as a psychological disorder characterized by sadness, loss of interest or pleasure, passions of guilt or low self- esteem, disturbed sleep or appetite, feelings of exhaustion and poor attention. This research will contribute in predicting using data mining, mostly using classification, the potential patients of depression.

PRESENTED TO: SIR TEHSEEN A. JILANI

DCS-UBIT, University of Karachi.

PROJECT DETAILS

Project	*Predicting Potential Patient of Depression using Data Mining*
Software Used	*SPSS, WEKA*
Course Name	*Data Mining and Data Warehousing*
Course Code	*CS – 626*
Course Supervisor	*Sir Tehseen A. Jilani*
Class	*MCS – Final*
Shift	*Morning*
Batch	*2020-2022*
Group Members	*P19101012 – Ayesha Abrar*
	P19101017 – Fakiha Khan
	P19101072 – Syeda Kisa Batool Zaidi
	P19101075 – Syied Dilawar Asad Zaidi

1. INTRODUCTION

Data mining is a vast filed spanned over a number of multiple fields including database management system, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge base systems, information retrieval and recovery, quantum computing and data visualization [1][2][3][4].

The application of data mining in healthcare is constantly increasing and becoming more popular. Data mining can play a vital role in healthcare allowing insurers uncover fraud and abuse, improving the decision making of customer relationship management, helping doctors identify effective treatments and best practices, identifying risk factors associated with the beginning of diabetes, and enabling patients to receive better and more affordable healthcare services [5]. Healthcare data mining provides countless opportunities for hidden pattern exploration from the huge healthcare data stores. These models can be used by doctors to establish diagnoses, prognoses and treatments for patients in healthcare establishments [6].

According to Obenshain [7], "Business and marketing organizations may be ahead of healthcare in applying data mining to derive knowledge from data. This is quickly changing. Successful mining applications have been implemented in the healthcare arena. Further exploration of data mining for research related to infection control and hospital epidemiology seems in order, especially where the data volume exceeds capabilities of traditional statistical techniques."

Depression (major depressive disorder) is a common and serious medical illness that negatively affects how you feel, the way you think and how you act. Fortunately, it is also treatable. Depression causes feelings of sadness and/or a loss of interest in activities you once enjoyed. It can lead to a variety of emotional and physical problems and can decrease your ability to function at work and at home [8].

Depression symptoms can vary from mild to severe and can include [8]:

- Feeling sad or having a depressed mood
- Loss of interest or pleasure in activities once enjoyed
- Changes in appetite — weight loss or gain unrelated to dieting
- Trouble sleeping or sleeping too much
- Loss of energy or increased fatigue
- Increase in purposeless physical activity (e.g., inability to sit still, pacing, handwringing) or slowed movements or speech (these actions must be severe enough to be observable by others)
- Feeling worthless or guilty
- Difficulty thinking, concentrating or making decisions
- Thoughts of death or suicide

Symptoms must last at least two weeks and must represent a change in your previous level of functioning for a diagnosis of depression [8].

Also, medical conditions (e.g., thyroid problems, a brain tumour or vitamin deficiency) can mimic symptoms of depression so it is important to rule out general medical causes [8].

Depression affects an estimated one in 15 adults (6.7%) in any given year. And one in six people (16.6%) will experience depression at some time in their life. Depression can occur at any time, but on average, first appears during the late teens to mid-20s. Women are more likely than men to experience depression. Some studies show that one-third of women will experience a major depressive episode in their lifetime. There is a high degree of heritability (approximately 40%) when first-degree relatives (parents/children/siblings) have depression [8].

2. ATTRIBUTES SELECTION

Attributes (symptoms regarding depression) selection is an important task in data mining. This includes solving an active set of relevant symptoms needed for the constructing a data model. Rushing this process can result in possibly selecting repetitive and unnecessary attributes, which could impact the created model and the results of the process.

This work is relied on a number of surveys and questionnaires [9][10]. The selected set was further tested and remodelled referencing to depression and predicting it effectively. After filtering out redundant attributes, a final set of attributes is presented in Table 1 below.

