International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 5, October 2021, pp. 4587~4596
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i5.pp4587-4596

Identification of important features and data mining classification techniques in predicting employee absenteeism at work

Amal Al-Rasheed
Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Kingdom of Saudi Arabia

Article Info

Article history:

Received Sep 21, 2020
Revised Mar 3, 2021
Accepted Mar 24, 2021

Keywords:

Absenteeism at work
Classification algorithms
Data mining
Feature selection
Prediction model

ABSTRACT

Employees absenteeism at the work costs organizations billions a year. Prediction of employees' absenteeism and the reasons behind their absence help organizations in reducing expenses and increasing productivity. Data mining turns the vast volume of human resources data into information that can help in decision-making and prediction. Although the selection of features is a critical step in data mining to enhance the efficiency of the final prediction, it is not yet known which method of feature selection is better. Therefore, this paper aims to compare the performance of three well-known feature selection methods in absenteeism prediction, which are relief-based feature selection, correlation-based feature selection and information-gain feature selection. In addition, this paper aims to find the best combination of feature selection method and data mining technique in enhancing the absenteeism prediction accuracy. Seven classification techniques were used as the prediction model. Additionally, cross-validation approach was utilized to assess the applied prediction models to have more realistic and reliable results. The used dataset was built at a courier company in Brazil with records of absenteeism at work. Regarding experimental results, correlation-based feature selection surpasses the other methods through the performance measurements. Furthermore, bagging classifier was the best-performing data mining technique when features were selected using correlation-based feature selection with an accuracy rate of (92%).

Corresponding Author:

Amal Al-Rasheed
Department of Information Systems
Princess Nourah bint Abdulrahman University
PO Box 84428, Riyadh, Kingdom of Saudi Arabia
Email: aaalrasheed@pnu.edu.sa

1. INTRODUCTION

Absenteeism at work can be described as a "habitual pattern of absence from a duty or obligation" [1]. This happens when employees do not show up or engage in events related either directly or indirectly to their jobs [2]. In general, absenteeism is believed to be a main indicator of poor performance [3]. Unpredicted absenteeism causes extra workload for other staff and reduces work efficiency. It also may results in low productivity and high direct and indirect costs [4]. It is therefore necessary for organizations that are heavily dependent on human resources to develop and implement absenteeism-prediction mechanism in order to help managers take preventative actions against the absence of employees to reduce financial costs [2], [5].

Organizations usually collect data about employees which could be used in improving decision-making processes. Data mining provides us a forum for predicting, analyzing, and grouping different problems of various genres without a subject matter expert. Data mining techniques are being steadily implemented now in different domains. One of the domains which still requires the interference of data mining is human resource management. Even though the pattern in human behavior is difficult to analyze, data mining techniques helps us to identify hidden and interesting pattern [6].

An in-depth analysis of the huge amount of data collected by organizations would take a long time and require a lot of human capital. When there is an abundance of irrelevant data, it is unlikely to be readily understood and absorbed [7]. Consequently, a very important issue for predicting human behavior, especially for predicting absenteeism at work, is how to filter and summarize a huge volume of data. As a preprocessing step, feature selection is one among the foremost essential steps in data mining process. It aims at filtering out the original data from irrelevant and redundant features [8], [9]. Irrelevant and redundant data inserted into a model could consume a great deal of cost and time and even reduce the degree of model accuracy [10], [11].

While there are many feature selection methods that might be used for absenteeism prediction, the paper's first research question is: what is the best method for enabling the prediction models to deliver the best performance?. The second research question is: what is the best combination of feature selection method and data mining technique to enhance absenteeism prediction accuracy?. In this paper, we take into consideration three feature selection methods to compare their prediction accuracy in absenteeism prediction, namely, relief-based feature selection, correlation-based feature selection and information-gain feature selection. To find the best combination of feature selection method and data mining technique to enhance the absenteeism prediction accuracy, seven classification techniques were used as the prediction model, namely, naive Bayes, logistic regression, multilayer perceptron, k-nearest neighbor, bagging, J48, and random forest. Additionally, cross-validation (CV) approach was used to evaluate the applied prediction models to have more realistic and reliable results. The dataset used was developed at a courier company in Brazil with records of absenteeism at work.

