[[1]](#footnote-1)

Project Report on Analysis of Absenteeism at Workplace using Machine Learning

Himangshu Pal, Mohammed Junaid Asif, Sankshiptha Arroju

*Abstract* — This is a report on the analysis performed on the Absenteeism at workplace dataset collected from UCI repository. The absenteeism in hours were grouped into three categories high, low and moderate so that the prediction could be made effectively using the machine learning. The analyses and predictions were made entirely using R programming and various packages that facilitated the training of the model using the commonly used Machine Learning algorithms. The models built using the popular algorithm were evaluated for performance against the test validation set split form the parent dataset and their accuracies are compared. Finally, the best performing model was validated for performance by testing the model on entire dataset. This study aims to predict the absenteeism and probably use the decision trees, classification models or clustering to identify the key attributes that influenced absenteeism at workplace.

*keywords*— absenteeism, classification, neural networks, pre-processing, clustering, predictors

# INTRODUCTION

From North to South, East to West, Europe to North America, absenteeism is a nightmare to organizations and managers need to regularly compute the cost of this unfortunate trend and how it affects profits and put in place measures to curb it.

Managers in Companies are required to manage absenteeism at the workplace, but it is crucial to evaluate this concept and define different types of absenteeism and also examine both direct and indirect costs of this trend.

First, employees are required by their employment contract to report to duty daily unless there are instances of sickness or annual leave that is approved by the human resource department. However, when employees fail to report duty states, it gives to cases of absenteeism, and this is very dangerous to business. In such instances, workers present ton of excuses on various reasons of absenteeism, like they had the subpoena to appear in court as a witness, the car had broken down or one of the family members was unwell. Therefore, line supervisors are required to enforce strict human employee code of conduct and evaluate the authenticity of reasons given before granting work leaves to workers. Also, every organization needs to have a dynamic policy that defines what absenteeism is and managers need to follow this policy to the letter and enforce as required.

The current report examines the various factors and reasons that influenced the absenteeism at the workplace using machine learning. We will use absenteeism data to analyze some of the variables that lead to absenteeism from work by using Machine Learning algorithms in predicting absentees from duty. In the end, we propose the best model based on performance that will predict whether a worker will be absent for a day, week or even a month.

# RELATED WORK

Ahn & Lee, (2011), examines how absenteeism from work has adversely affected performance in organizations, an aspect that is not only complex but also encompasses some behavior of individual characteristics and systems behavior of companies. According to Johns (2008) argues that absenteeism is failure of employees to report to work as scheduled and this can adversely affect the performance of organizations which is now a poorly understood concept among organization. Some of the organizations that are vulnerable to absenteeism include labor intensive like construction industry that is dependent on manpower. When absenteeism is manifest in construction projects lads to interruption of work accomplishment, increased overtime and loss of revenue thus there is strong dependency of work processes like labor force and the concept of right place at right time is very important aspect. Several academic researchers have been looking at the causal factors of absenteeism and many have narrowed down to aspects like job satisfaction and work attitudes. While others have established weak correlation when it comes to individual levels factors and absenteeism measures. Thus, absence cultures tend to cover some key concepts in every aspect of employee’s absence behavior like beliefs and perception. This kind of relationship can lead to deep understanding of the feedback relations with small actions within an organization which can impact overall organization performance of the whole company.

Therefore, Ahn & Lee, (2011), explores the concept of absenteeism using agent-based modeling (ABM) methodology in examining some of the mechanism that is absent among the workers. ABM is a research methodology within the organizational science that uses social variables and provides insight on how managers can lead organizations by using computer simulations to determine human behavior. Consequently, Ahn & Lee, (2011) comes up with model that uses mental processes that are related to absence and named as the behavior rue in the ABM model. The model takes into consideration several variables like the feedback relation between individuals , absence behavior and overall perception .It applies complex systems which entails simple and similar parts and can easily be understood and behavior of the overall system .The complex systems provide learning and adaptation using interactions and experiences in the same system .After the worker forms his absence own standard and history and this deans on some of the personal considerations , an iterative process comes into play in order to make observations. Ahn & Lee,(2011)concludes that if organization use the ABM modeling , it will enable one to understand some of the dynamic features of system behavior from interaction with worker agents in the organization by using self-standards and making of absence decisions.

