International Journal of Advance and Innovative Research
Volume 6, Issue 2 (VI): April - June, 2019 Part - 2

ISSN 2394 - 7780

PREDICTION OF FUTURE DEPRESSION USING DATA MINING

Prof. Bhakti N. Raut
Sonopant Dandekar College, Palghar

ABSTRACT

Depression is a complex mental health condition that causes a person to have low mood and may leave them feeling persistently sad or hopeless. Depression can take several forms, including bipolar disorder (formally called manic-depression), which is a condition that alternates between periods of euphoria and depression. The effects of depression may extend beyond a person's emotions and mental health. Depression can also affect a person's physical health.

Purpose of this paper is to diagnose through the application of data mining, namely classification, to predict patients who will most likely develop depression or are currently suffering from depression. To obtain result, data mining software WEKA was used.

Keywords: Depression Diagnosis; Data mining; Prediction; Basic Mental Health Problems.

I. INTRODUCTION

Depression is likely to strike many people to some degree in their lifetime. According to the Centres for Disease Control and Prevention, 9.1 percent of people reported current major or minor depression. Depression may lead to serious diseases like heart diseases, diabetes, blood pressure etc. Prediction of depression will helps us to prevent the affects and possibility of future diseases.

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). Prediction is one of the key property of data mining. Many forms of data mining are predictive. Predictions have an associated probability (How likely is this prediction to be true?). Prediction probabilities are also known as confidence (How confident can I be of this prediction?).Data mining provides the methodology and technology for healthcare organizations to evaluate treatment effectiveness, save lives of patients using predictive medicine, manage healthcare at different levels, manage customer relationship, detect waste, fraud and abuse.

It is stated in [1] that Data Mining in healthcare is used mainly for predicting various diseases, assisting with diagnosis and advising doctors in making clinical decisions. But, the potential of data mining is much bigger – it can provide question-based answers, anomaly-based discoveries, provide more informed decisions, probability measures, predictive modelling, and decision support.

[2] Stated that Prediction is nothing but finding out the knowledge or some pattern from the large amounts of data. In Data Mining, the term "Prediction" refers to calculated assumptions of certain turns of events made on the basis of available processed data. It is a cornerstone of predictive analytics. The prediction itself is calculated from the available data and modelled in accordance with the existing dynamics. The nature of prediction varies from the nature of the project. It can be simple correlation of sentiments and conversions out of which you can understand whether the user will engage with your piece of content in a productive manner or not.

In this paper we have used logistic regression to predict the result based on certain number of attributes and past data. To train and test the classification model sample data is used. Section 3 gives an overview of logistic regression. Section 4 presents the methodology and the data sets used in this research to predict depression. Section 5 evaluated the result of test data set. Finally, conclusion is covered in Section 6. The well-known WEKA tool is adopted for this study.

II LITERATURE REVIEW

[3] In this study author have analysed the performance of machine learning techniques to predict mental health disorder in children. Best First Search technique has used to eliminate redundant and irrelevant attributes. Author has compared three machine learning techniques (**Multilayer Perceptron, Multiclass Classifier, LAD Tree Technique**) based on dataset for different mental health problems. And then concluded that Multiclass Classifier produces much accurate results than other techniques on selected attributes.

ISSN 2394 - 7780

[4]In this study author have predicted the future depression possibility by using data mining tool WEKA. Synthetic data, created using JAVA program was used to train and test the classification model. He has used j4.8 algorithm for prediction. Author have selected various attributes through the questionnaire. He has used training data set to create model and testing data set to predict the result on unknown instances. 400 unknown instances were used to predict the result. Finally he concluded that the outcomes for the synthetic datasets were reasonable in terms of accuracy, precision, and recall of the training and testing processes.

[6] In this study author have predicted mental health problems among children using machine learning techniques. Author has built the model which can assists the professionals to identify the problem if the known evidences of the patient are given as input. She has compared eight classification algorithm and concluded that Multilayer Perceptron, Multiclass Classifier and LAD Tree produce more accurate results than the others.

