Prediction of Absenteeism at Work using Data Mining Techniques

Mikhail Skorikov*, Muhammad Abrar Hussain†, Mahfujur Rhaman Khan*, Mohammad Kaosain Akbar‡, Sifat Momen*, Nabeel Mohammed* and Taniya Nashin §
*Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh,
†Fujitsu Research Institute, Tokyo, Japan,
‡Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh,
§ Department of Business Administration, Victoria University of Bangladesh, Dhaka, Bangladesh
Email: *{mikhail.skorikov, mahfujur.rhaman, sifat.momen, nabeel.mohammed}@northsouth.edu,
†m.abrar.hussain@jp.fujitsu.com, ‡kaosain.cse@diu.edu.bd, §taniyanashin.vub@gmail.com

***Abstract*—High absenteeism among employees can be detrimental to an organization as it can result in productivity and economic loss. This paper looks into a case of absenteeism in a courier company in Brazil. Machine learning techniques have been employed to understand and predict absenteeism. Understanding this would provide human resource managers an excellent decision aid to create policies that can aim to reduce absenteeism. Data has been preprocessed, and several machine learning classification algorithms (such as zeroR, tree-based J48, naive Bayes, and KNN) have been applied. The paper reports models that can predict absenteeism with an accuracy of over 92%. Furthermore, from an initial of 20 attributes, disciplinary failure turns out to be a very prominent feature in predicting absenteeism.**

***Index Terms*—absenteeism, prediction, data mining, classification**

I. Introduction

In Human Resources and Management (HRM), employees are regarded as valuable assets to their organization. Employee productivity directly contributes to the work efficiency and the success of a company. The economic sustainability of a company lies in its revenue, which is generated only with sufficient support from the employees. If employees underperform or do not perform well, it results in an incurment of high costs to the organization. Consequently, in good organizations, HRM professionals have to work prudently to ensure that employee motivation stays high enough to yield target productivity [1].

One key indicator of low motivation in an organization is high employee absenteeism. Absence in work can arise due to an array of factors including age, health condition, bad work-life balance, low motivation, organizational culture, policies, pay scale, poor recognition in the organization, and many more. Understanding why employees stay absent and realizing the patterns in absenteeism behavior is an invaluable tool for human resource managers. Understanding this, in turn, would help organizations to shape the policies and organizational culture to transform the organization into a more people-driven organization, which is expected to result in higher productivity [2].

978-0-7381-4403-0/20/$31.00 ©2020 IEEE

In this paper, we present a prediction of absenteeism at work using data mining techniques in a courier company in Brazil. The dataset [3] is publicly available in Kaggle and comprises of a total of three years of data (from July 2007 to July 2010) containing details of absences in the company. When employees were absent, their details of absence have been recorded. The novelty of the paper lies in the use of data mining techniques (particularly classification) to predict absenteeism. Such a methodology can be applied to understand absenteeism behavior in other companies and develop a decision-aid system that can help managers in controlling it.

The rest of the paper is organized as follows: In section 2, a critical literature review has been conducted. This follows our research methodology in section 3. Experimental results are presented in section 4. Finally, the paper is concluded in section 5 with the highlights of the results obtained.

II. Background

Absenteeism is defined as the temporary withdrawal from work due to personal reasons, including illness and demise of close relatives [4]. Cuchiella and colleagues, on the other hand, remarks absenteeism as an employee's habitual absence from work [5]. Despite the differences in opinion, high absenteeism is always found to be correlated with high productivity loss and economic loss. Research in understanding why employees stay absent is a very active area partly due to the fact that it often has serious economic consequences. Different factors have been found to contribute towards absenteeism. Arai and Thoursie [6], for instance, found that incentives have a crucial role to play in absenteeism. Using industry-region panel data, they were able to establish a negative correlation between sick rates and shares of temporary contracts. Arai and Thoursie argue that this is because workers on temporary contracts have lower job security and hence, a higher possibility of being laid off compared to the employees on time-unlimited contracts. Hence, workers on temporary contracts tend to be more loyal in the hope of a renewal of contracts. Hesselius [7] found that absenteeism is negatively correlated to unemployment. This is due to the fact that when unemployment is high, the probability of an individual to find a new job is low. This acts as an

