A Critical Analysis Of The Relationship Between Depression And Smoking Using Machine Learning

Mohammad Kharabsheh, Ahmad Qawasmeh, Omar Megdadi, Nadera Jawabreh, Rola Mudallal, Sukaina Alzyoud

Abstract: Smoking has been a major concern for many decades due to its negative impact on societies. A study of the role of decision support system for student smokers in order to find the depression type is investigated in this research. In this work, we developed a hybrid machine learning model that consists of clustering and classification. The idea of this model is to predict the type of depression for youth smokers using numerous novel features such as father's job, how many narghile (Shisha) heads students usually smoke, and some other relevant features. Our model illustrated a significant relationship between smoking and depression. Our model demonstrated a prediction accuracy of 97% when applied on a dataset consisting of 993 student smokers in Jordan. Therefore, efficient solutions must be considered to find useful alternatives to smoking.

Index Terms: Classification, Clustering, Depression, Hybrid, Machine Learning, Prediction, Smokers.

1 INTRODUCTION

A Branch of Artificial Intelligence (AI) focused with "teaching" computers how to act without being explicitly programmed for each possible scenario is called Machine learning (ML) [17]. The major concept in ML is developing algorithms that can self-learn through training on a huge number of inputs [16] [18]. There are different types of ML, one of them is supervised learning which is consider the main type. This type relies primarily on estimating the future instances based on known instances. The purpose of supervised learning is to extract a pattern of the distribution of class labels which rely on predictor features. This pattern is used to select class labels to the testing instances. Class labels for these instances are selected based on the predictor features that are known. Since features can be large in number and some of them can be less informative than the others, Feature Selection (FS), which is also called attribute selection, is an essential phase to build any prediction model. As a multidisciplinary field to help people developing approaches and automated tools, researchers have focus on Bioinformatics studies and researches.

RQ1: Could we predict depression type through machine learning methods? The obtained results demonstrate that we could develop effective prediction models for classifying smoker student's depression type. We were able to develop classifier using machine learners that has a recall value of 97% and a precision value of 97%.RQ2: Which attributes are the most important as predictors of depression type? We next created a decision tree model and accomplished a Top Node analysis to recognize the main effective attribute for making a decision regarding the depression type. Our results show that the factor called age is the most important attribute for such decision. The remainder of this paper is as follows. Section II analyses related work. Section III presents the methodology that we follow in this paper. Section IV shows the achieved results of our paper. Section V presents the core threats to validity of this study. The conclusion of our study and some ideas for future work presents in Section VI with.

2 RELATED WORKS

The Machine learning algorithms can improve health care delivery and management via supporting decision making. Previous researches used machine learning for decision support of health care systems. Demner-Fushman et.al [1] discussed recent interest in developing fundamental natural language processing methods and for computerized clinical decision support. Karakülah et.al [2] developed an approach for automatic scanning and defining of the phenotypic factors from the case reports associated with congenital anomalies. The proposed approach represents text processing techniques, and also represents a framework for probable diagnostic decision support system for congenital anomalies. Jin et.al [3] developed a classification model for recognizing molecular units to infection concepts in order to decide if the primary probabilistic approach could be generalized to dissimilar concepts within model retraining. The developed approach uses Conditional Random Fields trained using some domain-specific features. The approach precision value of 0.85 with also 0.83 recall value. Moreover, the approach achieved a value of 0.84 with respect to F-measure metric. Skounakis et.al [4] applied learning methods to gain gene-disorder relatives [18]. They evaluated their method effectiveness by extracting binary relations in three biomedical domains. In [5], authors described a health data analytics engine that is based on machine learning algorithms. The engine is aimed at analyzing cloud based PHR health datasets for knowledge extraction that help healthcare decisions, such as disease prognosis and diagnosis, in an efficient way. The engine has been effectively applied to a dataset provided by Apache Hadoop. Several Classification Techniques have been

