Classifying Relevant Causes of Employee Absenteeism Using Machine Learning

I. Fosić*, A. Živković*, I. Fosić**
*Josip Juraj Strossmayer University of Osijek, Faculty of Economics in Osijek, Osijek, Croatia
**HEP-Telekomunikacije PS Osijek, Osijek, Croatia
ivana.fosic@efos.hr, ana.zivkovic@efos.hr, igor.fosic@hep.hr

***Abstract*: Employee absenteeism has a significant impact on the organization as it imposes high costs that can be reflected in the organization's economic results. The impact can be seen in productivity, sustainability, competitiveness, profitability, and inter-organizational relations. Considering the wide range of causes of absenteeism, this paper provides a systematic overview of the causes considering different theoretical approaches. A primary survey was conducted (N=420), and the causes of absenteeism were divided into two groups of causes (personal causes and organizational causes). Unsupervised and supervised machine learning methods were applied to the data set. In the unsupervised machine learning phase, the Elbow data grouping method was applied, and the K-Means algorithm was used to assign each record in the dataset to a specific cluster. In supervised machine learning, the data were divided into learning and testing parts in a 70/30 ratio, and the Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), and k-Nearest Neighbors (K-NN) classification algorithms were used. The best performance and area under the ROC curve (AUC) values are obtained using the RF classification algorithm. Using the machine learning algorithm RF, the main causes of absenteeism in both groups of causes of absenteeism were identified. From the group of personal causes (personal illness as well as illness of family members), from the group of organizational causes (mobbing, conflict with work colleagues) stand out the most important causes of absenteeism.**

Keywords: causes of absenteeism, employee absenteeism, unsupervised and supervised machine learningrt

I. INTRODUCTION

The absenteeism rate in Croatia, as in the countries of the European Union, varies over time. It depends primarily on the industry and the size of the business entity, as well as on various factors such as the economic situation, legislation and the environment. According to Eurostat data [1] on absenteeism in 2021, Croatia is not behind the European Union average. In Croatia, 10.6% of employees were absent from work in the first quarter (EU=9.4%), 8.1% in the second quarter (EU=7.6%) and 14% in the third quarter (EU=15%). The number of absences in 2021 is approaching the average for the period 2015 to 2019, while 2020 is characterized by numerous measures taken to contain the spread of COVID - 19, which consequently had a significant impact on the increase in absenteeism.

Frequent worker absences negatively impact the cost and productivity of the business system, which can result in the end product or service being uncompetitive [2]. Absenteeism represents a significant cost to employers, and according to Absence Insight [3] , it is the second largest cost after salary costs, suggesting that significant savings in total labor costs can be achieved by addressing this issue, as much as 15%.

The aim of this work is to identify which causes of absenteeism from the group of organizational and personal causes play the most important role in absenteeism, taking into account and highlighting the demographic characteristics of these workers. The above findings can be used in efforts to reduce employee absenteeism to the lowest possible level.

This thesis follows the following concept: after introductory remarks highlighting the aim of the thesis, Chapter 2 provides an overview of the theoretical basis and previous research. Chapter 3 describes the methodology used and the data analysis, including the presentation of the results of the research conducted. Finally, Chapter 4 draws conclusions, identifies the limitations of the research, and provides recommendations for future research.

II. THEORETICAL BASIS AND PREVIOUS RESEARCH

Workplace absenteeism or absenteeism is an important issue for all organizational stakeholders, not just HR professionals. Any justified or unjustified absence of an employee from the workplace is called absenteeism. Absenteeism is the absence of an employee from work, and the main causes are sick leave and abuse of sick leave by employees [2]. The slowness of public health procedures, the lack of an absence management process in organizations, a negative organizational culture, and a rigid labor law framework often lead to unnecessary costs [2]. The author also notes that sick leave is the most prevalent form of absenteeism, but that there are other forms, such as being late to work, taking frequent breaks, manipulating leave, and "apparent attendance." Absenteeism should be looked at from different angles to understand its impact at different levels: from the social to the individual level. In general, absenteeism is a major problem in companies, and a distinction must be made between regulated and unregulated absenteeism [4]. Worker absenteeism has negative consequences for the operation of the company, relationships between colleagues, and the ability to work as a team, and affects workers' families and the environment [5]. Identifying the causes of absenteeism allows for better management, recognition of uncontrolled, partially controlled, or fully controlled forms of absenteeism [5].

