7-1-2022

# Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism

Gopal Nath
*Murray State University Murray*

Yawei Wang
*Montclair State University*

Austin Coursey
*Murray State University Murray*

Krishna K. Saha
*Central Connecticut State University*

Srikanth Prabhu
*Manipal Institute of Technology*

*See next page for additional authors*



## Recommended Citation

Gopal Nath, Yawei Wang, Austin Coursey, Krishna K. Saha, Srikanth Prabhu, and Saptarshi Sengupta. "Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism" *Information (Switzerland)* (2022). https://doi.org/10.3390/info13070320

information

*Article*

# Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism

**Gopal Nath [1,*], Yawei Wang [2], Austin Coursey [3], Krishna K. Saha [4], Srikanth Prabhu [5] and Saptarshi Sengupta [6]**

[1] Department of Mathematics and Statistics, Murray State University, 6C-19 Faculty Hall, Murray, KY 42071, USA
[2] Feliciano School of Business, Montclair State University, Montclair, NJ 07043, USA; wangya@montclair.edu
[3] Department of Computer Science and Information Systems, Murray State University, Murray, KY 42071, USA; acoursey3@murraystate.edu
[4] Department of Mathematical Sciences, Central Connecticut State University, New Britain, CT 06050, USA; sahakrk@ccsu.edu
[5] Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal 576104, India; srikanth.prabhu@manipal.edu
[6] Department of Computer Science, San Jose' State University, 1 Washington Sq, San Jose, CA 95192, USA; sengupta.sap@gmail.com
* Correspondence: gnath@murraystate.edu; Tel.: +1-270-809-2503

**Abstract:** Productivity losses caused by absenteeism at work cost U.S. employers billions of dollars each year. In addition, employers typically spend a considerable amount of time managing employees who perform poorly. By using predictive analytics and machine learning algorithms, organizations can make better decisions, thereby increasing organizational productivity, reducing costs, and improving efficiency. Thus, in this paper we propose hybrid optimization methods in order to find the most parsimonious model for absenteeism classification. We utilized data from a Brazilian courier company. In order to categorize absenteeism classes, we preprocessed the data, selected the attributes via multiple methods, balanced the dataset using the synthetic minority over-sampling method, and then employed four methods of machine learning classification: Support Vector Machine (SVM), Multinomial Logistic Regression (MLR), Artificial Neural Network (ANN), and Random Forest (RF). We selected the best model based on several validation scores, and compared its performance against the existing model. Furthermore, project managers may lack experience in machine learning, or may not have the time to spend developing machine learning algorithms. Thus, we propose a web-based interactive tool supported by cognitive analytics management (CAM) theory. The web-based decision tool enables managers to make more informed decisions, and can be used without any prior knowledge of machine learning. Understanding absenteeism patterns can assist managers in revising policies or creating new arrangements to reduce absences in the workplace, financial losses, and the probability of economic insolvency.

**Keywords:** absenteeism; multi-class classifications; multinomial logistic regression; support vector machines; random forests; artificial neural networks



**Citation:** Nath, G.; Wang, Y.; Coursey, A.; Saha, K.K.; Prabhu, S.; Sengupta, S. Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism. *Information* **2022**, *13*, 320. https://doi.org/10.3390/info13070320

Academic Editors: Agnes Vathy-Fogarassy and János Abonyi

Received: 7 May 2022
Accepted: 25 June 2022
Published: 30 June 2022



## 1. Introduction

Absenteeism has been identified as an important factor in company performance and productivity losses. A study conducted by the Bureau of Labor Statistics suggests that nearly 2.8 million workdays are lost each year due to absenteeism at work [1]. Despite employers' expectations, excessive absences may reduce productivity and negatively impact the company's finances and other aspects [2]. The Gallup Wellbeing Index surveyed over 94,000 individuals from 14 major occupations and determined that absenteeism among

particular U.S. workers costs USD 153 billion annually [3]. It is alarming to learn that many employers in the United States are unaware of the extent of absenteeism in the workplace. Less than one half of all companies have a system for tracking absenteeism, and only 16 percent have measures to reduce absenteeism [1]. Consequently, absenteeism has a significant impact on a company's financial performance, as it has a direct and powerful effect on the organizational structure [1,3,4]. In order to effectively handle employee absenteeism, an organization first needs to understand the causes and patterns of absence from multiple perspectives in order to properly classify the reasons for employee absence. Absenteeism has generally been considered a significant challenge in human resource management in various industries and organizations [5]. Artificial intelligence and predictive analytics are perceived as key drivers in improving organizational performance and productivity [6]. In this study, we explore those factors that are closely related to absenteeism at work with a high degree of accuracy. Furthermore, we propose a decision support tool for human resource managers that provides a classification of a particular employee guided by the CAM theory. With this tool, managers without expertise in the field of machine learning can gather more informed information to revise policies or develop new plans to reduce absences from work, thus reducing the adverse effects of workplace absences on the productivity and the performance of the company. The paper is organized as follows. The proposed interactive web-based tool is described in Section 2, along with a brief review of the literature related to predicting work-related absences. A detailed examination of the dataset used in this study is presented in Section 2.1. Different multiclass classification models are compared in order to determine the most effective model for predicting workplace absenteeism of potential candidates, and are discussed in Section 3, along with the performance metrics used to select the most effective model for subsequent integration with the web-based interactive tool. In Section 4, we describe the results obtained by the multiclass classification models and provide recommendations for selecting the best model based on performance metrics. Incorporating the best model as found in Section 4, Section 5 describes the developed interactive absenteeism management tool. Moreover, we present a demonstration of how to use our proposed tool to identify the absenteeism class of a potential candidate. Finally, concluding remarks are provided in Section 6.

