Lecture Notes in Electrical Engineering 597

Pradeep Kumar Singh
Arpan Kumar Kar
Yashwant Singh
Maheshkumar H. Kolekar
Sudeep Tanwar *Editors*

Proceedings of ICRIC 2019

Recent Innovations in Computing

Springer

Editors

Pradeep Kumar Singh
Department of Computer Science
and Engineering
Jaypee University of Information
Technology
Waknaghat, Himachal Pradesh, India

Yashwant Singh
Central University of Jammu
Jammu, Jammu and Kashmir, India

Sudeep Tanwar
Institute of Technology
Nirma University
Ahmedabad, Gujarat, India

Arpan Kumar Kar
Indian Institute of Technology Delhi
New Delhi, Delhi, India

Maheshkumar H. Kolekar
Indian Institute of Technology Patna
Patna, Bihar, India

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-3-030-29406-9 ISBN 978-3-030-29407-6 (eBook)
https://doi.org/10.1007/978-3-030-29407-6

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Advanced Computing

Intellegent Networking

Image Processing and Computer Vision

E-Learning Cloud and Big Data

Security and Privacy

Digital India

Predictive Analysis of Absenteeism in MNCS Using Machine Learning Algorithm

Krittika Tewari, Shriya Vandita and Shruti Jain

Abstract Absenteeism has become a severe problem for many organizations. The problem posed in this paper was to build a predictive model to predict the absenteeism for MNCs by previously recorded data sets. This exercise not only leads to prevent or lower absenteeism but forecast future workforce requirements and suggests ways to meet those demands. For faster processing of massive data set, the data was analyzed efficiently so that we get the minimum response time and turn-around time, which is only possible when we use the right set of algorithms and by hard wiring of the program. Different machine learning algorithms are used in the paper that includes linear regression and support vector regression. By analyzing the results of each technique, we come across that the age parameter mainly affects the absenteeism that is linearly related to absenteeism.

Keywords Absenteeism · Machine learning · Linear regression · Support vector regression

1 Introduction

Absenteeism is a vital issue that requires immediate attention by both, the employee and the employer. If the employee enjoys the work they do, then they will not take leave. The employers are expected to keep their employees in good spirit and motivated so that the employees deliver their best to the organization for the benefit of both. It is a habitual pattern of the absence from duty. Absenteeism not only affects the cost but is also an indicator of the poor morale of the employees. Absenteeism can be a result of depression, personal stress, and anxiety which can lead to an employee being detached and unable to cope up with the work and interaction at workplace, burnout, heavy workloads, stressful meetings/presentations, bullying at workplace,

K. Tewari · S. Vandita · S. Jain (✉)
Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Solan, Himachal Pradesh, India
e-mail: jain.shruti15@gmail.com

S. Vandita
e-mail: simranshriya@gmail.com

P. K. Singh et al. (eds.), *Proceedings of ICRIC 2019*, Lecture Notes
in Electrical Engineering 597, https://doi.org/10.1007/978-3-030-29407-6_1

childcare, illness, and unequal treatment that leads to heart burning and feelings of being unappreciated (non-recognition of his/her contribution). Absenteeism has become a severe problem for many organizations. Obviously, it has been an undeniable issue faced by companies which can result in serious financial and non-financial losses [1]. Because of the negative consequences of employee absenteeism, it is important that the organization keep the absenteeism low [2]. The following may be adopted by the organization to check or reduce absenteeism:

1. Adopting a clear attendance policy.
2. Providing healthy and safe working environment to the employees.
3. Encouraging the employees by visiting the workplace during working hours by the higher management.
4. Celebrating the success of any project as a team.
5. Providing some reward to the employee for excess extra paid time off.
6. Maintaining a flexible work schedule if an employee comes late, he/she may be allowed but should be asked to put extra working hours in the week to compensate.

Employees can cope with stress much better when they are given some autonomy for self-governance, social support, and opportunities for personal growth. It has been seen that the rate of absenteeism does not follow a normal distribution, but it is skewed and truncated distribution that denies the beliefs of conventional statistical methods such as correlation analysis and ordinary least square (OLS) regression [3]. It has been seen that correlation and multiple regression dominate absence research. Managers and organizational practitioners need a detailed method for measuring absenteeism loss as well as other measures needed for managerial evaluation to decrease absenteeism rate and compare the effectiveness of absence/attendance policy from period to period. Since the 1980 s, the companies have collected abundant amount of customer data stored in databases [1]. The data is collected by the companies and the thought process on how to provide additional benefits or to improve the operations. This type of thought process formed a natural progression toward the use of improving estimates, forecasts, decisions, and efficiency [2]. These databases grew to such a large extent that it becomes difficult for humans to analyze on their own. Predictive analytics is an answer on how to handle such large databases. It is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. It was conceived from the study of related areas like artificial intelligence, statistics, machine learning, pattern recognition, and data mining [2].

