Using Data Mining to Predict Possible Future Depression Cases

Kevin Daimi, Shadi Banitaan
Computer Science and Software Engineering, University of Detroit Mercy, USA

Article Info

Article history:

Received Oct 28, 2014
Revised Nov 14, 2014
Accepted Nov 26, 2014

Keyword:

Classification
Data Mining
Depression
Diagnosis
Healthcare

ABSTRACT

Depression is a disorder characterized by misery and gloominess felt over a period of time. Some symptoms of depression overlap with other somatic illnesses implying considerable difficulty in diagnosing it. This paper contributes to its diagnosis through the application of data mining, namely classification, to predict patients who will most likely develop depression or are currently suffering from depression. Synthetic data was used for this study. To acquire the results, the popular suite of machine learning software, WEKA, was used.

Corresponding Author:

Kevin Daimi,
Department of Mathematics, Computer Science and Software Engineering,
University of Detroit Mercy,
4001 West McNicols Road, Detroit, MI 48334, USA.
Email: daimikj@udmercy.edu

1. INTRODUCTION

Data Mining is a multidisciplinary field that is based on various fields including database management systems, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization. In Data Mining, the extraction of implicit, previously unknown, and potentially useful information from data is dealt with [1]-[7].

Data mining applications in healthcare are constantly increasing and becoming more popular. Data mining can play a major role in healthcare allowing insurers uncover fraud and abuse, improving healthcare customer relationship management decisions, helping physicians identify effective treatments and best practices, identifying risk factors associated with the onset of diabetes, and enabling patients to receive better and more affordable healthcare services [8]. Healthcare data mining provides myriad opportunities for hidden pattern exploration from the huge healthcare data stores. These patterns can be used by physicians to establish diagnoses, prognoses and treatments for patients in healthcare organizations [9]. Wang et al [10] investigated the use of data mining in the healthcare industry. The enormous healthcare data are looked upon as one of the most challenging and most difficult of all data to work with. Suitable data mining practices offer the techniques and tools to transform the voluminous amounts of data into valuable information for decision making. Within healthcare, data mining can be employed to aid in discovering cures for current diseases, uncovering patterns for genetic diseases, and recognition of the causes of new diseases worldwide.

According to Obenshain [11], “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.” The techniques of data mining have a number of applications in healthcare. Agrawal et al [12] applied classification to analyze colon surgery data. They constructed risk prediction models for post-operative undesirable consequences in colon surgery using data mining techniques. Data mining was also applied to heart transplant data from the United Network for Organ Sharing (UNOS) program to predict risk of mortality within 1 year of heart transplant. The goal was to aid physicians in their decision making process by furnishing them with patient-specific risk assessments [13]. Tahsin et al [14] described a data mining application to develop systems for automated classification of drug pair mentioned in text into one of the following four classes: no interaction, advice, effect, mechanism and generic interaction.

Depression is a serious medical condition accompanied by a disruption in mood, deliberation, and body causing a person to feel very miserable, fruitless, and frequently lacking the ability to experience a normal live. The symptoms and impact of depression have been analyzed by a number of researchers. Biological symptoms of depression were studied by Matthew et al [15]. They studied the frequency of episodes of diverse biological symptoms in relation to the intensity of depression and neuroticism using 37 depressed patients. Studied individually, they observed that the superlative sign of depression severity and neuroticism were early waking up and excessive dreaming. They deployed stepwise multiple regression analyses of variance technique to find a group of biological symptoms predicting the severity of depression. The technique was not as successful in providing an insight on the severity of neuroticism. However, they further observed that neuroticism was a useful predictor of the biological symptoms when taken as a whole. Trivedi [16] investigated the role played by physical symptoms with regards to depression. The author stressed that unexplained aches and pains are repeatedly the pinpointing symptoms of depression. These symptoms include chronic joint pain, limb pain, back pain, gastrointestinal problems, tiredness, sleep disturbances, psychomotor activity changes, and appetite changes. A large percentage of patients, who suffer from depression, report only their physical symptoms. This can cause the diagnosis of depression to be a difficult task. The author emphasized that physical pain and depression exhibit stronger biological connection than simple cause and effect.

A number of studies treated the prediction of depression using various techniques. A study aimed at demonstrating how a risk prediction index would enable depression avoidance by pinpointing patients who would be most likely to benefit most from preventative procedures in primary care settings was carried out by Van Voorhees et al [17]. They adopted social and cognitive vulnerability and mood as baseline risk factors to predict onset of a depressive episode at 1-year follow-up. They relied on boosted classification and regression trees to develop a prediction index appropriate for a personal computer or hand-held device. De Choudhury et al [18] looked at the potential of using social media to identify and diagnose key depressive disorder in individuals. They first employed crowdsourcing to gather a set of Twitter users who reported being diagnosed with clinical depression, based on a standard psychometric vehicle. Using their social media postings over a year prior to the inception of depression, behavioral attributes relating to social engagement, emotion, language and linguistic styles, ego network, and indications of antidepressant medications were measured. These behavioral cues were the basis for building a statistical classifier that offered estimates of the risk of depression. A paper suggesting a statistical inference approach, named Negative Emotion Evaluation (NEE) Model, to explore the depression trend of web posts was introduced by Tung et al [19]. For this purpose, a Chinese forums post dataset was collected from PTT Prozac zone in Taiwan. Each post was classified and verified in terms of four depression tendency variables namely, negative emotion, triggering event, symptom, and negative thinking. Those were collected from the Diagnostic and Statistical Manual of Mental Disorder, Fourth Edition (DSM-IV-TR) based on the definition of major depressive episode.