## 2. Literature Review and Proposed Method

There are many types of employee leaves of absence, including short-term and long-term disability, workers' compensation, family and medical leave, and military leave. Additionally, there is evidence that companies with low morale suffer from higher rates and costs of unscheduled absences. Based on the 2005 CCH Survey, only 35% of unscheduled absences are attributed to personal illness, while 65% are attributed to other reasons, including family concerns (21%), personal needs (18%), entitlement attitude (14%), and stress (12%) [7]. Tunceli et al. [8] found that male employees with diabetes have a 7.1% lower likelihood of working and female employees with diabetes have a 4.4% lower likelihood of working than those without diabetes. Furthermore, the study conducted by Halpern et al. showed that smoking policies in the workplace affect absenteeism and productivity, concluding that smokers are more likely to be absent from work than former smokers and nonsmokers [9]. Researchers have increasingly used artificial intelligence (AI)-based methods in recent years to model absenteeism problems that have a detrimental impact on the company's infrastructure. Using Naive Bayes, Decision Trees, and Multilayer Perceptrons, Gayathri predicted absenteeism at the workplace and recommended multilayer perceptrons on the basis of validation score [10]. Using a multilayer perceptron with the error back-propagation algorithm, Martiniano et al. proposed a neuro-fuzzy network to predict workplace absenteeism [11].

In a recent study, Skorikov et al. [12] compared six different data mining techniques to predict multiclass absenteeism at work, concluding that KNN classifiers with the Chebyshev distance metric performed the best. Although the authors divided absenteeism into different degrees, all models showed significantly lower prediction power. Furthermore,

the authors did not follow the recommended performance metric guideline [13] when comparing methods for imbalanced data.

Thus, we have utilized most of the current absenteeism research and followed the methodology of Cognitive Analytics Management (CAM) theory in order to examine absenteeism issues in the workplace. CAM theory, which was proposed Osman et al., consists of the following three steps [14]:

(a) Cognitive process: acquire a thorough understanding of a problem, identifying the attributes and relationships that align with the desired goals and objectives determined in coordination with executive authorities.
(b) Analytical process: determine the most appropriate analytical models and methods for addressing the identified challenge and achieving the desired objectives, then validate the outcomes and convey the findings to executive authorities.
(c) Management process: A critical component of the CAM theory that is essential for the successful commencement and completion of any analytical project.

As part of the cognitive process, we preprocessed the data and explored the factors related to absenteeism through the use of ANOVA F-values and Random Forest feature selection. In the analytical process, we trained different machine learning classification models based on the attributes selected during the cognitive process, then determined the most significant model by comparing their performance metrics.

Furthermore, human resource managers may not have sufficient knowledge of the data mining techniques that used in existing models. In order to resolve this issue, we propose an automated decision support tool guided by the final stage (management process) of the CAM theory. For example, Delen, Sharda, & Kumar [15] developed an automated web-based tool integrating prediction models to provide Hollywood producers with a way of classifying a movie into one of nine success categories, ranging from flops to blockbusters. Simsek et al. [16] have developed an automated tool that uses artificial neural networks (ANN) to identify point velocity profiles on rivers with an accuracy level of 0.46. Figure 1 provides an overview of how the web-based interactive tool proposed here was developed and how it can be utilized. This tool ensures that the end-user does not need to know anything about machine learning in order to make predictions about absenteeism; they must simply input data into the tool and click ”predict”.

The methodology in this paper is organized as follows. To begin, we preprocess the data by performing feature selection and scaling, one-hot encoding, and classification of absenteeism hours. We use the Synthetic Minority Over-Sampling Technique (SMOTE) to improve the performance of the classification models. In the next step, we split the data into training and testing sets, train four models (MLR, SVM, ANN, and RF) using the training data, and predict absenteeism classes using the testing data. Utilizing performance metrics, we compare and choose the most suitable model, then integrate the selected model into a proposed web-based interactive tool. Last, we provide a brief description of how the user can access the interactive tool.

### *2.1. Preprocessing and Cross-Validation*

The dataset for this study was obtained from the UCI Machine Learning Repository and provided by Martiniano, Ferreira, Sassi & Affonso [11]. This database was created using absenteeism records from a Brazilian courier company from July 2007 to July 2010. The dataset consists of 740 instances and 21 attributes. For more detailed information regarding the dataset, please see Skorikov et al. [12]. In accordance with information provided by the data source (UCI Machine Learning Repository) and authors [11], the dataset permits several kinds of attributes to be combined and excluded, and permits modification of the type of attributes (categorical, integer, or real) depending on the purpose of research.

There are several categorical attributes in the Absenteeism dataset (month of absence, day of the week, season, and more). In machine learning models, it is assumed that two numbers that are very close will appear to be more similar than two numbers that are

further apart. However, this is not always the case for categorical attributes [17]. With One-Hot-Encoding, the machine learning algorithm does not assume that larger numbers are more significant. Thus, we applied One-Hot-Encoding (by creating a binary column) to the categorical attributes [18]. An important step in the development of a machine learning model is to determine the importance of features. A number of features are redundant or do not contain useful information. Features can be selected appropriately based on their importance. We utilized ANOVA- F-value attribute selection (ANOVA-FS) and Random Forest attribute importance (RFFI) to determine the feature importance. In ANOVA, each attribute is ranked by calculating the ratio of variances between and within groups [19]. Based on this ratio, it can be determined how strongly the $i$-th attribute is related to the group attribute [20].

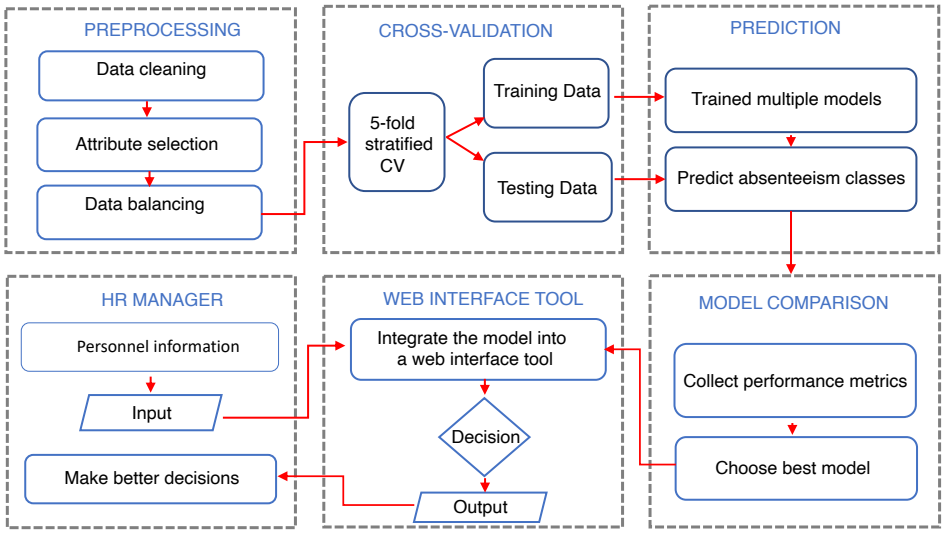


**Figure 1.** A flowchart outlining the proposed methodology.

During a Random Forest training process, the significance of each attribute is quantitatively measured by the Gini index (error rate). In this way, the importance of each attribute can be ranked [21]. Further information about random forest attributes can be found in Sarelaa et al. [22].