TABLE I
ATTRIBUTES AND THEIR DESCRIPTION

Attribute	Data Type	Description/Remarks	Min	Max	Mean	Distribution
ID	Integer	Identification number of the employee	x	x	x	x
Reason for absence	Integer	The reason for absence (range from 1 - 28)	17, 3, 2	23	x	x
Month of absence	Integer	Month of absence of the employee (range from 1 - 12)	12	3	x	x
Day of the Week	Integer	Range is from 1 - 7 where Sunday = 1 and so on	5	2	x	x
Seasons	Integer	There are four seasons (1 - 4)	1	4	3	x
Travel Expense	Integer	Transportation cost from home to work	118	388	221.3	negatively skewed
Distance	Integer	Distance from residence to work	5	52	29.6	positively skewed
Service time	Integer	Months of service time	1	29	12.6	negatively skewed
Age	Integer	Age of the employee in years	27	58	36.5	negatively skewed
Workload / day	Integer	Average workload per day	206	379	271.5	positively skewed
Hit target	Integer	Target for the employees	81	100	94.6	negatively skewed
Disciplinary failure	Boolean	=1 for past disciplinary failure, otherwise 0	1	0	x	x
Education	Integer	Range is from 1 - 4 depending on the education level	4	1	x	x
Children	Integer	Number of children of the employee	0	4	1	normally distributed
Social Drinker	Boolean	=1 for being a social drinker, otherwise 0	1	0	x	x
Social Smoker	Boolean	=1 for being a social smoker, otherwise 0	1	0	x	x
Pet	Integer	Number of pets that the employee has	0	8	0.7	positively skewed
Weight	Integer	Weight (in nearest Kg) of the employee	56	108	79	negatively skewed
Height	Integer	Height (in nearest cm) of the employee	163	196	172.1	positively skewed
BMI	Integer	Body mass index (to the nearest integer)	19	38	26.7	positively skewed
Absenteeism time	Integer	Absenteeism time in hours	0	120	6.9	positively skewed

incentive for the individual to be less absent in work and thus reducing his/her probability of being laid off. Winkelmann [8] found that absenteeism is dependent on factors such as wages and even the firm size. Organizational policies also impact the level of absenteeism [9]. For example, Halpern and colleagues [10] found that the smoking policy in the workplace affects absenteeism and productivity. Their research concludes that current smokers tend to have significantly higher absenteeism than former smokers and non-smokers. Absenteeism is repeatedly reported to be strongly correlated with employees' health status. For instance, Tunceli and colleagues [11] found that for employees with diabetes, the absolute probability of working for male and female employees is 7.1 % and 4.4 % points less compared to the individuals without diabetes. Gates and colleagues [12] found that moderately or extremely obese workers experience a 4.2% loss in productivity, which is tantamount to 1.18% more than all other employees. Shah and colleagues [13] used deep neural networks to predict absenteeism before employees are actually hired. Research in understanding absenteeism behavior reveals that it depends on various factors that act as incentives (directly or indirectly), including organizational culture, policies (both national and organizational), size of the organization, and many more.

III. RESEARCH WORK

This paper looks into the factors affecting absenteeism in a courier company in Brazil. Whenever an employee is absent, he/she needs to fill up a form detailing reasons for absence. This, in conjunction with other personal details, has been recorded in a dataset. The dataset was first published by [3] and is now publicly available in Kaggle.

A. Dataset

The dataset comprises of a total of 740 instances recorded over 21 attributes per instance. The dataset is a collection of 3 years of data from July 2007 to July 2010. Information pertaining to health, work, workload, habit, traveling, and details of absence have been incorporated into the dataset. Table I shows the details of the attributes recorded in the dataset. One particular attribute to note in the dataset is the reason for absence, which is an important attribute to investigate. Table II describes the various reasons for absence, along with the percentage of occurrence of each. The dataset size is small and as a consequence applying deep learning techniques will not be effective. Classical data mining techniques are able to predict with high accuracy. Hence we refrained from using any deep learning techniques. We split the dataset into training and test sets in an 80-20 ratio with a stratified class attribute distribution.

