AL- Rafidain
University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain
University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS
Journal of AL-Rafidain
University College for Sciences

Predicting Mental Disorders Using Data Mining Techniques

Lamyaa N. Khaleel	Karim H. Al-Saedi
lamyaa.naji@uomustansiriyah.edu.iq	karimnav6@gmail.com
Department of Computer Science - Al-Mustansiriyah University, Baghdad, Iraq	Department of Computer Science - Al-Mustansiriyah University, Baghdad, Iraq

Article Information

Article History:
Received: April, 11, 2021
Accepted: July, 4, 2021
Available Online: June, 25, 2022

Keywords:
Data mining, mental disorders, Naive Bayes, K-nearest neighbor classifier.

Correspondence:
Lamyaa N. Khaleel
lamyaa.naji@uomustansiriyah.edu.iq

Abstract

Over the last few years, data mining makes a transition to new applications in medicine. Many people suffer from mental disorders because of absence of care in the health care sector, especially in the mental care. Lack of health records for patients of mental health, has been caused by many reasons, for example, patients' hesitation to go the psychiatrist to follow-up their condition due high cost of examination fee, waiting time for their turn, and diagnosis inaccuracy. Classifiers are used to predict normal, addiction without disorder, addiction with disorder, or disorder without addiction patients' cases. However, a rise in the cases of personality disorder and substance abuse. We used Naive Bayes and K-Nearest Neighbor classification technique on dataset consists of opioid addicts patients derived from the United States of America to obtain high accuracy of mental disorder prediction. The precision of the NB classifier is (87%), and the precision of KNN classifier is (92%). This paper presents a new method for diagnosing mental disorders through data mining.

DOI: https://doi.org/10.55562/jrucs.v51i1.535

1. Introduction

1.1. preamble

Personality disorder is a type of mental disorder Whereas a person suffers of rigid and unhealthy pattern of thinking, functioning, A person with a personality disorder may faces a difficulties in understanding and dealing with situations and people. In which they may blame other people for their challenges. Personality disorders starts at the time of adolescent years or early adulthood. Kinds of personality disorders as follows: Few Kinds that less apparently in the middle ages[1].

A mental disorder is characterized by a mixture of thoughts, perceptions, emotions, abnormal behaviors, relationships with others, and psychological dysfunction. Usually, mental conditions are associated with extreme conditions such as depression or impairment in social, professional, or other essential tasks, expected or culturally acceptable response to shared stress or loss. The loved one loss , is not a mental illness, nor psychological disturbances as a result of an aberration, nor conflict as a result of product of Individual dysfunction [2].

Over the past few decades, mathematical sciences and analytical approaches found new uses in fields like medicine. Modern data collection and data processing protocols have been of great assistance to medical researchers and clinical scientists. New computational methods have been made available, especially in psychiatry, technology, and science, to support predictive modeling and recognize diseases more accurately. Data mining (or knowledge discovery) aims to extract information from large datasets and solve difficult tasks, such as patient evaluation, diagnosis of early mental illness, and evaluation of drug efficacy. Accurate and fast data analysis methods are very important, especially when working with severe psychological diseases such as personality disorders. We will concentrate on computer techniques relating to data processing and, more precisely, data mining. They are predicting valuable data from nominal and numerical data. Machine learning is a set of artificial intelligence computational algorithms that learn from the experience of specific tasks and data to enhance their performance—conceived with methodologies of machine learning. There are three types of machine learning: supervised, unattended, and strengthened learning [3]. The main idea is to combine results of algorithms for Data Mining to improve the quality of the affecting rules. The objective is to develop a solution to help experts make decisions, which can superimpose the results of the algorithms.

1.2. Related Works

Mahdi Mohammadi et al. [4] attempt to use a data extraction technique to identify the EEG of patients with MDD and VHS. It includes (A) Data pre-processing and application of linear discriminant analysis (LDA) for mapping characteristics in a new space of features through the application of genetic algorithms (GAs) for the classification of essential characteristics. (B) Building predictive models using a decision tree (DT) algorithm to classify hidden patterns and rules according to the micro and mapped characteristics. (C) Test samples based on false-positive and accuracy values based upon the MDD and HV EEG results for participants to use a decision tree and genetic algorithm to eliminate outcomes for depressed patients. This research performs the analysis and classification of the EEG signals (i.e., electroencephalogram) from 100 patients to 2 classes, like the healthy volunteers and major depressive disorders. A GA utilizes for the identification of the influencing traits and a DT to generate a predictive model. Data analysis executes by Matlab by using feature selection and machine learning methods.