We have listed the significant attributes for each of the methods in Table 1 with a checkmark. For the RFFI method, we set a significant level of 0.025 for all attributes; thus, any attributes with an importance level less than 0.025 were considered to have insufficient predictive power. The ANOVA-FS and RFFI both select ten attributes, although the selected attributes within each model are not identical. We selected the attributes from the union set of ANOVA-FS and RFFI.

As outlined in our proposed study, *absenteeism in hours* is a variable of interest. The absenteeism rate is divided into three categories: *none* means that an employee is never absent; *moderate* involves employees who are absent for 1 to 15 h per month; and *excessive* refers to employees who are absent for 16 to 120 h per month. For simplicity, we refer to these three groups as $A^+$, $B^+$, and $C^+$, respectively. Skorikov et al. [12] introduced the concept of the classification of absenteeism data, which is helpful when comparing groups within an organization. Furthermore, these predefined classes provide an opportunity to compare our findings with existing results.

**Table 1.** Feature importance as determined by Random Forest.

| Attribute | ANOVA-FS | RFFI | ANOVA-FS U RFFI |
|---|---|---|---|
| Age | ✓ | | ✓ |
| Body mass index | ✓ | | ✓ |
| Children | | | |
| Day of the week | ✓ | ✓ | ✓ |
| Disciplinary failure | ✓ | ✓ | ✓ |
| Distance from residence to work | | ✓ | ✓ |
| Education | | | |
| Height | ✓ | ✓ | ✓ |
| Hit target | ✓ | ✓ | ✓ |
| Month of absence | | ✓ | ✓ |
| Pet | | | |
| Reason for absence | ✓ | ✓ | ✓ |
| Seasons | ✓ | ✓ | ✓ |
| Service time | | | |
| Social drinker | ✓ | | ✓ |
| Social smoker | | | |
| Transportation expense | | | |
| Weight | | ✓ | ✓ |
| Workload per day | ✓ | ✓ | ✓ |
| Total | 10 | 10 | 13 |

We trained multiple classification models using selected attributes from each method and the dependent attribute *absenteeism in hours*. Moreover, we applied five-fold stratified cross validation before training with the absenteeism dataset in order to evaluate model performance. Five-fold stratified cross-validation involves randomly dividing the dataset into five equal folds, each of which contains the same number of classification classes. The test set consists of one fold and the training set consists of four folds. The test set changes every time, and the remaining four subsets are used as training sets to produce a total of five models. We considered the average of five performance metrics across five models [23,24]. Furthermore, due to the categorization of *absenteeism in hours*, three imbalanced classes were formed ($A+ = 6\%, B+ = 85\%$, and $C+ = 9\%$). Therefore, in order to increase the performance of the models, we only applied SMOTE to the training dataset, leaving the testing dataset unchanged. A major objective of SMOTE is to create data for the minority class to support the balance between classes. This process generates new data points using the k-nearest neighbors algorithm [25].

Having applied all of the above techniques to the dataset, we trained multiple models; for brevity, we included only the four best-performing models (SVM, MLR, ANN, and RF) in this study. The following section provides a brief description of each of the classification models.

## 3. Prediction Models

For our study, we analyzed four commonly used models to classify the absenteeism category of potential employees. The following subsections outline the method and corresponding optimal hyperparameters calculated by the grid search algorithm.

### *3.1. Support Vector Machine*

The support vector machine (SVM) has become very popular because it offers significant accuracy with minimal computational power. Support vector machines perform reasonably well with linear dependencies, have reasonable performance with sparse data sets, and can be used for a wide variety of data types [6]. The purpose of the support vector machine algorithm is to produce a hyperplane capable of differentiating between two dif-

ferent classes of data. A separating hyperplane with the largest margin defined by $d = \frac{2}{||a||}$ maximum distance between data points of two classes (where vector $a$ is perpendicular to the separating hyperplane specified) is shown in Figure 2 [26]. The hyperplane may not be readily available in certain cases due to the greater dimensions of the problem, in which case a kernel function helps in the smooth computation of the problem [27]. In order to apply SVM to multi-class classification problems, it is common to divide the problem into multiple binary classification subsets and then apply a standard SVM to each subset, which is called the one-versus-rest technique. In order to achieve optimal accuracy, we tested several different parameters, and found that the radial basis kernel function provides the highest level of accuracy.

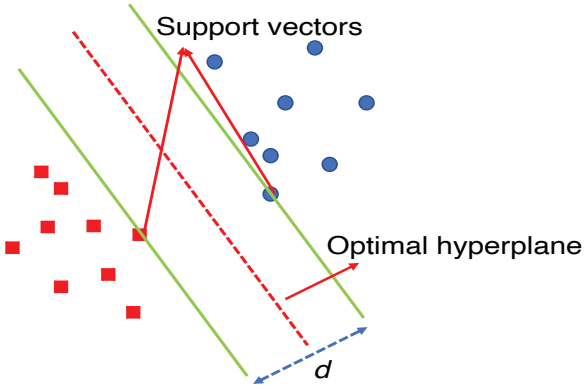


**Figure 2.** Optimal classification algorithm.

### *3.2. Multinomial Logistic Regression*

Multinomial logistic regression is an extension of binary logistic regression in which the outcome variable can be categorized into more than two categories. In order to extend logistic regression to multiclass classification problems, one commonly used approach is to divide the multiclass classification problem into a series of binary classification sets and fit a standard logistic regression model for each subset. Consider $h_{mn}$ as the success ($h_{mn} = 1$) or failure ($h_{mn} = 0$) of multinomial outcomes $n$, $n = 1, \ldots, N$ for observation $m$, $m = 1, \ldots, M$. Consider $x_m$ to denote observation $m's$ $K$-dimensional vector of the predictor variable, $k = 1, \ldots, K$. Based on reference outcome $N$, the multinomial logistic regression (MLR) can be defined to predict probabilities $\pi_{mn}(x_m)$ for outcomes $n = 1, \ldots, N-1$ as follows [28]:

$$\pi_{mn}(x_m) = \frac{\exp(\boldsymbol{\lambda}_n + \boldsymbol{\theta}_n' \boldsymbol{x}_m)}{1 + \sum_{k=1}^{J-1} \exp(\boldsymbol{\lambda}_k + \boldsymbol{\theta}_k' \boldsymbol{x}_m)}, \quad (1)$$

where $\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nK})'$ refers to the coefficients for the $n$th linear predictor, excluding its intercept, $\lambda_n$. The log-likelihood method is used to estimate $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, providing normal and consistent estimates. Here, the Newton method was used to optimize the problem for the most accurate prediction.