B. Research Methodology

Our prime objective is to develop a model that can predict absenteeism (with high accuracy) in the courier company for aid in decision-making by the managers. Since there exists different reasons for absence among employees and many factors contribute towards absenteeism, a deductive learning approach is an infeasible option. We instead use the inductive learning approach, where we use data mining techniques to predict absenteeism. Scikit-learn [14], a Python library for data science, has been used to carry out the data mining tasks.

The research methodology embraced in this work has been outlined in figure 1.

The raw data (the Absenteeism dataset) is first preprocessed to a form that is suitable for applying machine learning algorithms. The preprocessing phase includes data cleaning (removal of attribute and marking missing labels), data discretization (i.e., converting continuous data into discrete categories), and data transformation (converting data from a numeric form to categorical). After the preprocessing step,

TABLE II
PREPROCESSING FOR THE ATTRIBUTE "REASON OF ABSENCE"

Original Value	New Value	Description	Percentage in dataset
1	IPD	Certain infectious and parasitic diseases	2.1%
2	NP	Neoplasms	0.1%
3	BOI	Blood-forming organ & immune mechanism	0.1%
4	ENM	Endocrine, nutritional and metabolic diseasess	0.3%
5	MBD	Mental and behavioural disorders	0.4%
6	DNS	Diseases of the nervous system	1.1%
7	DEA	Diseases of the eye and adnexa	2.0%
8	DEM	Diseases of the ear and mastoid process	0.8%
9	DCS	Diseases of the circulatory system	0.5%
10	DRS	Diseases of the respiratory system	3.4%
11	DDS	Diseases of the digestive system	3.5%
12	DSST	Diseases of the skin and subcutaneous tissue	1.1%
13	DMSCT	Diseases of musculoskeletal system & tissue	7.4%
14	DGS	Diseases of the genitourinary system	2.6%
15	PCP	Pregnancy, childbirth and the puerperium	0.3%
16	CPP	Conditions originating in the perinatal period	0.4%
17	CMDCA	Congenital malformations and chromosomal abnormalities	15.1%
18	ACLF	Abnormal clinical and laboratory findings	2.8%
19	IPEC	Injury, poisoning and consequences of external causes	5.4%
20	ECMM	External causes of morbidity and mortality	0%
21	FHSHS	Factors to health status and health services	0.8%
22	PFU	Patient follow-up	5.1%
23	MC	Medical consultation	20.1%
24	BD	Blood Donation	0.4%
25	LE	Laboratory examination	4.2%
26	UA	Unjustified absence	4.5%
27	PTH	Physiotherapy	9.3%
28	DC	Dental Consultation	15.1%
?	null	Null values for employees with no absenteeism	5.8%

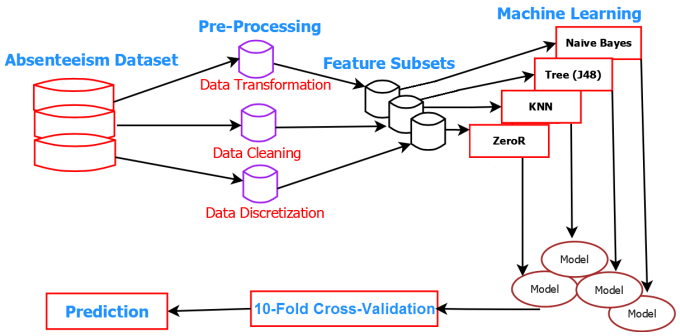

Fig. 1. Research Methodology

two subsets of features are found that would make predictions better. Machine learning algorithms are then applied to the new feature subsets and the full feature set to predict the absenteeism class (discussed later).

C. Preprocessing of Predictor Attributes

Preprocessing is required to transform the raw data into a form that would be suitable for data mining tasks. Missing data in the dataset is first marked by null value, because the missing values were originally marked with question marks (?). After this, several preprocessing techniques are applied to the dataset. All the attributes that were not real-valued were converted to categorical, including absenteeism time in hours, the class attribute.

ID works as a unique identification number for the employee and is irrelevant, thus has been removed. Other attributes are converted to categorical form since a numeric or boolean value is not appropriate for it. Furthermore, it may cause problems when applying machine learning algorithms. Seven is higher than one, but there is no quantitative difference if an employee is absent in July (7) or January (1). Hence, in places where a numeric or boolean value does not make sense, we have converted them to categorical values.