We can use either supervised learning based models or unsupervised learning models for prediction. The basic idea behind supervised learning based models is to predict a target variable. Supervised learning is also referred to as predictive modeling. Classification is a popular predictive modeling algorithm while dealing with categorical variable [4–8]. Another type of supervised learning is regression where we predict continuous outcomes [9, 10]. This procedure determines the computational methods and incorporates the patterns in large data. In descriptive modeling or unsupervised learning based approaches, a model is always constructed through clusters of the data [11].

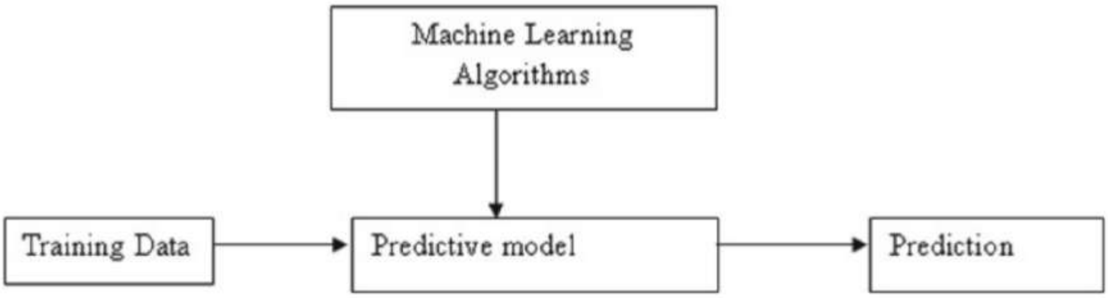

Fig. 1 Steps using machine learning algorithm

Authors in [12] explain machine learning is a scientific discipline which focuses on automatically recognizing complex patterns and making intelligent decisions based on available data. This branch of study evolves behavior that helps in developing an algorithm for the computer. Figure 1 signifies the typical machine learning algorithm [11, 12]. Machine learning focuses on the development of computer programs that can change when exposed to new data. It is the process of converting experience into expertise or knowledge [12]. There are different machine learning algorithm which includes linear regression, decision tree, logistic regression, Naïve Bayes, support vector machine (SVM), k-nearest neighbor (kNN), and random forest.

Predictive model is made to train data that helps in analyzing the parameters which affect the absenteeism in MNCs and how to reduce that absenteeism. Data is collected from online source which is preprocessed by removing outliers. Feature engineering is applied to the data. There are various parameters which affect the absenteeism. Using feature selection, best features were selected and different machine learning algorithms like linear regression and support vector regression are applied.

In this paper, Sect. 2 defines the predictive analysis using machine learning algorithm; Sect. 3 explains the different implementation steps which were concluded at the end of the paper.

2 Predictive Analysis Using Machine Learning Algorithm

Predictive analysis is used to predict unknown events or unobserved events by analyzing the existing data set with the help of machine learning techniques, statistical modeling, and data mining. For predictive analysis, objectives are defined, and then, the data set is prepared. Based upon the prepared data, a model is laid down for deployment and monitoring. Predictive analysis identifies the cause–effect relationship across the variables from the given data set and discovers hidden patterns with the help of data mining techniques. It may apply to the observed patterns to unknowns in the past, present, or the future. The predictive models have a clear focus on what they want to learn and how they want to learn. The models which are used for the prediction of target features of categorical values are known as classification models [4–7]. Predictive models may also be used to predict numerical values of the target feature based on the predictor features. The models which are used for the prediction of the numerical values of the target feature of a data instance are known as regression models. There are different types of regression models (linear, logistic, support vector regression). Mainly, they are categorized as linear and nonlinear. In linear regression model, the dependence of the response on the regressors is defined by a linear function, which makes their statistical analysis mathematically tractable. On the other hand, in nonlinear regression model, this dependence is defined by a nonlinear function, hence, the mathematical difficulty in their analysis.

In this paper, we are working on linear regression and support vector regression (SVR).

2.1 Linear Regression

For linear regression, dependent variable (y) is continuous and independent variable (x) may be continuous or discrete. Figure 2 shows the algorithm illustrating the linear regression.