### *3.3. Artificial Neural Networks*

In recent years, artificial neural networks (ANN) have been applied to various fields thanks to their ability to model highly challenging problems. ANNs are new and useful models when applied to problem solving and machine learning. This is a model of information management that is comparable to the function of the human nervous system. A key feature of the human brain is its ability to process information in a unique manner. Many interconnected neurons serve as components of the system, and work in concert to solve specific problems on a daily basis [29]. An ANN consists of nodes, representing neurons, and connections between nodes, representing axons and dendrites carrying information. There is a value or weight attached to every connection between two nodes for the purpose of assessing the strength of the signal [30]. The neurons are arranged in layers, with an

input layer representing one type of input data, an output layer representing the result of the classification, and one or more hidden layers. One of the most common and widely used forms of ANN is the perceptron, which is a fully connected feed-forward network [31]. The linear combination of weights and input values is passed through a non-linear function, known as an activation function [32]. Neural activation functions approximate the complex physical processes of neurons, which modulate their output in a non-linear way. The architecture of an artificial neural network is shown in Figure 3 [33]. We built a six-layer fully connected ANN, in which each neuron in one layer is connected to every neuron in the following layer. After one hot encoding and standardization, the first input layer consists of 42 input neurons. Our results indicated that the highest degree of accuracy was achieved using a network with four hidden layers, consisting of 400, 100, 50, and 20 neurons (nodes), respectively, and with the output layer consisting of three neurons using the *relu* activation function.

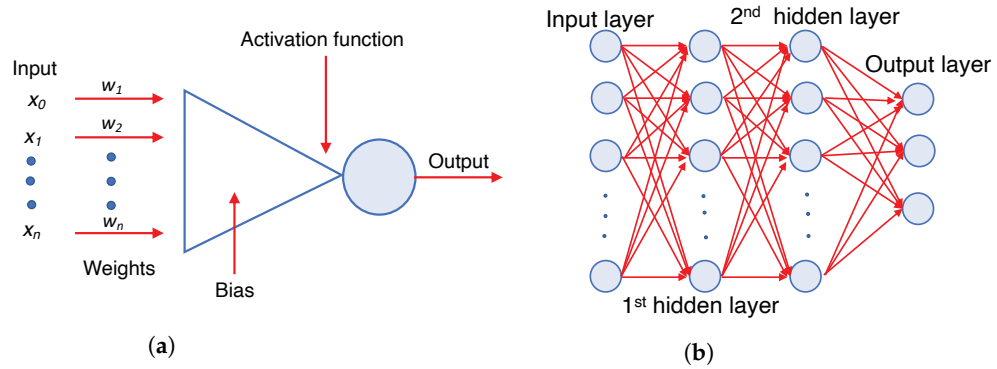


**Figure 3.** Architecture of artificial neural networks: (**a**) the input values are transformed in a perceptron by the weights, biases, and activation functions, then output values are sent to the next perceptron; (**b**) a multilayered perceptron is composed of several perceptrons.

### *3.4. Random Forest*

Decision trees are the core component of random forest classifiers. Using the features of a data set, a decision tree is built into a hierarchical structure. In the decision tree, each node represents a measure associated with a subset of features [34]. A random forest is composed of trees that produce class predictions for each tree, and the class that receives the most votes becomes the model prediction [35]. In this study, Gini impurity-supported criteria were used to determine most accurate predictions.

### *3.5. Performance Metrics*

The following section demonstrates several of the statistical assessment metrics that we used to validate our model's performance.

Accuracy: An important metric for evaluating classification models is accuracy. Accuracy can be determined based on binary classification, as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where $TP$ stands for true positives, $TN$ for true negatives, $FP$ for false positives, and $FN$ for false negatives [36].

With imbalanced datasets, accuracy can be misleading; therefore, there are additional metrics found in a confusion matrix that can be utilized to evaluate performance.

Precision: Precision is a popular metric in classification systems. A measure of how well a model is able to predict positive values is referred to as precision [36]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall: Recall addresses imbalances that may occur in a dataset, and is defined as follows [37]:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{4}$$

F1-score: In terms of precision and recall, the F-measure is defined as follows [38]:

$$F - \text{score} = (1 + \alpha^2)\frac{P \times R}{\alpha^2(P + R)}, \tag{5}$$

where $P$ and $R$ are the precision and recall, respectively, and $\alpha \geq 0$ represents the balance between $P$ and $R$. This is commonly referred to as the F1 score when $\alpha = 1$ [39].

ROC AUC score: Over the past few decades, receiver operating characteristic (ROC) curves have become popular, and have been widely used as a tool to evaluate the discrimination ability of various machine learning methods for predictive purposes [40]. Better models pass through the upper left corner and have greater overall testing accuracy. One of the most widely used metrics for assessing the performance of models is the Area Under Curve (AUC), which provides the ability of the classifier to distinguish between classes and is used as a summary of the ROC curve. AUC values are generally between 0.5 and 1.0, and a larger AUC indicates better performance. In the case of a perfect model, the AUC would be 1, indicating that all positive examples are always in front of all negative examples [39]. There are two popular methods for evaluating multi-class classification problems. The one-versus-one algorithm computes the average of the pairwise ROC AUC scores, while the one-versus-rest algorithm computes the average of the ROC AUC scores for each class compared to all other classes [41].

