IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Survey on Predicting Diseases of Employees under Work Pressure Using Data Mining Techniques

S. Anitha[1*], M. Vanitha[2]

[1,2]Dept. of Computer Applications, Alagappa University, Karaikudi, Tamilnadu, India

**Corresponding Author: nathan.anitha@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i4.616620 | Available online at: www.ijcseonline.org

Accepted: 18/Apr/2019, Published: 30/Apr/2019

Abstract— Employees working in various professions & occupations are prone to face varieties of health problems due to work pressure. The level of increasing work pressure as assessed by the perception of having little control but lots of demands have been demonstrated to be associated with increased rate of health issues such as hypertension, back pain, feeling fatigued, headaches, disorders and sometimes heart attack. Work pressure also causes accidents, diminished productivity, medical, legal and financial costs. In healthcare industry, data mining plays an essential role for predicting diseases of employees under work pressure. High volume of data that can be generated for the prediction of diseases of employees is analyzed traditionally and is too complicated along with voluminous to be processed. Data Mining provides the methods and techniques for transformation of the data into useful information for decision making. These techniques can make process fast and take less time to predict the diseases of employees under work pressure with more accuracy to improve their health in advance.

Keywords— Data Mining, Predicting Disease, Work Pressure, Healthcare, Decision Making

I. INTRODUCTION

Data Mining is a process of extracting knowledge from the vast amount of data. Data Mining can also be referred as Knowledge Discovery from Database (KDD). In recent years, Data mining is recognized as a powerful tool in various fields like information technology, healthcare industry, Financial data analysis and many more. Machine learning is an emerging field that has taken many of the methods and techniques from data mining.

Data Mining uses two methods: supervised and unsupervised learning. In supervised learning, a training data set and response variable are used whereas in unsupervised learning no train data and no response variable are used.

Data mining consists a lot of algorithms which can be applied separately or combined based on the need for the particular application. Data mining processes are generally classified into two types namely Predictive and Descriptive. Predictive process refers to building a model for predicting future behavior for a given input attributes. Examples for Predictive process are Classification, Prediction and Deviation detection . Descriptive process refers to describing the data in an understandable and effective form. Examples of descriptive process are data characterization, association rule discovery and clustering.

The objective of this paper is to analyze the existing works on data mining techniques/algorithms which have been used for disease prediction of employees under work pressure among different employees & occupations with more accuracy in less time.

II. LITERATURE SURVEY

Qasem A. Al-Radaideh[1] proposed to build a classification model CRISP-DM (Cross Industry Standard Process for Data Mining) to predict the performance of employees by utilizing data mining techniques. It included the following steps : Business understanding, data understanding, data preparation, modeling, evaluation and employment. Experiments were conducted on 130 employees from several companies for real data analysis. In that experiment, easier interpretation and understanding for decision makers using decision tree. And also author created an initial tree using the divide-and-conquer algorithm, and used WEKA toolkit package. Here compared and justify the classifier techniques such as C4.5/C5.0/J4.8, NBTree.

Feixiang Huang,et al,[2] In this paper, the author reported the experiments of using data mining techniques to predict diseases from a large number of real world medical records. Information of pathological indices is not used in our work. The raw data of our experiments consists of 309383 patients' historical medical records stored in MySQL database. Here derived and computed 284 statistical features such as number

of visits, presence of certain diseases and used a simple approach of considering only the presence or absence of diseases in patient's medical records with an overall prediction accuracy around 83.5% and not as an independent document. Their experimental results, ensemble approach seems to provide only minor improvement of prediction performance.

E Deepak Chowdary ,et al,[3] intended Pegasos which modified algorithm of stochastic gradient methods. This algorithm combined with Adaboost ensemble and achieved better classification results with a good optimization objective. Dataset collected from various software companies and the factors are identified and categorized into eight groups with twenty-three attributes. The collected dataset was partitioned into Training set and Testing set and finally the model was used for classification and achieved an accuracy of 96.19% by comparing with existing datasets and existing methods.