Meanwhile, (Trivedi, 2010) examines the issue of absenteeism by using Artificial Neural Networks that are building upon unique modeling of human brain as basis of designing algorithms to determine complex human patterns. (Trivedi, 2010), caries this study using a dataset from employees of a courier company in Brazil and looks some of the underlying reasons for absenteeism between 2007 and 2010.The dataset contains information that has drawn from UCI machine learning repository with 740 rows of data ad 20 distinctive features which are either related or not related to work that is collected form 36 workers within a span of 3 years .The features entails some reasons for absence like , age, BMI, number of children and service time. (Trivedi, 2010), comes up with back propagation neural network that can used in predicting workers absenteeism in hours within a given parameter and this can be used to analyze on the number of reasons that can be used to make corrective measures at the workplace.

In addition, Neural Networks uses machine learning algorithms which have ability to carry out work sing incomplete information but with some amount of fault tolerance in the data. However, it is important to note that Neural networks comes with drawbacks where there has heavy dependency on hardware on the system that requires some kind of parallel processing especially to large data sets and this kind of pruning can be applied on network to remove some unnecessary features and give overall output in the network. (Trivedi, 2010) used some heuristic criteria like minimum weight, activation and Average percentage of Zeros which are used to identify some of the least important parameters in the network. He then employees use of cross validation of the dataset in assigning some of appropriate actions to the network to provide some good way of increasing efficiency and accuracy in the network. Thus, in this study, (Trivedi, 2010) uses prune patterns and neural network in coming up with reasons for absenteeism at the workplace, the number of hours and give reasons. In the end, manager is required to take decisive actions that is contributes to productivity of work.

In another study, Ricardo et al, argue that today high competitiveness in the labor market, continuous professional development that is combined with organization and pressure in reaching high set goals leads to overburdening of workers that leads to emergence of disturbance of health that is related to type of work activity, depression decorates the working ethics of 21st century. Ricardo et al (2018) applies artificial neural networks in predicting absenteeism at work by using database of 38 attributes with 2243 records in determining some of the underlying factors that lead to absenteeism at work. Therefore, the ANN models are usually inspired by the structure of the human brain in simulation human behavior processes like learning, adaptation, association, some kind of fault tolerance and generalization. In this case, the learning process usually takes place by using simple processing units called artificial neurons and uses incomplete data that is subject to noise process. In normal computing process, when a system fails, the system deteriorates but in artificial neural network there is a fault tolerance that forms part and parcel of distributed processing.

Thus, Ricardo et al (2018) argues that the advantages of using ANN include it has high tolerance of noisy data and entails the ability to classify some patterns which are not trained in the process. Thus, the artificial neural networks are sued in establishing relationships between attributes and classes which are continuous value inputs and outputs with system algorithms at the end of the day. In the experiment by Ricardo et al (2018), with38 attributes and 2243 records are reduced by using Rough sets to 17attributes that includes database experiment like the reason for absenteeism by adopting international certified classification like diseases and other reasons like medical allowance and physiotherapy just to mention a few. Thus, the authors conclude that it is possible for one to predict absenteeism by applying neural networks by reducing number of attributes by using Rough sets and obtain good results.

Meanwhile, in other another study (Thurston & Glendon, 2018) looks at the occupation safety and critical factors in organizational psychology which are likely to trigger some of physical and psychosocial outcomes in the organization. Today, there are increasing regulatory market pressure that are now driving managers to investing in great resources to promote employee safety and organizational connections. But there has been little or no research in organizational identification and empowerment of workers. Thus, (Thurston & Glendon, 2018). uses hypothesized model and risk exposure of workers in organization to determine how they influence workers attitude and behaviors at the workplace. Thus, safety and culture are now an important factor for not only employer of choice but also for now workers in the company .in this case, (Thurston & Glendon, 2018), documented that there is a strong association between workers attitude and management commitment to safety, thus there is need for managers to engage in positive safety and behaviors of the organization.

# DATA

The absenteeism data was collected by Andrea Martiniano, Ricardo Pinto Ferreira and Renato Jose Sassi at a Courier Company in Brazil from records of absenteeism at work from July 2007 to July 2010. The dataset now resides in the UCI online repository and now is used for research purposes. The dataset contains 21 attributes/variables and 740 records. The target variable of concern that needs to be predicted in this study is the ‘Absenteeism time in hours’. The other notable predictor variables are the Reason for absence, Work load Average/Day, Hit target, Disciplinary failure, Education, Age, Social Smoker, Social Drinker, Son, Pet.

# METHODOLOGY

Methodologies involved data collection, preprocessing and transformation, application of machine learning algorithms to find a better model to predict absenteeism at work. Successive sections are the explanation of the methodologies involved in this analysis.

## Data Collection

Absenteeism data set is readily available in University course repository Moodle for Machine Learning course project. The same data set is used in this project. The format of the data set was Comma Separate Values. This data set had 740 observations and 21 variables. The variables presented in the data set are, ID, Reason for absence, Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, Weight, Height, Body mass index, Absenteeism time in hours.