There exists three boolean attributes in the dataset: (1) disciplinary failure, (2) social drinker, and (3) social smoker. We converted the initial 0 and 1 values to 'False' and 'True' respectively.

Table II shows the conversion of attribute *Reason of absence* from numeric to categorical. For the attribute *seasons*, the value from the numeric is also converted to categorical with 1, 2, 3, and 4 substituted to A, B, C, and D, respectively.

Values of the attribute "Education" are also converted from numeric to categorical. The values in the range 1-4 are substituted by "High School", "Graduate", "Postgraduate", and "Master and Doctor", respectively.

D. Preprocessing of the class attribute

The class attribute is the attribute that is intended to be predicted. The class attribute that we would like to predict is absenteeism time. Initially, the values in this attribute were continuous. However, it would make more sense if we classify

the absenteeism time in terms of categories. We do so because it would allow the model to predict different degrees of absence on test data.

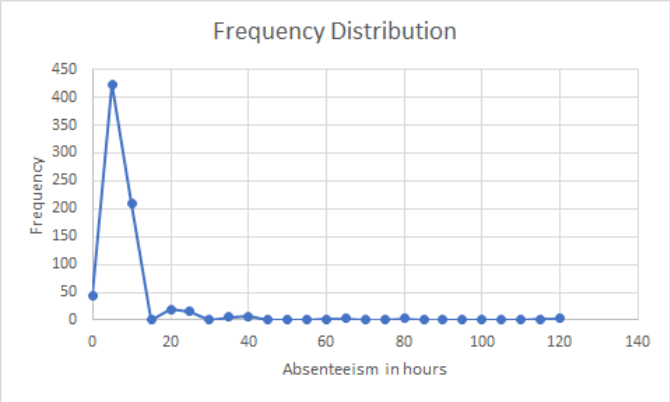

Fig. 2. Frequency Distribution of the class attribute

Figure 2 shows the frequency distribution of the class attribute. The x-axis of the graph indicates absenteeism time (in hours), whereas the y-axis signifies the corresponding frequency value. Figure 2 provides a conspicuous display of the existence of three classes in the class attribute.

TABLE III
DIFFERENT CLASSES OF CLASS ATTRIBUTE

Absenteeism time (hours)	Class
0	A
1 - 15	B
16 - 120	C

E. Class Imbalance

It is sensible that we categorize the class attribute as described in table III. However, this also results in an imbalance of the three classes with class B taking up about 85% of the dataset. To combat this problem, we applied an oversampling technique called Synthetic Minority Oversampling TEchnique (SMOTE) onto the training set. The resulting dataset was almost equally balanced in terms of the class attribute. We then split the training process into one tested using SMOTE applied, and one without.

F. Feature Selection

The selection of prominent features is crucial in data mining tasks for two main reasons: (1) irrelevant attributes act as noise, and this can degrade the predictability of the model. The removal of irrelevant attributes improves the predictability of the model. (2) It results in the reduction of the dimension of the dataset - thus allowing to avoid the curse of dimensionality.

Correlation Feature Set (CFS) [15], a well-known feature selection technique, has been used to find the prominent features that can be used to predict the class attribute. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature and the degree of redundancy between them. Using the CFS algorithm, it has been found that four attributes (the month of absence, age, disciplinary failure, social drinker) play as the most influential attributes to predict absenteeism.

TABLE IV
FEATURE SCORES BY RELIEF ATTRIBUTE EVALUATOR

Attribute	Feature Score
Disciplinary failure	0.3629
Reason for absence	0.3303
Month of absence	0.1816
Seasons	0.1685
Social drinker	0.1594
Distance from residence to work	0.137
Children	0.1225
Transportation Expense	0.119
Education	0.1121
Weight	0.1097
Age	0.1005
Body mass index	0.0993
Day of the Week	0.0783
Pet	0.0685
Service time	0.064
Height	0.0617
Work load / day	0.055
Hit target	0.0434
Social Smoker	0.0386

Another popular feature evaluator is the relief attribute evaluator, based on the relief algorithm [16]. The relief attribute evaluator evaluates each attribute's worth in terms of a score called the feature score. Each attribute's feature score lies between -1 and 1 with values going towards 1 indicating its prominence level as a feature for predicting the target attribute.