- *Mohammad Kharabsheh. Dept. of Computer Information System. The Hashemite University. Zarqa, Jordan. Email: mohkh86@hu.edu.jo*
- *Ahmad Qawasmeh. Dept. of Computer Science. The Hashemite University. Zarqa, Jordan. Email: ahmadr@hu.edu.jo*
- *Omar Megdadi. Dept. of Software Engineering. Jordan University of Science and Technology. Irbid, Jordan. Email: ommeqdadi@just.edu.jo*
- *Nadera Jawabreh. Dept of Software Engineering. University of Szeged. Hungary. Email: nadera@inf.u-szeged.hu*
- *Rola Mudallal. Dept. of Community & Mental Health Nursing. The Hashemite University. Zarqa, Jordan. Email: rula@hu.edu.jo*
- *Sukaina Alzyoud. Dept. of Community & Mental Health Nursing. The Hashemite University. Zarqa, Jordan. Email: sukaina-alzyoud@hu.edu.jo*

offered in the literature for Healthcare researches. Das et al. [6] proposed a neural networks method for the diagnosis of heat diseases. They evaluated their approach using a dataset from Cleveland heart disease database. The evaluation results show that the approach achieves 89.01% accuracy. Chien et al. [7] developed a hybrid decision tree approach that aims at classifying the activity of chronic disease patients in more accurate way. Zuoa et al. [8] proposed a Fuzzy K-NN method to help the healthcare asscociated with Parkinson disease. The introduced approach has achieved the highest classification results through the 10-fold cross-validation analysis, where the accuracy of 97.47% is reached. Jena et al. [9] used neural network and linear discriminate analysis to classify chronic diseases that is needed to generate prompt warning systems. The approach examines the relation between cardiovascular disease and hypertension along with the risk factors of numerous chronic diseases, where the early warning system developed to decrease the complication occurrence of such diseases. Our study is the unique examination in the area of exploring the idea of machine learning classifiers in identifying the depression type of youth smoking persons (Age: 11-17 Years old).

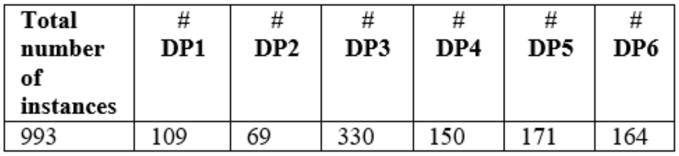
3 PROPOSED METHODOLOGY

The methodology that we followed in our study is presented in this section. Firstly, we discuss the dataset that is used for our evaluation. Then, we introduce the factors that are used in the learning of our classifier. Lastly, we present the developed model and the metrics that are used in the evaluation experiments.

A. Creating the Corpus

The main step of classification experiment is preparing the corpus from the dataset that next would be used for training a developed machine learning classifier. Here, for every instance belongs to our examined dataset, we computed the value related to every considered factor posed early. We then developed a model that combines supervised and unsupervised learning. First, we labeled each instance with the relevant depression level using K-Means clustering algorithm. Second, we trained a classifier based on the supervised dataset generated from the first step. TABLE 1, summarizes the corpus information. We used the types DP1-DP6 to refer to the depression types 1 to 6 in the rest of the paper.

TABLE 1. SUMMARY OF CORPUS INFORMATION



Total number of instances	# DP1	# DP2	# DP3	# DP4	# DP5	# DP6
993	109	69	330	150	171	164

B. Classification Algorithms

In our experiments, we employed the supervised classifiers in which the inputted corpus is distributed into two groups: a training set and a test set. The first group is the one that is used to train the developed machine learner. On the other hand, the performance of the learner is computed through the second group. Here, we used the widely popular 10-fold cross-validation [13] technique to obtain both the training and test sets. On the other hand, we employed the WEKA toolkit [14, 15] to perform the supervised classification in our work. Now, we discussed the various classification algorithms that were widely used in the literature [7, 8, 9, 21, 22] of decision support systems presented for healthcare domain and have been used to develop our classifiers.