In previous works, researchers attempted to use different data mining techniques or merge different models to deal with the absenteeism prediction problem. Martiniano *et al.* [12], built a neuro-fuzzy network utilizing the error backpropagation algorithm with multilayer perceptron in absenteeism at work prediction. Another study was conducted by Nunung *et al.* to develop a decision tree classifier to discover the common features of employees who were regularly absent from the workplace [13]. In a recent study, Gayathri used naive Bayes, multilayer perceptron, and J48 classifiers to create a classification model to predict employee absenteeism for a short or long period of time [14]. Ferreira *et al.* [15], the researchers used artificial neural networks. In a similar research [1], the authors suggested the use of neural networks and deep learning algorithms to predict employees' behaviors regarding adherence at their workplace.

Literature shows the lack of studies that have applied data mining techniques in absenteeism at work prediction and indicates that the critical process of feature selection is not carefully considered. According to Dogruyol *et al.* no research has focused on finding the appropriate methods to compare and evaluate the performance of different data mining classification techniques while using particular combinations of features [4]. Thus, a thorough analysis is needed to test various feature selection methods to define the most relevant features and data mining classification techniques that will enhance the performance of prediction and ensure that the results are accurate and acceptable.

The contributions of this paper can be summarized as follows: help organizations in finding the main reason behind employees' absence to reduce expenses and increase productivity, improve the accuracy prediction, understand the best method of feature selection for efficient prediction of absenteeism, and identify the baseline feature selection method for relevant research in the future. The outline of the paper is as follows: Section 2 demonstrates the research method followed by this paper. Sections 3 and 4 illustrate the experimental results and discussions. The conclusion of the research is outlined in section 5.

2. RESEARCH METHOD

The primary objective of this research is to compare three well-known feature selection methods used in absenteeism prediction to examine their prediction performance and to find the best combination of feature selection method with data mining technique in enhancing the absenteeism prediction accuracy. There are two research questions that will be answered by this study:

- RQ1: What are the important algorithms of feature selection to predict the absenteeism of employees at work?
- RQ2: what is the best combination of feature selection method and data mining technique in enhancing absenteeism prediction accuracy?

In order to perform the research goal and to address the abovementioned research questions, we began the process of data mining with data preprocessing, followed by feature extraction and then classification modeling. Classification modeling was repeated for the combination of attributes selected by each feature selection method. During each iteration, the output of each developed model was documented, depending on the selected features and the data mining techniques, and the results were presented after the full process had been completed. In this study we used 10-fold cross-validation because experimental results of previous studies proved that the optimum number of folds seems to be 10, as it optimizes the required time to perform the test and at the same time reduces the bias correlated with the validation process [16].

2.1. Dataset description

The used dataset in this research consists of records of employees' absenteeism at a courier company in Brazil. These records were collected from July 2007 to July 2010 and were later made available at the UCI machine learning repository [12]. The dataset consists of 740 records with 21 attributes. The attributes and their descriptions are shown in Table 1.

Table 1. Absenteeism-at-work dataset

No	Attribute	Description and Values
1	ID	Individual identification number
2	Reason for absence	Absences recorded by the International Code of Diseases (ICD) classified into 21 categories
3	Month of absence	Number represents real months (1-12)
4	Day of the week	Monday (2) Tuesday (3) Wednesday (4) Thursday (5) Friday (6)
5	Seasons	Spring (1) Summer (2) Autumn (3) Winter (4)
6	Transportation expense	Number
7	Distance from residence to work	In kilometers
8	Service time	In hours
9	Age	In years
10	Workload average/day	In hours
11	Hit target	Number
12	Disciplinary failure	Yes (1) No (2)
13	Education	High school (1) Graduate (2) Postgraduate (3) Master and Doctor (4)
14	Son	Number of children
15	Social drinker	Yes (1) No (0)
16	Social smoker	Yes (1) No (0)
17	Pet	Number of pets
18	Weight	Integer
19	Height	Integer
20	Body mass index	Integer
21	Absenteeism time in hours	Number of absenteeism hours