Lee et al. [8] investigated the association between the chronic obstructive pulmonary disease (COPD) assessment test (CAT) and depression in COPD patients. Their results indicated that the CAT scores are significantly associated with the presence of depression and have good accuracy for predicting depression in COPD patients. In addition, among the eight items of the CAT, the energy score revealed the best correlation with the presence of depression. Fuller et al. [20] investigated the association between migraine and depression. They found that migraine is associated with higher odds of current depression among Canadians. They also found that those with depression were younger, unmarried, and poorer and had activity limitations. Further work on predicting depression could be found in [4],[21]-[24].

In this paper, a data mining application based on classification is proposed to predict who would be a possible candidate for developing depression. Synthetic data is used to train and test the classification model. Section 2 introduces the attributes used for this study. In Section 3, training and testing the model are presented. Section 4 deals with making predictions on new unseen data. Finally, discussion and conclusion are covered in Section 5. The well-known WEKA tool is adopted for this study.

2. ATTRIBUTES SELECTION

Attributes (symptoms in the case of depression) selection is one of the most important processes in data mining. This process involves selecting an effective subset of relevant attributes or features needed for constructing the data mining model. Rushing this process can result in possibly selecting redundant and unnecessary attributes, which could heavily impact the constructed model and the outcomes of the mining process.

This work relied on a number of online surveys and questionnaires including those presented in [25]-[27] to select the attributes needed for classifying depression. The selected set of attributes was further enlarged by adding more attributes from the above mentioned references on depression and predicting depression. After filtering out redundancies, the number of attributes in the selected set was 50. Following consultation with faculty at the College of Health professions, this set was reduced to 31 attributes including the class variable "May Have Depression." The final set of attributes is presented in Table 1 below.

Table 1. Attributes Set

Attribute	Values
Sadness	None: 0, Mild: 1, Medium: 2, Serious: 3
Discouragement	None: 0, Mild: 1, Medium: 2, Serious: 3
Low self-esteem	None: 0, Mild: 1, Medium: 2, Serious: 3
Inferiority	None: 0, Mild: 1, Medium: 2, Serious: 3
Guilt	None: 0, Mild: 1, Medium: 2, Serious: 3
Indecisiveness	None: 0, Mild: 1, Medium: 2, Serious: 3
Irritability and frustration	None: 0, Mild: 1, Medium: 2, Serious: 3
Loss of interest in life	None: 0, Mild: 1, Medium: 2, Serious: 3
loss of motivation	None: 0, Mild: 1, Medium: 2, Serious: 3
Poor self-image	None: 0, Mild: 1, Medium: 2, Serious: 3
Poor memory	None: 0, Mild: 1, Medium: 2, Serious: 3
Lose libido	None: 0, Mild: 1, Medium: 2, Serious: 3
Hypochondriasis	None: 0, Mild: 1, Medium: 2, Serious: 3
Suicidal impulse	None: 0, Mild: 1, Medium: 2, Serious: 3
Sluggish	None: 0, Mild: 1, Medium: 2, Serious: 3
Crying spells	None: 0, Mild: 1, Medium: 2, Serious: 3
Lack of emotional responsiveness	None: 0, Mild: 1, Medium: 2, Serious: 3
Helplessness	None: 0, Mild: 1, Medium: 2, Serious: 3
Pessimism	None: 0, Mild: 1, Medium: 2, Serious: 3
Agitation	None: 0, Mild: 1, Medium: 2, Serious: 3
Past failure	None: 0, Mild: 1, Medium: 2, Serious: 3
Reduced pain tolerance	None: 0, Mild: 1, Medium: 2, Serious: 3
Desire for Social Support	None: 0, Mild: 1, Medium: 2, Serious: 3
Psychomotor retardation	None: 0, Mild: 1, Medium: 2, Serious: 3
Confusion	None: 0, Mild: 1, Medium: 2, Serious: 3
Scatterbrained	None: 0, Mild: 1, Medium: 2, Serious: 3
Cognitive impairment	None: 0, Mild: 1, Medium: 2, Serious: 3
Loss warm feeling toward family or friends	None: 0, Mild: 1, Medium: 2, Serious: 3
Substance Abuse	None: 0, Mild: 1, Medium: 2, Serious: 3
Childhood trauma	None: 0, Mild: 1, Medium: 2, Serious: 3
May Have Depression	Yes, No

3. MODEL CREATION AND TESTING

In this study, classification is deployed for finding hidden patterns in data. To create the model, a classification algorithm needs to be applied. To acquire the classification model, the C4.5 decision tree algorithm is employed. WEKA implements a later and slightly improved version namely, C4.5 revision 8. This is referred to as J4.8. The results of implementing the depression classification model are obtained using J4.8. Splitting a dataset into training and testing sets is a central part of assessing data mining models. Normally, when a data set is divided into a training set and testing set, the highest portion of the data is used for training, and a smaller fraction of the data is used for testing. The J4.8 algorithm is trained using 600 instance dataset and tested with 400 instance dataset. Table 2 depicts a sample of the training data. The numbers in the header row represent the symtoms (attributes) mentioned in Table 1. The numbers in the first column imply rows 301 to 320 out of the 600 training data rows (depression cases).