Ilić [2] highlights that employers often do not consider the indirect costs of absenteeism, such as the cost of a replacement employee, administrative costs, and costs associated with a drop-in productivity, which are difficult to assess without specialized knowledge and skills. For this reason, the costs of absenteeism are generally explained as unmeasurable, minor, and difficult to manage without affecting the organization and cost of work. In order to make

decisions on reducing absenteeism, data on the prevalence of absenteeism in the organization is needed, which is why it is important to have a system for detecting and recording absenteeism and analyzing trends. To achieve this, it is necessary to develop and implement a system for recording absenteeism, including time measurement [5]. Systematic and objective monitoring of absences can provide the organization with useful data on whether absences are above allowable and acceptable levels and whether they are characteristic of particular jobs and categories of workers. Absence data can provide clues to possible causes and factors of absenteeism. However, in order to take more serious action against absenteeism, the causes and conditions for its occurrence and increase must be accurately identified [6].

According to Čikeš and others, [7] the determinants of absenteeism are divided into several categories depending on their nature: Attitudes, personal, demographic, health, and organizational factors. The authors emphasize that understanding absenteeism behavior begins with understanding its determinants and outcomes.

Different authors have different approaches to classifying the factors that influence absenteeism and its consequences. The basic division of causes of absenteeism for the purposes of this paper is between personal and organizational causes of absenteeism.

Personal causes of absenteeism include absenteeism due to: personal illness, family issues, personal needs and problems, sense of entitlement, stress [8] but also illness of family members, own laziness, own work done by the employee in addition to his/her current job, due to difficulties in going to the workplace, social events during working hours, medical examinations, and problems with alcohol and/or drugs. As mentioned earlier, sick leave is one of the most common causes of absenteeism, while according to Kocakulah et al [9], family problems play an important role in absenteeism. Balancing work and family life can be difficult, especially because of the high cost of child care. One of the most common family problems is that adult offspring have the duty of caring for their elderly parents, which means that they have additional obligations and time constraints to meet their parents' needs, such as doctor's appointments or hospitalizations that require additional time from workers [9]. It can be very difficult to recognise alcohol or drug abuse in the workplace, and alcohol and drug dependence can lead to incapacity and absenteeism [10]. It is desirable for organizations to have a substance abuse policy in place to adequately address this problem [10]. Scheduling a necessary doctor's appointment is often a challenge for employees because they don't want it to interfere with their scheduled workday. This poses a problem for both the employee and the employer because time is lost that cannot be compensated for [9]. Workplace health promotion and education, in addition to laws and regulations, are very important and are achieved through the cooperation of employers, employees and society to improve the health and well-being of workers. In order to achieve better work organization and work environment, it is necessary to promote workers' active participation and personal development. It is important to keep in mind that presenteeism can turn into absenteeism and vice versa [11].

Organizational causes of absenteeism include: Dissatisfaction with work, lack of interest in work, poor working conditions, overtime, dissatisfaction with pay, conflicts with co-workers, conflicts with supervisors, and "bullying" (violent behavior) or mistreatment at work [12] Low employee autonomy, participation and responsibility lead to low satisfaction, which in turn leads to higher absenteeism [13]. When employees are dissatisfied with their work or their motivation wanes, they want to spend as little time as possible at work and avoid going to work whenever possible [2]. Nath Gangai [14] emphasizes that it is better to provide positive incentives to workers to reduce absenteeism than to just impose penalties. Rewards such as additional time off or monetary bonuses are more effective in reducing absenteeism than punishments such as loss of benefits or job. The author also notes that a combination of incentives and penalties, with an emphasis on motivational incentives, is the most effective approach to reducing absenteeism.

It is also necessary to emphasize the importance of demographic characteristics. Choi [15] suggests that there is a relationship between basic worker demographic characteristics such as age, gender, and tenure and various absenteeism reduction measures. It is known that younger workers are more likely to be absent due to lower levels of responsibility, job satisfaction, and poorer working conditions, while older workers are more likely to be absent due to greater job satisfaction and higher status and loyalty to the organization, despite having a higher risk of health problems [16].