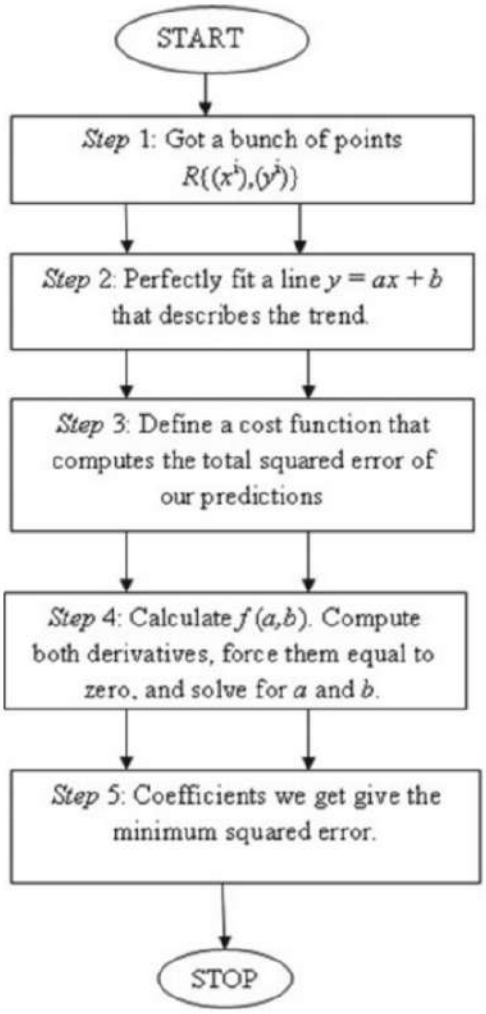
START
Step 1: Got a bunch of points $R\{(x^i),(y^i)\}$
Step 2: Perfectly fit a line $y = ax + b$ that describes the trend.
Step 3: Define a cost function that computes the total squared error of our predictions
Step 4: Calculate $f(a,b)$. Compute both derivatives, force them equal to zero, and solve for a and b.
Step 5: Coefficients we get give the minimum squared error.
STOP

Fig. 2 Algorithm illustrating linear regression model

For linear regression, the shape of regression line is linear whose slope of line is b and intercept is a. Linear regression is expressed as:

$$\underbrace{y}_{\text{actual (observed)}} = \underbrace{ax + b}_{\text{explained (prdeicted)}} + \underbrace{\varepsilon}_{\text{error}} \tag{1}$$

In Eq. (1), e is the error term. Linear regression can also be expressed by Eq. (2)

$$\underbrace{y}_{\text{observed}} = \underbrace{\hat{y}}_{\text{predicted}} + \underbrace{\varepsilon}_{\text{error}} \tag{2}$$

For predicted values, Eq. (1) can be written as:

$$\hat{y} = \hat{a}\,x + \hat{b} \tag{3}$$

where slope is represented as:

$$\hat{a} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \tag{4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means and intercept is represented by Eq. (5),

$$\hat{b} = \bar{y} - \hat{a}\,\bar{x} \tag{5}$$

If the independent variable (x) is more than one (i.e. $x_1, x_2, x_3 \ldots$) than regression is known as multiple regression. For multiple regressions, Eq. (1) can be expressed as

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots b_nx_n + e \tag{6}$$

Methods like partial least square analysis (PLS) and OLS are used for calculation of linear regression, while methods like stepwise approximation (SWA), forward selection (FS), and backward elimination (BE) are used in multiple regression analysis.

2.2 Support Vector Regression (SVR)

A nonlinear function is leaned by linear learning machine mapping into high-dimensional kernel-induced feature space containing all the main features that characterize maximum margin algorithm. The system capacity is controlled by parameters that are not affected by the dimensionality of feature space. The main idea is to optimize the generalization bounds and rely on defining the loss function that

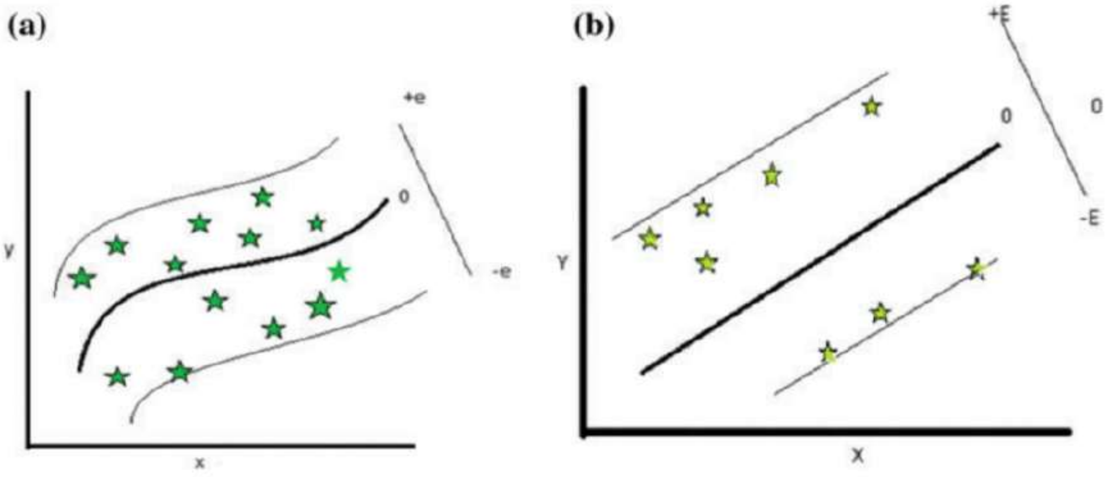

Fig. 3 SVM **a** nonlinear and **b** linear hyperplane

ignores errors that are located within the certain distance/range of the true value for regression. SVR is also used for making prediction models. Another widely used and powerful learning algorithm is a support vector machine (SVM). In SVM, the objective is to maximize the margin which is defined as the distance between the separating hyperplane and training samples that are closest to hyperplane. SVM can be applied as regression and not limited to classification only. SVR works on same principle as SVM works with only a few differences. Figure 3a shows the nonlinear plane of SVM, and Fig. 3b shows linear plane of SVM.

