Predicting Absenteeism at Work Using Machine Learning Algorithms

Samir Qaisar Ajmi

Al Muthanna University, Almuthanna, Alsamawa, Iraq

*Corresponding Author : samir@mu.edu.iq

Received 12/10/2019 , Accepted 1/11/2019 , published:17/12/2019

DOI: 10.52113/2/07.01.2020/1-12

Abstract: To work in the commercial environment, the company needs to be a major competitor in the business market, which depends mainly on the company's resources. One of the most important resources is the employees. Based on that, the absence of the employees from work leads to deterioration and reduce production in the institutions which leads to heavy losses. There are many reasons why employees are absent from work. Those may include health problems and social occasions. The purpose of this paper was to apply machine learning techniques to predict the absenteeism at work. There are four methods have been used in this research (neural network(NN) technique ,decision tree (DT) technique, support vector machine (SVM) technique and logistic regression (LR) technique. . decision tree model has the highest accuracy equals to 83.33% with AUC 0.834 and the support vector machine has the lowest accuracy equals to 68.47 % with AUC 0.760.

Keyword: Decision Tree; SVM; ANN; Logistic Regression; Predicting Absenteeism at Work

1. Introduction

Competitiveness in the market is increasing day by day and this development has led to institutionalized pressures on employees to develop and increase production quantity and production quality. Absenteeism at work is a very important factor that can negatively affect the company's production rate and profits. Each company depends on its employees to provide its services to the customers. When the number of employees is reduced due to unplanned absents, this can reduce the quality or the quantity of the services provided by the company and that will affect the credibility of the company. To overcome this problem, the company must find the causes that lead to the employees' absence and it must try to eliminate or reduce these causes. These causes can be identified by analyzing the employees' absence patterns and employees' profiles. One of the important techniques used in the research related to the engagement of the employees and the human resource research in organizations is machine learning. The database used in this paper has 20 Attributes which created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. The purpose of this paper was to apply machine learning techniques to predict the absenteeism at work. There are four methods have been used in this research (neural network(NN) technique ,decision tree (DT) technique, support vector machine (SVM) technique and logistic regression (LR) technique. . decision tree model has the highest accuracy equals to 83.33% with AUC 0.834 and the support vector machine has the lowest accuracy equals to 68.47 % with AUC 0.760.

2. Literature Review

Much work have been done on predicting the absenteeism of the employees using many different machine learning approaches. In this section, a review of these approaches is presented.

Ferreira et al. have developed a prediction model using artificial neural network, ANN, in order to predict the absenteeism of the employee at work. They used a dataset that has 2,243 data record and it has thirty eight features. The data have been collected from 2008 to 2016. They reduced the number of the features with the Rough Sets to be seventeen attributes which were used in building the model. They have obtained a good results. [1]

Another research has been done by Harshit Trivedi to predict the Absenteeism at Workplace using ANN. He used a feed forward network design trained by the back propagation algorithm. He used a dataset that was collected from a courier company from 2007 to 2010. In addition, he used a rule based filter to identify the underperforming neurons. This technique increased the accuracy of the network which reached 58%.

Wahid et al. have used four machine learning techniques to predict the time of the absenteeism at work. They used a dataset collected from a courier company at Brazil. They have used Random Forest technique, Decision Tree technique, Gradient Boosted Tree, and Tree Ensemble technique. The best model was the Gradient Boosted Tree model that achieved 82% accuracy. [2]

Qomariyah and Sucahyo have built a predictive model using Decision tree algorithm. This model used to predict the attendance patterns of the employees at private Indonesian company. The model finds the employees' characteristics that are related to the pattern of frequent absence at the company. The dataset collected for a duration of two years at the company. The result of their research was creating five classification rules that could be used to predict the attendance of the employees. [3]

Asiri and Abdullah have used three machine learning techniques to predict absenteeism at work. These models include Naïve Bayes technique, Decision Tree algorithm, and Random Forest technique. They found that the Random Forest prediction model has achieved the best accuracy rate which equals 91%. These models have identified the causes of absenteeism at work. [4]

Oliveira et al. have used six machine learning techniques in order to predict the absenteeism at work for the employee at a phone company at Brazil. These techniques include Multilayer Perceptron, Naive Bayes, XGBoost, Random Forest, Support Vector Machine, and Long Short Term Memory. They built the models and used the evolutionary algorithms to tune the models' parameters. The model that achieved the high precision, 72%, was the XGBoost model. They collected the data of 13.805 employees at the company. The dataset contains 241 attributes. [5]

Ferreira et al. have used the neuro fuzzy technique to predict the absenteeism at work. They collected a dataset from a Courier company from July 2007 to July 2010. The dataset contains 21 features that represent the employees information. Their work found partial results of using the neuro fuzzy network to predict the absenteeism at work. [6]

3. The Methodology

In this section, a brief review of the machine learning techniques that is used in this research is introduced.