S. NO.	QUESTION	DATA TYPE	POSSIBLE VALUE
1.	Gender	String	Male, Female
2.	Age	Numeric	Between 10 – 80
3.	Marital Status	String	Single, Married, Widowed, Divorced, Separated
4.	Education	String	High School, College, Undergraduate, Masters, Doctorate, Other
5.	Employment	String	Student, Full-time, Part-time, Freelancer, Own Business, Retired, Other
6.	All the tasks you have performed, are taking much more time than usual?	Numeric	On the scale of 1 – 5
7.	You are facing a lack of concentration?	Numeric	On the scale of 1 – 5
8.	You are feeling you have no future?	Numeric	On the scale of 1 – 5
9.	You are facing problems with making decisions?	Numeric	On the scale of 1 – 5
10.	You feel, your life is sad, as there is no joy in your life anymore?	Numeric	On the scale of 1 – 5
11.	You have lost interest in all things that were important to you once upon a time?	Numeric	On the scale of 1 – 5
12.	You have been feeling guilty for everything you do?	Numeric	On the scale of 1 – 5
13.	You have been very irritated and angry recently?	Numeric	On the scale of 1 – 5

14.	You have been feeling very fatigued?	Numeric	On the scale of 1 – 5
15.	You are feeling that everything you have done has been a failure?	Numeric	On the scale of 1 – 5
16.	You are having a lack of sleep?	Numeric	On the scale of 1 – 5
17.	You are having suicidal thoughts?	Numeric	On the scale of 1 – 5
18.	You have lost or gained weight without any diet programs?	Numeric	On the scale of 1 – 5
19.	You are having a loss of appetite?	Numeric	On the scale of 1 – 5
20.	You are having trust issues with everyone around you?	Numeric	On the scale of 1 – 5
21.	You are having trouble in all your relationships (home as well as professional)?	Numeric	On the scale of 1 – 5
22.	Q_Sum	Numeric	Sum of all factors
23.	May have depression?	String	If Q_Sum <= 50, YES; Else NO
24.	Depression Boolean	Numeric	If Depression = YES, 1; Else 0

Table 1 - Attributes Set to Predict Depression

1= Completely Agree, 2= Somewhat Agree, 3= Neutral, 4= Somewhat Disagree, 5= Completely Disagree

3. NORMAL DISTRIBUTION

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. The equation for Normal Distribution is as follow.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$f(x) = Probabilty\ Density\ Function, \sigma = Standard\ Deviation, \mu = Mean$$

We use SPSS to check our data for Normal Distribution. Following are the results of the output produced.

a. <u>CASE PROCESSING SUMMARY</u>

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
QSum	149	100.0%	0	0.0%	149	100.0%

Table 2 - Case Processing of Total Population

b. DESCRIPTIVES

			Statistic	Std. Error
QSum	Mean		46.2416	1.21484
	95% Confidence Interval for Mean	Lower Bound	43.8409	
		Upper Bound	48.6423	
	5% Trimmed Mean		46.1324	
	Median		45.0000	
	Variance		219.901	
	Std. Deviation		14.82905	
	Minimum		16.00	
	Maximum		76.00	
	Range		60.00	
	Interquartile Range		24.00	
	Skewness		.141	.199
	Kurtosis		-.929	.395

Table 3 - Descriptive Statistics of Population

c. TESTS OF NORMALITY

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
QSum	.071	149	.061	.973	149	.006