2.2. Data preprocessing

Data preprocessing is a significant step in data mining applications. Data collection methods are often poorly regulated, resulting in missing values, out-of-range values, and unlikely data combinations. These problems can result in misleading findings if they are not examined at the beginning of the data mining process. Therefore, the representation and quality of the data should be addressed before the analysis is carried out [17]. In reality, most of the time required by data processing is spent creating data mining applications [18].

The absenteeism-at-work dataset was structured with no missing values. However, through careful examination of the dataset, we found that we needed to do dimension reduction and grouping for some attributes. Since the attribute ID does not has influence on absenteeism, we removed it from the attributes list. As a result, twenty attributes were selected after the data was cleaned. We noticed that the values of some attributes in the dataset were scattered, which would make getting good prediction results difficult and

complicated. We found that the solution lay in grouping values of certain attributes in order to improve prediction results. Grouping values is necessary when the number of values of an attribute is too high, since dealing with each value individually can lead to problems during computation and interpretation [19]. Table 2 illustrates grouping categories and values of some attributes as presented in [11].

Table 2. Grouping some attributes of the absenteeism-at-work dataset

Attribute	Category	Values
Transportation expense	Cheap	100–200
	Expensive	200–300
	Very expensive	>300
Distance to work	Close	0–15
	Far	15–35
	Very far	>35
Age	Young	25–35
	Mid age	36–45
	Old	>45
Body mass index	Underweight	<18.5
	Normal weight	18.5–24.9
	Overweight	25–29.9
	Obesity	>=30
Absenteeism time in hours	No absence	0
	Moderate absence	1–10
	High absence	>10

2.3. Feature selection algorithms

Most real-world data contain more details than what is required to build a model. Such redundant details make extracting the most significant information more difficult [8]. Feature selection is the process of selecting the most important and most relevant features of a dataset [20], [21]. Feature selection enhances the performance of the prediction model, makes the modeling process more efficient, and provides better understanding of the data [9].

There are many feature selection methods available in literature. After this study's experiment, we used three methods to discover the most influential attributes-namely, correlation-based, information-gain, and relief. While the correlation-based feature selection is a greedy search method, the others are rank-based search methods [22]. By using these methods, we have identified ten attributes as the most influential features. The selected ten features were used in building prediction models, while the rest were omitted.

Relief-based feature selection (RFS). This method allocates weights to all dataset features, and these weights can be modified over time. The most-relevant features have a high weight value, and the rest of the features have low weights. The techniques used by relief are the same as those used by k-nearest neighbor, which assigns weights to features [23].

Correlation-based feature selection (CFS). This method ranks the subset of features in accordance to their association with other features and the label of class. Subsets of features that demonstrate robust association with the label of class and low association with other features are assigned a higher value. This method is considered multivariate, as it removes all the redundant and irrelevant features from the dataset [24].

Information-gain feature selection (IGFS). This method ranks the subset of features according to high information gain entropy in descending order. This algorithm specifies a threshold value, and the attributes whose values exceed the threshold require additional processing [25].

2.4. Experimental setup

In this study, we have chosen to perform the experiments using Waikato environment for knowledge analysis (WEKA). WEKA is a commonly used data mining method that applies most data mining techniques and provides visualization of the results [26]. WEKA offers a powerful and user-friendly visual design environment for creating and testing various feature selection and prediction models.

3. EXPERIMENTAL RESULTS

We conducted the experiment in three phases. First, we examined the performance of various data mining algorithms such as naive Bayes, logistic regression, multilayer perceptron, k-nearest neighbor, bagging, J48, and random forest on full features of the absenteeism-at-work dataset. Second, we used three feature selection algorithms-relief, correlation-based, and information-gain-to select important features.

