Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Prediction of Absenteeism at Work using Machine Learning Techniques

NAZMUL HASAN SHAIKOT[1], ABRAR RAIYAN[2], ATIK MAHMUD.[3], AND SHAMMAM AHMED[4]
Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
Email: (nazmul.hasan2, abrar.raiyan, atik.mahmud, shammam.ahmed)@northsouth.edu

ABSTRACT The success of any company depends on the labor and sincerity of the workers. This rate of labor is affected much by the absenteeism in work place. This paper looks into a case of absenteeism which will predict absenteeism in hours of the workers. Absenteeism will cause a very high role in reducing the production rate. So, it's important to solve this problem in a more scientific way. If the authority has the significant reasons behind the absenteeism, it's easier to solve this problem as it causes economic fall for any company or organization. It will help them to set some necessary steps which will reduce the rate of absenteeism. The novelty of this work is we used different techniques and algorithm to predict or target feature. We used train and testing as well as k-fold cross validation both which helps us to compare with each other precisely. We used our prepossessed data and applied several machine learning algorithm such as Decision Tree, KNN, Random Forest, SVM and ZeroR classifier. Overall, we got more than 90% accuracy. There are some features worked as very good predictors as well form our dataset.

INDEX TERMS Absenteeism, Decision Tree, Knn, Logistic Regression,Machine Learning,Prediction, Random Forest, SVM, ZeroR

I. INTRODUCTION

ALL the organizations consider their employees as a valuable asset as they are the main key to success of their organization. The rise and fall of an association mostly depend on the productivity of employees. It can negatively affect the production rate and profits. The economic sustainability and the revenue can only be achieved by the consistent labor of some industrious employees. One of the most important and effective indicator that can change the production and supply of any company or organization is absenteeism. If the number of employees are absent frequently in their specified work, this will reduce the quality and quantity of the services they provide to customers. Absenteeism can be defined in such a way that as an employee is habitually as well as frequently absent from work [1]. Generally, absenteeism is an important indicator to measure how regular they are in work. It is a very common phenomenon all over the world is facing in every sector of working. It is any failure to report on work. It can affect growth potential and overall performances [2]. So, it's much needed to figure out behind the reason of absenteeism so that those reasons can be solved to make the organization economically more stable. In this paper, we will present our prediction model of absenteeism using machine learning algorithm. This data we used is available in Kaggle containing 740 rows and 21 columns. This information was recorded through a real-life observation. We used several classifier to check accuracy where we got a significant accuracy score. The rest of this paper organized as follows: First of all we preprocessed our dataset. The next section is our literature review where reviewed some papers what they worked about and how they did. Then we wrote our methodology section and result. At last we analyzed our result and concluded our paper.

II. LITERATURE REVIEW

The absenteeism dataset is used in to investigate the cause of absence from work. Employee absence can reduce productivity in industries and organizations. Employees may be absent from work for a variety of reasons, including illness, depression, job dissatisfaction, and so on. The use of DM and ML approaches in human resource management is a relatively recent subject of study. Predicting and assessing workplace absenteeism is critical for proving a company's productive and lucrative capabilities.