Research to predict early-stage depression in patients suggests by Daimi et al. [5]. In this research, a data-set includes 1,000 patients with 31 chosen feature sets utilize. J48 algorithm performs with the use of the WEKA tool for the classification of the data set. (symptoms) according to the interviews and surveys conducted with the experts in the depression area, these are selected. Some of those traits are overlapping with a variety of physical ailments. Collectively, however, the set of approved features is adequate for the isolation of the depression from the other illness types. Synthetic data has been utilized for training and classification testing Models. Results for synthetic data sets have been reasonable in conditions for accuracy, call-to-train, and test operations. The system displays a predictive accuracy rate of 83.3 %.

Dipnall et al. [6] use a hybrid approach to classify key depression-related biomarkers using the ML algorithm of boosted regression and logistic regression. The authors use a data-set of 5230 samples from the National Health and Nutrition Review (2009-2010). The study chooses three bio-markers: Glucose, Serum, and Red Cell distribution width and total bilirubin out of 20 related to depression. Following the creation of 20 imputation datasets from several chained sequences of regression, ML boosted regression has initially identified 21 bio-markers related to depression.

Ben Youssef et al. [7] performs diagnostic research on Alzheimer's patients. The authors used three separate approaches of data mining, like the DT, discrimination analysis, and logistic regression, and conclude that the results of the classification that have been obtained from the discriminant analysis are superior to products of the other two methods. The highest predictive accuracy rate reaches by discriminant analysis up to 66%.

Warda et al. [8]. Use RFs for the prediction of drug abuse and several mental disorders. There are two data sets utilized, the first data-set includes substance abuse patients derived from the U.S., and the other data-set includes psychotic patients that are obtained from Egypt. RF classification method are utilized to increase the accuracy of the systems of mental disorders prediction. The analysis of the exploratory data mining created 2 RF models with two patient groups at a variety of the risks for psychotic diseases and substance abuse. This research presents a new method of data mining for the enhancement of mental disorder diagnoses. This method also serves as a therapeutic intervention.

1.3. Naïve Bayes Algorithm

Naïve Bayes (NB) can be considered a good learning algorithm in machine training and data mining. It is an easy probabilistic gradation that takes advantage of Bayes' theory, which works according to conditions or can be seen as a calculation method for posterior probability. The classification here could be called after Reverend Thomas Bayes, who presented the Bayes theory and was initially used to classify texts by Wallace and Mosteller. "Naïve" can be considered as if all the features of a document are separate. In other words, the theory of Bayes is an arithmetic formula that uses knowledge from previous events to predict future events and collect values in past information on a conditioned probability basis. The conditional probability may be regarded as likely to occur since another occurred [9].

Following the conditional probability definition:

P (B/A) = P (A and B)/P (A) As B: Dependent event, A: The last event. It makes the model built on training data, searching for novel data to what class label it follows via calculating probability.

Let D be a training group of tuples. An n-dimensional features vector can exemplify every tuple, X = (f_1,f_2,…, f_n), and every associated class label of theirs are of two classes (C_1, C_2). The rule of Bayes is as the posterior probability of X being in class C is given as:

$$P\ (C|X) = \frac{P(X|C)p(C)}{p(X)}$$

Where:

P (C|X): Posterior probability of class (c, target) given predictor (x, attributes).

P (C): Prior probability of class.

P (X|C): Likelihood or posterior probability of X conditioned on Ci.

P (X): Prior probability of X

1.4. K-Nearest Neighbor's Algorithm

It could be defined as follows: A method for classifying objects in feature space based on the nearest training specimen. The K-nearest neighbor algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve classification and regression problems. Every training specimen is a vector with a class label. The training step consists of simply saving each feature vector with its training specimen labels. In the classifying step, for discovering the specimen of test for the class's class, it is related. Step one calculates the distance for each training specimen individually. After that, maintain the k (k: positive whole number) nearest training specimen, secondly via a majority vote of its neighbors, assign the most superior class for the specimen of tests [10]. Relying on the dataset, the most successful selection of k can be specified. K's value is evaluated as:

$$k = sqrt(N)/2$$

Where:
N: number of the specimen in the training dataset of yours.
Nevertheless, the more superior way can be trying several K values and find out which one of them could give the most successful result [11]. There are many distance metrics utilized in this algorithm; the most successful ones are Euclidean distance [12] and;

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Where:
d= Euclidean distance
n= number of dimensions (2 for our data)
X= data point from the dataset
Y=new data point to be predicated

1.5. Performance Evaluation

The efficiency of any machine learning model is determined by using metrics such as accuracy, recall, and measurement F. These metrics are used to evaluate the performance of machine learning models in comparison to human judgments [13].