a. Lilliefors Significance Correction

Table 4 - Test of Normality

d. Q_SUM HISTOGRAM

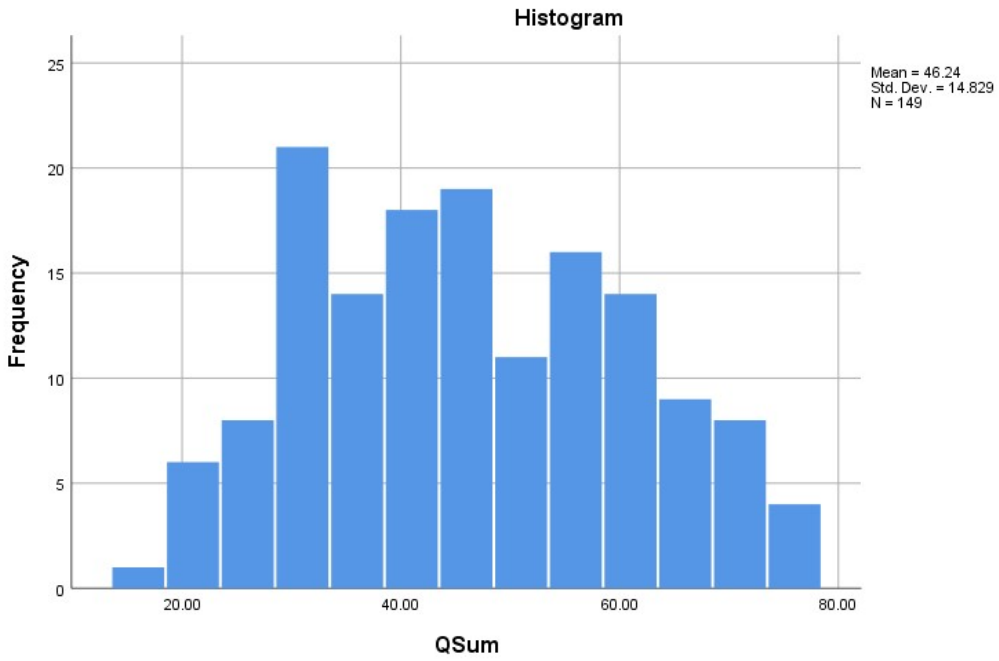

Figure 1 - Histogram (Distribution of Q_Sum Score)

e. QSUM STEM-AND-LEAF PLOT

```
Frequency     Stem &  Leaf

     2.00        1 .  69
     5.00        2 .  00223
    17.00        2 .  55577788999999999
    14.00        3 .  00000001123344
    16.00        3 .  5556777778889999
    19.00        4 .  0000011222223344444
    15.00        4 .  555566777788889
    15.00        5 .  011233333344444
    14.00        5 .  66666788888999
    12.00        6 .  011111233334
     9.00        6 .  566667789
     8.00        7 .  01222234
     3.00        7 .  566

Stem width:     10.00
Each leaf:       1 case(s)
```