- Support Vector Machine (SVM): this algorithm rises the dimensionality of training instances to achieve differentiable points in one of the dimension. This algorithm is very popular since it is efficient in high dimensional spaces and thus provides more accurate results [12].
- Trees.RandomTree: An active data mining algorithm that is used with large amounts of data. The technique employs several classification trees to a data set and next generates the prediction from all of the correlated trees
- Trees.J48: The algorithm creates a binary tree for classification problems. The approach splits the data into range using the values of attributes for that item that are recognized in the training set.
- Bayesian Learner (Naïve Bayes): This method assumes that all classification factors are independent. It shows great performance in terms of accuracy when it was applied in medical domain studies.
- Sequential Minimal Optimization (SMO): An approach that trains a support vector leaner using polynomial. It converts attributes from nominal to binary values.
- Logistic Regression: The approach uses regression models for classification tasks that models the posterior class probabilities for each of the needed n-classes from the dataset.
- K-Star: it represents a nearest neighbor method uses the distance calculations from the training set, such as the Mahalanobis metric, to classify the instances of the testing set.
- Decision Table: The table for a given dataset is generated using grouping-and-counting in order to apply classification over unknown sample.
- K-Nearest Neighbor (K-NN): This technique is based on discovering the unidentified instances using the formerly known instances (e.g., nearest neighbor) and hence classify other instances using the voting approach [12].
- IBk: is another nearest-neighbor algorithm that uses the distance metrics created from the training set as closest associated vectors that would be used to classify data instances of the testing set.

C. Evaluation Metrics

To evaluate the effectiveness of our proposed classification model, several performance metrics have been widely used in the literature. That is, we choose to use the following metrics.

- Precision: the ratio of retrieved instances that are truly relevant. It is calculated as (P = True Positives / (True Positives + False Positives)).
- Recall: the ratio of relevant instances that are retrieved by the classifier and hence it is computed as (R= True Positives / (True Positives + False Negatives)).
- F-Measure: a metric that depends on both recall and precision of a model and thus calculated by a combination of these two metrics as ((2 * Recall * Precision) / (Recall + Precision)). The value of this metric is between 0 and 1.
- ROC: the area calculated from the region below the Receiver Operating Characteristic (ROC) curve via the plotting of true positives against false positives obtained from a classifier.

4 STUDY RESULTS

Here, we present our obtained results from the undertaken classification experiments and hence we answer the research questions mentioned early. A.RQ1: Could we predict depression type through machine learning methods? To address this question, we developed machine learning classifiers that are based on the classification algorithms discussed before. When applying the classifiers, we distributed the instances of the inputted dataset into training and test sets by using the widely known 10-fold cross

TABLE **2**. SUMMARY OF CLASSIFICATION FACTORS

Classification Factor	Definition
Age	We selected 11-17 years old as the age range of youth participants in this study.
Gender	Both Six
Birth_Day	Day of Birth
Birth_Month	Month of Birth
Birth_Year	Year of Birth
Birth_Place	Place of Birth
Nationality	Country of Birth
Father_job	Father Job
No_Shisha_Month	Times have you smoked narghile (Shisha), even a puff, in the past 30 days
No_Times_Shisha_Mon	The past 30 days (month), on the days you smoked, how many narghile (Shisha) heads did you usually smoke
Deperssion_1	I had trouble keeping my mind on what I was doing
Deperssion_2	I felt lonely.
Deperssion_3	I had crying spells.
Deperssion_4	I felt sad.
Deperssion_5	I felt that people disliked me.
Deperssion_6	I could not get "going".

__TABLE 3__ provides the obtained evaluation results of our developed classifiers. That is, we would conclude that our models would accurately predict the depression type of youth smokers.

On the other hand, we noticed that SMV and KNN obtained better results in the term of accuracy when compared with the other classifiers. The main explanation of this observation is that KNN finds the relevant likelihood for each class by investigating the characteristics of one of the instances in order to expect it for its nearest neighbors given that nearest neighbor data points have similar trends. Thus, for each training instance, the preceding and the likelihood could be changed dynamically to gain strength against possible classification faults. Similarly, with the SMV learner, the instances of the training set are recognized in some specific dimensions by increasing the dimensionality of inputted dataset, in order to enhance the classification accuracy. Furthermore, SMO is appropriate to work with large datasets and gains greater accuracy because the memory required for SMV is lined with the size [12]. Moreover, the similarity between the recall and precision values indicates that our created dataset can efficiently be used for prediction.