Dogruyol et al [3] proved that using three neural network models and machine learning approaches, nonlinear issues such as forecasting absence in the workplace may be solved. Furthermore, a comparison of these three models is conducted to identify which one best fits this problem. Varalakshmi et al [4] studied work in this field centered on classification algorithms and highlighted the benefits of employing the Random Forest model for workforce prediction, demonstrating the value of absence data and its scope for workforce planning. Gayathri et al [5] demonstrated a comparison of three prediction models of the problem of absenteeism in a corporation using Prediction tree ANN, Nave Bayes – Classification. Demonstrating which of them is the best fit for the problem in terms of accuracy and mean quadratic error. Using Naive Bayes, Prediction tree, ANN, SVM – Regression, Olawale, O. et al. [6] used machine learning techniques to create four prediction models for absenteeism, demonstrating which of them had the best results. Furthermore, they emphasized the relevance of their use in any company's human resources department to weed out applicants who may stymie business flow due to a high absence rate. According to Wahid, Z. et al. [7], the key contribution of this paper is the investigation of four machine learning methods in the prediction of absenteeism and the comparison of these in terms of seven assessment criteria utilizing Prediction trees – Regression. The research that shows which of them achieves the best outcomes in solving this challenge in businesses. Asiri, A. et al [8] demonstrated the utility of human resources in employing computerized approaches such as auto-tuning apprentice techniques by implementing three prediction models based on these techniques that uncover the elements that predict absenteeism with high precision using Naive Bayes Prediction trees – Regression. Adaekalavan, S. et al [9] demonstrated that it is possible to take early efforts to resolve the absence problem by offering a methodology for predicting absenteeism at work using the K-means grouping algorithm, employing KNN, Nave Bayes, Prediction Trees – Grouping. The proposed methodology not only predicts absenteeism, but also recognizes the category of ab-senteeism in absentee, absent in days, ab-sent in months or absent in years Ali Shah et al [10] illustrated the application of a Deep Neural Network-based strategy for predicting absenteeism by comparing results from machine learning methods for the same goal using ANN, SVM Prediction Tree – Classification. The findings [11] of the examination of machine learning models for the prediction of absenteeism using KNN, Prediction Trees, and SVM – Grouping reveal the accuracy of the neural network in this problem and its usefulness in identifying the key psychosocial element that causes it. Nunung N. Qomariyah et al. [12] used a Decision Tree classifier to identify the unique characteristics of groups of employees who were absent frequently from work. During the years 2009 to 2011, they collected 14,400 records of employee attendance data from a private corporation in Jakarta, Indonesia. Based on the frequency of absence each month, the data was divided into three categories using HRD rules: "frequent absent employee," "rare absent employee," and "frequent present employee." With a test accuracy rate of 95.05 percent, they correctly identified 2936 occurrences as correct predictions and 153 occurrences as incorrect predictions. They discovered that a 33-39 year old female employee with at least three children and a working time of 12-14 years had more days off than other employees with similar characteristics. R. Schouteten et al. [13] employed logistic regression analysis to link workability, burnout, and job characteristics to absenteeism as an indicator of workplace health concerns. They performed a seven dimensional study of 242 university personnel in the Netherlands on workability, burnout, and job characteristics related to absenteeism data from the university's occupational health and safety database. It was discovered that rather than job qualities, "workers' own forecast of work abilities in two years from now," "mental resources/vitality," and "emotional tiredness" predict absenteeism. The less likely employees were to be absent in the next year, the better their own assessment of their work performance two years later. The mental resources and vitality dimension revealed that the more respondents enjoyed their work, felt fit, and had hope in the future, the lower their risk of exceptional absenteeism, whereas the more respondents who experienced emotional tiredness, the greater their risk of exceptional absenteeism. Majella J Albion et al. [14] provided a model of the links between organizational climate, psychological mediators, absenteeism, and leave intention. They used the Queensland Public Agency Staff Survey (QPASS) using IBM SPSS on 1097 employees of Queensland regional Health Service District (HSD) to conduct a statistical analysis to determine employee reactions to their work environment. The model identified a complex pattern in which psychological factors such as mood, stress, and fatigue drive psychological reactions and different types of withdrawal behavior. Absenteeism and turnover are the two most extreme forms. Only individual morale was found to have a significant relationship with absenteeism while the quality of work life, individual distress, individual morale, and job satisfaction all have significant relationships between with intentions to live. Another study by Shandizi in [15] predicts a pilot's absenteeism in an airline company. Crew costs are the second most important cost in the airline industry after fuel costs, and pilots are the most important airline crew. Having a system that can predict pilot absenteeism can help airline companies manage their operations. They are developing a decision support system that will use the Decision Tree algorithm to predict the number of hours a pilot will be absent and make the necessary arrangements to deal with the situation. This system is only applicable to pilot absenteeism and can be used in the airline industry.

Wahid et al. [16] used four machine learning techniques to predict when employees would be absent from work. They used a dataset obtained from a Brazilian courier company. They used the Random Forest technique, the Decision Tree technique, the Gradient Boosted Tree technique, and the Tree Ensemble technique. The best model was the Gradient Boosted Tree model, which had an accuracy of 82

Evandro.L et al. [17] used six machine learning techniques to predict absenteeism at work for a phone company employee in Brazil. Multilayer Perceptron, Naive Bayes, XG-Boost, Random Forest, Support Vector Machine, and Long Short Term Memory are among the techniques used. They created the models and used evolutionary algorithms to fine-tune their parameters. The model with the highest precision, 72 percent, was the XGBoost. They gathered information from 13.805 employees at the company. There are 241 attributes in the dataset.