Table 2. Sample training data

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
301	2	2	2	1	1	3	0	1	0	2	3	1	1	2	1	1	1	2	2	3	0	2	1	2	2	1	1	1	2	0	No
302	2	3	2	3	1	1	2	1	1	1	1	2	2	3	2	3	1	1	2	2	1	3	3	3	2	1	3	3	2	3	Yes
303	1	1	1	2	3	2	2	3	2	1	2	2	1	1	0	3	1	1	3	3	2	3	1	1	2	2	3	2	3	3	Yes
304	2	3	2	2	0	3	2	1	1	1	1	1	1	1	1	1	2	0	1	1	1	1	1	1	2	1	3	1	3	0	No
305	2	1	1	2	1	1	1	1	2	3	0	3	3	3	3	3	1	3	3	1	1	3	3	2	1	3	1	3	1	1	No
306	1	2	2	3	3	1	1	2	1	3	3	1	2	2	2	2	1	2	3	1	2	1	1	2	2	2	3	1	2	3	Yes
307	2	1	2	2	1	3	1	3	2	3	3	1	3	3	3	1	1	1	3	1	3	3	2	2	2	1	2	1	3	3	Yes
308	2	3	2	2	3	2	3	2	2	1	2	3	1	2	3	1	2	2	3	3	2	2	1	3	1	1	2	1	3	2	Yes
309	3	3	3	1	1	2	1	2	1	2	2	2	2	2	2	3	1	3	2	1	2	1	2	1	2	1	3	2	2	3	Yes
310	3	1	2	1	3	1	3	3	2	1	1	1	1	3	1	2	2	2	2	1	0	1	3	3	3	2	3	2	3	3	Yes
311	1	1	2	2	1	3	2	1	2	1	1	3	2	2	1	2	0	1	3	3	2	2	3	2	1	2	3	3	3	0	No
312	1	2	3	2	3	1	1	1	1	0	3	1	1	2	1	2	2	1	2	1	1	1	0	1	1	1	1	3	1	0	No
313	0	0	3	1	1	0	3	0	1	1	3	3	3	0	3	3	0	0	2	3	2	2	0	3	3	0	3	0	0	0	No
314	2	2	2	2	2	1	1	3	3	2	2	1	1	1	2	3	1	2	1	1	2	3	1	2	1	2	1	3	3	3	Yes
315	1	1	2	2	3	0	1	2	1	3	2	3	3	1	1	0	2	3	1	2	3	1	2	3	2	0	2	2	1	3	No
316	0	3	3	1	1	3	2	3	2	2	1	3	2	2	3	1	1	3	3	1	2	2	3	2	1	2	2	3	3	3	Yes
317	1	2	1	2	1	1	3	1	2	1	2	1	2	2	3	1	1	1	0	3	3	2	2	2	1	1	1	2	2	3	No
318	1	1	2	1	3	1	1	3	2	2	2	2	3	3	1	2	2	3	3	2	3	1	2	1	2	2	1	2	3	1	Yes
319	1	3	1	2	1	3	2	1	3	2	1	2	1	2	2	1	1	3	1	1	0	1	2	2	0	3	3	1	2	0	No
320	1	3	3	1	1	2	0	2	3	3	1	3	3	3	2	1	1	1	1	3	2	1	2	3	3	3	1	3	2	2	Yes

As mentioned above, synthetic data was used. This data was created using a Java program. The training of the J4.8 algorithm provided encouraging results based on the attributes subset (30 attributes) that were included after careful consideration and consultation with experts in the field of depression. The WEKA output of training J4.8 is summarized in Figure 1.

As can be observed in Figure 1, 555 instances were correctly classified and 45 instances were incorrectly classified. This resulted in 92.5% of the instances being correctly classified. There was a relative absolute error of 24.94%, and a root relative squared error of 49.94%. Figure 1 also shows the Confusion matrix below indicating the true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$\begin{bmatrix} \begin{array}{l} 263 \\ \text{(TP)} \end{array} & \begin{array}{l} 34 \\ \text{(FN)} \end{array} \\ \begin{array}{l} 11 \\ \text{(FP)} \end{array} & \begin{array}{l} 292 \\ \text{(TN)} \end{array} \end{bmatrix}$$

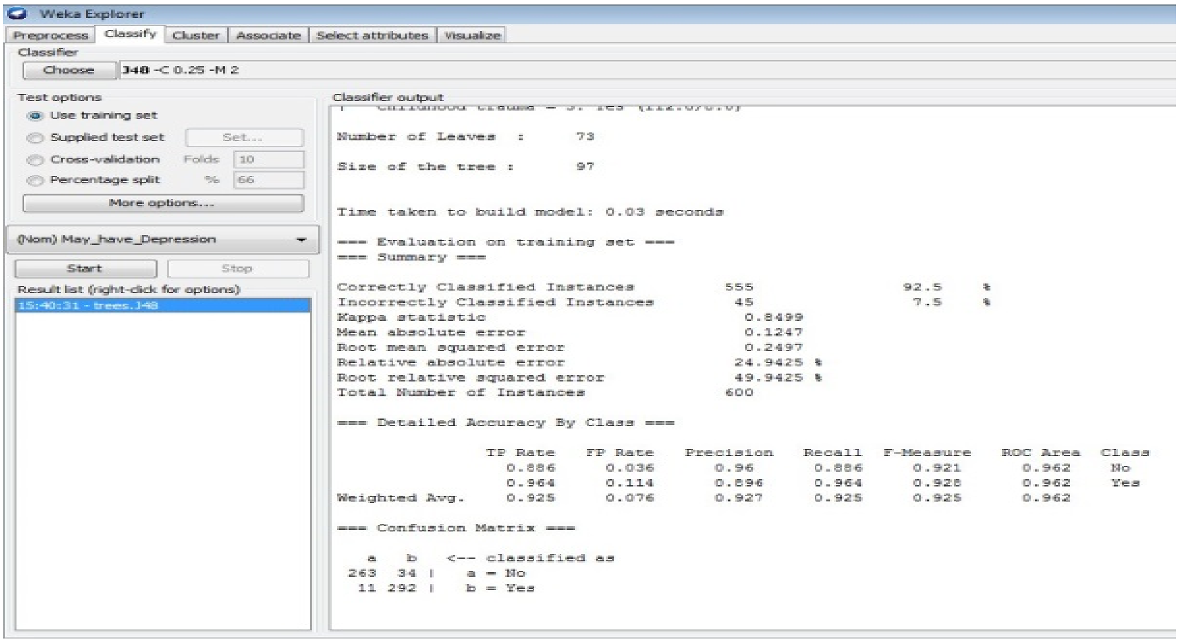
Several classification metrics were used for evaluation namely accuracy, precision, and recall. These metrics are defined as follows:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Based on the training data, we notice that accuracy = (263+292)/600 = 0.925, precision = 263/ (263+11) = 0.959,and recall=263/ (263+34) = 0.885.