III. OVERVIEW OF LOGISTIC REGRESSION

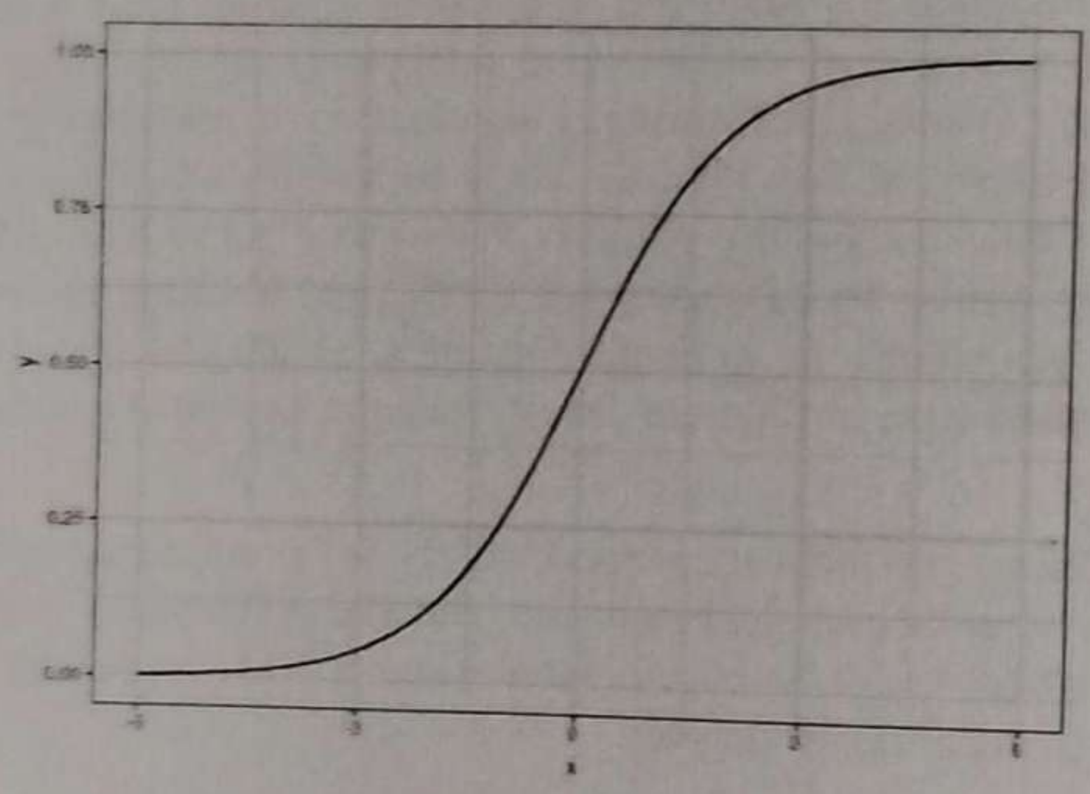
Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables.

Logistic regression is a solution for classification. Instead of fitting a straight line the logistic regression uses logit function to give the output of linear equation between 0 and 1.

Logistic function is defined as follows:

logistic(n)=1/1+exp(−n)

and it looks like this



IV METHODOLOGY

4.1 Data Collection

To collect the data, set of attributes are selected. Attribute selection is very important part of this research. Data set are of two types. One is training data set, to create model and another one is testing data set, to test the result. We have collected data through the questionnaire. To prepare data set we have collected some attribute from online surveys. This set has **22** attributes including the class variable 'Depression Possibility'. The final set of attributes is presented in Table below.