1.5.1. Precision: represents the number of true positives that are divided by the number of true positives and the number of false positives, which are instances the model precisely labels as positive, which are in fact negative, or in our example, individuals the model classifies as terrorists that are not[14]

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

Where:
TP: True Positives
FP: False Positives

1.5.2. 1.5.2 Recall The capability of finding every relevant instance in a data set, precision represents the ratio of the data points this model states is
in fact, relevant [15].

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

Where:
TP: True Positives
FN: False Negatives

True Positive Rate: it is a measure of the proportion of real correctly recognized negatives
True Negative Rate: Measures the proportion of the real correctly recognized positives.

1.5.3. F-measure: The F (or F-measure) is both precisions, and the test is called to calculate a grade. The accuracy of p is the number of positive results divided by the number of all positive outcomes, and the retrieval r is the number of positive results divided by the number of positive outcomes which are supposed to be returned. The traditional or balanced FF (mean F1) is the harmonic average of accuracy and recall, with F1 being the optimal value at one and worse at zero. The general formula contains a real positive β so that F-score measures the effectiveness of recovery for the user who hangs β importance times remember as accuracy [16]

$$F = \frac{2 * (\text{Precision} \; * \; \text{Recall})}{(\text{Precision} \; + \; \text{Recall})}$$

2. Research Objectives

The objective of this proposal is to present a

1. Designing and building a system to order medical care predicts when a person is "suffering from personality disorders" and determines the severity of the disorder. Tracking the patient's health record by building a database that stores the patient's details and any test he conducts.
2. Give the patient anonymity by encrypting the patient's data or not recognizing that others see a psychiatrist.
3. Detection of flicker at the lowest cost, time, and accuracy.
4. Knowing the number of troubled people in the country every year.

3. Data Set

Treatment Loop Data Set - Admissions Processes (TEDS-A) can be defined as a national annual admission data system for substance abuse treatment facilities. Then, the states reported that data from the state's administrative systems to SAMHSA. The resulting data system is known as TEDS-A. For this reason, TEDS does not cover all admissions for drug abuse treatment .. TEDS-A includes records of admissions ages 12 and over, and contains information on admission demographics (age, race / ethnicity, gender, employment status, etc. And drug use characteristics (age at first use, substances used, course of use, number of previous admissions, frequency of use, etc.). TEDS-A records admissions represented rather than individuals, as a person may be accepted for treatment more than once. The sibling data system, referred to as the Treatment-Discharge Loop Data Set (TEDS-D), obtains data on leaks from drug abuse treatment facilities. Consisting of 2,005,395 cases[17].

3.1. Data Pre-Processing

Normalization is a technique of scaling, mapping or pre-processing. Where can we find an existing domain. For predicting or for predicting many things it can be useful. As we know, there are several predictive or predictive methods, but all of them can differ a lot. A normalization technique must therefore be approximated to maintain the great difference in prediction and prediction. As shown in the equation, according to the Min-Max normalization technique[18].

$$Xi,_{0 \text{ to } 1} = \frac{Xi - XMin}{XMax - XMin}$$

Where:
Xi= Each data point
XMin = The minima among all the data points
XMax = The maxima among all the data points
Xi, 0 to 1= The data point i normalization between 0 and 1

The technique which gives the normalized values or range of data from the original unstructured data using the concepts like mean and standard deviation.

4. Proposed System

This section provides work methods and designs for the development of a predictive system for mental disorders. The system is a detected application for addiction and disorder that is used. In order to learn the system, we used machine learner algorithms, like Naive Bayes and the K Nearest Neighbour.is illustrated in the block diagram as shown in Figure 1. The first step in construct prediction system was evaluated two machine learning algorithms these are (NB, KNN) and then select KNN algorithm that achieve high accurse and lowest error as the classifier for the proposed system.