The main idea is to maximize the margin by individualizing the hyperplane. The results of linear regression and linear SVR are mostly similar. The main focus of linear regression is to maximize the conditional likelihoods of the training data from the source, which makes it more susceptible to outliers as compared to SVM. Furthermore, linear regression models can be easily updated, and the SVM focuses on the points that are closest to the decision boundary (support vectors).

3 Implementation of Predictive Model for Absenteeism

Our aim is to predict the absenteeism for MNCs by the previous recorded data sets. We analyzed the data from online source www.kaggle.com, which is a platform for predictive modeling and analytics competitions. Kaggle is an online community of machine learners and data scientists. It allows users to find, explore, and publish data sets. The availability of advanced machines and special tools has led to the analysis of big data. It also broadened our horizons of looking at an unknown data and trying to find useful features and patterns. We programmed our analysis on Python language. Figure 4 shows the proposed methodology employed in the paper to make a predictive model that helps in predicting the absenteeism. A significant amount of time and effort was spent in organizing, cleaning, and redefining variables in the data.

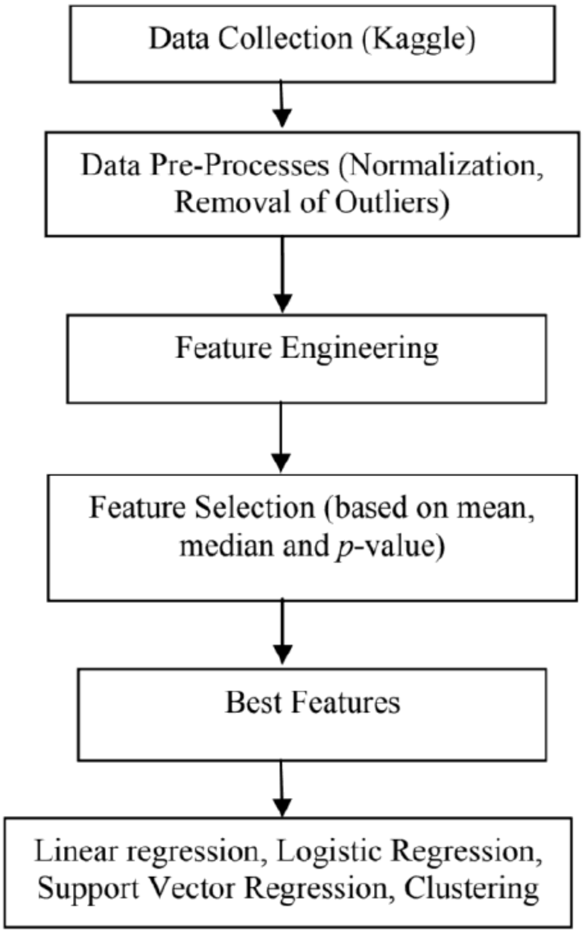
Data Collection (Kaggle)
Data Pre-Processes (Normalization, Removal of Outliers)
Feature Engineering
Feature Selection (based on mean, median and p-value)
Best Features
Linear regression, Logistic Regression, Support Vector Regression, Clustering

Fig. 4 Proposed methodology

Different steps were followed for implementation of our algorithm:

1. *Data Exploration*: In this section, data set is explored. This is an important step in the machine learning process as firstly we need to know more information about the data we are using and secondly we need to make a few alterations to the data itself. In this paper, we are using 741 observations and 21 parameters before data extraction and selection out of which we are considering the following different parameters: absent (1 = YES, 0 = NO), employee class (1 = 1st, 2 = 2nd, 3 = 3rd), name, age, sex, distance from home, seasons transport, service time, day shifts, workload, hit target, and education.
2. *Data Preprocessing*: Raw data can be transformed into understandable format is called data preprocessing. Data preprocessing is required because real-world data is often inconsistent, noisy, incomplete, or lacking in certain behaviors or trends. The missing values can be identified through data preprocessing. The identification of missing values is important in successful management of data. Missing values, if not handled properly, lead to inaccurate inference about the data. The result obtained will differ from ones where the missing values are present due to improper handling of data.
3. *Feature Engineering*: Since the data can have missing fields, incomplete fields, or fields containing unknown information, a fundamental step in building any

Table 1 Title from the field Name

Index	Title	No of occurrence
1.	MR	757
2.	MRS	198
3.	MS	2

prediction system is feature engineering. Many times the data set contains highly varying features in magnitudes, units, and range. The field "Name" contained employee's title: Mr., Mrs., Ms. Since name is unique for each employee, it is not useful for our prediction system. However, an employee's title can be extracted from his or her name. In our data set, we found three titles which are shown in Table 1. Title indicates employee's sex (Mr. and Mrs.) and age (Ms. and Mrs.).