f. NORMAL Q-Q PLOT

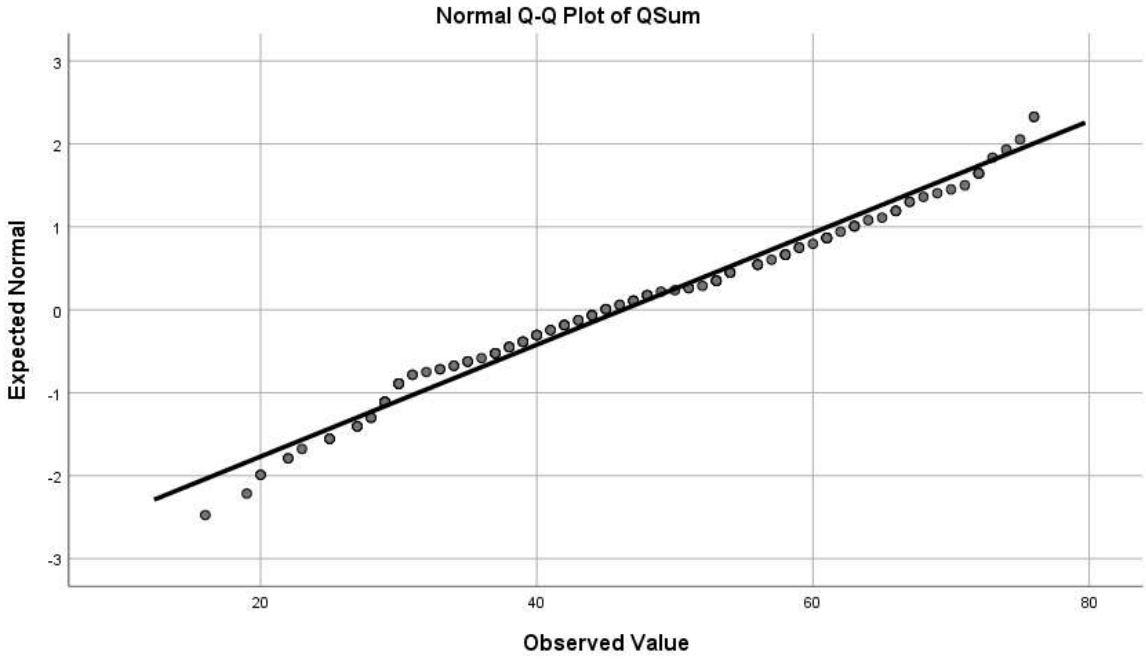

Figure 2 - Q-Q Plot

g. DETRENDED NORMAL Q-Q PLOT

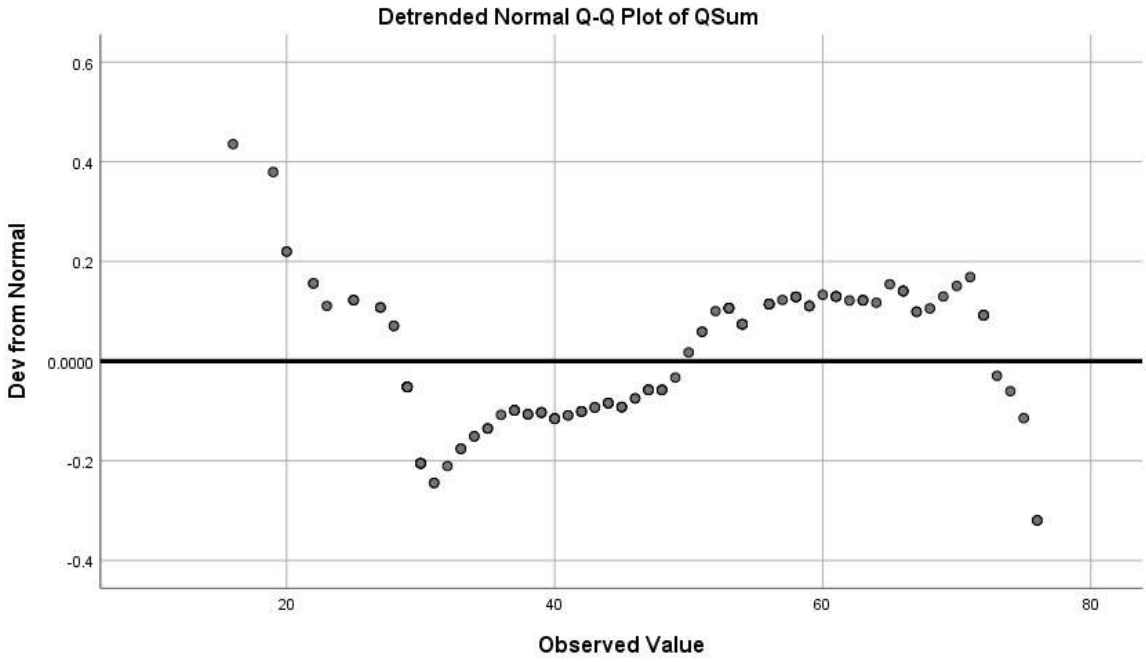

Figure 3 – Detrended Normal Q-Q Plot

h. BOX PLOT

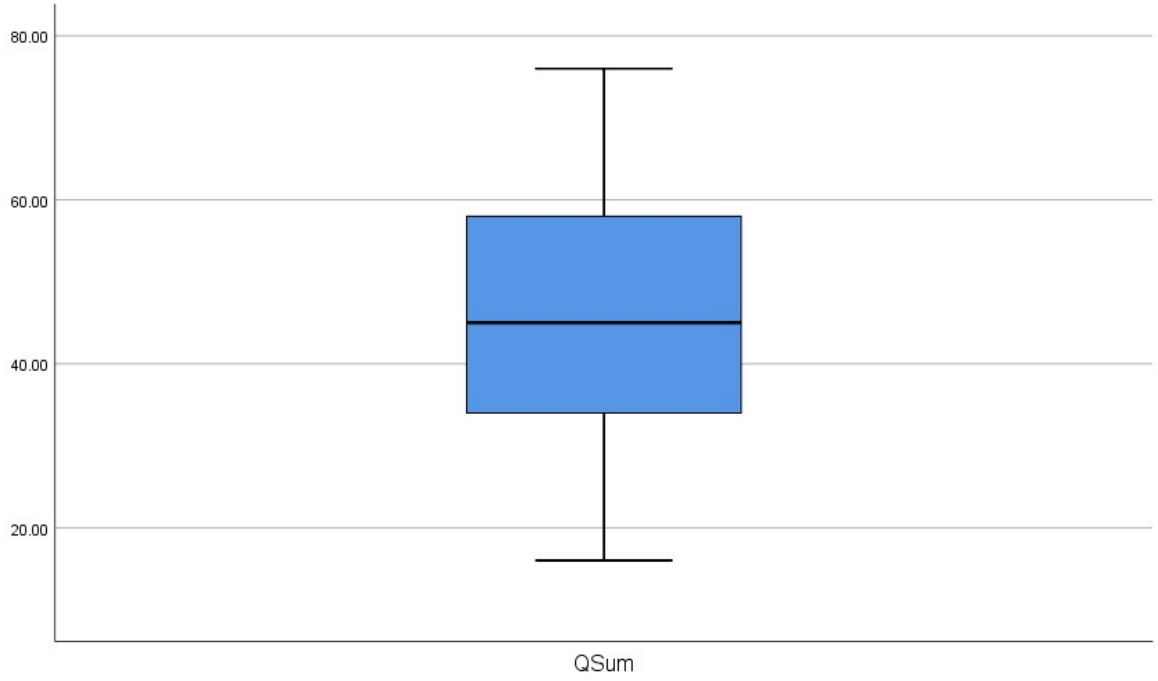

Figure 4 - Box Plot

4. ORDINAL REGRESSION

In statistics, ordinal regression (also called "ordinal classification") is a type of regression analysis used for predicting an ordinal variable, i.e., a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant. It can be considered an intermediate problem between regression and classification.

We use SPSS to check our data for Ordinal Regression. We take MayHaveDepression as the Dependent Variable, (Age, Gender, Marital Status, Education Employment) as and Q1-Q16 (all 16 Question Variables) as Covariates. Following are the results of the output produced.

a. CASE PROCESSING SUMMARY

		N	Marginal Percentage
May_Have_Depression	NO	61	40.9%
	YES	88	59.1%
Gender	Female	99	66.4%

	Male	50	33.6%
Age	18.00	8	5.4%
	19.00	3	2.0%
	20.00	11	7.4%
	21.00	12	8.1%
	22.00	22	14.8%
	23.00	18	12.1%
	24.00	24	16.1%