3.1 Artificial Neural Networks

Artificial Neural Network, ANN, can be represented by a set of input unites and output unites. These unites are connected to one another by connections that are weighted. In order to build an ANN model that can be used in prediction, the right weights has to be found. These weights can be learned from data tuples by altering the initial connections' weights so the model can predicts the correct target class for some input data tuples. The ANN model can be trained using learning algorithms such as the Backpropagation Algorithm. The ANN model is known by its favorable features such as resisting the noise that can be found in the data; its ability to produce a well-trained classification models when the relation between the target label and the dataset attributes is not obvious. Artificial Neural Network is a very popular machine learning technique which is used at many different tasks such as image recognition, handwritten recognition, and speech recognition. Artificial Neural Networks can be designed in many ways. One design is the fully connected multilayer feed forward network design in which the network has an input layer, hidden layers, and the output layer. In addition, the connections in the network never cycle back to an input unit or to an output unite that is located in the previous layer. Also in this design, each unit that is located in a layer L provides input to each unit located in the layer L+1. [8]. A three layer fully connected feed forward ANN has been used in this research. The network consists of an input layer, one hidden layers, and the output layer. The input layer has nineteen input unites while the hidden layer has twelve hidden unites that use sigmoid activation function. The third layer is the output layer which has only one output unite.

3.2 Logistic Regression

Logistic Regression considers as a mathematical modeling technique that describes the relationship between several independent variables, $X_1...X_K$, and a dependent variable, D. The logistic model uses the logistic function as a mathematical form which has the range between 0 and 1 for any given input. The logistic model can describe a probability of an event which is always a value between 0 and 1. The following formula represents the logistic model.

$$(D = 1|X1,\ X2,\ \dots\ ,\ Xk\ ) = 1\ 1 + e - (\alpha + \sum \beta i\ Xi\ k\ 1\ )$$

Where α and β are the model's parameters that can be learned from a set of labeled instances in the training dataset. Gradient Descent Algorithm can be used to find the best values of the model's parameters during the training phase [8].

3.3 Support Vector Machine

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

3.4 Decision Tree

The decision tree model can be represented as a graph which contains nodes and branches .there are two type of nodes which are known as (internal node , leaf node). The test on the data set features is represented on the internal node of tree. The result of the test can be represented by the branch. The target can be represented by the leaf node. The node on the top of tree is known as the root node .decision tree is one of the most famous classification algorithms because the work with this algorithm does not require a previous knowledge of the problem and it does not require a tedious parameters' configurations. DT can be easily comprehended and easily transformed to classification rules. Decision tree algorithms have been used in number of applications such as medical applications, manufacturing production applications, financial analysis applications, molecular biology applications and astronomy applications. [7].

4. The Experiment

4.1 Dataset and Data Sources

In this research the dataset used has been obtained from the UCI Machine Learning Repository which is created with records of absenteeism at work collected from July 2007 to July 2010 at a courier company in Brazil [9]. The dataset contains 721 instances and 21 attributes. The dataset records have been labeled based on the total number of hours of absents. Less than six hours of absent is considered as a normal absent rate while more than six hours is considered as a not normal rate. Based on that, the dataset has (314) employees with normal rate and it has (460) employees with not normal rate.

4.2 Machine Learning Software

In this research, the RapidMiner Studio machine learning software is used to train and test the models.

4.3 Validation Method and Accuracy and Performance Measures

In this research, we proposed using a decision tree model that is configured using a maximum depth equals 20 and we applied tree pruning. In addition, we set the decision tree confidence to be 0.5, the minimal gain equals to 0.1, and the minimal leaf size to 2. In addition, we configured the ANN model to use 600 training cycles, 0.01 learning rate, and a momentum that equals 0.3. In training the logistic regression model, we set the program to automatically select solving method. The support vector machine has been configured to use a neural kernel type with 0.001 convergence epsilon .The dataset has been divided into two parts, 70% of the data is used to train the classification models and the remaining 30% of the data is used to test the models. We used two measures to find out the efficiency of the models. These measures include the classification accuracy and the area under the curve, ROC index.

Table 1: The accuracy and ROC index measures for the Models

Model	**tp**	**fp**	**tn**	**fn**	**Accuracy**	**Roc index**
ANN	131	30	45	16	79.28	0.798
DT	123	15	62	22	83.33	0.834
LR	171	28	111	60	79.22	0.832
SVM	134	57	18	13	68.47	0.760

5. The Results

Four machine learning techniques have been used to create four models (neural network technique, Logistic Regression technique, support vector machine (S.V.M) technique, decision tree technique). Table1 shows the AUC and accuracy measured for each model. As shown in the table 1, the decision tree model has the highest accuracy equals to 83.33% with AUC 0.834 and the support vector machine has the lowest accuracy equals to 68.47 % with AUC 0.760. The decision tree model considered a good classification model due to its Inexpensive training and constriction, its efficient in excluding the unimportant data features, and it can handle unknown data records accurately. The following figures show the ROC index of each model.