## Data Pre-processing

Data pre-procession entails data cleaning or data transformation (GayathriT, 2018). Absenteeism data was unstructured. As GayathriT, (2018) mentioned in her study, absenteeism data was given proper tabular structure for the extraction of data insights. Since this research involves the use of machine learning algorithms, as advisable in the study of GayathriT, (2018), data was cleaned, transformed, reduced before presented to the machine learning algorithms for better accuracy and performance of the selected models.

Provided data set did not have so called missing values, but there was visible existence on zero values in the data for different attributes. Zero values in the reason code was meaningless, so removed. Observations with zero in Absenteeism in time with non-zero positive reason codes were contradictory, so those observation were amputated too. Data anomaly like observations with valid reason codes but zero hours in absenteeism time were removed. Any noise such as Disciplinary failure had only one value zero in it, so, removed Disciplinary failure attribute. Identified outliers in Transportation.expense, Service.time, Age, Work.load.Average.day, Hit.target, Height, Absenteeism.time.in.hours were capped using percentile values. Outliers which were above 75 percentile were replaced by 75 percentile values and the outliers which were below 25 percentile were replaced with 25 percentile values. This would help in keeping the data and maintaining the data variability. Variables: Reason.for.absence, Month.of.absence, Day.of.the.week, Seasons", Education, Social.drinker, Social.smoker, Son, Pet were discrete in nature, so, transformed into factors.

Multicollinearity in the data were tested using correlation plot for continuous variables. Weight and Body.mass.index were found high correlated. Highly correlation between independent variable does not satisfy data independence and thus may problem for linear regression and simple classification trees. So, attribute weight was removed. The relationship between categorical variables were tests using Chi-square test. Significance relationship between with Month.of.absence, Day.of.the.week, Seasons", Education, Social.drinker, Social.smoker, Son, Pet and Reason.for.absence were identified by Chi-square test, P-value << 0.05. Variables Month.of.absence, Day.of.the.week, Seasons", Education, Social.drinker, Social.smoker, Son, Pet were removed to increase independence in the attributes.

Absenteeism data did not have class. To use this data set in classification problem, the dependent variable Absenteeism.time.in.hours was transformed into categorical variable. The class labeling was done unsupervised learning way using K Nearest Mean clustering method. K-means clustering method concluded 3 clusters on the provided data set. So, Abenteeism.time.in.hours was used to create three classes. Low absenteeism group where absenteeism hours within 4 hours, moderate absenteeism group in which absenteeism hours within 4 to 8 hours range and high absenteeism greater than 8 hours.

## Analysis Methods

Because of non-linearity nature of the data, classification methods were considered for the prediction of absenteeism classes. Clustering, classification and prediction were utilized to find the best possible prediction model for the provided data.

Classification employs a set of pre-classified classes to develop a model that can explain the large variability in the population. Classification techniques used were basic decision tree, C5.0 classification tree, Ensemble learning decision tree Random Forest and Support Vector Machine. Linear classification technique Linear Discriminant Analysis was utilized. Clustering method was used to identify dense and sparse regions in the data and to discover overall correlation and distribution pattern in the data. K-means clustering was used to partition the dataset.

# RESULTS

## Clustering Model

The initial analysis for determine the number of clusters utilizing the Within sum of squares (wssplot) ‘elbow’ method (Fig. 1) and ‘*NbClust’*  yielded plots (Fig.2 (a) and (b)) and analysis that helped us narrow down to k-value of 3.

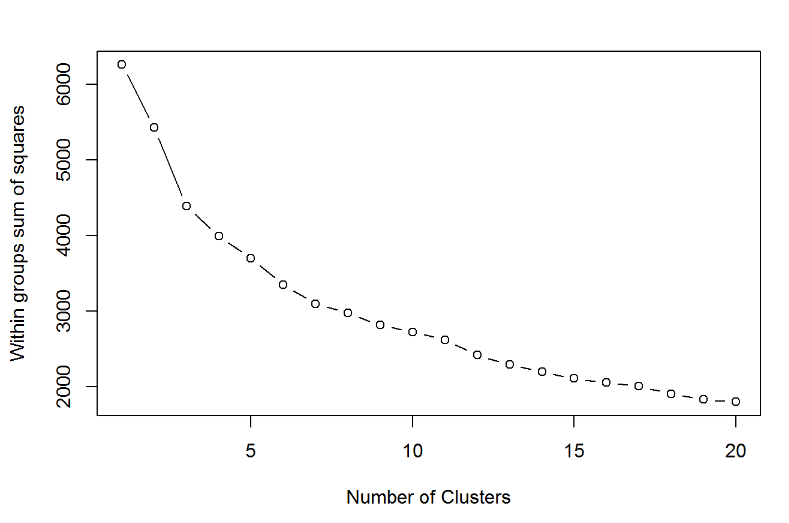


Fig. 1. Within sum of squares plot data for number of clusters.