Third, we checked the performance of classifiers on the selected features. The effectiveness of each classifier was evaluated in terms of accuracy, precision, recall, and time to build the model. We used 10-fold cross-validation because the number of selected features was small [27].

3.1. Experiment 1: Comparison of classifiers performance on full features (n=19)

For this experiment, cross-validation was used to examine the performance of the seven data mining classifiers on the full features of the dataset where specified values of parameters were passed over the classifiers. For this experiment, cross-validation was used to check the performance of the seven data mining classifiers on the full features of the dataset were specified values of parameters were passed over the classifiers. In Table 3, the random forest shows good performance that has 91% classification accuracy, 88% precision, and 91% recall.

Table 3. The performance evaluation of different classifiers on full features of Absenteeism at work dataset.

	Metrics of classifiers performance evaluation			
Predictive model	Accuracy	Precision	Recall	Processing time (s)
Naive Bayes	88%	87%	88%	0.01
Logistic Regression (C=10)	89%	86%	90%	0.24
Multilayer Perceptron (13, 16, 2)	88%	86%	88%	15.35
K-Nearest Neighbor (K-NN, K=7)	88%	89%	88%	0.01
Bagging	91%	88%	91%	0.62
J48	90%	87%	90%	0.08
Random Forest (50)	91%	88%	91%	0.25

3.2. Experiment 2: Comparisons of feature selection methods

Three feature selection methods were chosen in this experiment to select the most relevant features to be used with the classifiers and compare their performance. We conducted experiments on different number of selected features, but in our simulation results we recorded the performance of classifiers on only 10 features as we found that the performance of classifiers on 10 features was very good:

3.2.1. Results of relief-based feature-selection algorithm

The relief-based feature-selection algorithm designates weights to features and selects significant features based on their weights [28]. According to the results, disciplinary failure and reason for absence were the most important features selected by relief for the prediction of absenteeism at work. Figure 1 displays weights assigned to all features using relief and Table 4 shows the selected significant features.

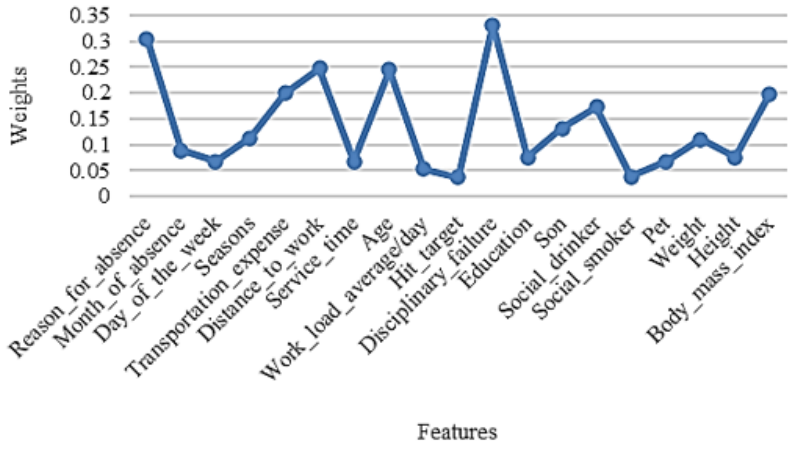

Figure 1. Weights assigned to features using relief FS

3.2.2. Results of correlation-based feature-selection algorithm

The CFS approach suggests that important features show a strong correlation with the label of class and a low correlation with other features, so a high weight should be given to such features [29]. After examining the results, we found that disciplinary failure and reason for absence were the most important

features selected by CFS for the prediction of absenteeism at work. Figure 2 displays weights assigned to all features using CFS and Table 5 shows the selected significant features.