4. *Machine Learning Algorithm*: It explains how we have applied different machine learning algorithms (linear and SVR) to our data set in order to build a model. Figure 5 elaborates the machine learning algorithm that we have incorporated in our paper. We have split the data set into training data set and testing data set in the ratio of 70:30, and using training data set, we came up with a model to be deployed. To this model, we applied test data set that yields predictive model.

In this paper, we have used linear regression and SVR to get predictive model. Linear regression requires minimal adjustments to the data provided by Kaggle. We have performed linear regression of all the parameters (shown in Fig. 6) and find that absenteeism is linearly varying with age parameter only.

Figure 7 shows that the *age of the employee* being our predictor variable and *number of days our employee is absent* as our response variable using linear regression.

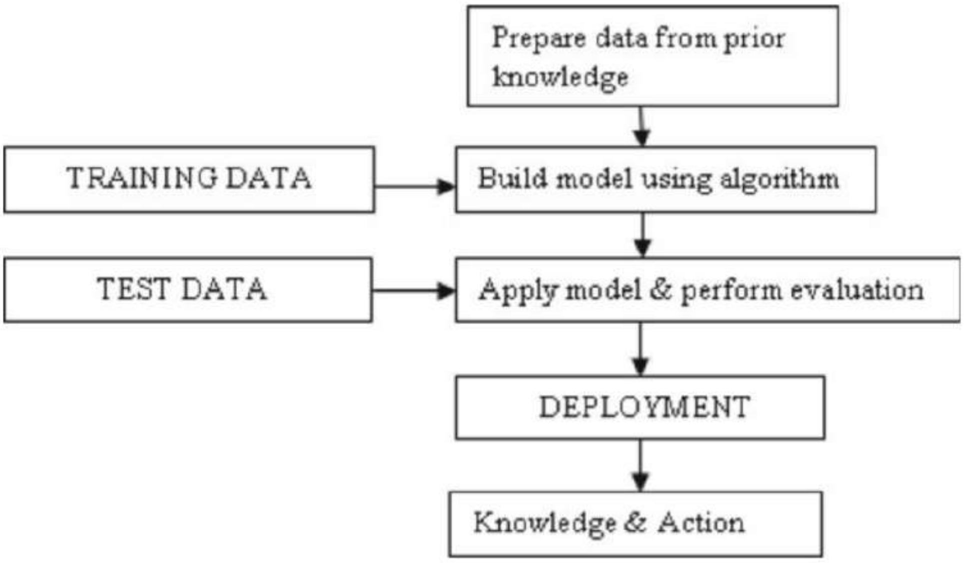
Prepare data from prior knowledge
TRAINING DATA
Build model using algorithm
TEST DATA
Apply model & perform evaluation
DEPLOYMENT
Knowledge & Action