IV. EXPERIMENTS AND RESULTS

A comparison cannot be made with relevant previous works due to the class value ranges for each work being different, but our model outperforms similar models [17].

This section discusses the experimental methodology and corresponding results. Three types of experiments, as outlined in table VI, are devised to assess how well the absenteeism class can be predicted. Experiment A uses the four prominent features (the month of absence, age, disciplinary failure, social drinker) as found using the CFS method to train data. Experiment B uses all the 19 attributes. Since initial experiments show that the attribute *disciplinary failure* has the highest information gain as well as the highest feature score from the relief algorithm, experiment C is conducted with only one attribute - the disciplinary failure.

Each experiment is run using a 10-fold stratified cross-validation strategy and several different classifiers (including ZeroR, naive Bayes, KNN, and tree-based J48 classifiers). The ZeroR classifier does not have any predictability power as it merely predicts the majority class for every query input. However, the ZeroR classifier has been selected as a baseline classifier. The naive Bayes classifier uses the Bayes theorem (equation 1) to predict the absenteeism class.

$$P(Y|X_1, X_2, ..., X_n) = \frac{P(X_1, X_2, ..., X_n|Y)P(Y)}{P(X_1, X_2, ..., X_n)} \quad (1)$$

TABLE V
WEIGHTED AVERAGE OUTPUT OF ORIGINAL DATASET WITHOUT SMOTE

Expt. & Classifier	Precision	Recall	F-measure	ROC area	Accuracy
A (zeroR)	0.74	0.86	0.79	0.50	85.5 +/- 0.6 %
A (naive Bayes)	0.83	0.91	0.87	0.76	**90.1** +/- 2.0 %
A (J48)	0.83	0.91	0.87	0.74	89.7 +/- 2.9 %
A (KNN-Euclidean)	0.83	0.91	0.87	0.74	88.0 +/- 5.3 %
A (KNN-Manhattan)	0.83	0.91	0.87	0.74	88.0 +/- 5.3 %
A (KNN-Chebyshev)	0.78	0.74	0.75	0.69	81.2 +/- 5.9 %
B (zeroR)	0.74	0.86	0.79	0.50	85.5 +/- 0.6 %
B (naive Bayes)	0.85	0.80	0.82	0.80	66.5 +/- 13.8 %
B (J48)	0.86	0.88	0.87	0.80	**89.1** +/- 4.3 %
B (KNN-Euclidean)	0.82	0.89	0.84	0.77	85.4 +/- 1.3 %
B (KNN-Manhattan)	0.82	0.89	0.84	0.77	85.4 +/- 1.3 %
B (KNN-Chebyshev)	0.83	0.91	0.87	0.67	86.9 +/- 2.3 %
C (zeroR)	0.74	0.86	0.79	0.50	85.5 +/- 0.6 %
C (naive Bayes)	0.83	0.91	0.87	0.69	**90.9** +/- 1.6 %
C (J48)	0.83	0.91	0.87	0.69	**90.9** +/- 1.6 %
C (KNN-Euclidean)	0.83	0.91	0.87	0.69	**90.9** +/- 1.6 %
C (KNN-Manhattan)	0.83	0.91	0.87	0.69	**90.9** +/- 1.6 %
C (KNN-Chebyshev)	0.83	0.91	0.87	0.69	**90.9** +/- 1.6 %

TABLE VI
TYPES OF EXPERIMENTS

Experiment Name	Features used
A	As found using the CFS method
B	All 19 attributes
C	Disciplinary failure

A lazy classifier, KNN, is used with a K value of 5. In order to find nearest neighbors, distance is measured using different metrics, including Euclidean, Manhattan, and Chebyshev. Equations 2, 3, and 4 shows the ways Euclidean, Manhattan and Chebyshev distances are calculated.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{3}$$

$$D(x,y) = max(x_i - y_i) \tag{4}$$

Finally, a decision tree (J48) has been used as a classifier. J48 uses entropy-based mutual information gain to construct the decision tree.