TABLE 3: OBTAINED CLASSIFICATION RESULTS OF **DEPRSSION TYPES**

Learner	Accuracy	Recall	Precision	F-measure	ROC
RandomTree	0.857	0.857	0.856	0.856	0.916
J48	0.889	0.889	0.888	0.888	0.947
NaiveBayes	0.913	0.913	0.916	0.913	0.980
SMO	**0.979**	**0.979**	**0.979**	**0.979**	**0.994**
Logistic	0.968	0.968	0.968	0.968	0.998
IBK	0.910	0.910	0.910	0.909	0.942
KStar	0.789	0.789	0.789	0.783	0.961
DecisionTable	0.816	0.816	0.810	0.809	0.962

The confusion matrix of SMO classifier is shown in TABLE 4. That is, we observed more details about the performance of SMO classifier in our model of prediction the depression type.

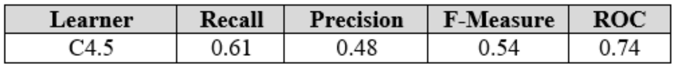
Table 4: CONFUSION MATRIX OF SMO FOR **DEPRSSION TYPE1 (DT1) TO DEPRESSION TYPE6 (DT1)**

Confusion Matrix of SMO									
True Class	**DT 1**	103	1	0	2	0	0		97.17%
	DT 2	0	74	2	1	0	1		94.87%
	DT 3	0	2	330	0	0	0		99.40%
	DT 4	1	0	0	144	0	0		99.31%
	DT 5	1	0	0	0	170	0		99.42%
	DT 6	1	2	5	0	2	151		93.79%
		97.17%	93.67%	97.92%	97.96%	98.84%	99.34%		
		DT 1	**DT 2**	**DT 3**	**DT 4**	**DT 5**	**DT 6**		
Predicted Class									

nt

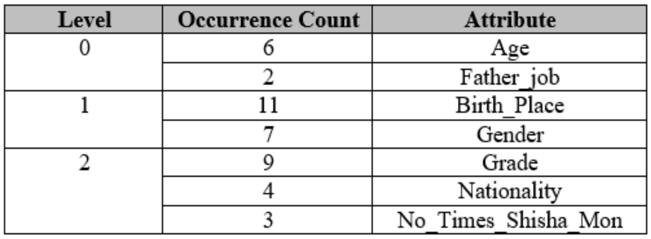
To answer our current question, we need to assess the usefulness of every attribute individually as a predictor of the depression type. To do so, we developed our classifiers using decision trees that are trained using all classification factor discussed early. Using decision trees, it is possible to rank attributes based on their usefulness in our prediction experiments by performing the Top Node analysis [11] associated with the decision tree approaches. This node analysis approach counts the presence of each factor under the consideration through the inspection of the structure and levels of a developed decision tree [10]. Next, the tree level where an attribute occurs and the computed count of the attribute are used to decide the usefulness rank of that attribute. Explicitly, the most influential factor would be the root node of the constructed decision tree, while a factor effectiveness reduced as we move toward the leaves of the tree [23]. Thus, in our study, we developed a decision tree using the C4.5 algorithm [10], which was trained by using all the factors under consideration in this work. C4.5 is a greedy technique that adds decision nodes at each level of the constructed tree by following the divide and conquer algorithm on the training set. At each stage of the running algorithm, the information observed from each attribute is computed, and next the attribute that has the top ranked value is selected among other attributes to be added to the tree. The number of running stages of the greedy algorithm depends on a given cut-off value, which is used to determine the number of record in the leaf nodes while constructing the tree. The performance results obtained from our decision tree classifier is shown in TABLE 5. Additionally, the outcomes of the mentioned analysis are illustrated in TABLE 6. Specifically, for each effective factor, the table provides the level where it appears in the constructed tree (e.g., column one) and occurrence frequency associated with the factor (e.g., column two). As we could observe, the age represents the root node of our resultant tree and hence it is the most influential factor in our experiments.