	25.00	11	7.4%
	26.00	11	7.4%
	27.00	5	3.4%
	28.00	2	1.3%
	29.00	2	1.3%
	30.00	4	2.7%
	31.00	2	1.3%
	32.00	2	1.3%
	33.00	1	0.7%
	34.00	1	0.7%
	35.00	1	0.7%
	40.00	3	2.0%
	45.00	2	1.3%
	46.00	1	0.7%
	49.00	1	0.7%
	50.00	1	0.7%
	70.00	1	0.7%
Marital_Status	Divorced	1	0.7%
	Married	27	18.1%

	Seperated	1	0.7%
	Single	120	80.5%
Education	College	12	8.1%
	Doctorate	1	0.7%
	Graduate	4	2.7%
	High School	5	3.4%
	In MSc continue	1	0.7%
	Masters	62	41.6%
	Undergraduate	64	43.0%

Table 5 - Case Processing of Depression

b. **MODEL FITTING INFORMATION**

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	201.638			
Final	.033	201.606	56	.000

Table 6 - Model Fitting

c. **GOODNESS OF FIT**

	Chi-Square	df	Sig.
Pearson	.016	82	1.000
Deviance	.033	82	1.000

Table 7 - Goodness of Fit Test

5. MODEL CREATION

In this research, classification is implemented for finding patterns in the data. To create a model, we need a classification algorithm. To achieve this, we are using C4.5 decision tree. WEKA implements a later released revised version C4.5 rev 8, referred to as J4.8 in its own library.