III. RESEARCH METHODOLOGY

DATA ACQUISITION

Data acquisition is a process of collecting data to work on. We found our dataset in Kaggle. This dataset was made based on a factory workers for 3-4 years in the 2010.

PREPROCESSING

Preprocessing is a process which is applied in dataset to make it more suitable so that it can be trained well. It is done before training a model. Preprocessing is needed to transform the data into more suitable one. There are a little bit of missing values in our dataset. First of all, we replaced those missing values by the mean values. There are some features which are in continuous values. We converted them into discrete values zero or one using a process called discretization to apply the model [18]. It makes the values into a several classification values.

FEATURE SELECTION

There are more than 20 features in our dataset whereas all of them are not equally important and some of them are very less significant to be a good predictor. In this step, we removed some features from our dataset which are less important such as id, name, department etc. to predict absenteeism. Several techniques are used in this process. To find out the most significant features from our dataset, first we did feature selection. Feature selection make the dataset more suitable for applying machine learning algorithms. Feature selection decries the number of features and keeps those which are more important for predicting. We used cfssubseteval algorithm by using WEKA- a collection of machine learning algorithms. This algorithm drops some less significant column form the dataset. There are some other algorithms available in WEKA also.

NORMALIZATION

Normalization is another part of data preprocessing process. The goal of this process is to turn the dataset into a common scale. Sometimes, there different types of values from different scales are available in dataset which could affect the accuracy score and other scores as well. So, it's important to make all the values into same scales. Normally, there is much difference between two values in dataset. But after applying normalization, it will reduce the differences between values. It will set them from a range 0 to 1 which is easier to train a model. We converted our target feature into 0 and 1 two classes which was continuous before. We used MinMax scaler to scale the values.

MACHINE LEARNING ALGORITHM

In this section, we will discuss the machine learning algorithms that we have used in this research.

Logistic Regression : Logistic Regression considers as a mathematical modeling technique that describes the relationship between several independent variables, X1...XK, and a dependent variable, D. The logistic model uses the logistic function as a mathematical form which has the range between 0 and 1 for any given input. The logistic model can describe a probability of an event which is always a value between 0 and 1. The following formula represents the logistic model. (D = 1 | $\chi_1, \chi_2, \ldots, \chi_\kappa$) = 1 1 + e ($\alpha + \sum \beta_i \, \chi_i \, \kappa$ 1), Where α and β are the model's parameters that can be learned from a set of labeled instances in the training dataset. Gradient Descent Algorithm can be used to find the best values of the model's parameters during the training phase[19].

Support Vector Machine : SVM takes a set of input data and predicts, for each given input, which of the two possible classes comprises the input. By that, SVM can be represented as non-probabilistic binary linear classifier. Using training examples that are labeled to one of two categories, the SVM training algorithm creates a model that is used to assign the new examples to one category or the other category. The Support Vector Machine model can be viewed as a representation of the examples as points in the space, mapped so that the examples of the separate categories are divided by using clear gap. Making this gab as wide as possible. The new examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Decision Tree : The decision tree model can be represented as a graph which contains nodes and branches .there are two type of nodes which are known as internal node and leaf node. The test on the data set features is represented on the internal node of tree. The result of the test can be represented by the branch. The target can be represented by the leaf node. The node on the top of tree is known as the root node .decision tree is one of the most famous classification algorithms because the work with this algorithm does not require a previous knowledge of the problem and it does not require a tedious parameters' configurations. DT can be easily comprehended and easily transformed to classification rules. Decision tree algorithms have been used in number of applications such as medical applications, manufacturing production applications, financial analysis applications, molecular biology applications and astronomy applications[20].

ZeroR : ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category(class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. It constructs a frequency table for the target and select its most frequent value.

KNN : k-nearest-neighbors algorithm is a non-parametric classification method which is used for classification and regression In both cases, the input consists of the k closest training examples in dataset. The output depends on whether k-NN is used for classification or regression: In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

Random Forest : Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

METHOD

First, we took the dataset and then pre-processed the dataset. After pre-processing, the dataset was split into train and test sets. ML algorithms were applied on the training set to create ML models. The performance of the models were then evaluated using the test set from the dataset.