TABLE5: CLASSIFICATION RESULTS OF THE C4.5 TREE



Learner	Recall	Precision	F-Measure	ROC
C4.5	0.61	0.48	0.54	0.74

TABLE 6: OUTCOMES OF TOP NODE ANALYSIS



Level	Occurrence Count	Attribute
0	6	Age
	2	Father_job
1	11	Birth_Place
	7	Gender
2	9	Grade
	4	Nationality
	3	No_Times_Shisha_Mon

5 THREATS TO VALIDITY

As with any case study that is based on a sample of smokers, we suffer from some potential threats that inhibit us from generalizing our obtained results to datasets from different environments. We use datasets of 993 students where their age is between 11 and 17 years, and hence the inputted dataset could not be demonstrative of all students, and so we could not generalize our results for a variety of datasets. Moreover, there could be further attributes that have not been used in this study (e.g., smoking sessions and psychological status like depressive modes). These factor might positively impact our obtained results. Our developed classifiers are based on machine learning techniques that were successfully and widely used in the literature. However, there are some drawbacks for each classification technique that might negatively impact the correctness of our experiments.

Therefore, developing classifiers using other machine learners would be our consideration in the future. Also, our future plan would be the assessment of the usefulness of other factors as predictors for the depression type of youth smokers.

6 CONCLUSION AND FUTURE WORK

In this study, we offer an investigation of the effectiveness for machine learning models in categorizing the depression type of youth smoker. Here, we have accomplished a study on a dataset of 993 students age is between 11 and 17 years, using a group of attributes, for example age, gender, and father's job. We developed classifiers based on numerous techniques including support vector machine, and nearest-neighbor algorithms. The classifiers aim at predicting depression type of a set of youth students. We observed from the undertaken evaluation experiments that the developed classifiers have the best equitable recall of 97% and the worst recall value is 78%. On the other hand, our model achieved the best precision value of 97% and the worst value of 78%. By performing a Top Node analysis, we found that the attribute age is the most influential attribute in predicting the depression type of youth smokers. In future, we plan to examine additional factors and other machine learning techniques to enhance the prediction of depression types for datasets from a variety of domains and countries.

ACKNOWLEDGMENT

The current study was funded by the Deanship of Scientific Research at the Hashemite University. We would also like to thank Hiba Privet Hospital for approving the study and providing access to their out-patient's clinics.

REFERENCES

[1] Demner-Fushman, D., Chapman, W., Mcdonald, C., "What can Natural Language Processing do for Clinical Decision Support?", Journal of Biomedical Informatics Volume 42 issue 5, pp.760-72, 2009.

[2] Karakülah, G., Koşaner, O., Birant, C., Berber,T., Karakülah, A., Karakulah, G., Suner, A., Dicle, O., "Computer Based Extraction Of Phenoptypic Features Of Human Congenital Anomalies From The Digital Literature With Natural Language Processing Techniques", Studies In Health Technology And Informatics, Volume 205, pp. 570-574, 2014.

[3] Jin, Y., McDonald, R., Lerman, K., Mandel, M., Carroll, S., Liberman M., F., Winters,, R., White, P., "Automated recognition of malignancy mentions in biomedical literature", BMC Bioinformatics,Volume 7: 492, 2006.

[4] Skounakis,M., Craven, M., Ray, S. "Hierarchical hidden Markov models for information extraction", in Proceedings of the 18th international joint conference on Artificial intelligence (IJCAI'03), pp. 427- 433, 2003.

[5] Poulymenopoulou, M., Malamateniou, F., Vassilacopoulos, G., "Machine Learning for Knowledge Extraction from PHR Big Data", Studies in health technology and informatics, Volume 202, pp.36-39,2014.