## 4. Results and Model Comparison

In this article, we explored four popular machine learning algorithms for predicting employee absenteeism. We used different methods of selecting the most significant set of attributes, as discussed in Section 2.1. Our primary objective was to determine the most appropriate model using performance metrics and develop a decision support tool for human resource managers. We employed four different classification models across three different sets of attributes. In comparing the results shown in Table 2, we found that the AUC score for the attribute set selected by ANOVA-FS $\cup$ RFFI was the highest of all models. Therefore, SVM and MLR had the highest AUC scores overall compared to other models. It should be noted that as concerns predictions, there is no universally accepted best method that is applicable in all problems. Trial and error and experimentation are required in order to find the best model for each scenario [42]. Furthermore, Delen et al. recommend incorporating the knowledge of multiple experts in the development and training of a model [15]. To develop an appropriate model, companies should consult with experts in the relevant field and choose the appropriate attributes based on their experience and geographic location. Saidene et al., for instance, investigated the factors leading to work absenteeism in Tunisia and concluded that excessive work hours, poor posture, workplace stress, and insufficient rest time are significant attributes leading to absenteeism [43].

As the ANOVA-FS $\cup$ RFFI set provided the overall best performance, we selected these thirteen attributes for the final model. The relative importance of the attribute set is presented in Figure 4. It can be seen that *reason for absence* has the highest importance compared to others. Almost all levels of *reason for absence* are caused by health-related absences. The establishment of an exercise center at the workplace may be one solution to reducing absenteeism. In an analysis of 517 employees selected randomly, Baun et al. explored the differences between exercisers and non-exercisers in terms of health care costs and absenteeism [44]. They found that exercise reduced illness absence among exercisers and increased illness absence among nonexercisers. The findings of our study are therefore consistent with those of their experiments, demonstrating a significant association between illness and absenteeism at work. In addition, Dula et al. conducted an experimental study of absenteeism among medical staff at Arba Minch General Hospital and concluded that

workload has a positive relationship with absenteeism [45]. From Figure 4, it can be seen that *workload* is an important factor. In this case, a practical solution would be to decrease workload by hiring more employees. Moreover, as identifying the maximum workload for an employee is a critical decision, our proposed web-based tool can help managers in finding possible maximum workloads; this issue is discussed in detail in the next section.

Categorizing employees' absenteeism in hours generates imbalanced classes. Numerous suggestions have been made as to how to overcome the negative effects of such an imbalance on performance metrics (for further details, see Luque, Carrasco, Martin & Heras [46]). On the basis of our experimental results, and as suggested by Johnson, Halbesleben, Marilyn & Khoshgoftaar [47] and Simsek et al. [6], we determined the best model based on the AUC score and confusion matrix. In Table 2, we report the accuracy, weighted F1-score, weighted precision, and the one-versus-one macro ROC AUC scores. A previous approach Skorikov et al. [12] proposed a KNN classifier with a Chebyshev distance metric that achieved an AUC score of only 0.69.

Table 2 demonstrates that, under all of the scenarios assessed, our models performed better than their proposed model. This may be due to the fact that we applied One-Hot-Encoding to the categorize attributes and standardized the continuous attributes before using them to train the models.

We selected SVM for the web-based supporting tool because of its accuracy, which was 100% for class $A^+$, 85% for class $B^+$ is 85%, and 77% for class $C^+$, the best overall performance in comparison to the other models. Skorikov et al. proposed KNN classifiers with the Chebyshev distance metric, obtaining an accuracy of 67% for class $A^+$, 92% for class $B^+$, and 8.3% for class $C^+$ [12]. Thus, our proposed model for absenteeism data shows a substantial improvement in terms of both AUC score and confusion matrix.

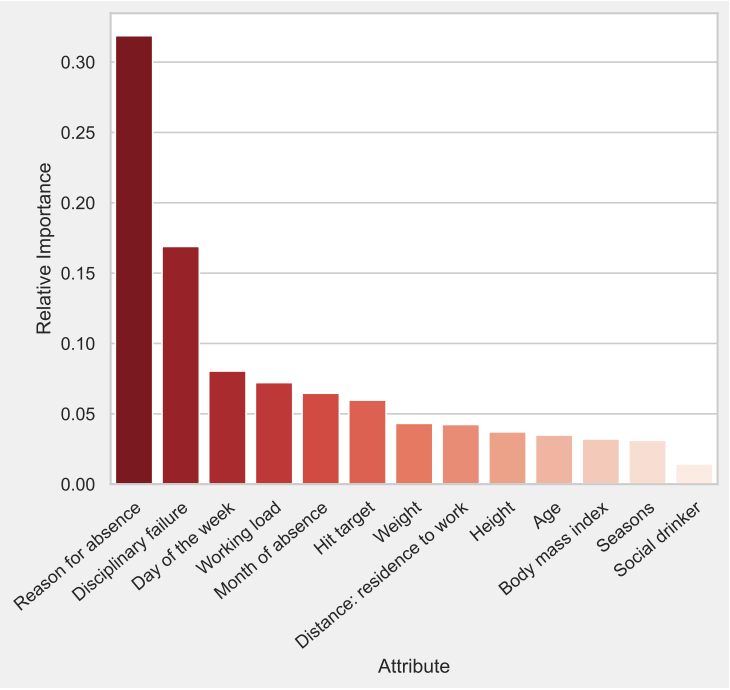


**Figure 4.** Relative importance of the individual attributes.

**Table 2.** Performance of the trained models when applied to the test data.

| Model | Method | Accuracy | Precision | Recall | F1-Score | AUC Score |
|---|---|---|---|---|---|---|
| SVM | ANOVA-FS | 0.831 | 0.915 | 0.831 | 0.859 | 0.948 |
| | RFFI | 0.810 | 0.912 | 0.810 | 0.845 | 0.946 |
| | ANOVA-FS ∪ RFFI | 0.790 | 0.909 | 0.790 | 0.830 | 0.951 |
| MLR | ANOVA-FS | 0.797 | 0.929 | 0.797 | 0.837 | 0.946 |
| | RFFI | 0.797 | 0.929 | 0.797 | 0.837 | 0.947 |
| | ANOVA-FS ∪ RFFI | 0.783 | 0.927 | 0.783 | 0.847 | 0.951 |
| ANN | ANOVA-FS | 0.864 | 0.866 | 0.864 | 0.864 | 0.930 |
| | RFFI | 0.858 | 0.880 | 0.858 | 0.867 | 0.928 |
| | ANOVA-FS ∪ RFFI | 0.878 | 0.887 | 0.878 | 0.881 | 0.941 |
| RF | ANOVA-FS | 0.885 | 0.868 | 0.885 | 0.874 | 0.934 |
| | RFFI | 0.905 | 0.892 | 0.905 | 0.894 | 0.938 |
| | ANOVA-FS ∪ RFFI | 0.891 | 0.886 | 0.891 | 0.889 | 0.939 |

## 5. Absenteeism Interactive Tool

As discussed earlier, this study followed the CAM theory and current research on absenteeism to identify the factors that are significantly associated with absenteeism at work. In this study, we predicted absenteeism classes using these factors in order to explore possible ways for organizations to reduce absenteeism. Here, we propose an interactive web application for absenteeism classification guided by the final step of CAM theory, namely, the management process.