Pinky Saikia Dutta,et al,[4] discussed elaborately in detail on the existing Data mining techniques and also application tools which are most important for healthcare services. In this survey, applied Apriori algorithm and FP growth algorithm and compared dataset. Using Association Technique showed the final report. Here the methodology worked on real historical data, it provided accurate and efficient results, which helped the patients, get diagnosis instantly. Author intended to predict the diseases using input data sets and suggested the best doctors and remedial guidelines for Effective Treatment.

P.Manivannan,[5] proposed clustering technique based partitioning method. In this technique, identical objects are grouped in one cluster and non-identical objects are grouped in another cluster. This paper predicted dengue fever using one of the unsupervised algorithm. The data size which was collected 1910 records and 171 attributes from urban Ho Chi Minh City,Vietnam. Generally, Dengue is a viral disease responsible for most of the illness. The author has intended to predict household clustering of dengue by using dengue serotypes depending upon the age group through applying the k-means clustering algorithm.

D.Umanandhini[6] has found the level of academic stress among higher secondary school students by academic stress scale. It contained of 40 items and each item has the various responses such as No stress ,Slightly stress, Moderate stress, Highly stress and Extremely High stress. High scores indicated high stress and low scores on the scale indicated low stress.

Divya Sharma, et al,[7] highlighted the important role played by data mining tools in analysis of huge volume of healthcare related data in prediction and diagnosis of lifestyle diseases. This survey defined various techniques and data mining classifiers for efficient and effective heart disease and type II diabetes diagnosis. And also showed the table that different data mining techniques used heart disease diagnosis over various data sets. In that experiment easily diagnosis and predict the lifestyle diseases and provides better treatment for affected disease.

Huijie Lin,et al, [8] analysed that users stress state is closely related to that of their friends in social media, and a huge dataset from real-world social platforms is used to analyze the relation of users stress states and social interactions. Stress-related attributes like textual, visual, and social attributes are defined, and then a factor graph model combined with Convolutional Neural Network is proposed to make use of the tweet content and social interaction information for stress detection. The paper revealed that proposed model can improve the detection performance by 6-9 percent. The number of social structures of sparse connections of stressed users is around 14 percent higher than that of non-stressed users, indicating that the social structure of stressed users tends to be less connected and less complicated than that of non-stressed users.

Durga Kinge, et al,[9] used diverse machine learning algorithms with WEKA tool for categorization the disease. In this experiment, using Naive Bayes, Random Forest, simple logistic, Bagging, MLP and Adaboost classifiers. Finally, The classification results will be showed by various representation like 2D diagrams, pie graphs, and different techniques. It showed that some of the classifiers such as random forest, simple logistic and Naive bayes perform better for heart disease prediction.

J. S. Kanchana,et al,[10] In this paper, the author present a model for detecting the psychological stress level of the users by leveraging the tweets of each user and their social behavior. The classification revealed the number of stressed and non-stressed users, the number of tweets posted per week by each user and the number of stressed and non-stressed posts per week. Support vector machine and Naive Bayes algorithm algorithm are used to classify stressed and non stressed user. Ranking is based on the number of stressed posts tweeted by each user. Then the user with highest stress is found.