Fig. 5 Machine learning algorithm

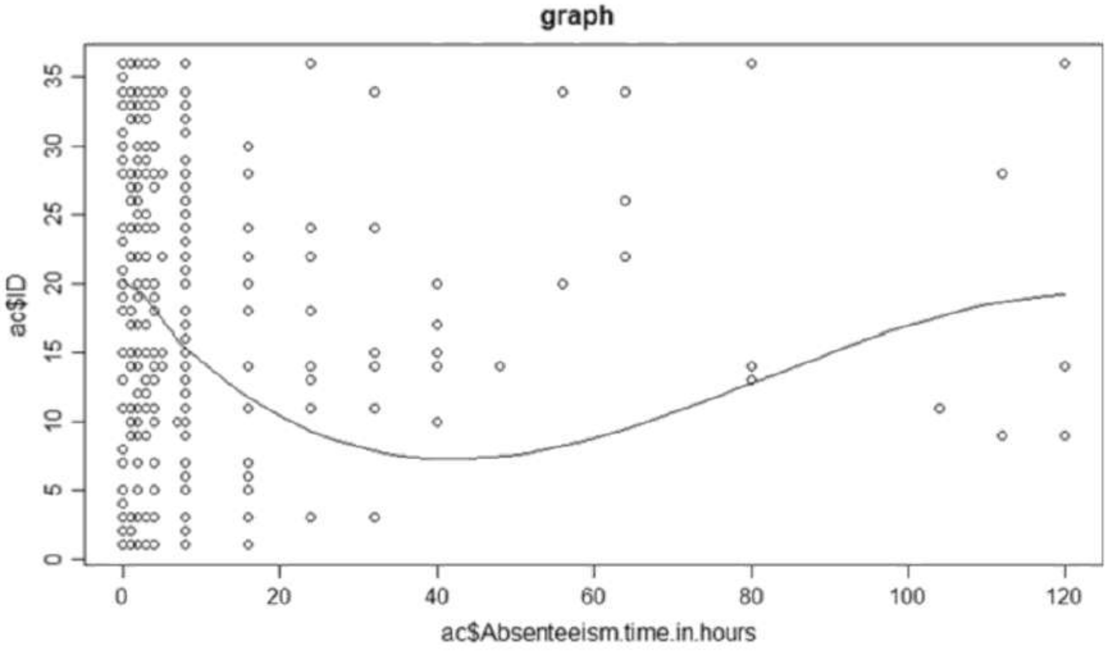

Fig. 6 Linear regression on different variables our data set

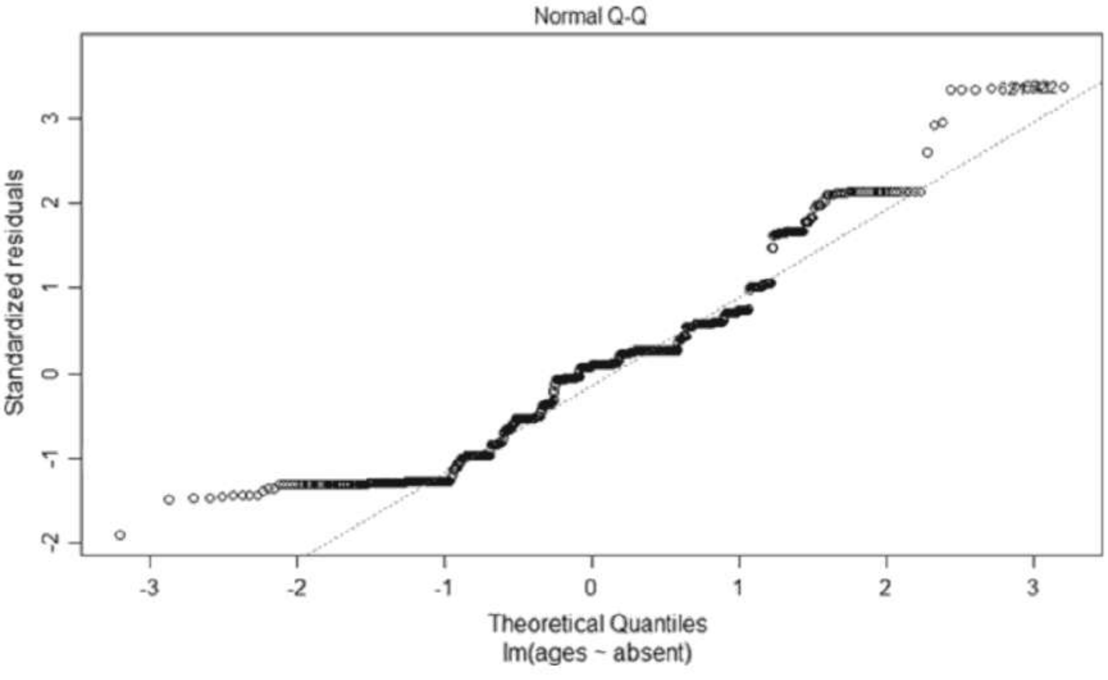

Fig. 7 Residuals versus age using linear regression

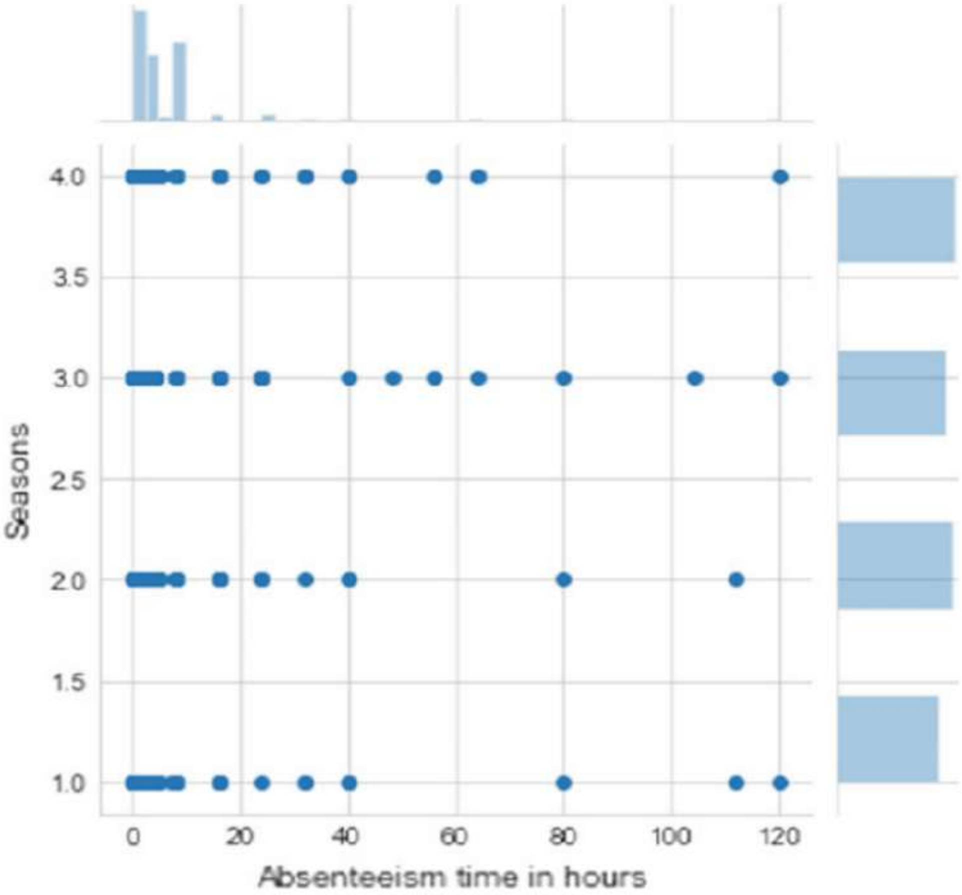

Fig. 8 Graph between seasons of the year versus absenteeism using SVR

We have also applied SVR on our data sets. For SVR, we have considered two parameters age and seasons of the year. Figures 8 and 9 depict the graph for season of the year and age versus absenteeism respectively using SVR.

We have divided the year into four major seasons with an interval of 0.5 but as shown in Fig. 8, we can infer that absenteeism rate does not get much affected by a particular season of the year. The absenteeism rate remains unaffected by this factor as the absenteeism rate is almost equal in all seasons. From Fig. 9, we can infer that the age group of 35–40 years has the highest absenteeism rate as compared to the age group of 50–60. With age parameter, days of the week with hours is also considered.

Figure 10 depicts the absenteeism on different days of the week. Here, each day starting from Monday to Saturday is assigned a different color and the gradual decrease in absenteeism on 2nd day of the week is observed as the age increases, and the absenteeism on the 6th day (Saturday) is fairly high than any other day, reason being the end of the week.

From the experiments, we come to know that age parameter of the age group of 35–40 is linearly related to absenteeism and the maximum absenteeism is on the 6th day of the week, and minimum absenteeism is on the 2nd day of the week. There is no much affect of season on absenteeism.