Study by Choi [15] wanted to investigate what factors managers should consider to reduce absenteeism among their employees. His research relied on worker demographics as potential causes of absenteeism using machine learning predictions, identifying which worker demographic characteristics might predict absenteeism in the workplace.

III. METHODOLOGY AND DATA ANLYSIS

In the research part of the paper, field primary research was conducted in written form at selected organizations to investigate the opinions and attitudes of the respondents. The test method was applied, more precisely the probing with group test. The test was conducted in the premises of the respondents' work organizations. The prepared questionnaire was anonymous and the respondents filled it out independently. The sample size after subtracting missing values is 420 people. For the research part of the work, the collected demographic data and the data on the causes of absenteeism are used. A good combination of unsupervised and supervised machine learning and their complementarity can be applied to methods in various fields of scientific research such as medicine [17], mathematics [18], banking [19] etc. A very widespread application of these methods is the analysis of absenteeism in various organizations such as

schools [20], industries [21] and in general in business [22]. A proposed approach in school absenteeism research [20] involves using both unsupervised and supervised machine learning techniques to examine data pertaining to school absenteeism. First, group detection is solved using unsupervised techniques. Then, supervised learning methods are used to classify the remaining data. The proposed method is a combination of unsupervised and supervised models, resulting in a model for classifying students into high-risk groups, medium-risk groups, and low-risk groups in terms of truancy. The research [21] compares various ML classifiers and finds that they exhibit a high level of accuracy, making ML techniques a viable option for predicting and analyzing absenteeism. The best ML algorithms are used for forecasting and analyzing school truancy. The study [22] analyzes a situation of employees missing work in a delivery company with the aim of assisting HR executives in forming plans and policies to decrease absenteeism. The data was cleaned and several machine learning methods were utilized to categorize the data (zeroR, J48 based on decision trees, naive Bayes, and KNN).

This paper describes the use of machine learning techniques for the analysis of data related to absenteeism for personal and organizational reasons on the collected data. In the first phase, within the unsupervised machine learning, the Elbow method of data clustering was applied as a basic step to determine the optimal value k, i.e., the number of clusters of the observed data set, and the K-Means algorithm was used to assign each data set to a specific cluster. In supervised machine learning, a classification success test was performed on the previously labeled dataset using four different classification algorithms: RF, DT, SVM, K-NN.

A. Unsupervised machine learning

Unsupervised machine learning algorithms are useful for creating labels in data used in performing supervised learning tasks. Unsupervised algorithms allow intrinsic groupings within unlabeled data and assign a label to each data value [23]. To determine the optimal value of k, i.e., the number of clusters of the observed dataset, the Elbow method was used as a fundamental step for the process of unsupervised machine learning. The Elbow method is used to determine the number of clusters in the dataset by calculating the sum of squared errors (SSE eng. Sum of Squared errors). The number of clusters is determined by observing the graph and the position of the point representing the "elbow" [24]. For the personal and organizational cause datasets, graphs were created using the Elbow method, from which the possible four values for the number of clusters are read: 2, 3, 4, and 5. Since no value has a position at the "elbow point," an analysis

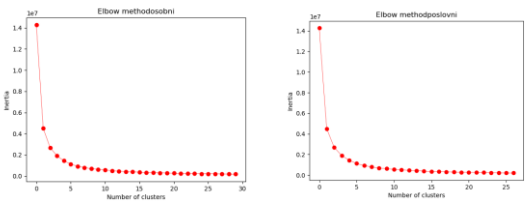

Figure 2. Elbow method applied for both datasets

Figure 1. Silhouette method applied for both datasets

of the Silhouette coefficients was therefore performed for each proposed number of clusters. Determining the silhouette value is a very useful method for finding the optimal number of clusters when the Elbow method does not show a clear "elbow" point. Silhouette coefficient values vary from -1 to 1, with 1 being the best value when samples are perfectly distributed in easily distinguishable clusters. The silhouette coefficient is calculated from the mean distance within the cluster - a - and the mean distance of the nearest cluster - b - for each sample using the following formula [25]:

$$\text{Silhouette score} = \frac{b-a}{\max(a,b)} \qquad (1)$$

According to Figures 1 and 2 and Table I, the optimal number of clusters k = 2 was chosen for both data sets.