Partitioning a dataset in an essential step in creating a data model. We divide a data set into two set, namely, Train Set and Test Set. The former is quite big in size as compared to the later.

To collect the data for this model, we shared an online survey form among friends and family. The results are then classified in WEKA using J4.8 implementation. This resulted in 98.6% of the instance being classified correctly. Below equation shows a Confusion Matrix, indicating True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) [18].

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} = \begin{bmatrix} 88 & 0 \\ 2 & 59 \end{bmatrix}$$

Several other classification metrics are also used for evaluation, namely Accuracy, Precision and Recall. These are defined as;

$$Accuracy = \frac{(\,TP + TN\,)}{(\,TP + TN + FP + FN\,)} = \frac{(\,88 + 59\,)}{(\,88 + 59 + 2 + 0\,)} = 0.986$$

$$Precision = \frac{TP}{(\,TP + FP)} = \frac{88}{(88 + 2)} = 0.977$$

$$Recall = \frac{TP}{(\,TP + FN)} = \frac{88}{(88 + 0)} = 1.000$$

A test set is resulted in an Accuracy of 97.72%, with Precision of 96.15% and Recall of 100% [20].

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} = \begin{bmatrix} 25 & 0 \\ 1 & 18 \end{bmatrix}$$

6. MODEL USAGE

A total of about 149 data instances are collected during the survey [1. We use this dataset for training with the model. Then we take another dataset comprise of 44 data instances to test the model.

After training and testing, this model should be used for classifying unknown cases provided that the accuracy of classification is acceptable. The classification model should be capable of predicting unseen instances using the model it has learned.

7. RESULT INTERPRETATION

S. No.	Gender	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Depression
1	F	3	1	2	3	4	1	1	1	2	2	1	3	2	2	1	1	YES
2	F	2	2	4	1	5	2	3	1	2	4	1	5	5	4	2	2	YES
3	F	3	3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	NO
4	F	1	1	2	2	2	1	2	1	2	4	1	5	1	1	1	3	YES
5	F	3	2	2	1	1	1	1	1	2	2	2	1	2	2	1	1	YES
6	F	3	2	2	1	1	1	1	1	2	2	2	1	2	2	1	1	YES

7	M	5	5	5	4	3	2	5	1	1	5	5	5	5	3	2	2	NO
8	M	1	2	2	2	3	3	3	1	2	1	2	2	2	2	1	1	YES
9	M	1	1	3	5	5	5	5	4	3	3	3	5	3	4	1	5	NO
10	F	3	2	3	3	2	1	3	1	3	2	3	3	4	4	1	3	YES
11	M	3	2	4	4	5	5	4	4	2	4	2	5	4	3	5	4	NO
12	F	2	4	5	5	5	4	5	5	4	5	5	5	2	5	2	2	NO
13	F	1	1	1	1	1	1	1	3	2	1	1	5	5	4	1	1	YES
14	F	2	3	1	1	1	1	4	1	2	2	3	1	5	5	3	4	YES
15	M	3	2	5	5	4	5	2	1	2	3	5	3	3	5	2	3	NO
16	M	2	2	5	1	5	5	4	4	2	2	2	5	1	5	1	5	NO
17	F	2	2	5	2	2	1	5	1	2	5	5	5	3	2	2	5	YES
18	M	2	3	1	3	2	4	4	1	3	2	1	1	5	3	2	2	YES
19	F	2	5	5	2	5	5	5	4	2	5	1	5	5	4	4	2	NO
20	F	5	3	1	3	1	1	4	1	2	4	4	1	2	2	1	3	YES
21	F	2	5	1	1	2	4	3	1	2	4	2	5	2	3	1	1	YES
22	F	3	4	1	3	2	2	4	2	3	3	3	3	3	3	3	3	YES
23	F	2	3	5	3	5	2	5	2	3	4	3	5	4	3	3	2	NO
24	F	5	5	3	2	4	2	4	2	4	4	4	5	5	5	5	5	NO
25	F	5	5	5	4	5	3	5	5	3	5	5	5	4	5	4	5	NO
26	F	3	1	2	1	1	1	5	2	4	2	1	2	1	2	1	1	YES
27	F	1	3	2	1	1	1	2	1	3	3	3	4	1	1	1	1	YES
28	F	2	3	4	4	5	2	5	2	3	5	5	5	2	4	5	5	NO
29	F	4	5	5	5	5	5	5	5	5	5	5	5	1	5	5	5	NO
30	M	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	YES
31	F	2	2	5	1	5	4	2	2	4	5	4	5	1	2	4	4	NO
32	F	1	2	3	1	1	1	3	1	1	3	1	4	2	1	3	1	YES
33	F	1	4	5	2	4	5	5	5	3	3	5	5	5	5	5	5	NO
34	F	3	2	5	4	5	1	5	2	3	5	2	5	3	4	2	2	NO
35	F	1	3	2	1	1	1	2	1	3	3	3	4	1	1	1	1	YES
36	F	1	4	1	1	1	1	4	1	2	1	1	5	3	2	1	1	YES
37	F	1	3	2	3	3	3	3	2	3	2	3	5	3	3	2	3	YES
38	F	3	4	5	3	2	2	4	4	4	5	4	5	4	3	2	2	NO
39	M	2	1	5	3	3	2	2	1	2	5	4	5	1	3	3	2	YES
40	M	2	1	1	2	4	2	2	1	3	5	2	5	3	3	3	1	YES
41	F	2	1	5	2	5	3	2	2	5	4	5	5	5	5	2	5	NO
42	F	3	4	5	4	5	5	5	5	4	4	5	5	5	4	4	5	NO
43	M	2	3	3	2	3	3	4	3	3	3	3	3	3	3	3	3	YES
44	M	2	1	1	2	4	2	2	1	3	5	2	5	3	3	3	1	YES