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) May_have_Depression

15:40:31 - trees.J48

```
|   Childhood trauma = 5: Yes (112.0/0.0)

Number of Leaves  :     73

Size of the tree :      97


Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         555               92.5    %
Incorrectly Classified Instances        45                7.5    %
Kappa statistic                          0.8499
Mean absolute error                      0.1247
Root mean squared error                  0.2497
Relative absolute error                 24.9425 %
Root relative squared error             49.9425 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.886     0.036     0.96      0.886     0.921      0.962    No
                 0.964     0.114     0.896     0.964     0.928      0.962    Yes
Weighted Avg.    0.925     0.076     0.927     0.925     0.925      0.962

=== Confusion Matrix ===

   a   b   <-- classified as
 263  34 |   a = No
  11 292 |   b = Yes
```

Figure 1. Training outcome

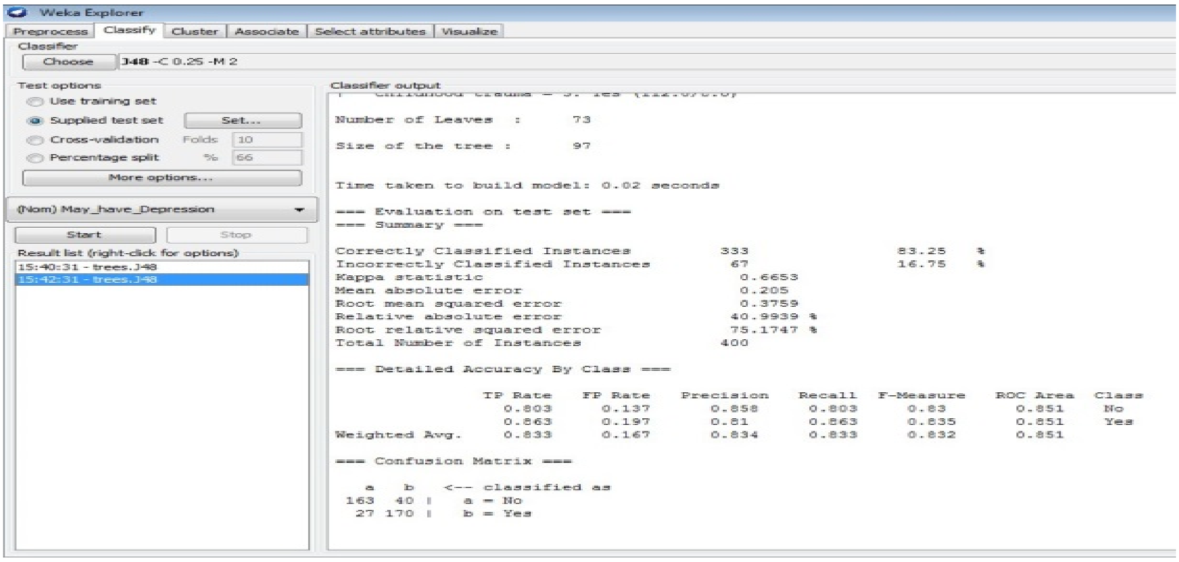
A test set is used to determine the accuracy (validation) of the model. The resulting model is applied to the testing instances. After completing the training phase satisfactorily, 400 rows (depression cases) of data were used to test the created model. Table 3 shows the partial testing instances (cases 101 to 120), and Figure 2 depicts the testing outcome.

Figure 2, illustrates that 333 instances were correctly classified and 67 instances were incorrectly classified. This led to the conclusion that 83.25% of the instances were correctly classified. The relative absolute error was 24.94%, and the root relative squared error was 49.94%. These relative error values aim to offset for the basic predictability or unpredictability of the class variable. The Confusion Matrix for testing is given below. The accuracy of the testing for the model is given by: accuracy = (163 + 170) / (163 + 170 + 40 + 27) = 0.833, precision = 163/(163+27) = 0.858, recall= 163/(163+40) = 0.803.

$$\begin{bmatrix} 163 & 40 \\ \text{(TP)} & \text{(FN)} \\ & \\ 27 & 170 \\ \text{(FP)} & \text{(TN)} \end{bmatrix}$$