TABLE I. SILLHOUETE VALUES FOR BOTH CLUSTERS

Number of clusters	*Dataset - personal cause*	*Dataset - organizational cause*
2	**0.591**	**0.591**
3	0.499	0.499
4	0.481	0.481
5	0.480	0.482

After determining the optimal number of clusters for both datasets, all records in both datasets were assigned to a specific cluster using the K-means algorithm. The K-means algorithm is one of the most well-known algorithms in the field of unsupervised machine learning. This algorithm finds non-overlapping clusters in which each data set is assigned to a particular cluster. This algorithm groups the data by attempting to divide them into n groups with equal variance by appropriately reducing a criterion known as inertia or sum of squares within clusters [26][27]:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2) \quad (2)$$

where x_i is a single record in cluster C and μ_j is the mean of the records in the cluster. Using the K-Means algorithm for both datasets, all records were labeled and assigned to one of the clusters. Of the total 420 records in each dataset, 272 records belong to cluster class 0, and 148 records are assigned to cluster class 1. Records in cluster class 1 indicate a more frequent personal or organizational cause of absenteeism, depending on which record is considered. Records labeled class 0 indicate weak absenteeism or some other cause of absenteeism that is not personal or organizational in nature. The distribution of records by the appropriate class in the personal and organizational cause records is shown in Figure 3. Because the datasets are multidimensional, containing 31 and 28 features, respectively, an nD transformation (n is the number of dataset features) was applied to 2D space using the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to represent the datasets in a 2D coordinate system.

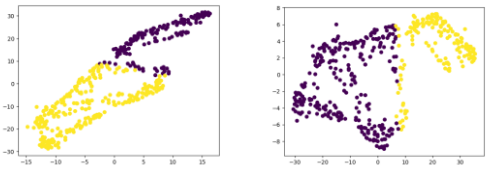

Figure 3. Visualisations of clusters using t-SNE

When applied to high-dimensional but well clustered data, t-SNE shows a visualization with clearly isolated clusters that match well with clusters derived with special clustering algorithms [28].

B. *Supervised machine learning*

The objective of supervised learning is to construct a model from labeled data in order to make predictions on future data. There are two predominant supervised machine learning techniques: classification and regression. Classification involves predicting discrete values such as categories, classes, or labels for new data. Regression, on the other hand, involves predicting the continuous value of a response variable [29]. By applying the classification algorithms to the data clustered in this way from both data sets, it is possible to determine which features contribute most to the classification or to membership in a particular cluster. In this way, the influence of certain questions and answers in the questionnaire used for data collection is indirectly revealed. For classification, four machine learning algorithms were tested and their classification performance was assessed by comparing accuracy, AUC and F1 results. Prior to the actual classification, the data were processed so that they were standardized. Standardization of datasets is a common procedure for many machine learning estimators. Machine learning results could be misinterpreted if the features do not look like normally distributed subplots by default [30]. In practice, the shape of the distribution is often ignored and the data are simply transformed to be centered by removing the mean of each characteristic and then dividing by the standard deviation. The min-max normalization technique involves a linear transformation of the raw data, converting the features in any range to a new range, usually on a scale between [0,1] or [-1,1]. The equation used in this method is:

$$x' = \frac{x_i - x_n}{x_m - x_n} \quad (3)$$

where x_n denotes the minimum value, x_m the maximum value, x_i the input value, and x' the normalized value.

When this normalization method is applied, each feature retains all relational properties in the data [31]. To compare the performance of different classification algorithms, the data for learning and testing must be defined. Prior to the application of machine learning classification, the data was split 70/30, meaning that 70% of the entire dataset is designated for learning and the classification success is tested on the remaining 30% of the data. The split of the dataset into a training and a testing part was done using the train_test_split() function in the Python programming language, which allows random selection of records from the dataset but maintains an equal distribution of classes so that the training and testing data is a representative sample of the original dataset. Classification was performed using four algorithms: RF, DT, SVM, and K-NN , and the results are presented in Table II. for the personal causes of absenteeism dataset and in Table III. for the organizational causes of absenteeism dataset.

