Predicting Depression Level of Youth Smokers Using Machine Learning

Sukaina Alzyoud, Mohammad Kharabsheh, Rola Mudallal

Abstract— Tobacco smoking is an alarming public health concern on a global scale due to its negative impact on the future generations wellbeing. This study aim to demonstrate the role of decision support system in predicting the depression level of youth using their smoking habits and related factors. In this work, we developed a hybrid machine learning model that consisted of clustering and classification. The idea of this model is to provide health care providers with a tool to predict the level of depression for youth smokers using a set of novel factors including: father's job, number of Aarghile (Shisha) heads smoked, and other relevant factors. Our model illustrated a significant relationship between smoking and level of depression. Our model demonstrated a prediction accuracy of 94% when applied on a dataset consisting of 993 student smokers in Jordan. Age was found as the most influential attribute in predicting the depression level of youth smokers. Therefore, efficient solutions must be considered to find useful alternatives to smoking.

Index Terms— Classification, Depression, Machine Learning, Smokers.

1 INTRODUCTION

MACHINE LEARNING (ML) is a branch of Artificial Intelligence (AI) concerned with "teaching" computers how to act without being explicitly programmed for every possible scenario [17]. The main concept in ML is developing algorithms that can self-learn by training on a very large number of inputs (possibly with known results) [16] [18].One of the major types of ML is supervised learning. This type relies fundamentally on estimating the future instances based on known instances. The aim 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 a fundamental phase to build any prediction model. In recent years public health researchers adapted a multidisciplinary approach with the help of Bioinformatics studies to develop an approaches with the use of automated tools to assist in identifying and assessing individuals with health problems. Bioinformatics is based on investigating and examining biological data for many goals to solve main problems under consideration by human beings through deducing biological data via computerized methodologies such as machine learning and artificial intelligence. Currently, smoking is a major health risk behavior that is affecting many aspects of the individual life including personal, community, and social factors. This problem is currently an epidemic that needs research, policy, and program resourcefulness [20]. That is, solutions to such problem might be supplemented by building models that classify smoking behaviors and thus enable health providers and researchers with tools to improve the wellbeing of youth. However, the effectiveness of these models relay on clinicians and researchers preciseness in the selection of factors that are intensely associated with the smoker behavior.Early research has no noteworthy development of automated depression level categorization from clinical datasets using machine learning classifiers. Thus, we aim at investigating the possibility of a decision support system that could contribute in classifying depression levels that youth smoker could experience. As such, we apply different techniques of machine learning to a dataset of youth smoker questionnaires, which were originally undertaken by the lead researcher of this work [20], to determine and detect depression level of Jordanian school students. The researchers hope to offer a tool that could enable and enhances the effectiveness of clinical decisions via developing a hybrid machine learning model.

The remainder of this paper is organized as follows. Section II analyses related work. Section III presents the methodology that we follow in this study. Section IV shows the obtained results of our study. Section V introduces the main threats to validity of this work followed by Section VI with the conclusion of our study and some ideas for future research.

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. More than a decade ago Demner-Fushman et.al [1] recommended that in order to move clinical and health care with the future it is imperative that both sides (i.e., Information Technology and health care) Particularly, our

- *Sukaina Alzyoud. Community & Mental Health Nursing. The Hashemite University, Zarqa, Jordan. E-mail: sukaina-alzyoud@hu.edu.jo*
- *Mohammad Kharabsheh. Computer Science. The Hashemite University, Zarqa, Jordan. E-mail: mohkh86@hu.edu.jo*
- *Rola Mudallal. Community & Mental Health Nursing. The Hashemite University, Zarqa, Jordan. E-mail: rula@hu.edu.jo*

team-up in developing fundamental natural language processing methods to provide computerized clinical decision support for health care. 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 is based on text and natural language processing techniques, and represents a framework for probable diagnostic decision support system for congenital anomalies. Jin et.al [3] developed a machine-learning approach for recognizing molecular entities to disease concepts in order to decide if the primary probabilistic model could be generalized to unrelated concepts within model retraining. The developed approach uses Conditional Random Fields trained using some domain-specific features. The approach achieves 0.85 precision, 0.83 recall, and 0.84 F-measure evaluation results. Skounakis et.al [4] applied machine-learning methods to gain gene-disorder relations [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 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 associated 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.To the best of our knowledge, our study is the first examination in the area of exploring the role of machine learning classifiers in identifying the depression level 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 each instance of our examined dataset, we extracted 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. **Error! Reference source not found.** summarizes the corpus information. We used the levels L1-L4 to refer to the depression levels 1 to 4 in the rest of the paper.

TABLE 1. SUMMARY OF CORPUS INFORMATION

Total number of instances	# L1	# L2	# L3	# L4
993	183	169	394	247

B. Classification Algorithms

In our experiments, we employed the supervised classifiers in which the inputted dataset is distributed into two groups: a training set and a test set. The training set is the one that is used to train the classifier, while the performance of the classifier is computed through the test set. 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 listed 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.*
- *Trees.RandomTree.*
- *Trees.J48.*
- *Bayesian Learner (Naïve Bayes).*
- *Sequential Minimal Optimization (SMO).*
- *Logistic Regression.*
- *K-Star.*
- *Decision Table.*
- *K-Nearest Neighbor (K-NN).*
- *IBk.*

C. Evaluation Metrics

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

- *Precision*: the ratio of retrieved instances which are relevant.*Recall*: the ratio of relevant instances

thatare retrieved by the classifier. *F-Measure*: is a metric epends on both recall andprecision of a model.