Table 3. Sample testing data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
101	3	3	3	1	1	3	3	1	2	2	2	3	2	2	3	1	1	2	2	3	2	1	3	3	2	2	1	1	3	3	Yes
102	2	1	1	3	1	1	1	1	1	3	1	2	1	3	3	2	1	1	2	3	3	1	1	1	1	1	1	1	3	0	No
103	2	1	2	2	2	1	3	2	2	1	1	2	3	2	1	1	1	2	3	1	2	3	1	1	1	3	2	1	2	3	No
104	2	2	3	3	2	3	2	2	2	3	1	2	3	2	3	0	2	1	3	3	1	2	1	2	1	2	0	3	2	1	No
105	2	1	2	2	3	3	3	3	3	2	2	2	2	2	1	2	2	1	1	3	1	2	3	2	2	2	1	3	3	2	Yes
106	2	2	2	2	2	2	2	3	3	1	1	1	2	2	2	1	2	1	2	3	2	2	1	3	1	2	1	1	1	2	No
107	1	0	3	2	1	1	3	1	0	1	1	2	1	2	2	1	3	1	1	3	2	1	3	3	3	3	1	2	0	1	No
108	3	1	2	3	3	1	1	1	1	2	1	2	1	1	1	1	1	2	2	3	1	1	2	1	1	3	1	1	1	0	No
109	2	1	1	3	1	2	3	2	2	2	1	2	1	3	3	3	2	2	3	2	2	2	1	3	2	2	3	3	2	3	Yes
110	1	1	3	1	2	3	0	2	1	1	3	3	2	1	3	2	3	3	2	2	2	3	1	1	3	2	2	1	3	3	Yes
111	2	1	1	1	3	1	2	3	3	3	2	3	1	1	1	3	3	3	1	1	1	3	2	1	2	1	3	3	3	2	Yes
112	0	3	1	3	1	2	3	3	2	1	2	2	2	3	1	2	3	2	3	2	2	3	1	3	2	3	3	2	2	2	Yes
113	3	3	3	3	2	3	1	3	1	1	2	3	0	2	3	2	3	1	2	2	3	3	2	3	0	2	1	2	1	1	No
114	2	0	1	3	0	1	2	3	2	1	3	3	3	2	1	2	3	1	1	1	3	0	1	1	1	1	2	3	2	0	No
115	3	1	2	3	1	2	1	3	3	1	2	3	3	2	2	1	2	2	2	1	3	1	1	1	3	1	1	2	1	1	No
116	1	3	3	2	3	2	3	2	1	2	3	0	2	3	1	2	3	1	2	2	2	1	1	0	3	2	3	1	3	2	Yes
117	3	1	3	1	2	3	3	3	1	2	2	2	1	2	2	1	1	1	2	2	2	1	2	1	2	2	1	3	3	3	Yes
118	1	1	3	2	2	1	2	1	3	2	1	2	2	1	3	1	1	1	1	1	2	2	3	1	1	3	1	2	3	3	No
119	1	3	2	1	3	1	2	2	2	1	1	0	3	3	1	2	3	1	3	2	2	1	3	2	1	1	1	2	3	3	Yes
120	3	3	2	3	2	1	1	3	2	3	3	0	3	2	2	1	3	1	1	0	2	1	3	3	1	2	0	1	0	1	No



J48 -C 0.25 -M 2

Folds 10

% 66

(Nom) May_have_Depression

15:40:31 - trees.J48
15:42:31 - trees.J48

```
|   Childhood trauma = 3: Yes (112.0/0.0)

Number of Leaves  :      73

Size of the tree :       97


Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         333               83.25   %
Incorrectly Classified Instances        67               16.75   %
Kappa statistic                          0.6653
Mean absolute error                      0.205
Root mean squared error                  0.3759
Relative absolute error                 40.9939 %
Root relative squared error             75.1747 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.803     0.137      0.858     0.803     0.83       0.851    No
                 0.863     0.197      0.81      0.863     0.835      0.851    Yes
Weighted Avg.    0.833     0.167      0.834     0.833     0.832      0.851

=== Confusion Matrix ===

   a   b   <-- classified as
 163  40 |   a = No
  27 170 |   b = Yes
```

Figure 2. Testing outcome

4. MODEL USAGE

After training and testing, the created model should be used for classifying unknown instances provided that the accuracy of classification is adequate. The classification model should be capable of predicting unseen instances using the model it has learned. Certainly, it is desirable to re-train periodically using new training data. The depression classification model was used to predict 20 unseen instances through re-evaluating the model on these unseen instances. Table4 depicts these predictions. Out of the 20 instances (unkown depression cases), 13 were classified as “No” and 7 instances as “Yes.” Column 31 represents the diagnosis.

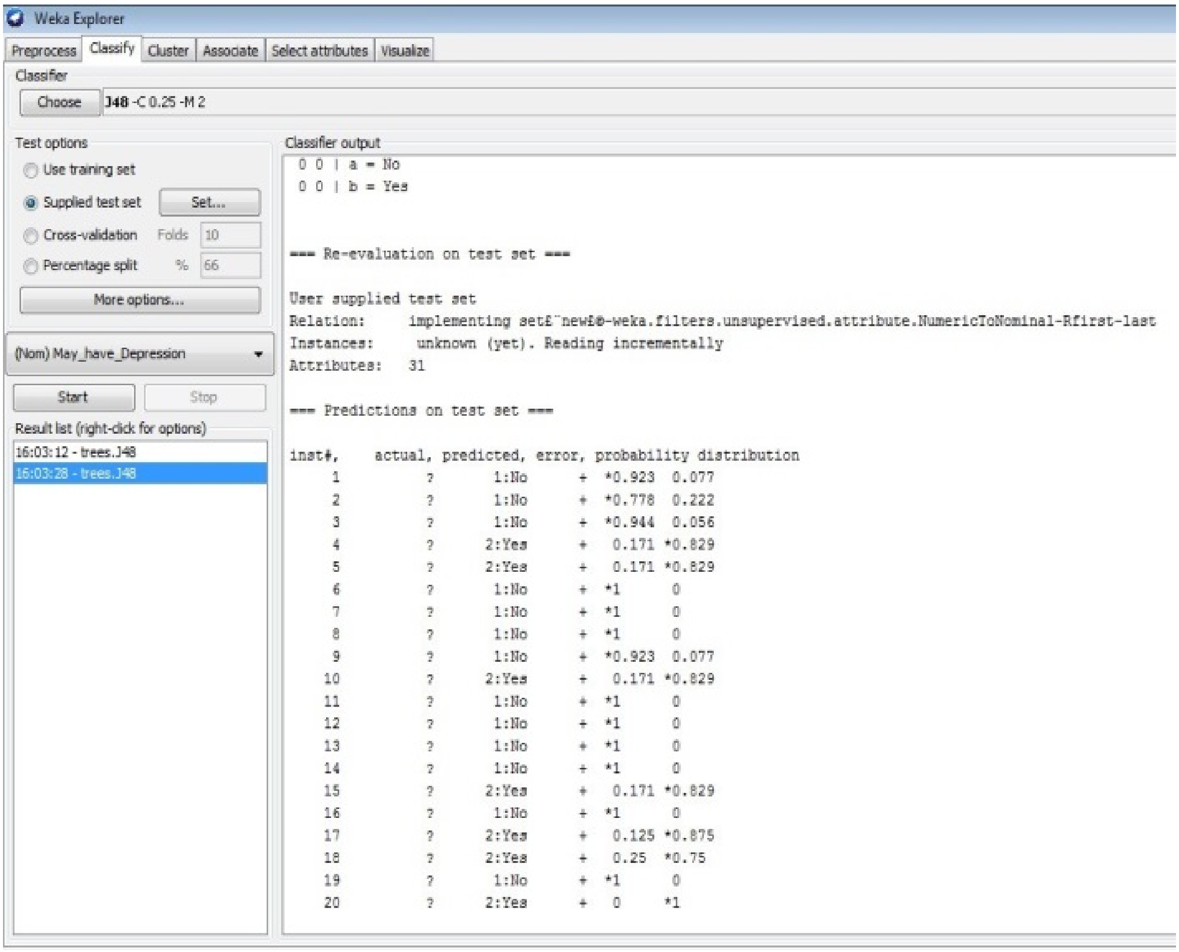
Another way of showing the results involves providing the probability distribution for the predictions. This is illustrated in Figure 3. The actual classes are unknown and therefore ‘?’ is displayed under the “actual” column. The “predicted” column contains the predictions (classe labels). The ‘+’ under the “error” column implies the actual and predicted classes are not the same. Since the actual is unknown (?), all the predicted classes do not match the actual classes. In other words, there are no errors because no actual classes exit or are known. On the other hand, if the ‘+’ appeared during the testing of the model, then the ‘+’ symbol is significant. There are two probability distribution columns. The first column is for class 1 (No), and the second for class 2 (Yes). The ‘*’ next to the probability distribution implies the correct class’s probability.