TABLE III. CLASSIFICATION PERFORMANCE – DATASET - PERSONAL CAUSE

ML algorithm	*Accuracy*	*F1*	*AUC*
RF	**0.98412**	**0.97777**	**0.98780**
DT	0.97619	0.96551	0.97117
SVM	0.97619	0.96703	0.98170
K-NN	0.93650	0.90476	0.91962

TABLE II. CLASSIFICATION PERFORMANCE – DATASET - ORGANIZATIONAL CAUSE

ML algorithm	*Accuracy*	*F1*	*AUC*
RF	**0.99206**	**0.98876**	**0.99390**
DT	0.97619	0.96551	0.97117
SVM	0.96031	0.94382	0.95898
K-NN	0.93650	0.90476	0.91962

The best results in classifying data with personal and organizational causes of absenteeism are obtained with the algorithm RF, both in terms of AUC, F1, and accuracy. Differences in certain features of the two datasets were examined for their impact on classification results. The popularity of the algorithm RF is based on its ability to make successful predictions. However, it is also important that it provides a complete nonparametric measure of feature importance in prediction/classification [32]. Feature importance allows users to identify those features that play a key role in prediction/classification. In many applications, a good predictive model is only one goal; another, often more important goal, is to identify variables that enable good

prediction/classification. The algorithm RF allows the measurement of the importance of variables, which can be used to rank the variables according to their predictive importance [33]. In this study, the feature_importances_ property and its values were used for each record feature in the algorithm RF. Figure 4 shows the order of importance of the extracted features from both datasets of personal and organizational causes of absenteeism.

When analyzing the data on personal causes of absenteeism, the characteristic of personal illness (A08) was found to be the most important for classification, followed by family responsibilities (A07), illness of family members (A09), and medical examination (A24), while other

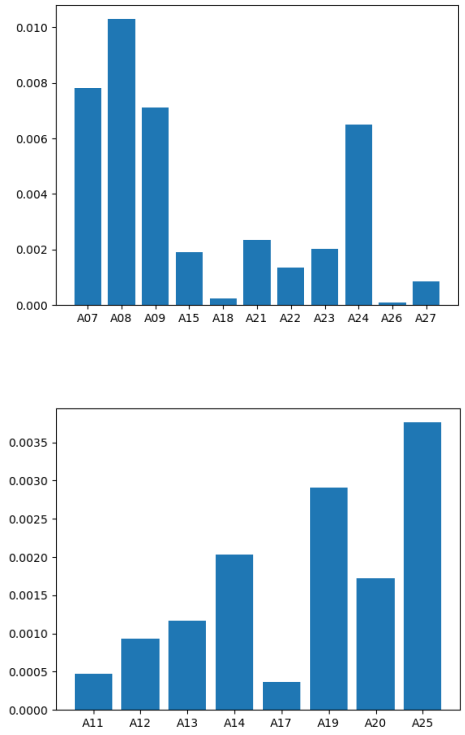

Figure 4. Feature importances

characteristics were less important. For the organizational causes of absenteeism dataset, the most important characteristics affecting classification are bullying (A25), conflicts with work colleagues (A19), overtime (A14), and conflicts with supervisors (A20).

Table IV. shows the order of importance of all characteristics in the classification, including demographic characteristics. In both data sets, respondents' demographic characteristics had the greatest importance in the classification. In the data set of absenteeism for personal reasons, including demographic characteristics, the following characteristics had the greatest importance in classification: total work experience (DM07), work experience in current organization (DM06). In the long case, in the dataset of organizational reasons for absenteeism, along with

TABLE IV. FEATURE IMPORTANCES – BEST DEMOGRAPHIC SCORES

feature importance - demographic scores in personal cause		**feature importance - demographic scores in organizational cause**	
feature	***importance score***	***feature***	***importance score***
DM07	0.367247	DM09a	0.651428
DM06	0.268832	DM10	0.04614
DM02	0.176321	DM12	0.045125

demographic data, the following characteristics had the greatest influence in classification: ownership type of organization (DM09a), level of job in the organization (DM10).