A. Experimental results without the application of SMOTE filter

The experimental results without applying the SMOTE filter are illustrated in figure 3 , with table V detailing the particulars.

B. Experimental results after applying the SMOTE filter

Figure 4 illustrates the summary of the results with SMOTE applied. Table VII denotes the performance scores of each classifier.

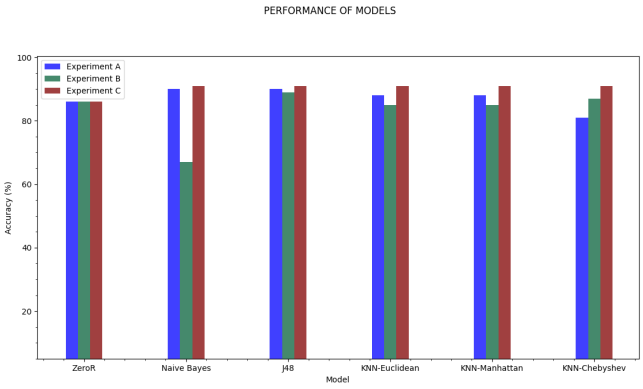

Fig. 3. Experimental results - original dataset

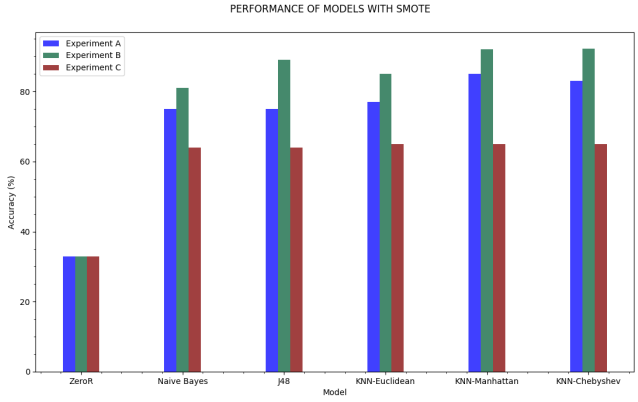

Fig. 4. Experimental results - SMOTE

C. Discussion

Experimental results from tables V and VII indicate that an accuracy of 90.1% is obtained using the naive Bayes classifier without oversampling, and it is the highest achieved

TABLE VII
WEIGHTED AVERAGE OUTPUT WITH SMOTE

Expt. & Classifier	Precision	Recall	F-measure	ROC area	Accuracy
A (zeroR)	0.00	0.06	0.01	0.50	33.2 +/- 0.1 %
A (naive Bayes)	0.85	0.62	0.70	0.77	75.0 +/- 4.0 %
A (J48)	0.86	0.41	0.51	0.72	75.1 +/- 3.3 %
A (KNN-Euclidean)	0.85	0.64	0.72	0.73	76.9 +/- 5.1 %
A (KNN-Manhattan)	0.84	0.83	0.84	0.70	**85.4** +/- 4.7 %
A (KNN-Chebyshev)	0.80	0.76	0.77	0.64	83.0 +/- 4.7 %
B (zeroR)	0.00	0.06	0.01	0.50	33.2 +/- 0.1 %
B (naive Bayes)	0.87	0.75	0.79	0.80	80.5 +/- 6.1 %
B (J48)	0.86	0.89	0.87	0.76	89.3 +/- 6.5 %
B (KNN-Euclidean)	0.85	0.67	0.72	0.81	84.5 +/- 5.0 %
B (KNN-Manhattan)	0.79	0.75	0.77	0.76	92.1 +/- 4.3 %
B (KNN-Chebyshev)	0.87	0.68	0.75	0.70	**92.3** +/- 9.5 %
C (zeroR)	0.00	0.06	0.01	0.50	33.2 +/- 0.1 %
C (naive Bayes)	0.83	0.91	0.87	0.69	64.0 +/- 1.1 %
C (J48)	0.83	0.91	0.87	0.69	64.3 +/- 1.0 %
C (KNN-Euclidean)	0.83	0.91	0.87	0.69	**64.5** +/- 1.0 %
C (KNN-Manhattan)	0.83	0.91	0.87	0.69	**64.5** +/- 1.0 %
C (KNN-Chebyshev)	0.83	0.91	0.87	0.69	**64.5** +/- 1.0 %