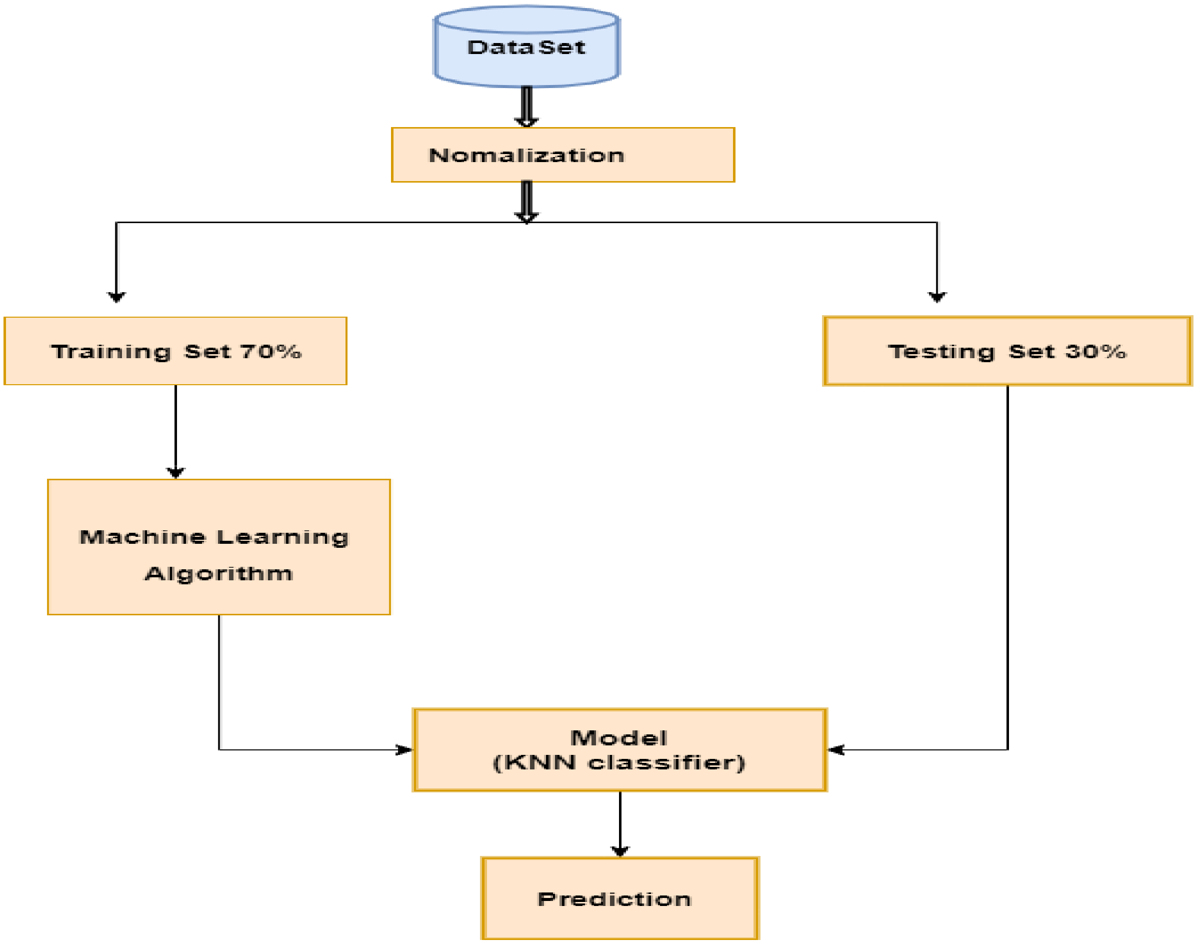

Figure (1) : Block diagram of Machine Learning Classifier

The design of any system is very important because it shows how the system works and explains the exact and practical steps that will be carried out to obtain the required need from it. The system consists of two parts: The first part is the client part responsible for collecting patient data and sending it to the server after it is encrypted by the Advanced Encryption Standard (AES) algorithm. While the second part is the server part in which the data received from the client side is decoded and the patient's status is classified if he suffers from personality disorders and its severity is done using KNN classifier. Then the prediction result is sent by message to the client part. Also, the server stores patient information and test results in a database to track the patient's health history. Moreover, Transmission

Control Protocol (TCP) has been used to ensure that data is delivered over the network as shown in Figure (2)