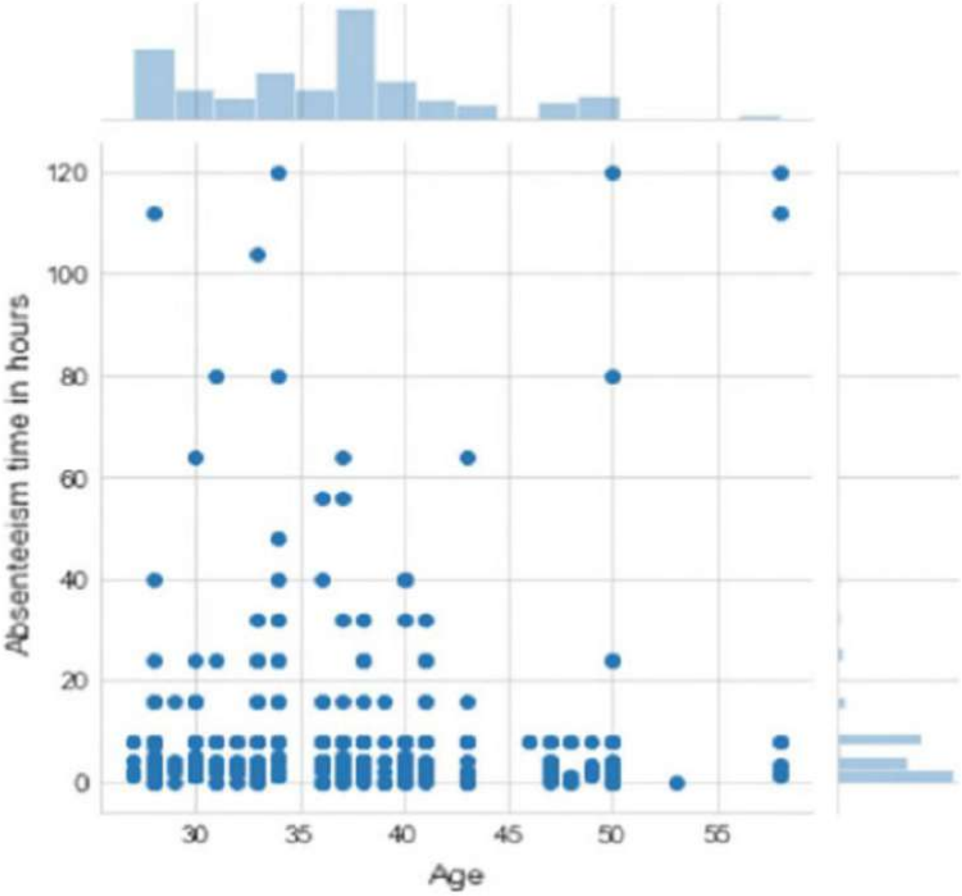

Fig. 9 Graph between age versus absenteeism using SVR

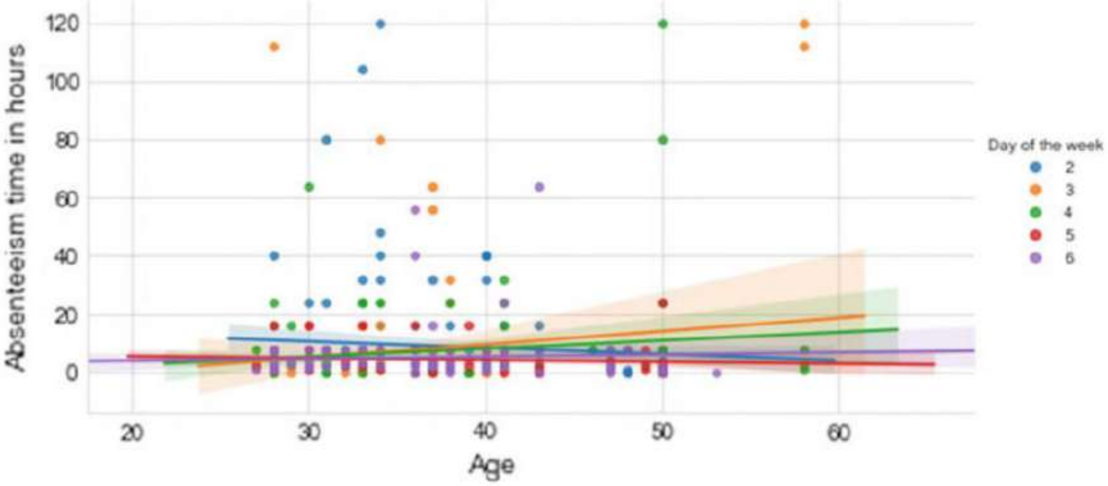

Fig. 10 Age and absenteeism in hours considering days of the week. *Note* 2 defines 2nd day, 3 defines the 3rd day of the week

4 Conclusion and Future Work

Predictive analysis of absenteeism in MNCs served as a framework for introductory predictive analytic methods. Managers and organizational practitioners need a detailed method for measuring absenteeism loss as well as other measures needed for managerial evaluation to decrease absenteeism rate and compare the effectiveness of absence/attendance policy from period to period. The motivation of this study developed from a desire to learn, understand, and apply linear, logistic, and SVM regression. It is apparent that the effort put forth when working on the absenteeism in MNCs problem has achieved our aims and goal of this study by linear regression and SVR. In the future, we will encode categorical values of parameters such as month of the absence.

References

1. Delen, D., Zaim, H., Kuzey, C., Zaim, S.: A comparative analysis of machine learning systems for measuring the impact of knowledge management practices. Decis Support Syst **54**(2), 1150–1160 (2013)
2. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques, p. 560. Morgan Kaufmann Publishers, San Francisco (2005)
3. Faber, F.A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S.S., Dahl, G.E., Vinyals, O., Kearnes, S., Riley, P.F., von Lilienfeld, O.A.: Prediction errors of molecular machine learning models lower than hybrid DFT error. J. Chem. Theory Comput. **13**(11), 5255–5264 (2017)
4. Jain, S.: Classification of protein kinase B using discrete wavelet transform. Int. J. Inf. Technol. **10**(2), 211–216 (2018)
5. Jain, S., Chauhan, D.S.: Mathematical analysis of receptors for survival proteins. Int. J. Pharma Bio Sci. **6**(3), 164–176 (2015)
6. Bhusri, S., Jain, S., Virmani, J.: Classification of breast lesions using the difference of statistical features. Res. J. Pharm., Biol. Chem. Sci. (RJPBCS), 1366 (2016)
7. Rana, S., Jain, S., Virmani, J.: SVM-based characterization of focal kidney lesions from B-mode ultrasound images. Res. J. Pharm., Biol. Chem. Sci. (RJPBCS) **7**(4), 83 (2016)
8. Sharma, S., Jain, S., Bhusri, S.: Two class classification of breast lesions using statistical and transform domain features. J. Glob. Pharma Technol. **9**(7), 18–24 (2017)
9. Jain, S.: Regression analysis on different mitogenic pathways. Netw. Biol. **6**(2), 40–46 (2016)
10. Jain, S.: System modeling of AkT using linear and robust regression analysis. Curr. Trends Biotechnol. Pharm. **12**(2), 177–186 (2018)
11. Zhang, L., Tan, J., Han, D., Zhu, H.: From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov. Today **22**(11), 1680–1685 (2017)
12. Borchers, M.R., Chang, Y.M., Proudfoot, K.L., Wadsworth, B.A., Stone, A.E., Bewley, J.M.: Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. J. Dairy Sci. **100**(7), 5664–5674 (2017)