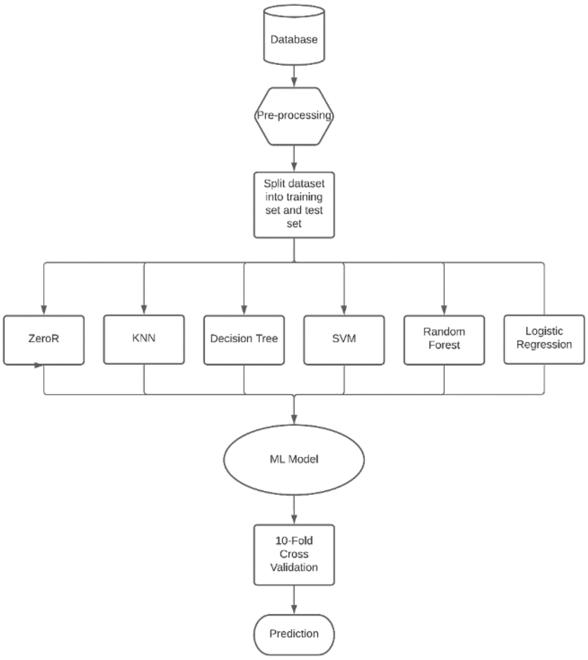
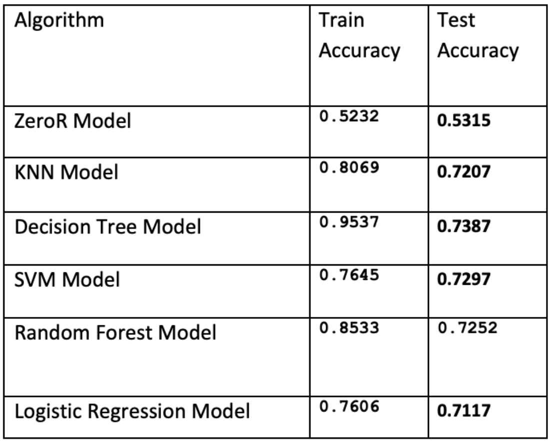


FIGURE 1. Flowchart for predicting absenteeism using Machine Learning

After applying these algorithms in the dataset, we found some test and train accuracy which is given below:



Algorithm	Train Accuracy	Test Accuracy
ZeroR Model	0.5232	**0.5315**
KNN Model	0.8069	**0.7207**
Decision Tree Model	0.9537	**0.7387**
SVM Model	0.7645	**0.7297**
Random Forest Model	0.8533	0.7252
Logistic Regression Model	0.7606	**0.7117**

FIGURE 2. Train and Test Accuracy

Here is the HeatMap we get:

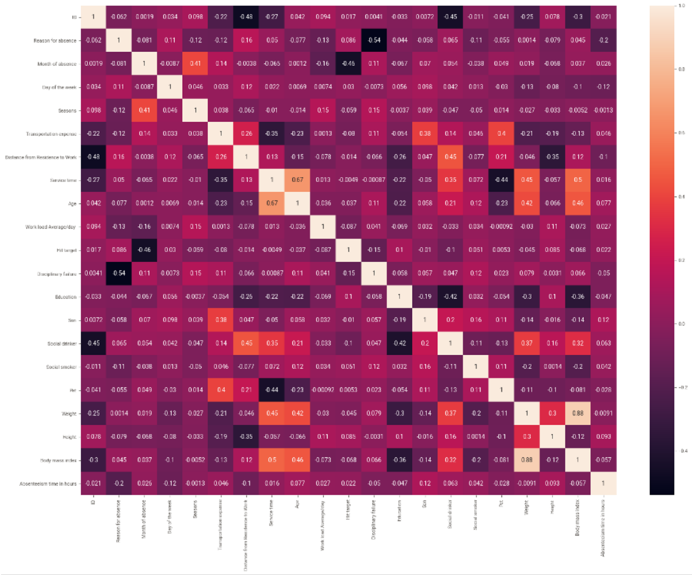


FIGURE 3. Correlation matrix of absenteeism Dataset

IV. RESULT

This section talks about the comparison and experimental methodology. We applied different types of models in our dataset to see which one suits well with high accuracy, precision and recall measurement. We applied machine learning algorithm without any preprocessing and with proper preprocessing both at a time. Then we figured out which one is more feasible. About six classification machine learning algorithms we applied in our dataset Decision Tree, Support Vector Machine, Random Forest etc. Here is our accuracy graph we got form our model. This accuracy is based on only train and test split method.