Table 4. Features selected by the relief algorithm and their rankings

Order	Feature	Feature name	Scores
1	11	Disciplinary failure	0.33
2	1	Reason of absence	0.30
3	6	Distance from residence to work	0.25
4	8	Age	0.24
5	5	Transportation expense	0.20
6	19	Body mass index	0.20
7	14	Social drinker	0.17
8	13	Son	0.13
9	4	Seasons	0.11
10	17	Weight	0.11

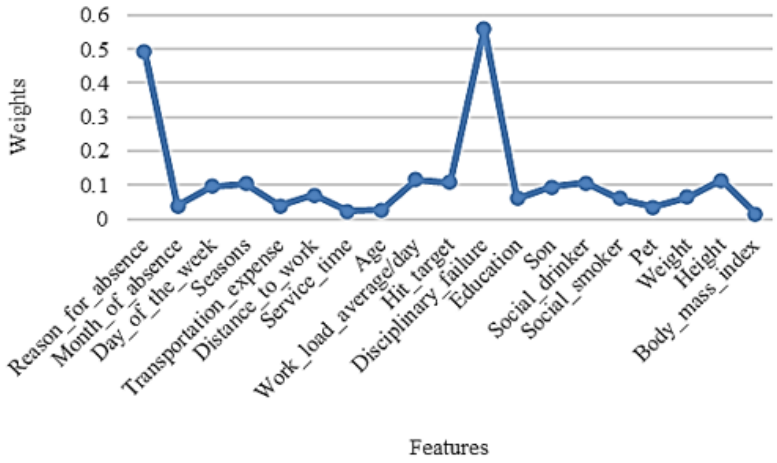

Figure 2. Weights assigned to features using CFS

Table 5. Features selected by CFS and their rankings

Order	Feature	Feature name	Scores
1	11	Disciplinary failure	0.56
2	1	Reason of absence	0.49
3	9	Workload average/day	0.12
4	18	Height	0.11
5	10	Hit target	0.11
6	14	Social drinker	0.11
7	4	Seasons	0.10
8	3	Day of the week	0.10
9	13	Son	0.09
10	6	Distance from residence to work	0.07

3.2.3. Results of information-gain feature-selection algorithm

IGSF ranks features in descending order, depending on the high information entropy gain. When examining the results, we found that reason for absence and disciplinary failure were the most important features selected by IGSF for the prediction of absenteeism at work. Figure 3 displays weights assigned to all features using IGSF and Table 6 shows the selected significant features. Tables 4-6 show the significant features selected by three feature-selection algorithms for the prediction of absenteeism at work. The first features selected by CFS have high scores, which means that CFS features have a strong influence on the prediction of absenteeism at work.

3.3. Experiment 3: Comparison of classifiers' performance on selected features (n=10)

In this experiment, we applied seven classification methods on the 10 attributes selected from the main dataset. We repeated the experiment with each feature-selection method. Once a predictive model was

developed, we could test how effective it was. For testing the predictive models, we compared the measures of accuracy for data mining algorithms with each feature-selection method. The results from different classifiers are presented in Tables 7-9. The results generated with classifiers optimized by the CFS method were the best reported results. The bagging model showed the best results compared to the other classifier algorithms, with an accuracy of 92%, precision of 90%, and recall of 92%. Logistic regression, k-nearest neighbor, and J48 classifiers achieved a competitive result, with accuracies of 91% when features were selected using the CFS extraction method. Although the performance of using random forest with relief-based and information-gain feature selection methods could achieve a high result with an accuracy of 90%, the achieved performance is still lower than when using the classifier with the full dataset. This is because random forest is great with high dimensional data, and it provides estimates of the important variables in the classification [30]. Training neural network models (multilayer perceptron) takes more time than training other data mining models [31]. Generally, the main issues related to inductive learning are the time required to build the model and the accuracy of classifying new examples [32].