- *ROC*: it is the area under the Receiver Operating Characteristic (ROC) curve.

4 Study Results

Here, we present our obtained results from the undertaken classification experiments We developed machine learning classifiers that are based on the classification algorithm discussed before. When applying the classifiers, we employee the 10-fold cross validation technique in order to split the inputted dataset into training and test sets. As we discussed formerly, the performance of our classifiers are evaluated using the precision, recall, F-measure, and ROC metrics. Moreover, we trained our classifiers using the factors that are listed in **TABLE2**.

TABLE 2. SUMMARY OF CLASSIFICATION FACTORS

Classification Factor	Definition
Age	Participants age should be between 11-17 years old
Gender	Both Sexs
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 Level_1	Rarely or none of the time(less than once a day)
Deperssion Level_2	Some or a little of the time (1 to 2 days)
Deperssion Level_3	Occasionally (3 to 4 days)
Deperssion Level_4	Most or all of the time (5 to 7 days)

The performance results of our classifiers are shown in **Error! Reference source not found.**. That is, we would conclude that our models would accurately predict the depression levelOn the other hand, we noticed that SMV and KNN obtained better accuracy when compared with the other machine learning classifiers. For instance, KNN computes a possibility for each class based on the investigating 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, the SMV learner reaches better accuracy since it increases the dimensionality of inputted dataset till the instances are recognized in specific dimension. Furthermore, SMO is appropriate to work with large datasets and gains greater accuracy because the space usage required for SMV is linear in size [12]. Moreover, the similarity between the recall and precision values indicates that our created dataset can efficiently be used for prediction.

TABLE 3: CLASSIFICATION RESULTS OF **DEPRSSION LEVELS** USING MACHINE LEARNING ALGORITHMS

Learner	Accuracy	Recall	Precision	F-measure	ROC
RandomTree	0.823	0.823	0.823	0.823	0.906
J48	0.869	0.869	0.868	0.868	0.927
NaiveBayes	0.900	0.900	0.901	0.903	0.961
SMO	**0.944**	**0.944**	**0.944**	**0.944**	**0.954**
Logistic	0.945	0.945	0.945	0.945	0.978
IBK	0.921	0.921	0.921	0.929	0.952
KStar	0.759	0.759	0.759	0.753	0.951
DecisionTable	0.810	0.810	0.810	0.800	0.950

We need to assess the usefulness of each factor independently as a predictor of the depression level. 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 factors 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 inspects the structure of a developed decision tree to count the presence of each factor under the consideration, at each level of the tree [10]. Next, the tree level where a factor appears and the occurrence count of the factor are used together to decide the usefulness rank of that factor. Explicitly, the most influential factor would be the root node of the constructed decision tree. Additionally, the factors recognized as less important as we move down 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 employs the greedy divide and conquer approach on the training data to add decision nodes at each level of the constructed tree. The information observed from each attribute is computed, and next the attribute that gains the highest information is selected. This process is a repetitive task at each level of the tree up until the count of records in the leaf nodes equals a given cut-off value. The performance results obtained from our decision tree classifier is shown in **Table 4**. Additionally, the results of the Top Node analysis are illustrated **in Table 5.** Specifically, the table provides the factors that are in the first three levels (e.g., levels 0, 1, and 2) of the developed tree beside the occurrence count associated with each factor. As we could observe, the *age* is the most influential factor compared with the other factors under our consideration.

TABLE4: CLASSIFICATION RESULTS USING C4.5 ALGORITHM

Learner	Recall	Precision	F-Measure	ROC
C4.5	0.59	0.46	0.57	0.76

5 Threats to Validity

As with any study our work has its limitations, one of which not to generalize these results to datasets of none-smokers. We used datasets of 993 school students with an age range of 11 to 17 years, which will not represent students beyond this age range making our dataset not be demonstrative of all school students. Moreover, the current study did not include students

smoking cessation efforts and other psychological status such as anxiety which the researchers recommend to be included in future studies.In this study, our developed classifiers are based on machine learning techniques that were successfully and widely used in the literature. But, each of the used machine learner has its own drawbacks that might negatively impact the validation of the obtained results. Therefore, developing classifiers using other machine learners would be our consideration in the future.

6 Conclusion And Future Work

In this study, we investigated the effectiveness of machine learning models in predicting and categorizing the depression level of youth smoker. Here, we have accomplished a study on a dataset of 993 students, with an age range of 11 to 17 years, using a set of factors such as age, gender, and father's job. The work offers machine learning classifiers developed using support vector machine, nearest neighbor algorithm, Bayesian learner, and decision trees. Developed classifiers aim at predicting depression level of Jordanian students. The evaluation experiments show that the developed classification models have equitable accuracy with 94% recall in the best case and 76% recall in the worst case. On the other hand, a precision of 94% is reached in the best case and 76% in the worst case. By performing a Top Node analysis, we found that the age factor is the most influential attribute in predicting the depression level of youth smokers.In future, we plan to examine additional factors and explore the usefulness of other machine learning techniques in predicting depression levels with the aim of achieving better prediction performance. Also, we plan to extend our study by studying diverges datasets from other 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.