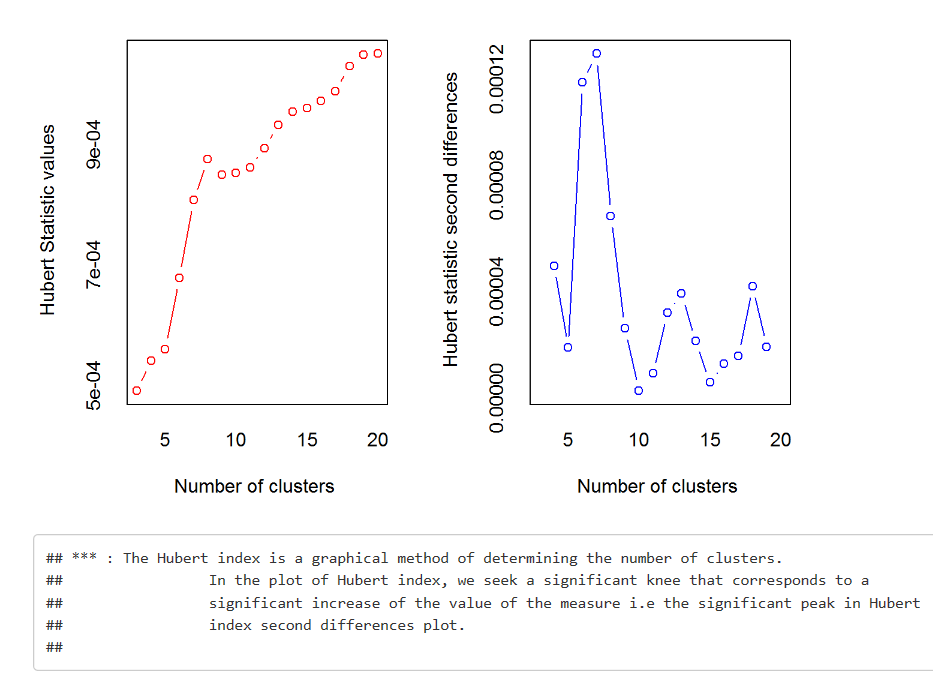


Fig. 2 NbClust Analysis yielded k-value to be three as the best number of clusters (a) Hubert index plots for determining the number of clusters

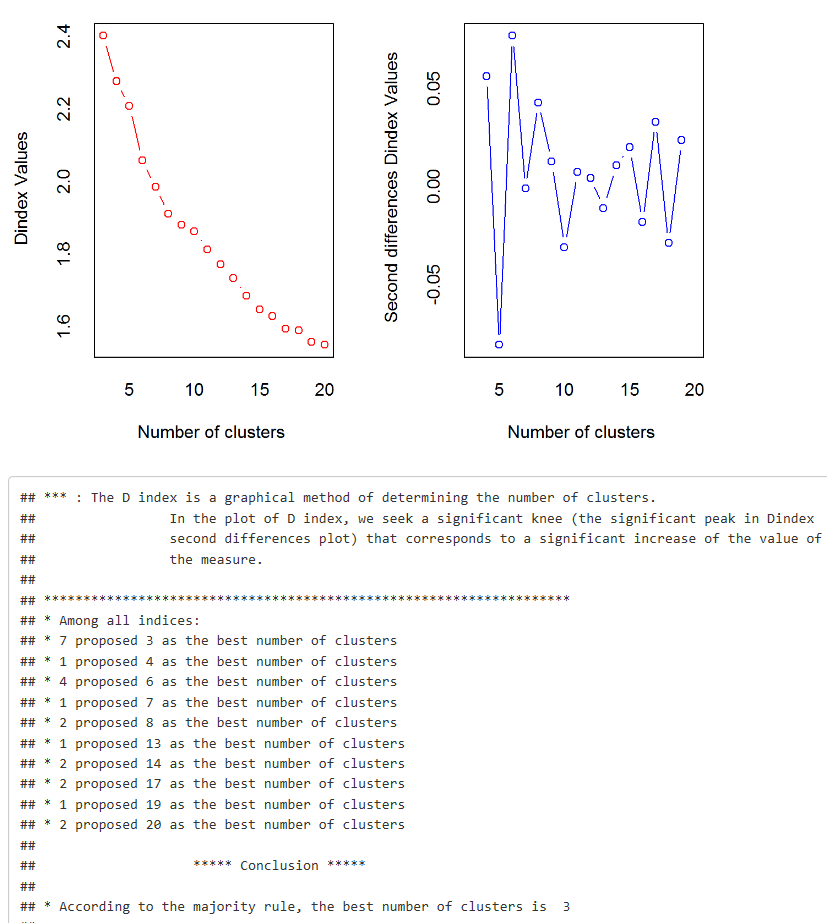


Fig. 2 NbClust Analysis (b) D index plots for determining the number of clusters, see the end line: “the best of number of clusters is 3”.