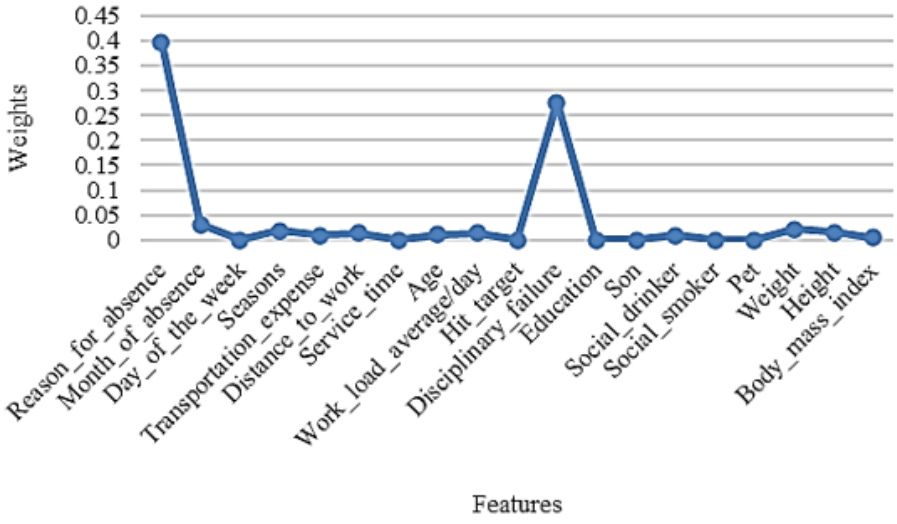

Figure 3. Weights assigned to features using IGFS

Table 6. Features selected by IGFS and their rankings

Order	Feature	Feature name	Scores
1	1	Reason for absence	0.40
2	11	Disciplinary failure	0.28
3	2	Month of absence	0.03
4	17	Weight	0.02
5	4	Seasons	0.02
6	14	Height	0.02
7	9	Workload average/day	0.01
8	6	Distance from residence to work	0.01
9	8	Age	0.01
10	14	Social drinker	0.01

Table 7. Performance evaluation of different classifiers on the absenteeism-at-work dataset using 10 features selected by the relief-based feature-selection algorithm

	Classifiers' performance evaluation metrics			
Predictive model	Accuracy	Precision	Recall	Processing time (s)
Naive Bayes	90%	84%	90%	0.01
Logistic regression (C=10)	90%	86%	90%	0.13
Multilayer perceptron (13, 16, 2)	87%	85%	87%	13.12
K-nearest neighbor (K-NN, K = 7)	87%	83%	87%	0.01
Bagging	90%	88%	90%	0.27
J48	91%	84%	91%	0.02
Random forest (50)	90%	87%	90%	0.34

Table 8. Performance evaluation of different classifiers on the absenteeism-at-work dataset using 10 features selected by the correlation-based feature-selection algorithm

	Classifiers' performance evaluation metrics			
Predictive model	Accuracy	Precision	Recall	Processing time (s)
Naive Bayes	89%	87%	89%	1%
Logistic regression (C=10)	91%	87%	91%	0.11
Multilayer perceptron (13, 16, 2)	88%	85%	88%	10.41
K-nearest neighbor (K-NN, K=7)	91%	90%	91%	0.01
Bagging	92%	90%	92%	0.36
J48	91%	87%	91%	0.02
Random forest (50)	91%	88%	91%	0.28

Table 9. Performance evaluation of different classifiers on the absenteeism-at-work dataset using 10 features selected by the information-gain feature-selection algorithm

	Classifiers' performance evaluation metrics			
Predictive model	Accuracy	Precision	Recall	Processing time (s)
Naive Bayes	88%	87%	88%	0.01
Logistic Regression (C=10)	90%	86%	90%	0.14
Multilayer perceptron (13, 16, 2)	88%	83%	88%	11.3
K-Nearest Neighbor (K-NN, K=7)	91%	89%	91%	0.01
Bagging	91%	88%	91%	0.38
J48	90%	86%	90%	0.03
Random Forest (50)	90%	86%	90%	0.22

4. DISCUSSION

In this paper, we applied data mining algorithms on the absenteeism at work dataset to predict absenteeism hours, based on the data of employees' attributes. Our aim was to compare various classification models using some feature-extraction methods and to identify the most effective model. From the tables above, we note that various algorithms performed better based on which feature-selection method was used. Each algorithm is capable of outperforming another algorithm based on the situation. For example, random forest performs better with a large number of features than with a subset of features, while bagging performs better with a select number of features. In general, the efficiency of the algorithms increased by applying feature-selection methods, and the correlation-based feature-selection method gave the best results. This demonstrates the need for feature selection before a classification is applied to the date. After applying feature-selection methods, we used performance metrics to compare the different data mining algorithms, since this is a standard process in evaluating algorithms. Without the optimization of feature selection, the best average precision value was from k-nearest neighbor with 89%. After optimizing by using the correlation-based feature-extraction method, we found the best precision was from bagging and k-nearest neighbor with 90%. Finally, when we compared the accuracy of the various classification algorithms with the feature-selection methods, we found that the best one was bagging with 92% accuracy.