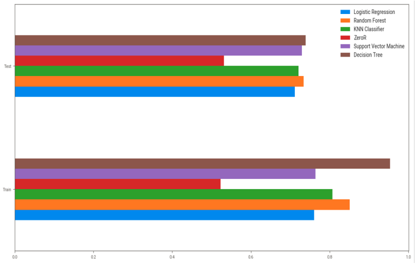


FIGURE 4. Accuracy Graph

In testing phase we got our highest accuracy using Decision Tree classifier as well as in training phase the highest accuracy is form Decision tree. However, there are some problems with train test split method that is why we also implemented K-Fold cross validation method. It will take all types of data from dataset which is distributed normally. We used different type of n values for cross validation but 10-fold cross validation works well for us.

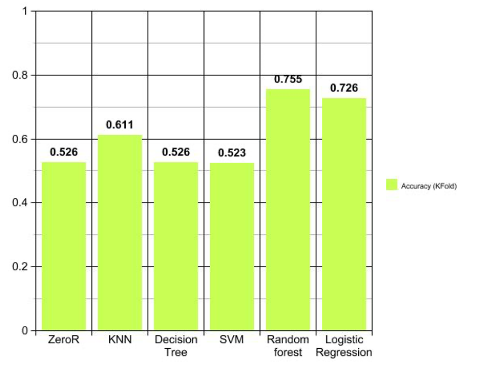


FIGURE 5. Accuracy using K-Fold

Roc curve is a performance measuring graph based on area under the curve. Here is the Roc Curve we got form our model using different classifier.

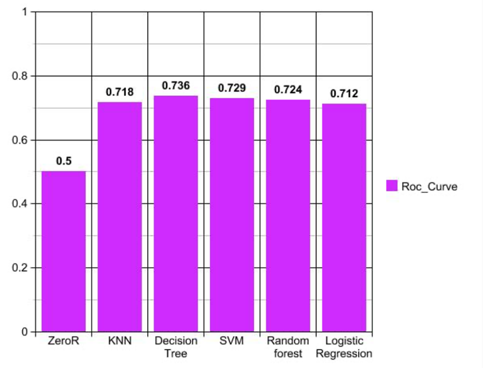
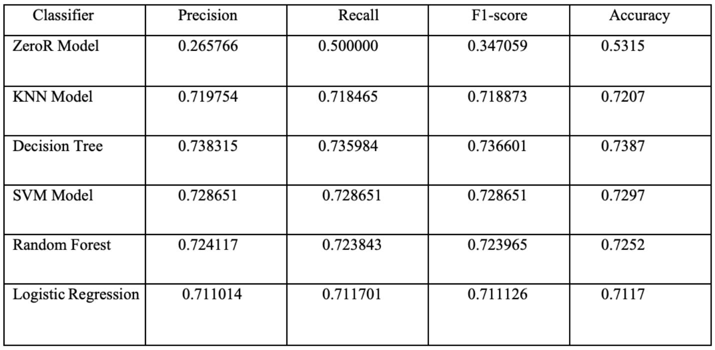


FIGURE 6. Roc Curve Accuracy graph

Here is our weighted average output we got from our applied model.



Classifier	Precision	Recall	F1-score	Accuracy
ZeroR Model	0.265766	0.500000	0.347059	0.5315
KNN Model	0.719754	0.718465	0.718873	0.7207
Decision Tree	0.738315	0.735984	0.736601	0.7387
SVM Model	0.728651	0.728651	0.728651	0.7297
Random Forest	0.724117	0.723843	0.723965	0.7252
Logistic Regression	0.711014	0.711701	0.711126	0.7117

FIGURE 7. Precision, Recall, F1-score, Accuracy Table

V. DISCUSSION

From our experimental results, we can see that if we apply train test split method, then Decision Tree classifier produces highest accuracy. It's also applicable for roc curve area where Decision tree covers more are under the roc curve which we can see from our fig no 6. But when we used K-Fold cross validation method, the accuracy score of Decision Tree decreased from the previous one. In K-Fold cross validation, Random Forest produces the highest accuracy. But overall, if we consider other accuracy also such as precision, recall or f-measure with accuracy score, Decision Tree is better one for our model

VI. CONCLUSION

Absenteeism at work has a negative impact on an organization's bottom line. Employers all over the world believe that employee absenteeism has a significant impact on company finances, morale, and other factors. They do not expect employees to be excessively absent from work, which reduces productivity and thus costs the company. Our goal in this study was to predict employee absenteeism at work using four tree-based machine learning algorithms: Decision Tree, KNN, Random Forest, SVM and ZeroR classifier. In the future, we hope to use feature engineering on the dataset to achieve the highest possible accuracy score for predicting employee absenteeism at work. By conducting a comparison study, we also intend to apply some other prominent machine learning algorithms to identify the best machine learning classifier that predicts absenteeism. The use of machine learning techniques is a good way to predict and analyze workplace absenteeism. These techniques make a significant contribution to an organization's human resource department. Management can gain a better understanding of employee activities and behavior, which can eventually aid in critical decision making regarding both monitoring employees at work and recruiting potential employees. This research could be used as a relevant tool for future investigations at other companies.