With the Clustering approach, we could determine that the data could be visualized into three clusters (fig. 3).

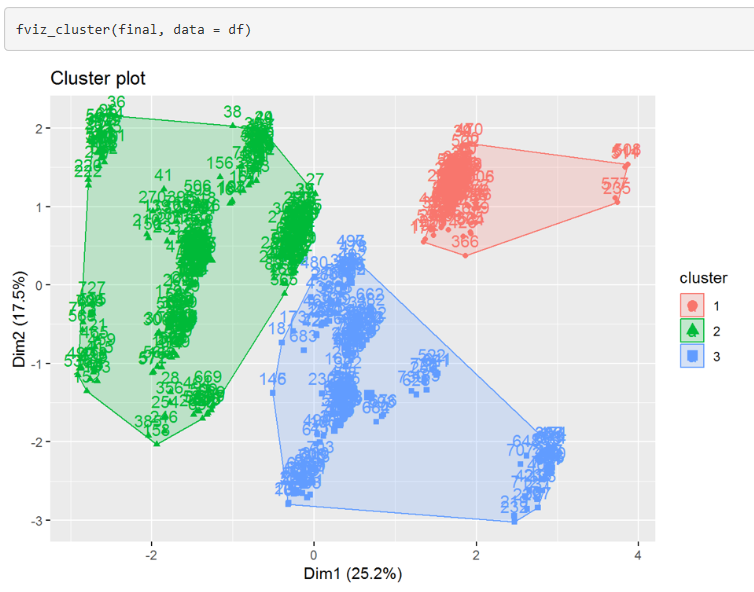


Fig. 3. K-means clustering of Absenteeism data set produced three clusters.

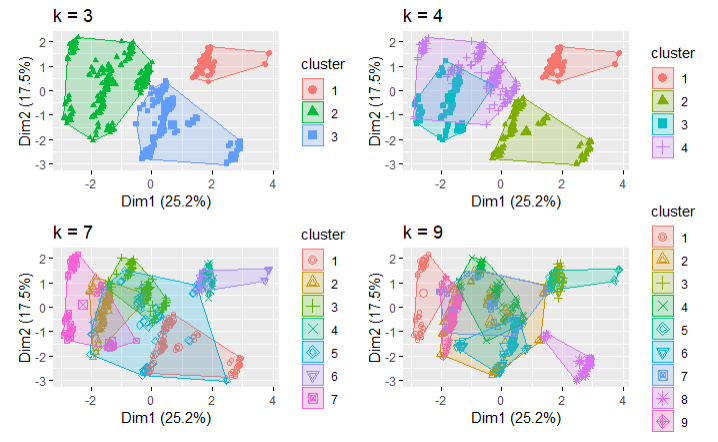


Fig.3.1 K mean clustering of absenteeism dataset produced when 3,4,7 and 9 clusters

## Classification Models

1. C5.0 Classification model

The C5.0 classification model gave us the full attribute usage (fig. 4) for building the decision tree and the models are really complex but gives us the branching, splitting information in visual pattern (fig. 5) for the model trained on the c5.0 classification algorithm. The confusion matrix built yielded an accuracy of 70.71% for the test dataset.

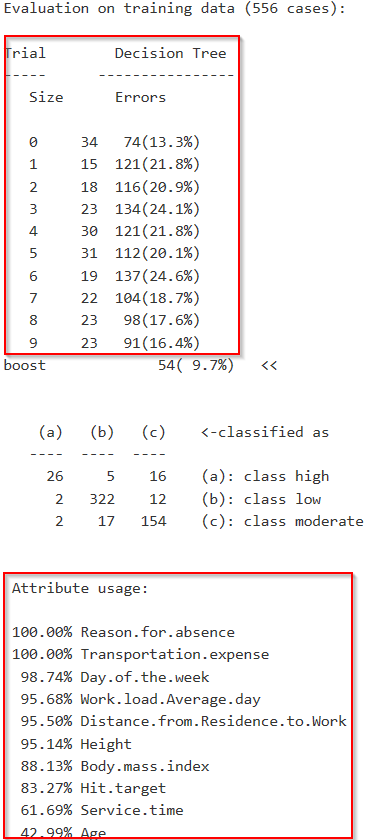


Fig. 4. C5.0 classification results displays the trials (used Trial =10), error rate on each trial and the attribute usage when the model is trained.