The correct class refers to the class shown under “predicted.” For each row, the probability distribution for the two classes sum up to 1. Taking row 1 as an example, it is noticeable that 0.923 and 0.077 add to 1. The given values indicate that class 1 was predicted with a probability of 0.923, and there is a very small probability, 0.077, to conclude it is class 2. Out of the twenty classes, ten were predicted with a probability of 1. This includes 9 No’s and 1 Yes. The smallest probability for predicting class 1 (No) is 0.778. For class 2 (yes), the smallest probability is 0.829. Therefore, the predictions made are trustworthy and reliable.

The above mentioned probability distributions are normally important if further analysis and research is needed by human. When dealing with depression, physicians will definitely pursue further examination prior to adopting the recommendation (prediction). Hence, these probability distributions are essential for medical applications of data mining and it would be appropriate to take them into consideration.

Table 4. Unkown cases with their diagnosis

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
1	2	1	2	2	3	2	1	2	2	2	2	1	1	2	1	3	0	2	2	2	3	1	1	2	1	1	2	1	1	1	No
2	3	2	1	2	1	0	1	3	2	2	1	1	3	2	3	2	3	0	1	1	2	1	1	2	2	2	2	1	1	2	No
3	2	3	2	3	1	3	1	3	2	2	3	1	1	0	1	1	1	1	1	2	1	2	1	1	1	3	3	1	3	0	No
4	1	1	2	3	3	3	1	3	1	1	3	2	3	1	1	1	3	2	1	2	2	2	3	3	2	3	3	2	3	2	Yes
5	2	1	0	2	2	2	2	3	1	2	1	2	2	1	1	1	3	0	1	3	2	1	3	1	3	1	2	1	3	2	Yes
6	0	2	2	1	1	3	3	2	1	2	2	2	2	1	2	3	1	2	1	2	1	1	1	2	1	1	2	2	2	1	No
7	1	3	1	3	3	2	1	1	2	3	2	1	2	2	1	2	1	0	2	2	1	1	1	1	2	1	2	2	2	2	No
8	2	2	0	2	1	0	2	2	2	2	3	1	3	3	2	2	1	3	3	1	1	3	1	2	1	3	1	1	0	1	No
9	3	2	1	3	2	1	3	1	3	1	1	2	1	2	1	2	2	2	1	3	2	3	3	3	3	2	3	2	1	1	No
10	1	1	2	2	1	2	1	3	3	2	2	3	0	3	1	1	1	3	1	3	2	3	1	2	2	2	1	2	3	2	Yes
11	3	2	3	3	2	3	1	1	3	2	1	3	3	2	2	1	1	2	1	3	1	3	2	3	2	1	1	1	2	1	No
12	1	1	2	1	2	0	1	1	0	2	3	0	1	0	1	2	0	3	1	2	0	3	3	2	2	3	0	3	0	0	No
13	2	2	1	2	2	1	3	0	2	1	1	3	3	2	3	3	2	2	2	3	2	1	3	1	2	3	1	1	0	3	No
14	3	1	1	2	3	2	1	2	3	2	3	3	3	1	2	1	1	1	2	3	2	2	2	1	1	2	3	3	0	2	No
15	1	1	2	1	2	1	2	2	3	3	2	1	2	1	1	1	1	3	2	3	0	3	1	2	1	2	3	3	3	2	Yes
16	3	2	2	3	3	1	3	3	2	3	2	2	3	2	2	2	3	1	2	3	1	1	1	2	1	2	1	2	1	2	No
17	0	1	1	1	0	3	3	2	1	3	1	1	3	1	2	3	1	2	3	1	1	1	2	2	3	1	0	2	2	2	Yes
18	3	1	1	1	3	1	2	2	3	2	1	3	1	1	1	3	2	1	1	2	3	1	3	3	2	2	3	2	2	3	Yes
19	1	0	3	3	3	0	2	0	3	0	1	3	3	2	1	3	0	0	2	2	2	2	2	1	2	1	3	0	0	0	No
20	1	2	2	2	3	3	2	3	3	1	1	2	1	1	2	2	3	3	1	1	2	3	1	3	2	2	2	1	2	3	Yes



Classifier: Choose J48 -C 0.25 -M 2

Test options
○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) May_have_Depression