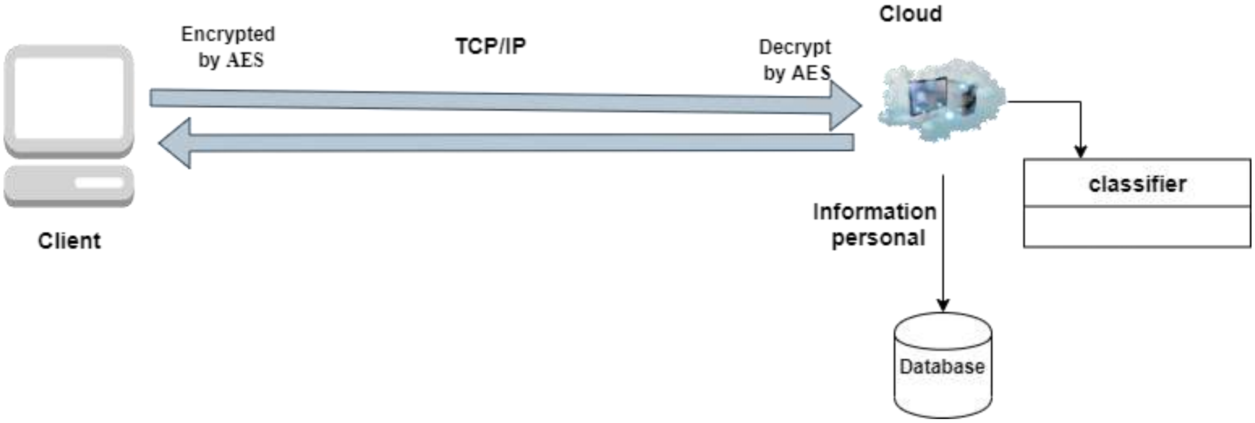

Figure (2) : Shows how the proposed system works

By comparing our results with "A Random Forest Model for Mental Disorders Diagnostic Systems" [8]Using the random forest classifier on the same data set, the result shows that the KNN classifier is more precise than the rest of the classifiers, as shown in the table.

Algorithm Type	Precision measure	Recall measure	F- measure
Random Forest (Warda et al.,2017)	87.72 %	88.87 %	88.29%
NB	87%	80%	81%
KNN	**92%**	**92%**	**92%**

5. Conclusions

This work offers the following conclusion:

A set of data mining techniques using a machine learning algorithm has been provided—a practical, structured approach to diagnosing mental disorders. The methodology will lead to more accurate results in diagnostic systems for mental disorders. It highlights implementing mixed data mining by comparing the random forest classifier with no missing ratios. Rate the learner on the mental disorder dataset. Clear of the results is that the random forest classifier is combined with the missing values learner Get more accurate results with the KNN Classifier. It turns out to be powerful for noisy data and loses data according to the experiments conducted with different classifiers, KNN we get the highest accuracy. NB, KNN, and Random Forest Ratings followed with an accuracy of 87%, 92%, and 87.72%, respectively. The significance of these findings is to verify data mining. The approach may be helpful in diagnostic practices. Also, forms can be trained useful in building diagnostic models of mental illness. Future trends include more trials with a larger group of patients and with different data methods mining analysis. Moreover, this study can be applied to expert mental systems diagnostic systems for disorders such as desktop or mobile applications.

References

[1] N. Dogra and S. Cooper, Disorders of personality. 2017.

[2] P. Kaur and M. Sharma, "Diagnosis of Human Psychological Disorders using Supervised Learning and Nature-Inspired Computing Techniques: A Meta-Analysis," J. Med. Syst., vol. 43, no. 7, 2019, doi: 10.1007/s10916-019-1341-2.

[3] M. Mohammadi et al., "Data mining EEG signals in depression for their diagnostic value Clinical decision-making, knowledge support systems, and theory," BMC Med. Inform. Decis. Mak., vol. 15, no. 1, pp. 1–14, 2015, doi: 10.1186/s12911-015-0227-6.

[4] K. Daimi and S. Banitaan, "Using Data Mining to Predict Possible Future Depression Cases," Int. J. Public Heal. Sci., vol. 3, no. 4, p. 231, 2014, doi: 10.11591/ijphs.v3i4.4697.

[5] J. F. Dipnall et al., "Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression," PLoS One, vol. 11, no. 2, pp. 1–23, 2016, doi: 10.1371/journal.pone.0148195.

[6] E. M. Benyoussef, A. Elbyed, and H. El Hadiri, "Data mining approaches for Alzheimer's disease diagnosis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10542 LNCS, pp. 619–631, 2017, doi: 10.1007/978-3-319-68179-5_54.