III. TABLE - MERITS AND DEMERITS OF THE DATA MINING ALGORITHMS

S.No	Algorithm	Merits	Demerits

1	Iterative Dichotomiser ID3	From the training data, understandable prediction rules are created. Builds the fastest and short tree . Number of tests will be reduced. Whole dataset is searched to create tree.	When small sample is tested,data may be over-fitted or over-classified. Multiple attribute can not be tested at a time for decision making.
2	C4.5 Algorithm	Most suitable for real world problems as it deals with numeric attributes and missing values. Build smaller or larger, more accurate decision trees. More time efficient and reduces ambiguity indecision-making.	C4.5 constructs empty branches and many nodes with zero values or close to zero values. so it makes the tree bigger and more complex. when algorithm model picks up data with uncommon characteristics, data may be over fitted. Susceptible to noise.
3	K Nearest Neighbors	Easy implementation for simple technique. This model is cheap. Suitable for multi-modal classes, records with multiple class labels.	This model can not be interpreted. It is computationally expensive, when the dataset is very large. Performance is calculated depending on the number of dimensions.
4	Naïve Bayes	Super simple because just doing a bunch of counts. Naive Bayes classifier will converge quicker than discriminative models like logistic regression, when NB conditional holds independence assumption. so need less training data.	It makes a very strong assumption on the shape of data distribution. Another problem happens due to data scarcity.
5	Support Vector Machine (SVM)	When a boundary is established, most of the training data is redundant. It can be applied to both linear and non linear problems.	For any given problem, It has several key parameters that need to be set correctly to achieve the best classification results. More time complexity. Sometimes it may result in a poor classification accuracy for problem.
6	Artificial Neural Network (ANN)	ANN has the ability to learn and model non-linear and complex relationships, which is really important. Because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex.	Most hardware dependency. Eg.Artificial neural networks require processors with parallel processing. If there is any issue, difficulty of showing the problem to the network.
7	J48 Decision Tree	Implement Univariate Decision Tree approach.It rectifies the disadvantages of ID3. Handling training data with missing values of attributes.	The run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the number of attributes. Tree depth is linked to tree size.
8	Classification and Regression Trees (CART)	Handling both numerical and categorical variables easily. This algorithm, itself identify the most significant variables and eliminate non-significant ones.	It may have unstable decision tree. Tree complexity may be increase or decrease depending the changes in splitting variables and values. It splits only by one variable.
9	Random Forest	Most accurate learning algorithm. It produces a highly accurate classifier For many data sets. It runs on large databases efficiently.	More complexity. It much harder and time-consuming to construct than decision trees. It also requires more computational resources and are also less intuitive.
10	Hierarchical clustering	Easy to understand and easy to implement. In addition, it works well to larger datasets.	But rarely provides the best solution.
11	Apriori	Multiple scans are generated for candidate sets.	When producing candidate set, every time wasted the execution time. it also needs more search space and computational cost is too high.
12	AdaBoost	It constructs strong classifier than linear.	More sensitive to noisy data and outliers. it has less

			susceptible to the overfitting problem than most learning algorithms.
13	Expectation Maximization (EM)	Guaranteed to increase for each iteration. it does not require an optimizer and very popular for fitting mixture distributions.	Requires both forward and backward probabilities. Slow convergence.
14	Back Propogation	Easy to implement. Mathematical Formula used in the algorithm can be applied to any network. If the weights chosen are small at the beginning, reduce the computing time.	No speed and efficient. A large amount of input/ output data is available, but not sure how to relate it to the output.
15	K-means	When variables are huge, then K-Means most of the times computationally faster than hierarchical clustering. Especially if the clusters are globular, It produces tighter clusters than hierarchical clustering.	To predict K-Value is not easy. It didn't work properly, when cluster is global.If it have different initial partitions, its result in different final clusters.

IV. CONCLUSION

In this paper, Classification techniques are attained of processing a large amount of data. In Healthcare Industry, classification is one of the most broadly used methods of Data Mining. The widespread classification techniques used in prediction of diseases are Decision Tree(J48), Support Vector Machines, Neural Network, Naïve Bayes and Random forest. And also its some of the disadvantages are run-time complexity, less classification accuracy, most hardware dependency, strong assumption of data distribution and more complexity respectively. But based on observations and review, Decision tree and Neural Network algorithms are currently used and produced satisfactory results in classification. It is better to propose new approach with features of above algorithms for reducing complexity as well as increasing effectiveness and accuracy.

ACKNOWLEDGMENT

This research work has been supported by RUSA PHASE 2.0, Alagappa University, Karaikudi.

REFERENCES

[1] Qasem A. Al-Radaideh , Eman Al Nagi ," Using Data MiningTechniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No.2,2012 .

[2] Feixiang Huang, Shengyong Wang and Chien-Chung Chan," Predicting Disease By Using Data Mining Based on Healthcare Information System", IEEE International Conference on Granular Computing,2012.

[3] E.Deepak Chowdary, K.Anusha Devi , D.Mounika , K.V.KrishnaKishore and S.Venkatramaphanikumar ," Ensemble Classification technique to detect stress in IT-Professionals ", IEEE 2016.