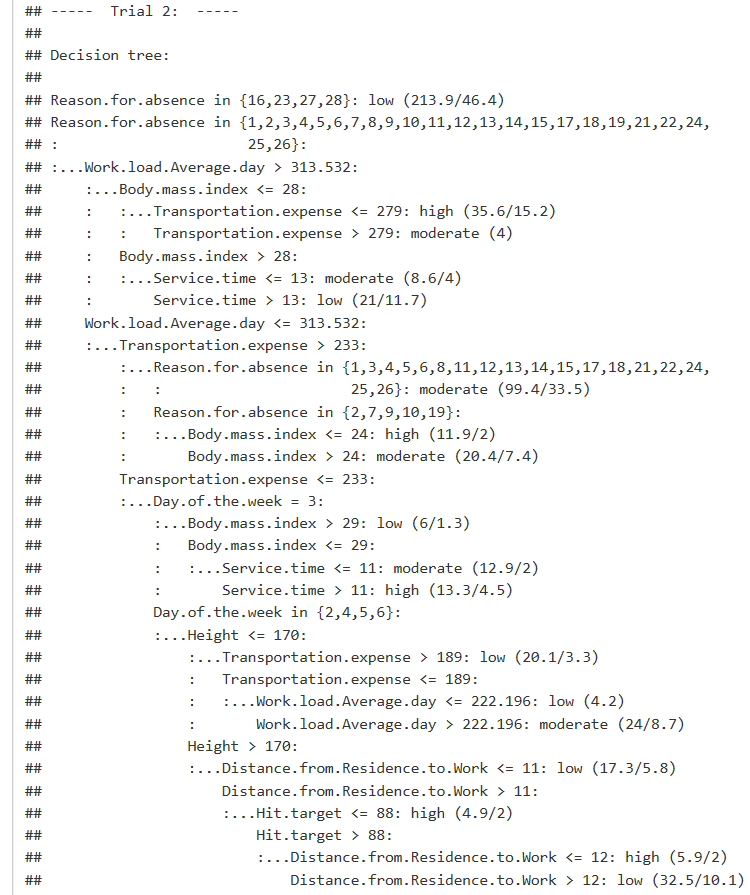


Fig. 5. C5.0 Decision tree from the Trial 2.

Supporting Vector Machines

The model built using the SVM method yielded an accuracy of 69.29% with an error rate of 0.2086 which could have been due to the high number of misclassifications for moderate absenteeism group.

Recursive Partitioning Regression Trees

The simplified binary trees built using the Recursive Partitioning model yielded an accuracy of 67% for the test dataset.

The regression trees (Fig. 6) produced by the Recursive Partitioning are very easy to understand if the accuracy is comparatively similar to the other classification models. The variable character length is trimmed to make it visually better.

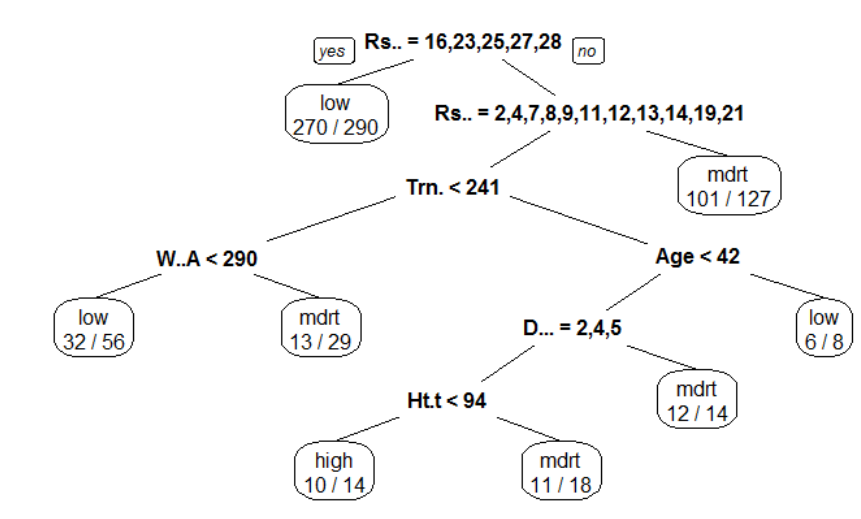


Fig. 6 Recursive Partitioning Regression Trees built in R using *rpart.plot* or *prp.*

Random Forest Method

Random Forest method produced the model with the highest accuracy of all the Classification models used in this study. The accuracy of this model was found to be 72.14% and low false positives misclassification compared to other algorithms.