5. CONCLUSION

Absenteeism at work is perceived as one of the most significant problems for organizations, as it may raise their expenses and pose a barrier to the accomplishment of organizational goals and priorities. It is important to build and incorporate methods for predicting absenteeism at work in organizations to enable management to take action against the shortage of employees and reduce financial costs. The goal of this study was to identify important features and the best-performing classification methods that enhance the accuracy of absenteeism-at-work prediction. We took advantage of feature-selection methods with a goal of enhancing the quality of absenteeism prediction. An experiment was first conducted to find the high-influence attributes. Then, classification was performed based on different classification algorithms such as naive Bayes, logistic regression, multilayer perceptron, k-nearest neighbor, bagging, J48, and random forest. The experimental results reconfirmed the significance of the selected features. Additionally, among the top seven methods, bagging has outperformed the other methods with 92% accuracy when features were selected using CFS. This research could be improved in the future in several ways.

There are many ways to enhance this research and address the limitations of this study. The same experiment could be carried out on a broad scale with real-world datasets to expand this work and generalize the findings. In addition, the performance of other data mining techniques in absenteeism prediction could be tested. Furthermore, new methods for selecting features may be used to obtain a broader perspective on the critical features used to enhance prediction accuracy.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah Bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

[1] S. Shah, I. Uddin, F. Aziz, S. Ahmad, M. Al-Khasawneh, and M. Sharaf, "An enhanced deep neural network for predicting workplace absenteeism," *Complexity*, vol. 2, pp. 1-12, 2020, doi: 10.1155/2020/5843932.

[2] V. Araujo, T. Rezende, A. Guimarães, V. Araujo e P. Souza, "A hybrid approach of intelligent systems to help predict absenteeism at work in companies," *SN Applied Sciences*, vol. 1, no. 6, 2019, Art. No. 536, doi: 10.1007/s42452-019-0536-y.

[3] A. Cohen, and R. Golan, "Predicting absenteeism and turnover intentions by past absenteeism and work attitudes," *Career Development International*, vol. 12, no. 5, pp. 416-432, 2007, doi: 10.1108/13620430710773745.

[4] K. Dogruyol, and B. Sekeroglu, "Absenteeism Prediction: A Comparative Study Using Machine Learning Models," *International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions*, 2020, pp. 728-734, doi: 10.1007/978-3-030-35249-3_94.

[5] E. L. de Oliveira, J. Torres, R. Moreira, and R. Lima, "Absenteeism prediction in call center using machine learning algorithms," *em World Conference on Information Systems and Technologies*, 2019, pp. 958-968, doi: 10.1007/978-3-030-16181-1_90.

[6] Z. Wahid, Z. Satter, A. Imran, and T. Bhuiyan, "Predicting absenteeism at work using tree-based learners," *3rd International Conference on Machine Learning and Soft Computing (ICMLSC 2019)*, 2019, pp. 7-11, doi: 10.1145/3310986.3310994.

[7] R. Mythily, and D. Mavaluru, "An efficient feature selection algorithm for health care data analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 3, pp. 877-885, 2020, doi: 10.11591/eei.v9i3.1744.

[8] A. Silva, "Feature Selection," Berlin, Germany, *Springer*, vol. 13, pp. 1-13, 2015, doi: 10.1007/978-981-287-411-5_2.