[7] H. A. Warda, N. A. Belal, Y. El-Sonbaty, and S. Darwish, "A random forest model for mental disorders diagnostic systems," Adv. Intell. Syst. Comput., vol. 533, pp. 670–680, 2017, doi: 10.1007/978-3-319-48308-5_64.

[8] S. Shah and A. Bhise, "Fast Speaker Recognition using Efficient Feature Extraction Technique 1," 2013.

[9] S. Shastri et al., "Development of a Data Mining Based Model for Classification of Child Immunization Data," Int. J. Comput. Eng. Res., vol. 8, no. 6, pp. 41–49, 2018, [Online]. Available: www.ijceronline.com.

[10] Shraddha Pandit and Suchita Gupta, "A Comparative Study on Distance Measuring Approaches for Clustering," Int. J. Res. Comput. Sci., vol. 2, no. 1, pp. 29–31, 2011, [Online]. Available: http://www.ijorcs.org/uploads/ijorcs/distance-measuring-approaches-for-clustering.pdf.

[11] D. Tasche, "A plug-in approach to maximising precision at the top and recall at the top," arXiv, pp. 1–10, 2018.

[12] J. Hua, "Study on the performance measure of information retrieval models," 2009 Int. Symp. Intell. Ubiquitous Comput. Educ. IUCE 2009, pp. 436–439, 2009, doi: 10.1109/IUCE.2009.105.

[13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS One, vol. 10, no. 3, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.

[14] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," pp. 37–63, 2020, [Online]. Available: http://arxiv.org/abs/2010.16061.

[15] E. Technologies, "Combined Substance Use and Mental Health Treatment Episode Data Set (TEDS) State Instruction Manual," no. 0930, 2019.

[16] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," Iarjset, pp. 20–22, 2015, doi: 10.17148/iarjset.2015.2305.

مجلة كلية الرافدين الجامعة للعلوم (2022)؛ العدد 51؛ 228- 236

AL- Rafidain
University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

التنبؤ بالاضطرابات النفسية باستخدام تقنيات تعدين البيانات

د. كريم هاشم الساعدي	لمياء ناجي خليل
karimnav6@gmail.com	lamyaa.naji@uomustansiriyah.edu.iq
قسم علوم الحاسوب - كلية العلوم - الجامعة المستنصرية بغداد، العراق	قسم علوم الحاسوب - كلية العلوم - الجامعة المستنصرية بغداد، العراق

المستخلص

على مدى السنوات القليلة الماضية، أدى التنقيب عن البيانات الى ظهور تطبيقات جديدة في مجال الطب. يعاني العديد من الناس اضطرابات نفسية غير واضحة بسبب عدم وجود اهتمام في قطاع الرعاية الصحية، وخاصة النفسية منها. تؤدي العديد من الاسباب الى نقص سجلات الصحة العقلية للمرضى فعلى سبيل المثال، يتردد المرضى في الذهاب الى الطبيب النفسي لمتابعة حالتهم بسبب ارتفاع تكلفة رسوم الفحص وطول فترة الانتظار حتى يأتي دورهم؛ وعدم الدقة في التشخيص. تم استخدم مصنفات للتنبؤ (طبيعي، ادمان بدون اضطراب، اضطراب بدون ادمان، أو ادمان مع اضطراب). لذلك يوجد هنالك ارتفاع في نسبة وجود الشخصية المضطربة وتعاطي المخدرات. استخدمنا مجموعة بيانات تتكون من مدمني المواد الافيونية من جميع انحاء الولايات المتحدة. تم استخدام تقنيات التصنيف Naive Bayes and K-Nearest Neighbor's للحصول على اعلى دقة نظام للتنبؤ بالاضطراب العقلي ومدمني المواد الافيونية بلغت دقة المصنف (NB) 87% وبلغت دقة المصنف (KNN) 92%. تقدم هذه الورقة طريقة جديدة لتشخيص الاضطرابات النفسية من خلال التنقيب عن البيانات.

معلومات البحث

تواريخ البحث:
تاريخ تقديم البحث: 2021/4/11
تاريخ قبول البحث: 2021/7/4
تاريخ رفع البحث على الموقع: 2022/6/25

الكلمات المفتاحية:
التنقيب في البيانات، الاضطرابات العقلية، Naive Baye، مصنف الجار الأقرب – K

للمراسلة:
لمياء ناجي خليل
lamyaa.naji@uomustansiriyah.edu.iq

DOI: https://doi.org/10.55562/jrucs.v51i1.535