Result list (right-click for options)
16:03:12 - trees.J48
16:03:28 - trees.J48

Classifier output

```
 0 0 | a = No
 0 0 | b = Yes



=== Re-evaluation on test set ===

User supplied test set
Relation:     implementing set£¨new£©-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:     unknown (yet). Reading incrementally
Attributes:   31

=== Predictions on test set ===

inst#,    actual, predicted, error, probability distribution
     1         ?      1:No      +  *0.923  0.077
     2         ?      1:No      +  *0.778  0.222
     3         ?      1:No      +  *0.944  0.056
     4         ?     2:Yes      +   0.171 *0.829
     5         ?     2:Yes      +   0.171 *0.829
     6         ?      1:No      +  *1      0
     7         ?      1:No      +  *1      0
     8         ?      1:No      +  *1      0
     9         ?      1:No      +  *0.923  0.077
    10         ?     2:Yes      +   0.171 *0.829
    11         ?      1:No      +  *1      0
    12         ?      1:No      +  *1      0
    13         ?      1:No      +  *1      0
    14         ?      1:No      +  *1      0
    15         ?     2:Yes      +   0.171 *0.829
    16         ?      1:No      +  *1      0
    17         ?     2:Yes      +   0.125 *0.875
    18         ?     2:Yes      +   0.25  *0.75
    19         ?      1:No      +  *1      0
    20         ?     2:Yes      +   0     *1
```

Figure 3. WEKA output on unseen instances

5. RESULTS INTERPRETATION

To interpret the results, Table 4 will beutilized. This table contains all the data used for the 20 unkown cases that need to be diagnosed. Each row or instance represents a case to be diagnosed. The numbers in the firstcolumn represent the case number. The numbers in the header row represent the symptoms (attributes) as mentioned in Table 1. To interpret the diagnosis of cases (rows) 5 and 8, IF-THEN rules (If conditions then conclusion) will be employed. Any symptom that has a value of '0' will not appear in the conditions of the rules. Table 1 in section 2 indicates that the value '0' stands for "None." Table 4 matches the outcomes of Figure 3 and was generated by the WEKA system. The following are two examples of using the prediction model.

Row 5 (case# 5)

IF

Sadness is medium & Discouragement is mild & Inferiority is medium & Guilt is medium & Indecisiveness is medium & Irritability and frustration is medium & Loss of interest in life is serious & loss of motivation is mild & Poor self-image is medium & Poor memory is mild & Lose libido is medium & Hypochondriasis is medium & Suicidal impulse is mild & Sluggish is mild & Crying spells is mild & Lack of emotional responsiveness is serious & Pessimism is mild & Agitation is serious & Past failure is medium & Reduced pain tolerance is mild & Desire for Social Support is serious & Psychomotor retardation is mild & Confusion is serious & Scatterbrained is mild & Cognitive impairment is medium & Loss warm feeling toward family or friends is mild & Substance Abuse is serious & Childhood trauma is medium

THEN*patient will develop depression* (**Yes**)

Row 8 (case# 8)
IF

Sadness is medium & Discouragement is medium & Inferiority is medium & Guilt is mild & Irritability and frustration is medium & Loss of interest in life is medium & loss of motivation is medium & Poor self-image is medium & Poor memory is serious & Lose libido is mild & Hypochondriasis is serious & Suicidal impulse is serious & Sluggish is medium & Crying spells is medium & Lack of emotional responsiveness is mild & Helplessness is serious & Pessimism is serious & Agitation is mild & Past failure is mild & Reduced pain tolerance is serious & Desire for Social Support is mild & Psychomotor retardation is medium & Confusion is mild & Scatterbrained is serious & Cognitive impairment is mild & Loss warm feeling toward family or friends is mild & Childhood trauma is mild

THEN*patient will not develop depression* (**No**)

6. CONCLUSION

Depression is an exponentially growing medical illness. It is hard to diagnose depression due to a number of its symptoms being shared with other somatic illness. In this paper, a large set of attributes (symptoms) were selected based on surveys and interviews with experts in the field of depression. Some of these attributes overlap with various somatic illnesses. However, taken together, the adopted attribute set is sufficient to isolate depression from other illnesses. Synthetic data was used to train and test the classification model. As can be observed in the figures above, the outcomes for the synthetic datasets were reasonable in terms of accuracy, precision, and recall of the training and testing processes.

The above depression classification application will be further improved in the future. First, the selected attributes will be further discussed with more experts in the field to derive the most effectual attributes set. Having done that, a survey will be created. The real data will be used to train and test the model. Later, the model will be applied to unseen instances and the outcomes will be compared with the outcomes that were obtained using the synthetic data.

ACKNOWLEDGMENTS

We would like to thank Dr. Carla Groh, professor of nursing at the McAuley School of Nursing, University of Detroit Mercy, for her valuable comments and suggestions regarding the attributes (symptoms) used in this study.

REFERENCES

[1] Dunham MH., “Data Mining: Introductory and Advanced Topics”, Prentice Hall, 2003.

[2] Shapiro, G., Smyth, P., “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, Vol.17, pp. 37-54, 1996.

[3] Han, J., Kamber, M., “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2006.

[4] Levin, HS., McCauley, SR., Josic, CP., Boake, C., Brown, SA., Goodman, HS., Merritt, SG., Brundage, SI., “Predicting Depression Following Mild Traumatic Brain Injury,” *Archives of General Psychiatry*, vol/issue: 62(5), pp. 523-528, 2005.

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

[6] Parthasarathy, S., “Data Mining at the Crossroads: Successes, Failures and Learning From Them”, The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, pp. 1053-1055, 2007.

[7] Tan, P., Steinbach, M., V. Kumar, “Introduction to Data Mining”, Addison-Wesley, 2006.