IV. CONCLUSION

In reviewing the literature, numerous classifications of the causes of absenteeism were found, each confirming the other, but also differing from each other, which means that it is necessary to create unique tools both for predicting absenteeism and for predicting the causes that lead to it. Because of the complexity and multifaceted nature of the causes of absenteeism, authors often use certain groups of causes in their research while ignoring others. Using unsupervised and supervised machine learning techniques, it is possible to predict or classify respondents depending on the frequency of their absence from work, which can be personal or organizational in nature. Considering the proposed number of clusters by unsupervised machine learning and the results of the most efficient classification of the Random Forest algorithm of supervised machine learning.

The results show that personal illness and family responsibilities are the most important absences from work for personal reasons. Considering the demographic characteristics of the personal reasons for absenteeism, the most important feature is the total number of years of service and the length of service in the current organization. From the group of organizational causes, bullying and conflicts with work colleagues, as well as the form of ownership in the organization and the level of position in the organization, have the greatest influence on the prediction of absenteeism.

One of the obvious limitations of the conducted research is also the above-mentioned classification that considers only a certain number of organizational and personal causes of absenteeism, which leads to the neglect of other causes that can also have a significant impact on absenteeism in the workplace.

The recommendation for future research is reflected in the increase of the number of respondents, which will contribute to the generality of the obtained results when other possible factors for the cause of absenteeism in workers are included. In the field of machine learning, techniques for reducing and selecting optimal features for prediction or classification can be applied to confirm or improve the adoption of measures to reduce worker absenteeism.

REFERENCES

[1] "Eurostat." Hours of work and absences from work - quarterly statistics - Statistics Explained (europa.eu) (accessed Jan. 30, 2023).

[2] Đ. Ilić, "Measuring absenteeism as a precondition of quality management of absenteeism," Trendovi u poslovanju, vol. 8, no. 1, pp. 66–74, 2020, doi: 10.5937/trendpos2001066i.

[3] "Absence Insight ." https://absenceinsight.eu/ (accessed Jan. 30, 2023).

[4] I. Katić and A. Nešić, "Socio-psychological effects of stress in organizations' absenteeism problems," Work, vol. 66, no. 3, pp. 689–697, 2020, doi: 10.3233/WOR-203211.

[5] D. A. Endovitskiy, "Developing A Classifier Of Relevant Causes Of Absenteeism In An Organisation," in Global Challenges and Prospects of The Modern Economic Development Proceedings of Global Challenges and Prospects of The Modern Economic Development (GCPMED 2020), 15-16 December, 2020, Samara State University of Economics, Samara, Russia, Apr. 2021, vol. 106, pp. 699–706. doi: 10.15405/epsbs.2021.04.02.84.

[6] Đ. Ilić, G. Mrdak, and M. Bojić, "Sociological aspect of labor force absentism," Oditor, vol. 7, no. 1, pp. 195–224, 2021, doi: 10.5937/oditor2101195i.

[7] V. Čikeš, H. M. Ribarić, and K. Črnjar, "The determinants and outcomes of absence behavior: A systematic literature review," Social Sciences, vol. 7, no. 8. MDPI AG, Jul. 24, 2018. doi: 10.3390/socsci7080120.

[8] S. G. Aldana and N. P. Pronk, "Health Promotion Programs, Modifiable Health Risks, and Employee Absenteeism," J Occup Environ Med, vol. 43, no. 1, pp. 36–46, Jan. 2001, doi: 10.1097/00043764-200101000-00009.

[9] M. C. Kocakulah, A. G. Kelley, K. M. Mitchell, and M. P. Ruggieri, "Absenteeism Problems And Costs: Causes, Effects And Cures," International Business & Economics Research Journal (IBER), vol. 15, no. 3, pp. 89–96, May 2016, doi: 10.19030/iber.v15i3.9673.

[10] R. M. Badubi, "A Critical Risk Analysis of Absenteeism in the Work Place," Journal Of International Business Research and Marketing, vol. 2, no. 6, pp. 32–36, 2017, doi: 10.18775/jibrm.1849-8558.2015.26.3004.