Basic Tree

Basic Trees produced by the *‘tree’* function in R produced a branched tree (Fig. 7.) and model predicted had an accuracy of 65%. The tree unlike the other decision tree is disordered and does not have proper splitting and branching for visual analysis.

Linear Discriminant Analysis

The model produced via the Linear Discriminant Analysis (LDA) methods had an accuracy of 70% and gave the overall distribution of the absenteeism hours in the plot (Fig. 8.) for each group high, low and moderate.

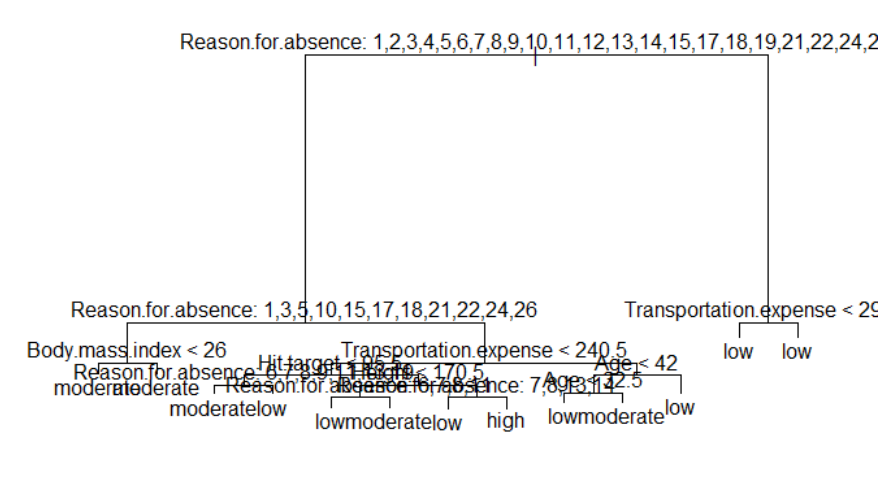


Fig. 7. Classification Model Tree produced the ‘Tree’ function. There

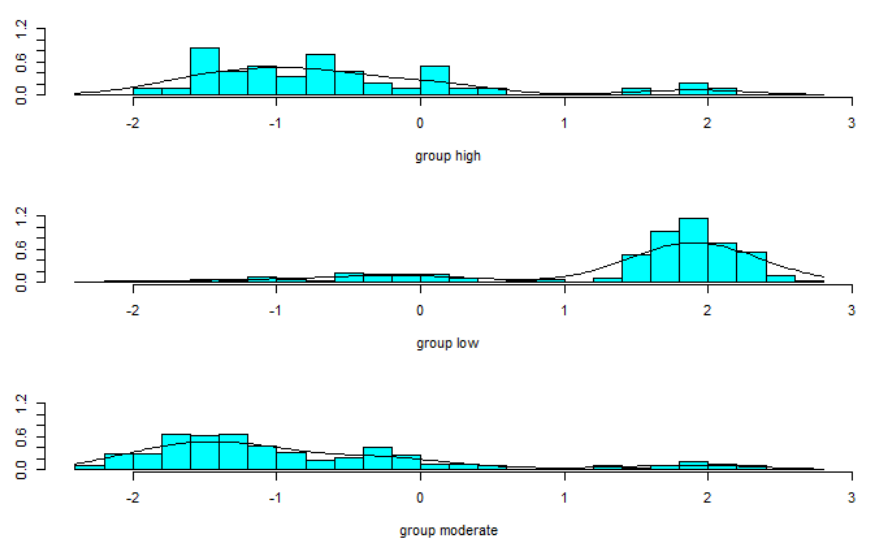


Fig. 8. LDA plot for the absenteeism data.

## Model Performance Comparison

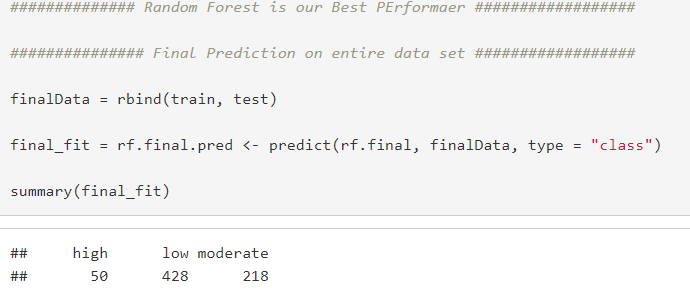
## The results produced by all the models were compared and it was found that the model produced by the Random Forest Classification method yielded the highest accuracy in prediction as compared to the other approaches. The other classification methods yielded accuracy in varying percentages but were almost in less than 3% accuracy range which makes the performance of each mode between 65- 72% range.