[6] Das, R., Turkoglub, I., Sengur, A., "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, Volume 36, pp. 7675-7680, 2009.

[7] Chien, C., Pottie, G., "A universal hybrid decision tree classifier design for human activity classification," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1065-1068, 2012.

[8] Zuoa, W., Wanga, Z., Liua, T., Chenc, H., "Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach", Biomedical Signal Processing and Control, Elsevier, Volume 8, Issue 4, pp. 364-373, 2013.

[9] Jena, H., Wang, C., Jiangc, B., Chub, Y., Chen, M., "Application of classification techniques on development an early-warning systemfor chronic illnesses", Expert Systems with Applications, Volume 39, pp. 8852-8858, 2012.

[10] Garcia, H., Shihab, E.,"Characterizing and predicting blocking bugs in open source projects," in

Proceedings of the 11th Working Conference on Mining Software Repositories (MSR'14), New York, NY, USA, pp. 72 - 81, 2014.

[11] Hassan, A., Zhang, K.," Using decision trees to predict the certification result of a build," In Proceedings of the 21st IEEE/ACM International Conference on Automated Software Engineering (ASE '06), pp. 189–198, 2006.

[12] Ahmad, P., Qamar, S.,Rizvi, S., "Techniques of Data Mining In Healthcare: A Review", International Journal of Computer Applications, Volume 120, pp. 38–50, 2015.

[13] Efron, B., " Estimating the error rate of a prediction rule: improvement on cross-validation ", Technical Journal of the American Statistical Association, vol. 78, no. 382, pp. 316–331, 1983.

[14] D. Mays, K. P. Tercyak, K. Rehberg, M.-K. Crane, and I. M. Lipkus, "Young adult waterpipe tobacco users' perceived addictiveness of waterpipe tobacco," Tobacco Prevention & Cessation, vol. 3, no. December, 2017 2017.

[15] https://www.cs.waikato.ac.nz/ml/weka/

[16] M. Bkassiny, Y. Li ,SK. Jayaweera, " A survey on machine-learning techniques in cognitive radios," IEEE Communications Surveys & Tutorials, Vol. 15, No. 3, pp. 1136-59, 2013.

[17] F. Thung, S. Wang, D. Lo and L. Jiang, "An Empirical Study of Bugs in Machine Learning Systems," IEEE 23rd International Symposium on Software Reliability Engineering, Dallas, pp. 271-280, 2012.

[18] J. S. Di Stefano and T. Menzies, "Machine learning for software engineering: case studies in software reuse, ", 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 246-251, 2002.

[19] K. A. Gunes, and L. Hongfang, " Building effective defect-prediction models in practice," IEEE Software, Vol. 22, No. 6, pp. 23-29 , 2005.

[20] Alzyoud, S., Kheirallah, K. A., Weglicki, L. S., Ward, K. D., Al-Khawaldeh, A., & Shotar, A. (2014). Tobacco smoking status and perception of health among a sample of Jordanian students. International journal of environmental research and public health, 11(7), 7022-7035

[21] Mohammad Kharabsheh, Omar Meqdadi, Mohammad Alabed, Sreenivas Veeranki, Ahmad Abbadi and Sukaina Alzyoud, "A Machine Learning Approach for Predicting Nicotine Dependence" International Journal of Advanced Computer Science and Applications(IJACSA), 10(3), 2019.

[22] Alaa Al-Nusirat, Feras Hanandeh, Mohammad Kamel Kharabsheh, Mahmoud Al-Ayyoub, Nahla Al-dhfairi: Dynamic Detection of Software Defects Using Supervised Learning Techniques. International Journal of Communication Networks and Information Security (IJCNIS) 11(1) (2019) 2017.

[23] Shihab, Emad, Akinori Ihara, Yasutaka Kamei, Walid M. Ibrahim, Masao Ohira, Bram Adams, Ahmed E. Hassan, and Ken-ichi Matsumoto. "Predicting Re-opened Bugs: A Case Study on the Eclipse Project", in Proceedings of the 17th Working Conference on Reverse Engineering, 2010.