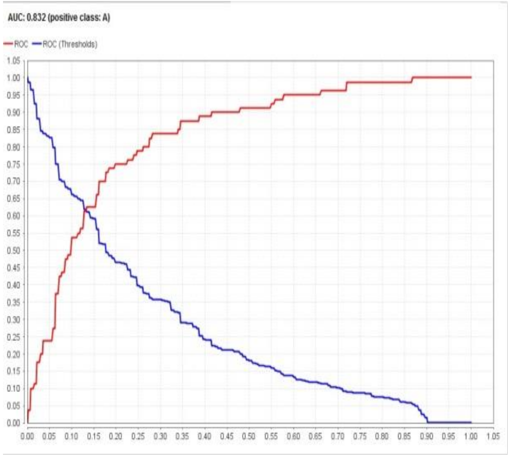

Fig. (1): LR ROC Index

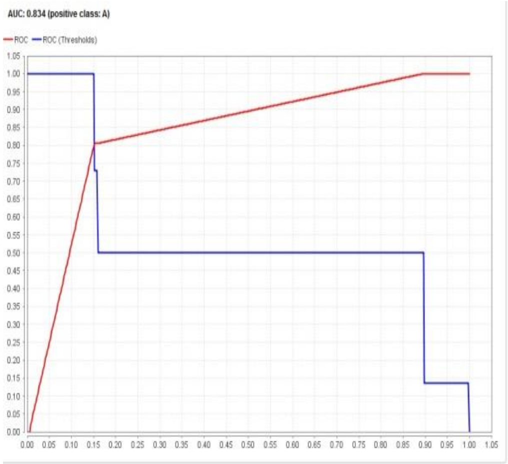

Fig. (2): DT ROC Index

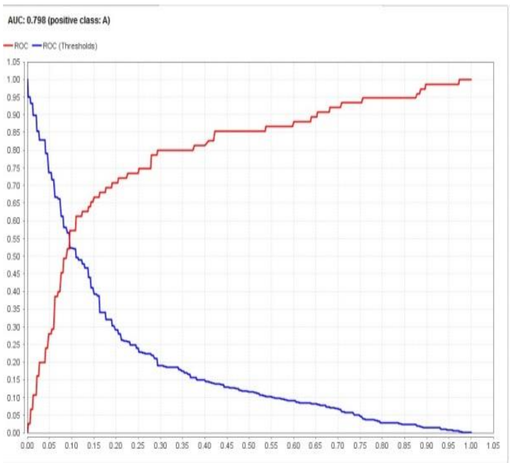

Fig. (3): ANN ROC Index

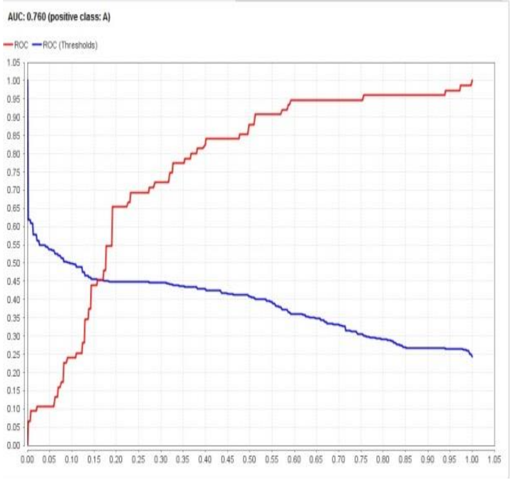

Fig. (4): SVM ROC Index

6. The Conclusion

To solve the problem of absenteeism of the employees at work in the companies, four models have been created in order to predict the absenteeism. In this research four machine learning algorithms have been used that include; neural network algorithm, support vector machine algorithm, decision tree algorithm and logistic regression algorithm. The database used in this paper has 20 Attributes which created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. To compare the models' performance, The ROC index measure and accuracy measure have been used. The Decision tree model has the highest accuracy which equals to 83.33% with AUC 0.834 and the support vector machine has the lowest accuracy which equals to 68.47 % with AUC 0.760.

References

[1] Ricardo ,P.f, Andréa,m, Domingos,n, Edquel,b.p, Renato,J.S,2018, artificial neural network and their application in the prediction of absenteeism at work, International Journal of Recent Scientific Research Vol. 9, Issue, 1(G), pp. 23332-23334

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

[3] Nunung.n. Yudho,g,2014, Employees' Attendance Patterns Prediction using Classification Algorithm, Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 1, Issue 1

[4] Afefa.A, Manal.A, Employees Absenteeism Factors Based on Data Analysis and Classifi cation ,2019, Thomson Reuters ISI ESC / Clarivate Analytics USA and Crossref Indexed Journal , Special Issue Vol 12 No (1)

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

[6] Martiniano.A, Ferreira.R, Sassi.R, Affonso.C,2012, Application of a neuro fuzzy network in prediction of absenteeism at work, Information Systems and Technologies (CISTI), Iberian Conference on,IEEE

[7] Jiawei.H, Micheline.K, Jian.P, 2012, Data Mining Concepts and

Techniques, Third Edition, Morgan Kaufmann publications

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

[9]http://www.uninove.br/curso/informatica-e-gestao-do-conhecimento