[9] G. Chandrashekar, and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[10] R. Kamala, and R. Thangaiah, "An improved hybrid feature selection method for huge dimensional datasets," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 1, pp. 77-86, 2019, doi: 10.11591/ijai.v8.i1.pp77-86.

[11] A. Asiri, and M. Abdullah, "Employees Absenteeism Factors Based on Data Analysis and Classification," *Special Issue in Communication and Information Technology*, vol. 12, no. 1, pp. 119-127, 2019, doi: 10.21786/bbrc/12.1/14.

[12] A. Martiniano, R. Ferreira, R. Sassi, and C. Affonso, "Application of a neuro fuzzy network in prediction of absenteeism at work," *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*, IEEE, Madrid, Spain, 2012, pp. 1-4.

[13] N. N. Qomariyah, and Y. G. Sucahyo, "Employees' attendance patterns prediction using classification algorithm case study: A private company in Indonesia," *International Journal of Computing, Communications and Instrumentation Engineering (IJCCIE)*, vol. 1, no. 1, pp. 2349-1477, 2014.

[14] T. Gayathri, "Data mining of absentee data to increase productivity," *International Journal of Engineering and Techniques*, vol. 4, no. 3, pp. 478-480, 2018.

[15] R. Ferreira, A. Martiniano, D. Napolitano, E. Farias, and R. Sassi, "Artificial neural network and their application in the prediction of absenteeism at work," *International Journal of Recent Scientific Research*, vol. 9, no. 1, pp. 2332-2334, 2018, doi: 10.24327/ijrsr.2018.0901.1447.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Ijcai*, vol. 4, no. 2, pp. 137-1145, 1995.

[17] D. Pyle, "Data preparation for data mining," Morgan Kaufmann, 1999.

[18] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques," *Elsevier*, 2011, doi: 10.1016/C2009-0-61819-5.

[19] P. Berka, and I. Bruha, "Discretization and Grouping: Preprocessing Steps for Data Mining," *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98*, Nantes, France, 1998.

[20] N. Kwak and Chong-Ho Choi, "Input feature selection for classification problems," in *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143-159, Jan. 2002, doi: 10.1109/72.977291.

[21] H. Esmaeel, "Analysis of classification learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 2, pp. 1029-1039, 2020, doi: 10.11591/ijeecs.v17.i2.pp1029-1039.

[22] A. Zakrani, M. Hain, and A. Idri, "Improving Software Development effort estimating using Support Vector Regression and Feature Selection," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 399-410, 2019, doi: 10.11591/ijai.v8.i4.pp399-410.

[23] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1-45, 2017, doi: 10.1145/3136625.

[24] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.

[25] M. Hall, and L. Smith, "Practical feature subset selection for machine learning," *Springer*, 1998.

[26] G. Holmes, A. Donkin e Ian Witten, "Weka: A machine learning workbench," In *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference, IEEE*, 1994, pp. 357-361, doi: 10.1109/ANZIIS.1994.396988.

[27] Z. Yang, and M. Zhou, "Kappa statistic for clustered matched-pair data," *Statistics in medicine*, vol. 33, no. 15, pp. 2612-2633, 2014, doi: 10.1002/sim.6113.
[28] U. wicz, M. Meeker, W. Cava, R. Olson, and J. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189-203, 2018, doi: 10.1016/j.jbi.2018.07.014.
[29] A. Wosiak, and D. Zakrzewska, "Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis," *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/2520706.
[30] M. Zakariah, "Classification of large datasets using Random Forest Algorithm in various applications: Survey," *International Journal of Engineering and Innovative Technology (IJJEIT)*, vol. 4, no. 3, pp. 189-198, 2014.
[31] N. Cuppens-Bouhlahia, F. Cuppens e J. Alfaro, "Data and Applications Security and Privacy XXVI," *26th Annual IFIP WG 11.3 Conference Springer*, 2012.
[32] J. Shavlik, R. Mooney, and G. Towell, "Symbolic and Neural Learning Algorithms: An Experimental Comparison," *Machine learning*, vol. 6, no. 2, pp. 111-143, 1991, doi: 10.1023/A:1022602303196.