It can be difficult for a human resource manager or project manager to predict employee absenteeism classes using a machine learning algorithm, especially if he or she lacks programming expertise. Our proposed interactive web-based tool enables human resource or project managers to predict absenteeism by analyzing all current employees and estimating how many will be absent, allowing them to make appropriate adjustments and preparations ahead of time. In addition, managers will be able to save significant amounts of time by eliminating the need to train a machine learning algorithm. We present a prototype of the proposed tool based on a publicly available open-access dataset collected by Martiniano, Ferreira, Sassi & Affonso [11] from a Brazilian courier company in 2012. This dataset is widely used, and there has been a significant amount of research published on absenteeism using this dataset. A prototype of our proposed web-based tool is accessible at the following link: https://share.streamlit.io/gopalnath1926/app_absenteeism_new/main/app_absenteeism.py (accessed on 24 June 2022).

It is important to note that this dataset represents specific human behaviors from the area in which it was collected. Related behaviors may be different in other areas of the world [48] or in different companies. Therefore, companies should consult with experts and select appropriate attributes (possibly based on internal data from their own company) for training the model and developing an effective version of the proposed web-based tool for human resources and project managers in order to improve workplace absenteeism problems.

It should be noted that this prototype is presented only for the benefit of the reader as a demonstration of the proposed tool. One of the referees remarked that this prototype tool, having been trained by the Brazilian courier dataset, may not be applicable or useful to other organizations. Perhaps a fully implemented version of the tool could streamline this process by allowing a company to upload its absentee data, automatically develop a model based on the process proposed in this study, and quickly create a unique version of the tool for their own human resource or projects managers.

Furthermore, based on the performance analysis discussed in the previous section, SVM was chosen for the prototype. We developed the prototype using the Streamlit, an open-source Python framework for developing web applications. By collecting an employee's information, the end-user can then input all of the information and, by clicking

on the predict button, predict the class of the employee. The end-user does not need to know anything about machine learning to make predictions; they simply input data into the tool and click predict. As shown in Figure 5, the output (predicted) class for the desired candidate is $A^+$ on a scale of $A^+B^+C^+$, indicating that this particular employee has a high degree of sincerity in the workplace and is less likely to miss work. Furthermore, as a follow-up to the problem we mentioned above, Dula et al. concluded that workload was a significant factor in predicting absenteeism in an experimental study on medical staff at Arba Minch General Hospital [45]. Our proposed interactive tool can help significantly in adjusting the workload of an employee. According to Figure 5, if we change the *workload average/day* from 70 to 73, the prediction class changes from $A^+$ to $C^+$, assuming that all other attributes remain constant. Therefore, human resources or project managers can determine the maximum workload an employee can handle based on all of the input information, thereby avoiding absenteeism. Moreover, we conclude that there is a positive association between workload and absenteeism, which is consistent with the experimental results of Dula et al.

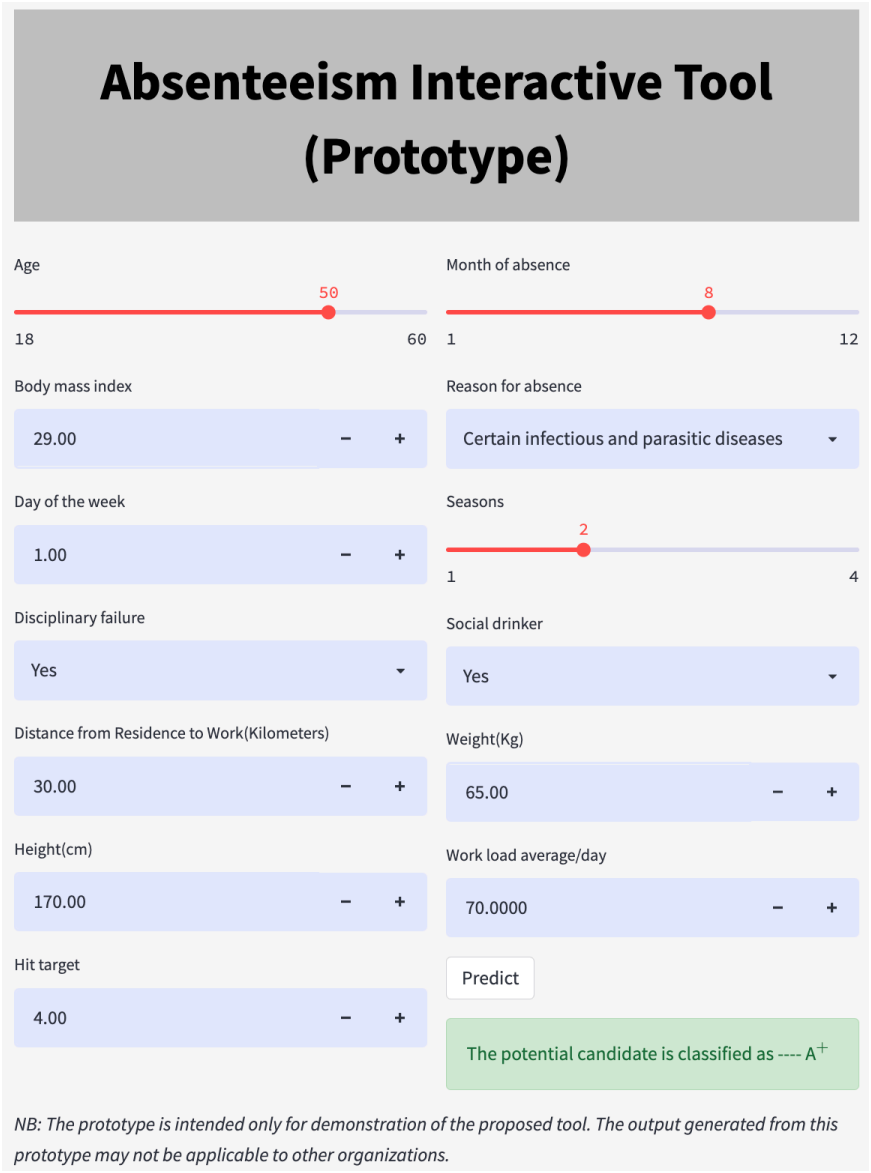


**Figure 5.** A screenshot of the prototype. Based on input values, the tool classifies the employee as $A^+$ on a scale of $A^+B^+C^+$. The web-based tool is accessible at https://share.streamlit.io/gopalnath1926/app_absenteeism_new/main/app_absenteeism.py (accessed on 24 June 2022).