[4] Pinky Saikia Dutta, Sunayana Dutta, Tridisha Das, Sweety Buragohain, SusmitaSarma," A Survey on Smart Health Care Using Data Mining", International Journal of Computer Sciences and Engineering, Volume-4, Special Issue-7, Dec 2016, ISSN: 2347-2693

[5] P.Manivannan, "Dengue Fever Prediction Using K-means Clustering Algorithm" ,IEEE International Conference of Intelligent Techniques in Control, Optimization and Signal Processing – Year 2017.

[6] Umanandhini.D, Kalpana.G "Survey on stress types using data mining algorithms", IJIRAE-International journal of innovative research in advanced Engineering, April 2017 ISSN:2349-2163.

[7] Divya Sharma, Anand Sharma, Vibhakar Mansotra," A Literature Survey on Data Mining Techniques to Predict Lifestyle Diseases", International Journal for Research in Applied Science & Engineering Technology (IJRASET), , June 2017,Volume 5 Issue VI, ISSN: 2321-9653

[8] Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen,Lexing Xie,Jie Tang, Ling Feng, and Tat-Seng Chua, " Detecting Stress Based on Social Interactions in Social Networks " in IEEE 2017.

[9] Durga Kinge, S. K. Gaikwad, "Survey on data mining techniques for disease prediction", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 01 , p-ISSN: 2395-0072, Jan 2018.

[10] J. S. Kanchana, H. Thaqneem Fathima, R. Surya, R. Sandhiya, "Stress Detection Using Classification Algorithm", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol.7 Issue 04, April-2018 .

[11] Xue-Hui Meng , Yi-Xiang Huang , Dong-Ping Rao , Qiu Zhang , Qing Liu," Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung Journal of Medical Sciences (2013).

[12] Kevin Daimi, Shadi Banitaan," Using Data Mining to Predict Possible Future Depression Cases", International Journal of Public Health Science (IJPHS), ISSN: 2252-8806, Vol.3, No.4, December 2014.

[13] Manimaran.R and Vanitha.M (2017) "Prediction of Diabetes Disease Using Classification Data Mining Techniques", International Journal of Engineering And Technology(IJET), Volume 9, Issue 5, Nov 2017, PP: 3610-3614.

[14] Silviya D'monte, Dakshata Panchal," Data Mining Approach for Diagnose of Anxiety Disorder", International Conference on Computing, Communication and Automation (ICCCA2015), ISBN:978-1-4799-8890-7, 2015 IEEE.

[15] Ms.Mehdi Khundmir, Prof. Vikas Maral," A Survey on Usage of Data Mining Techniques for Prediction of Heart Disease **",** IOSR Journal of Computer Engineering, e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, May-June 2017.

[16] S.Neelamegam, Dr.E.Ramaraj, "Classification algorithm in Data Mining:An Overview ", published on Sep-2013, International Journal of P2P Network Trends and Technology , Vol 3 Issue 5.

[17] E.Gokulakannan and Dr.K.V.Venkatachalapathy, "Comparison Study of Severaldata Mining Algorithms to Predict Employee

Stress",published in Advances in Natural and Applied Sciences, 2015.

[18] M.A.Nishara Banu, B.Gomathy, "Disease Predicting System Using Data Mining Techniques", published in International Journal of Technical Research And Applications, Volume 1,Issue 5, Nov 2013,PP41-45.

[19] Alexandra F.DA.S.Cordeiro, Irenilza De A.Naas, Stanley R.DE.M.Oliveira,Fabio Violaro, Andreia C.M.DE. Almeida, " Efficiency of Distinct Data Mining Algorithms for Classifying Stress Level in Piglets from other Vocalization", V 32,p.208-216 Apr 2012.

[20] Shubpreet Kaur and Dr.R.K.Bawa, " Future Trends of Data Mining in Predicting the various Diseases in Medical Healthcare System", published in International Journal of Enrgy,Information and Communications, Vol.6Issue 4, 2015. pp.17-34.

[21] K.Gomathi, Dr.D.Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", published online at http://www.publishinginda.com, sep 2017.