SET OF ATTRIBUTES

Attribute	Values
Fatigue	negative:0 mild:1 medium:2 positive:3
Mood swing	negative:0 mild:1 medium:2 positive:3
Dry Mouth	negative:0 mild:1 medium:2 positive:3
Vision Problems	negative:0 mild:1 medium:2 positive:3
Dizziness	negative:0 mild:1 medium:2 positive:3
Irritability	negative:0 mild:1 medium:2 positive:3
Constipation	negative:0 mild:1 medium:2 positive:3
Feelings of guilt	negative:0 mild:1 medium:2 positive:3
Worthlessness	negative:0 mild:1 medium:2 positive:3
Loss of interest	negative:0 mild:1 medium:2 positive:3

ISSN 2394 - 7780

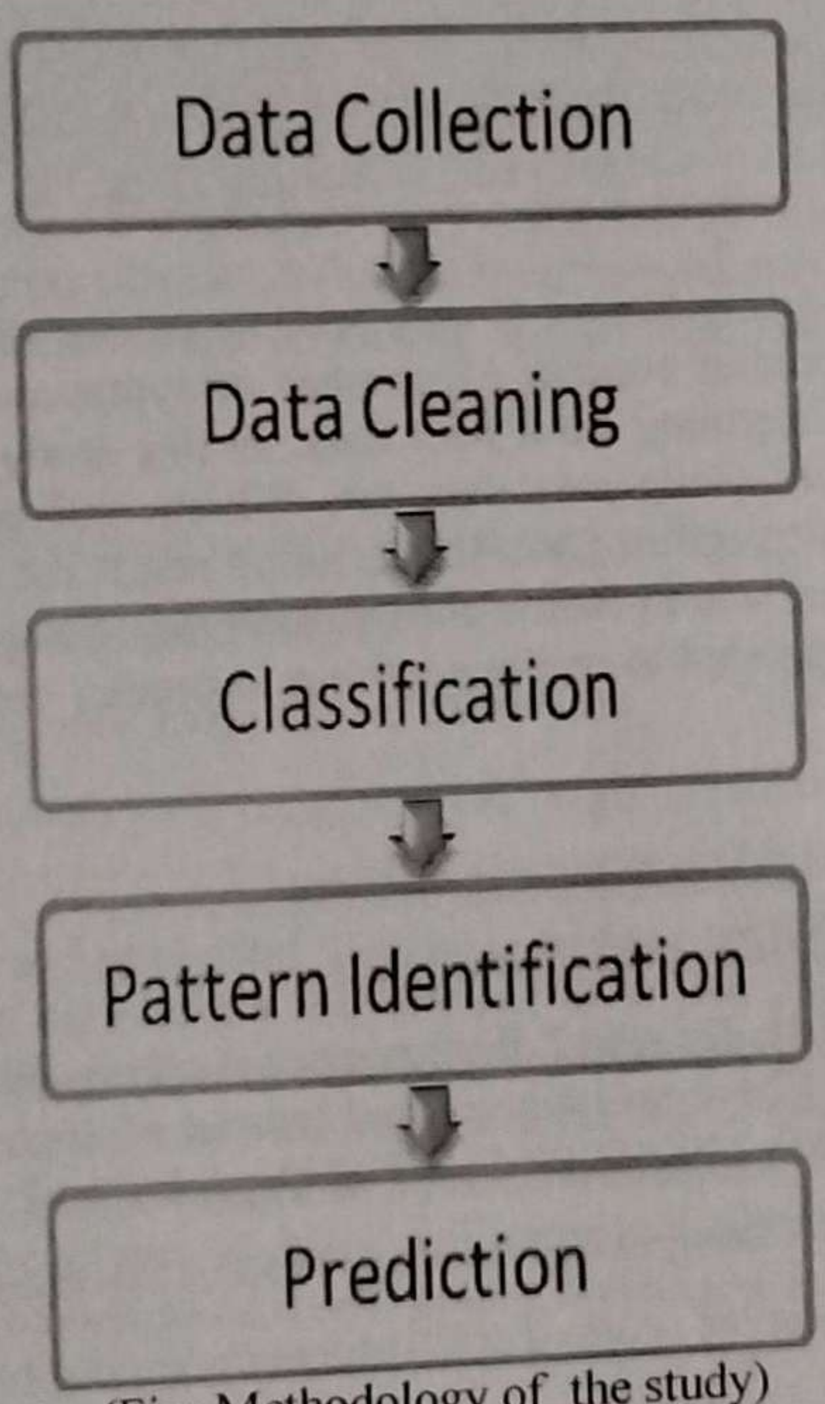
Difficulty concentrating	negative:0	mild:1	medium:2	positive:3
Difficulty Remembering	negative:0	mild:1	medium:2	positive:3
Insomnia	negative:0	mild:1	medium:2	positive:3
Oversleeping	negative:0	mild:1	medium:2	positive:3
Low appetite	negative:0	mild:1	medium:2	positive:3
Weight loss	negative:0	mild:1	medium:2	positive:3
Thoughts of death	negative:0	mild:1	medium:2	positive:3
Headache	negative:0	mild:1	medium:2	positive:3
Pessimism	negative:0	mild:1	medium:2	positive:3
Early morning awakening	negative:0	mild:1	medium:2	positive:3
Slowed thinking	negative:0	mild:1	medium:2	positive:3
Depression_Possibility	{ tested_negative,	tested_positive }		