[11] H. Brborović and J. Mustajbegović, "Mogućnost prevencije prezentizma i apsentizma zdravstvenih djelatnika," Sigurnost, vol. 58, no. 2, Jun. 2016, doi: 10.31306/s.58.2.2.

[12] 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, Aug. 2007, doi: 10.1108/13620430710773745.

[13] G. Egan, "An Investigation Into The Causes Of Absenteeism In 'Company X.'"

[14] K. Nath Gangai, "Absenteeism At Workplace: What Are The Factors Influencing To It?"

[15] J. W. Choi*, "Prediction of Workplace Absenteeism Time Using Machine Learning," International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2, pp. 3489–3493, Dec. 2019, doi: 10.35940/ijitee.B6571.129219.

[16] M. Chaudhury and I. Ng, "Absenteeism Predictors: Least Squares, Rank Regression, and Model Selection Results," 1992.

[17] T. Chauhan, S. Rawat, S. Malik, and P. Singh, "Supervised and Unsupervised Machine Learning based Review on Diabetes Care," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2021, pp. 581–585. doi: 10.1109/ICACCS51430.2021.9442021.

[18] J. Shen, W. Li, S. Deng, and T. Zhang, "Supervised and unsupervised learning of directed percolation," Phys Rev E, vol. 103, no. 5, p. 052140, May 2021, doi: 10.1103/PhysRevE.103.052140.

[19] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," Expert Syst Appl, vol. 128, pp. 301–315, Aug. 2019, doi: 10.1016/j.eswa.2019.02.033.

[20] F. Bowen, C. Gentle-Genitty, J. Siegler, and M. Jackson, "Revealing underlying factors of absenteeism: A machine learning approach," Front Psychol, vol. 13, Dec. 2022, doi: 10.3389/fpsyg.2022.958748.

[21] A. Rista, J. Ajdari, and X. Zenuni, "Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms," in 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Sep. 2020, pp. 485–490. doi: 10.23919/MIPRO48935.2020.9245118.

[22] M. Skorikov et al., "Prediction of Absenteeism at Work using Data Mining Techniques," in 2020 5th International Conference on Information Technology Research (ICITR), Dec. 2020, pp. 1–6. doi: 10.1109/ICITR51448.2020.9310913.

[23] P. Ebrahimi, M. Basirat, A. Yousefi, Md. Nekmahmud, A. Gholampour, and M. Fekete-Farkas, "Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning Approaches," Big Data and Cognitive Computing, vol. 6, no. 2, p. 35, Mar. 2022, doi: 10.3390/bdcc6020035.

[24] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," in Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia, 2020. doi: 10.4108/eai.24-1-2018.2292388.

[25] "silhouette_score." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#sklearn.metrics.silhouette_score (accessed Jan. 17, 2023).

[26] "K-means clustering." https://scikit-learn.org/stable/modules/clustering.html#k-means (accessed Jan. 17, 2023).

[27] Md. Zubair, MD. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, "An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling," Annals of Data Science, Jun. 2022, doi: 10.1007/s40745-022-00428-2.

[28] D. Kobak and P. Berens, "The art of using t-SNE for single-cell transcriptomics," Nat Commun, vol. 10, no. 1, p. 5416, Nov. 2019, doi: 10.1038/s41467-019-13056-x.

[29] I. Fosic, D. Zagar, and K. Grgic, "Network traffic verification based on a public dataset for IDS systems and machine learning classification algorithms," in 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), May 2022, pp. 1037–1041. doi: 10.23919/MIPRO55190.2022.9803674.

[30] "Standardization, or mean removal and variance scaling." https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler (accessed Jan. 18, 2023).

[31] G. Aksu, C. O. Güzeller, and M. T. Eser, "The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model," International Journal of Assessment Tools in Education, pp. 170–192, Apr. 2019, doi: 10.21449/ijate.479404.

[32] H. Ishwaran and M. Lu, "Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival," Stat Med, vol. 38, no. 4, pp. 558–582, Feb. 2019, doi: 10.1002/sim.7803.

[33] F. Degenhardt, S. Seifert, and S. Szymczak, "Evaluation of variable selection methods for random forests and omics data sets," Brief Bioinform, vol. 20, no. 2, pp. 492–503, Mar. 2019, doi: 10.1093/bib/bbx124.