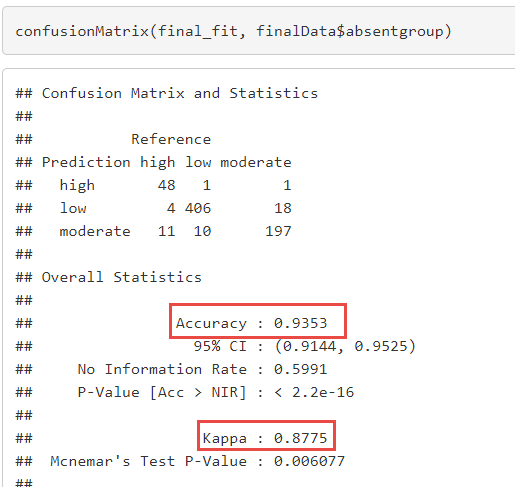
|  |  |  |
| --- | --- | --- |
| Classification Method | Accuracy | Kappa |
| Random Forest | 0.72 | 0.47 |
| C5.0 | 0.71 | 0.44 |
| LDA | 0.70 | 0.46 |
| SVM | 0.69 | 0.42 |
| Tree | 0.66 | 0.38 |

Table 1. Accuracy and Performance of all the Classification Models.

## Final Prediction

The final analysis performed by using the entire dataset as the validation method for Random Forest classification yielded an accuracy of 93.53%. This was performed to evaluate the performance of the model trained using the Random Forest classification algorithm. 88% Kappa signifies that the model explained the variability in the data almost perfectly. So, this model is fit for this data set given the most influencing predictors which are Reason for absence, Transportation Expense, Distance from residence to work, Service time, Age, Work load average day, Hit target, Height and High body mass index.



(a)

(b)

Fig. 9. Final prediction made to validate the Random Forest model.

# CONCLUSION

In our entire study, we have the seen that a number of attributes can influence the Absenteeism but only a few could be called as key predictors in the decision making or predictions.

The Machine learning algorithms used in the study center around the ability to generate a model that could make accurate predictions of absenteeism given the key predictors are used for making the classifications and allow the Decision maker to understand the reason for high, low and moderate absenteeism could be due to one of the predictors that produced that response frequency.

The Random Forest algorithm which is so far the robust classification method in supervised learning with low error rates, false positives leading to misclassifications in the predictions. The decision trees produced by other methods can visually produce the analysis results but can sometimes have higher error rate which is a major drawback in few machine learning algorithms. Also, the overfitting when large number of attributes are taken for visualizations is another problem in this dataset.

It is obvious from the identified predictors that, if Travel expense is high, that means distance from residence to work is also large. If the person is aged, or middle age, then daily travel to work sometimes could be issue for different reason like illness, demotivation, anxiety, physical problems etc. Also, if the work load is high in such cases that could be a real issue in terms of performance. Thus, we can conclude that the predictors identify through this analysis are best influential factors to predict absenteeism at workplace. Also, this research opened a doorway for future studies where the identified predictors can be analyzed more towards controlling of absenteeism at work place.

# REFERENCES

1. Ahn, S., & Lee, S. H. (2011). Development of model of workers’ mental processes related to absence norm as behavior rule in agent-based simulation. *Proceedings - Winter Simulation Conference*, 3479–3487. https://doi.org/10.1109/WSC.2011.6148043
2. Ferreira, R. P., Martiniano, A., Napolitano, D., Farias, E. B. P., & Sassi, R. J. (2018). Artificial Neural Network and their application in the prediction of absenteeism at work. *International Journal of Recent Scientific Research*, *9*(1), 1–12. https://doi.org/10.24327/IJRSR
3. GayathriT. (2018). Data mining of Absentee data to increase productivity. *International Journal of Engineering and Techniques*, *4*(3), 478–480. https://doi.org/10.1016/j.scitotenv.2003.07.004
4. Thurston, E., & Glendon, A. I. (2018). Association of risk exposure, organizational identification, and empowerment, with safety participation, intention to quit, and absenteeism. *Safety Science*, *105*(January 2017), 212–221. https://doi.org/10.1016/j.ssci.2018.02.012
5. Trivedi, H. (2018). *Explaining Absenteeism at Workplace Predicted by a Neural Network*. Canberra.

1. [↑](#footnote-ref-1)