(Table: Attribute set)

4.2 MODEL BUILDING

In this study , classification is deployed for finding hidden patterns in data set. Logistic regression is used to predict the result. To build this model training data set has used. After collecting data , data cleaning takes place. In data cleaning we replace the null values and then proceed for model creation.

As weka accepts csv or .arff files, so to create model data file for creating model is created with **.arff** extension. The depression classification model was used to predict 48 unseen instances through re-evaluating the model on these unseen instance. We have built model using training data set and further we will use the same model to predict the result. Following flowchart will depict the flow of this study.



(Fig: Methodology of the study)

V. RESULT

In this section, the performance analysis of logistic regression algorithms is carried out with a common dataset using WEKA tool. First, the model has loaded by using weka explorer and then we set the test data set i.e testing file(.arff) with unknown instances. Then classifiers were executed by including all the attributes (22) identified from the test data file(.arff) and then they were executed. The WEKA tool provides various measures to understand the classification. The value of actual class is unknown therefore '?' is used. As a result, it will give us either Positive or Negative value according to the prediction. The accuracy of the classifier depends on how well the classifier classifies the data set being tested. We have applied our model on 48 unknown instances and the result is as below.

ISSN 2394 - 7780

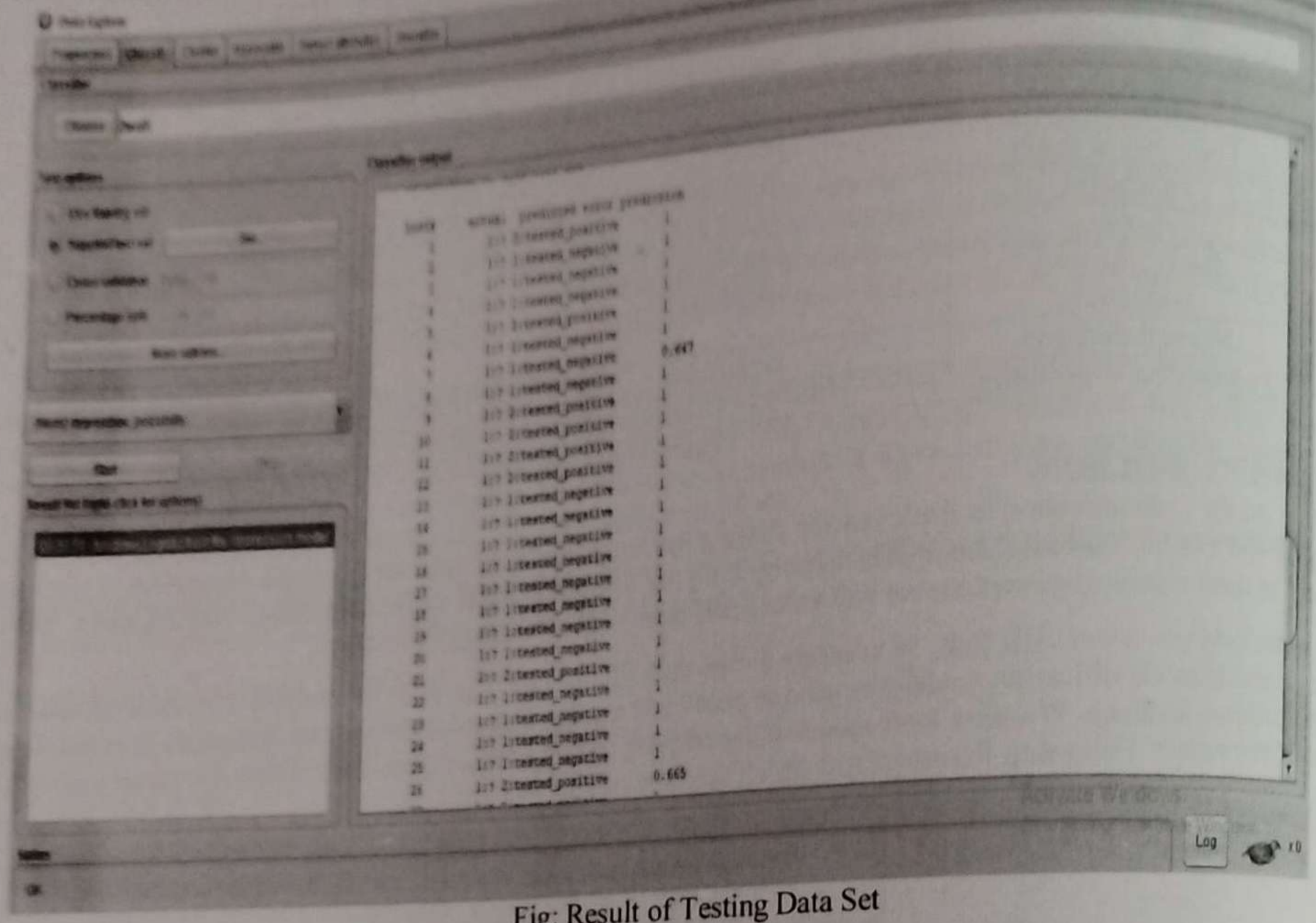