## 6. Conclusions

The cost of absenteeism is in the billions of dollars, and this does not include the time spent managing employees who are not performing well. Thus, absenteeism has a highly negative effect on a company's organizational structure. Companies differ depending on their working environment, type of work, geographical location, etc. Therefore, in order to effectively handle employee absenteeism, which adversely affects a company's financial stability, the company needs to identify the cause of the problem. Consequently, companies spend large amounts of money hiring additional management staff to identify causes and find suitable solutions.

To avoid negative impacts on companies related to employee absenteeism, machine learning algorithms can determine the potential class of employees in terms of absenteeism. We have utilized current studies on absenteeism and followed the guidelines of CAM theory in order to address the issue of absenteeism. Identifying absenteeism classes with a greater degree of accuracy is essential. We applied a hybrid optimization approach to identify a set of attributes that are significantly associated with absenteeism. Subsequently, a model was developed to predict absenteeism classes with a greater degree of accuracy. Furthermore, using machine learning algorithms can be quite challenging, as they often require a high level of knowledge, a significant amount of time, and strong programming abilities. Human resource managers may not have enough experience with machine learning algorithms, or may not have the time to develop such an algorithm. To address these issues, we have proposed a web-based interactive tool. As a proof-of-concept of the proposed tool, we developed a prototype using the Python Streamlit framework with an integrated SVM model (which was our best-performing model overall, based on experimental study). This proposed web-based tool serves as a link between machine learning algorithms and human resource managers. The proposed web-based interactive tool is very useful for human resource or project managers, allowing them to predict employee absences, determine how many will be absent, and adjust plans in advance. Additionally, managers can identify a possible threshold value of workload or hit a target workload for particular employees in order to prevent potential absenteeism. End users do not need to have any previous knowledge about how machine learning algorithms work in order to use our proposed tool. They can determine the absenteeism class by inputting the relevant information into the tool and clicking the predict button. Thus, human resource managers can utilize this simple tool to save time, decrease workloads, and make better decisions, preventing companies from experiencing adverse financial consequences.

**Author Contributions:** Data curation, G.N.; formal analysis, G.N.; methodology, G.N.; software, G.N.; validation, G.N., Y.W., A.C., S.P. and S.S.; writing—original draft, G.N.; writing—review and editing, G.N., Y.W., A.C., K.K.S., S.P. and S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data that support the findings of this study are publicly available in the UCI Machine Learning repository at https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work (accessed on 1 November 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kocakulah, M.C.; Kelley, A.G.; Mitchell, K.M.; Ruggieri, M.P. Absenteeism problems and costs: Causes, effects and cures. *Int. Bus. Econ. Res. J. (IBER)* **2016**, *15*, 89–96. [CrossRef]
2. Prater, T.; Smith, K. Underlying Factors Contributing To Presenteeism And Absenteeism. *J. Bus. Econ. Res.* **2011**, *9*, 1–14. [CrossRef]
3. Simpson, I. Ailing and Overweight Americans Cost Billions in Productivity. *Reuters*. Available online: https://www.reuters.com/article/us-absenteeism-idUSTRE79G17X20111017 (accessed on 20 November 2021).
4. Kim, J.; Sorhaindo, B.; Garman, E.T. Relationship between Financial Stress and Workplace Absenteeism of Credit Counseling Clients. *J. Fam. Econ.* **2006**, *27*, 458–478. [CrossRef]