VII. REFERENCES

1. E. Collier, "Workplace Absenteeism," Reducing Absenteeism in the Workplace, 07 May 2018.

2. Andrew, "DETRIMENTAL EFFECTS OF EMPLOYEE ABSENTEEISM ON THE WORKPLACE," 26 5 2017.

3. Dogruyol, K., and Sekeroglu, B., Absenteeism prediction: a comparative study using machine learning models. In: International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions, pp. 728–734; Springer, 2019.

4. Varalakshmi, R., and Dhivya, R.S., A survey on big data applica-bility in prediction using absence information for workforce man-agement. Int. J. Recent Technol. Eng. (IJRTE). 7: 97–100, 2019.

5. Gayathri, T., Data mining of absentee data to increase productivity. Int. J. Eng. Tech. 4: 478–480, 2018.

6. Olawale, O., Exploration of absenteeism with machine learning, https://medium.com/@ojoolawalejulius2016/ exploration-of-absenteeism- with-machine-learning-1f01a8f9357e, last accessed 2020/03/21.

7. Wahid, Z., Satter, A. K. M. Z., Al Imran, A., and Bhuiyan, T., Predicting absenteeism at work using tree-based learners. In: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing - ICMLSC 2019, pp. 7–11: ACM Press, Da Lat, Viet Nam, 2019. https://doi.org/10.1145/3310986. 3310994.

8. Asiri, A., and Abdullah, M., Employees absenteeism factors based on data analysis and classification. Biosci. Biotechnol. Res. Commun. 12: 119–127, 2019. https://doi.org/10.21786/bbrc/12.1/14

9. Adaekalavan, S., Enhancing the prediction of absenteeism by deci-sion cluster based rule generation. Int. J. Comput. Sci. Eng. 7: 326– 330, 2019. https://doi.org/10.26438/ijcse/v7i5.326330.

10. Ali Shah, S. A., Uddin, I., Aziz, F., Ahmad, S., Al-Khasawneh, M. A., and Sharaf, M., An enhanced deep neural network for predicting workplace absenteeism. Complexity. 2020, 2020.

11. Priyanka, D., and Nayak, J., Empirical analysis of absenteeism at work place using machine learning. In: International Conference on Application of Robotics in Industry using Advanced Mechanisms, pp. 150–160: Springer, 2019.

12. Nunung N. Qomariyah, Yudho G. Sucahyo (2014). Employees attendance patterns prediction using classification algorithm case study: a private company in Indonesia. Intl Journal of Computing, Communications and Instrumentation Engg.(IJCCIE) Vol. 1, Issue 1(2014) ISSN 2349-1469 EISSN 2349-1477.

13. R. Schouteten, Institute for Management Research, Radboud University(2017). Predicting absenteeism: screening for work ability or burnout. Advance Access publication Occupational Medicine 2017;67:52–57

14. Majella J Albion et al (2008). Predicting absenteeism and turnover intentions in the health professions. Australian Health Review Vol 32 No.2

15. A. H. H. Shandizi, Prediction of Pilot's Absenteeism in an Airline Company, Universite De Montreal, Montreal, Canada, 2014.

16. Zaman.w, Abdullah.i, Zaidi.s, Touhid.b, 2019, Predicting Absenteeism at Work Using Tree-Based Learners , ICMLSC , 25–28

17. Evandro.L, Jos´e.M, Rui.S, Rafael.A,2019, Absenteeism Prediction in Call Center Using Machine Learning Algorithms, AISC 930, pp. 958–968

18. Stephanie, "Statistics How To," 7 January 2018. [Online]. Available: https://www.statisticshowto.com/discretization/.

19. David.G, Mitchel.K, Logistic Regression A Self-Learning Text, Third Edition, Springer.

20. Jiawei.H, Micheline.K, Jian.P, 2012, Data Mining Concepts and ,P .f, artificial neural Andréa,m, Edquel,b.p,Techniques, Third Edition, Morgan Kaufmann publications