Fig: Result of Testing Data Set

VI. CONCLUSION

Now a days depression is rapidly growing illness. It may lead to serious diseases. So better to diagnose it as early as possible. But it is hard to predict because of number of symptoms. Data mining has various set of algorithms which can use to predict anything from past data. In this study we have used logistic regression algorithm to predict future depression using past data set. Which will predict, may the person will face depression in future or not. Logistic regression gives the accurate result for categorical outcomes than normal regression. The data set is very minimal and in future, the research may be applied for a large data set to obtain more accuracy. This model can be expanded to create a system to predict depression among people with more set of attributes.

REFERENCES

[1] https://www.archer-soft.com/en/blog/data-mining-healthcare

[2] https://www.quora.com/What-is-prediction-in-data-mining

[3] Anjume S, Amandeep K, Aijaz Ah M, Kulsum F Performance Analysis of Machine Learning Techniques to Predict Mental Health Disorders in Children- International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)

[4] Kevin Daimi, Shadi Banitaan, 2014 -Using Data Mining to Predict Possible Future Depression Cases

[5] Anxiety and depression association of America(US)-https://adaa.org/understanding-anxiety/depression/symptoms

[6] Ms. Sumathi M. R and Dr. B. Poorna, 2016 , Prediction of Mental Health Problems Among Children Using Machine Learning Techniques

[7] Oslon, D., Shi, Y., Kumar, V., "Introduction to Business Data Mining", McGraw Hill, 2007.

[8] http://blogs.fortishealthcare.com/mental-health-india-wake-up-call/