5. Halbesleben, R.B.H.; Whitman, M.V.; Crawford, W.S. A dialectical theory of the decision to go to work: Bringing together absenteeism and presenteeism. *Hum. Resour. Manag. Rev.* **2014**, *24*, 172–192. [CrossRef]
6. Simsek, S.; Albizri, A.; Johnson, M.; Custis, T.; Weikert, S. Predictive data analytics for contract renewals: A decision support tool for managerial decision-making. *J. Enterp. Inf. Manag.* **2021**, *34*, 718–732. [CrossRef]
7. Navarro, C.; Bass, C. The Cost of Employee Absenteeism. *Sage J.* **2016**, *38*, 26–30. [CrossRef]
8. Tunceli, K.; Bradley, C.J.; Nerenz, D.; Williams, L.K.; Pladevall, M.; Lafata, J.E. The impact of diabetes on employment and work productivity. *Diabetes Care* **2005**, *28*, 2662–2667. [CrossRef]
9. Halpern, M.T.; Shikiar, R.; Rentz, A.M.; Khan, Z.M. Impact of smoking status on workplace absenteeism and productivity. *Tob. Control* **2001**, *10*, 233–238. [CrossRef]
10. Gayathri, T. Data mining of absentee data to increase productivity. *Int. J. Eng. Tech.* **2018**, *4*, 478–480.
11. Martiniano, A.; Ferreira, R.P.; Sassi, R.J.; Affonso, C. Application of a neuro fuzzy network in prediction of absenteeism at work. In Proceedings of the 7th Iberian Conference on Information Systems and Technologies (CISTI 2012), Madrid, Spain, 20–23 June 2012; pp. 1–4.
12. Skorikov, M.; Hussain, M.R.; Akbar, M.K.; Momen, S.; Mohammed, N.; Nashin, T. Prediction of absenteeism at work using data mining techniques. In Proceedings of the 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2–4 December 2020; pp. 1–6. [CrossRef]
13. Williams, D.P.; Myers, V.; Silvious, M.S. Mine Classification with Imbalanced Data. *EEE Geosci. Remote Sens. Lett.* **2009**, *6*, 528–532. [CrossRef]
14. Osman, I.H.; Anouze, A.L.; Irani, Z.; Lee, H.; Medeni, T.D.; Weerakkod, V. A cognitive analytics management framework for the transformation of electronic government services from users' perspective to create sustainable shared values. *Eur. J. Oper. Res.* **2021**, *278*, 514–532. [CrossRef]
15. Delen, D.; Sharda, R.; Kumar, P. Movie forecast guru: A web-based DSS for hollywood managers. *Decis. Support Syst.* **2007**, *43*, 1151–1170. [CrossRef]
16. Simsek, S.; Genc, O.; Albizri, A.; Dinc, S.; Gonen, B. Artificial neural network incorporated decision support tool for point velocity prediction. *J. Bus. Anal.* **2020**, *3*, 67–78. [CrossRef]
17. Zhang, H.; Zheng, G.; Xu, X.; Yao, X. Research on the Construction and Realization of Data Pipeline in Machine Learning Regression Prediction. *Math. Probl. Eng.* **2022**, *2022*, 7924335. [CrossRef]
18. Imran, A.A.; Amin, M.N.; Rifat, M.R.; Mehreen, S. Deep Neural Network Approach for predicting the productivity of garment employees. In Proceedings of the IEEE, 6th International Conference on Control, Decision and Information Technologies (CoDIT), Paris, France, 23–26 April 2019; pp. 1402–1407. [CrossRef]
19. Johnson, J.K.; Synovec, R.E. Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 225–237. [CrossRef]
20. Nasiri, H.; Alavi, S.A. A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-ray Images. *Comput. Intell. Neurosci.* **2021**, *2022*, 4694567. [CrossRef]
21. Fei, H.; Fan, Z.; Wang, C.; Zhang, N.; Wang, T.; Chen, R.; Bai, T. Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier. *Remote Sens.* **2022**, *14*, 829. [CrossRef]
22. Saarela, M.; Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl. Sci.* **2021**, *3*, 272. [CrossRef]
23. Jiang, H.J.; Huang, Y.; You, Z.H. Predicting Drug-Disease Associations via Using Gaussian Interaction Profile and Kernel-Based Autoencoder. *BioMed Res. Int.* **2019**, *2019*, 11. [CrossRef]
24. May, R.J.; Maier, H.R.; Dandy, G.C. Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Netw.* **2010**, *23*, 283–294. [CrossRef]
25. Chawla, N.V.; Bower, K.; Hall, L.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
26. Cervantes, J.; Lamont, G.F.; Mazahua, R.M.A.L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
27. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, NY, USA, 2000; pp. 1–314, ISBN 978-0-387-98780-4.
28. Agresti, A. *Categorical Data Analysis*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2002; pp. 1–372, ISBN 978-0-471-22618-5.
29. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, 1–41. [CrossRef] [PubMed]
30. Yadav, A.M.; Chaurasia, R.C.; Suresh, N.; Gajbhiye, P. Application of artificial neural networks and response surface methodology approaches for the prediction of oil agglomeration process. *Fuel* **2018**, *220*, 826–836. [CrossRef]
31. Hallinan, S.J. *Computational Intelligence in the Design of Synthetic Microbial Genetic Systems*; Harwood, C., Wipat, A., Eds.; Elsevier: Amsterdam, The Netherlands, 2013; Volume 40, pp. 1–37, ISBN 978-0-1241-7029-2.
32. Young, W.A., II; Bihl, T.J.; Weckman, G.R. Artificial Neural Networks for Business Analytics. *Encycl. Bus. Anal. Optim.* **2014**, *40*, 193–208. [CrossRef]
33. Kim, J.; Hong, J.; Park, H. Prospects of deep learning for medical imaging. *Precis. Future Med.* **2018**, *2*, 37–52. [CrossRef]

34. Suthaharan, S. A cognitive random forest: An intra- and intercognitive computing for big data classification under cune condition. *Handb. Stat.* **2016**, *35*, 207–227. [CrossRef]
35. Sarica, A.; Cerasa, A.; Quattrone, A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Front. Aging Neurosci.* **2017**, *9*, 329. [CrossRef]
36. Kulkarni, A.; Chong, D.; Batarseh, F.A. *Foundations of Data Imbalance and Solutions for a Data Democracy*; Batarseh, F.A., Yang, R., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 83–106, ISBN 978-0-1281-8366-3.
37. Leonard, L.C. Web-based behavioral modeling for continuous user authentication (CUA). *Adv. Comput.* **2017**, *105*, 1–44. [CrossRef]
38. Sun, J.; Li, H.; Fujita, H.; Fu, B.; Ai, W. Class-imbalanced dynamic financial distress prediction based on adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf. Fusion* **2020**, *54*, 128–144. [CrossRef]
39. Zhang, K.; Su, H.; Dou, Y. Beyond AP: A new evaluation index for multiclass classification task accuracy. *Appl. Intell.* **2021**, *51*, 7166–7176. [CrossRef]
40. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
41. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
42. Ruiz, E.; Nieto, F.H. A note on linear combination of predictors. *Stat. Probab. Lett.* **2000**, *47*, 351–356. [CrossRef]
43. Saidane, O.; Mahmoud, I.; Gafsi, L.; Houda, A.; Tekaya, R.; Abdelmoula, L. Factors leading to work absenteeism in Tunisian ankylosing spondylitis patients. *Egypt. Rheumatol.* **2017**, *40*, 183–185. [CrossRef]
44. Baun, W.B.; Bernacki, E.J.; Tsai, S.P. A Preliminary Investigation: Effect of a Corporate Fitness Program on Absenteeism and Health Care Cost. *J. Occup. Med.* **1986**, *28*, 18–22. [CrossRef]
45. Dula, T.; Abara, G.; Reddy, P.A.K. The Assessment of Causes and Consequences of Medical Staff Absenteeism and Turnover in Arba Minch General Hospital. *J. Health Med. Nurs.* **2018**, *57*, 64–71.
46. Luque, A.; Carrasco, A.; Martin, A.; Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
47. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
48. Shah, S.A.A.; Uddin, I.; Aziz, F.; Ahmad, S.; Al-Khasawneh, M.A.; Sharaf, M. An enhanced deep neural network for predicting workplace absenteeism. *Complexity* **2020**, *2020*, 5843932. [CrossRef]