[8] Koh, H., Tan, G., “Data Mining Applications in Healthcare,” *Journal of Healthcare Information Management*, vol/issue: 19(2), pp. 64-72, 2005.

[9] Milovic, B., Milovic, M., ”Prediction and Decision Making in Health Care using Data Mining*”, International Journal of Public Health Science (IJPHS)*, vol/issue: 1(2), pp. 69-78, 2012.

[10] Wang, J., Zhou, Z., Yan, R., "Benefits and Barriers in Mining the Healthcare Industry Data", *International Journal of Strategic Decision Sciences (IJSDS)*, vol/issue: 3(4), pp. 51-67, 2012.

[11] Obenshain, MK., “Application of Data Mining Techniques to Healthcare Data”, *Infection Control and Hospital Epidemiology*, vol/issue: 25(8), pp. 690-695, 2004.

[12] Fayyad, U., Piatetsky-Agrawal, A., Al-Bahrani, R., Merkow, R., Bilimoria, K., Choudhary, A., “Colon Surgery Outcome Prediction Using ACS NSQIP Data”, KDD Workshop on Data Mining for Healthcare (DMH), Chicago, IL, Aug. 2013.

[13] Agrawal, A., Russo, M., Raman, J., Choudhary, A., “Heart Transplant Outcome Prediction using UNOS Data”, KDD Workshop on Data Mining for Healthcare (DMH), Chicago, IL, Aug. 2013.

[14] Tahsin, T., Emadzadeh, E., Gonzalez, G., “Automated Extraction and Classification of Drug-Drug Interactions from Text”, in *KDD Workshop on Data Mining for Healthcare (DMH)*, Chicago, IL, Aug. 2013.

[15] Mathew, RJ., Largen, J., Claghorn, JL., “Biological Symptoms of Depression”, *Psychosomatic Medicine*, vol/issue: 41(6), pp. 439-443, 1979.

[16] Trivedi, MH., “The Link between Depression and Physical Symptoms”, *The Primary Care Companion to the Journal of Clinical Psychiatry*, vol/issue: 6(1), pp. 12–16, 2004.

[17] Van Voorhees, BW., Paunesku, D., Gollan, J., Kuwabara, S., Reinecke, M., Basu, A., “Predicting Future Risk of Depressive Episode in Adolescents: The Chicago Adolescent Depression Risk Assessment (CADRA)”, *Annals of Family Medicine*, vol/issue: 6(6), pp 503-511, 2008.

[18] De Choudhury, M., Gamon M., Counts S., Horvitz, E., “Predicting Depression via Social Media”, The 7th International AAAI Conference on Weblogs and Social Media, Boston, Massachusetts, 2013.

[19] Tung, C., Lu, W., “Predict Depression Tendency of Web Posts using Negative Emotion Evaluation Model”, ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012), Beijing, China, 2012.

[20] Fuller-Thomson, Esme, Meghan Schrumm, Sarah Brennenstuhl, "Migraine and despair: factors associated with depression and suicidal ideation among Canadian migraineurs in a population-based study", Depression research and treatment, 2013.

[21] Abdel-Khalek, AM., “Can Somatic Symptoms Predict Depression?”, *International Journal ofSocial Behavior and Personality*, vol/issue: 32(7), pp. 657-666, 2004.

[22] Cloninger, CR., Svrakic, DM., Przybeck, TR., “Can Personality Assessment Predict Future Depression? A Twelve-Month Follow-Up of 631 Subjects”, *Journal of Affective Disorders*, vol. 92, pp. 35-44, 2006.

[23] Robinson, MS., Alloy, L.B., “Negative Cognitive Styles and Stress-Reactive Rumination Interact to Predict Depression: A Prospective Study”, *Cognitive Therapy and Research*, vol/issue: 27(3), pp. 275-291, 2003.

[24] Rude, SS., Valdez, CR., Odom, S., Ebrahimi, A., “Negative Cognitive Biases Predict Subsequent Depression”, *Cognitive Therapy and Research*, vol/issue: 27(4), pp. 415-429, 2003.

[25] Beck Depression Inventory, Mood/Depression Assessment Questionnaire, Available: http://www.ibogaine.desk.nl/graphics/3639b1c_23.pdf.

[26] Burns Depression Checklist, University Health services, University of California, Berkeley, 2010, Available: http://uhs.berkeley.edu/home/healthtopics/PDF%20Handouts/Depression%20Check%20List.pdf.

[27] Surveys of Adult U.S. Women and Doctors Gauge Perceptions about Depression through Hormonal Transitions, Society for Women Health research, 2007, Available: http://www.womenshealthresearch.org/site/DocServer/DepressionSurveyAnalysis.pdf?docID=1801.

BIOGRAPHIES OF AUTHORS

Kevin Daimi is a full professor and director of Computer Science and Software Engineering programs at the University of Detroit Mercy, USA. He joined the University of Detroit Mercy in 1998 after working in industry for a number of years. Kevin received a Master of Science in Applied Computation (1980) and a Ph.D. in Computational Optimal Control (1983) from University of Cranfield, England. He is a fellow of the British Computer Society (BCS), a senior member of the Association for Computing Machinery (ACM), a senior member of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the IEEE Computer Society. His research interests include computer and network security, software engineering, data mining, and computer science and software engineering education.

Shadi Banitaan is currently an assistant professor at the Mathematics, Computer Science, and Software Engineering department at the University of Detroit Mercy. He teaches classes in Software Engineering and Computer Science. His research interests include software engineering and data mining. He is a member of the Association for Computing Machinery (ACM), a member of the Institute of Electrical and Electronic Engineers (IEEE), and a member of the IEEE Computer Society. He received a B.S. degree in Computer Science from Yarmouk University, an M.S. degree in Computer and Information Sciences from Yarmouk University, and a Ph.D. degree in Computer Science from North Dakota State University. He taught for five years at the University of Nizwa, Oman. He joined the University of Detroit Mercy in 2013.