for experiment A. The KNN, naive Bayes, and J48 classifiers yield the highest accuracy in experiment C if applied to the dataset without the SMOTE filter. Applying the SMOTE filter to the data and conducting experiment C results in a significant reduction of performance. For experiment B, the performance of the naive Bayes classifier falls with or without SMOTE. However, the J48 classifier's performance stays more or less the same regardless of the sampling strategy. The highest measure of accuracy overall is 92.3%, achieved by the KNN classifier with the Chebyshev distance metric in experiment B after applying the SMOTE filter to the train set. For this model, the accuracy of class A is 67%, for class B it is 92.1%, and for class C is 8.3%. On another note, the experiment indicates that disciplinary failure is a very influential attribute for determining absenteeism as all classifiers other than the baseline result in over 90% accuracy for it without SMOTE applied. Experiment B, comprising of all attributes, is a good indicator of the performance of the classifiers and reflect a more realistic view due to the class imbalance problem causing overestimations in performance otherwise. An imbalanced dataset can sometimes lead to high bias. When the bias is removed owing to SMOTE, the performance naturally decreases.

V. CONCLUSION

This paper looks into the absenteeism dataset, a dataset detailing information about absence records. There were a total of 740 instances and 21 initial attributes. After careful preprocessing and feature selection, machine learning algorithms were applied. Three kinds of experiments with different subsets of features were devised. It has been found that the model can predict absenteeism with over 92% accuracy.

REFERENCES

[1] C. Navarro and C. Bass, "The cost of employee absenteeism," *Compensation & Benefits Review*, vol. 38, no. 6, pp. 26–30, 2006.

[2] M. Mayfield, J. Mayfield, and K. Q. Ma, "Innovation matters: creative environment, absenteeism, and job satisfaction," *Journal of Organizational Change Management*, 2020.

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

[4] R. L. Mathis and J. H. Jackson, *Human resource management: Essential perspectives*. Cengage Learning, 2011.

[5] F. Cucchiella, M. Gastaldi, and L. Ranieri, "Managing absenteeism in the workplace: the case of an italian multiutility company," *Procedia-Social and Behavioral Sciences*, vol. 150, pp. 1157–1166, 2014.

[6] M. Arai and P. S. Thoursie, "Incentives and selection in cyclical absenteeism," *Labour Economics*, vol. 12, no. 2, pp. 269–280, 2005.

[7] P. Hesselius, "Does sickness absence increase the risk of unemployment?" *The Journal of Socio-Economics*, vol. 36, no. 2, pp. 288–310, 2007.

[8] R. Winkelmann, "Wages, firm size and absenteeism," *Applied Economics Letters*, vol. 6, no. 6, pp. 337–341, 1999.

[9] S. A. Ruhle and S. Süß, "Presenteeism and absenteeism at work—an analysis of archetypes of sickness attendance cultures," *Journal of Business and Psychology*, vol. 35, no. 2, pp. 241–255, 2020.

[10] M. T. Halpern, R. Shikiar, A. M. Rentz, and Z. M. Khan, "Impact of smoking status on workplace absenteeism and productivity," *Tobacco control*, vol. 10, no. 3, pp. 233–238, 2001.

[11] K. Tunceli, C. J. Bradley, D. Nerenz, L. K. Williams, M. Pladevall, and J. E. Lafata, "The impact of diabetes on employment and work productivity," *Diabetes care*, vol. 28, no. 11, pp. 2662–2667, 2005.

[12] D. M. Gates, P. Succop, B. J. Brehm, G. L. Gillespie, and B. D. Sommers, "Obesity and presenteeism: the impact of body mass index on workplace productivity," *Journal of Occupational and Environmental Medicine*, vol. 50, no. 1, pp. 39–45, 2008.

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

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[16] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Aaai*, vol. 2, 1992, pp. 129–134.

[17] Z. Wahid, A. Z. Satter, A. Al Imran, and T. Bhuiyan, "Predicting absenteeism at work using tree-based learners," in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, 2019, pp. 7–11.