Table 8 - Sample Set

Table 8 contains all the data used for the unknown cases that need to be diagnosed. Each entry represents a case to be analysed. The column Q1 through Q16 represents the attribute questions numbered 6 to 21. The values mentioned are also defined earlier in the attribute table representing the value of scale.

To analyse the result, we use IF-THEN rule on row # 18 and 26.

Row # 18

IF

Getting easily irritated and angry (1)
And having problems maintaining any relation (2)
And having some problem making a decision (3)

THEN

There is potential that this person is suffering from Depression (YES).

Row # 28

IF

Is not getting irritated and angry (5)
Easily managing their sleep (5)

THEN

There is no suffering from Depression (NO).

8. CONCLUSION

Depression is an exponentially growing medical illness. It's hard to analyse depression due to a number of its symptoms being shared with other illnesses. In this research, a set of symptoms attributes were selected based on online surveying. Some of these attributes overlap with various mental disorders. In the end, the attribute set is sufficient to isolate depression from other illnesses. Collected data was used to train and test the classification model.

The classification technique will further improve over the time. Firstly, the symptoms attributes will be discussed with an expert in the field of mental health to set a more suitable set for classification. After that, a large data set will be used to train and test the algorithm further to deduce results with great efficiency.

9. REFERENCES

1. Data Mining: Concepts and Techniques – By Jiawei Han, Jian Pei, Micheline Kamber
2. Data Mining and Knowledge Discovery Technologies – edited by Taniar, David
3. Introduction to Data Mining and its Applications – By S. Sumathi, S.N. Sivanandam
4. Data Mining in Healthcare: Current Applications and Issues – By Ruben D. Canlas Jr., MSIT, MBA
5. Data-Mining Technologies for Diabetes: A Systematic Review – Miroslav Marinov, M.S, Abu Saleh Mohammad Mosa, M.S, Illhoi Yoo, Ph.D, and Suzanne Austin Boren, Ph.D., MHA
6. Medical Diagnosis Data Mining Based on Improved Apriori Algorithm – Wenjing Zhang, Donglai Ma, Wei Yao
7. Application of Data Mining Techniques to Healthcare Data – https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1080.8658&rep=rep1&type=pdf
8. https://www.psychiatry.org/patients-families/depression/what-is-depression
9. Depression Questionnaire – https://www.questionpro.com/blog/depression-questionnaire/
10. Patient Health Questionnaire – https://patient.info/doctor/patient-health-questionnaire-phq-9