International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Prediction of Employee Absenteeism Using Machine Learning Techniques

P. R. Nivetha, Student, Dr.N.G.P. Arts and Science College, Coimbatore

N. Vanitha, Assistant Professor, Dr.N.G.P. Arts and Science College, Coimbatore

ABSTRACT:

Absenteeism is a serious workplace problem and an expensive occurrence for employers and seemingly unpredictable in nature. A satisfactory level of attendance by employees at work is necessary to allow the achievement of objectives and targets by the department. There will be a high financial loss due to employee and because of the resultant reduction in productivity and the cost of sick leave benefits or others are paid as wages for no work. Absenteeism reduces the satisfaction level of the employee and makes employees unsecured about their job in the organization. Most researches had concluded that absence is a complex variable and that it is influenced by multiple causes, both personal and organizational.

The project "Prediction of Employee Absenteeism Using Machine Learning Techniques" is used to predict the absenteeism according the past records and reasons for absent. The term absenteeism is used to refer to unauthorized absence of worker from the job. The objective of the research is to identify the reason of absenteeism and its causes of manpower and to suggest various measures to reduce the absenteeism. The key issue for the growth of organization is absenteeism of employees, because of its adverse impact on work place productivity and long-term growth strategies. To solve this problem, organizations use machine learning techniques to predict employee absenteeism. To take action for retention or succession planning of employee accurate predictions are enabled. This is the key challenge that is the focus of this paper, and one that has not been addressed historically.

INTRODUCTION:

The absence of an employee from work is known as employee absenteeism. It is a major problem faced by almost all of the employees today. The work suffers due to absenteeism of employers. Back logs, piling of work are due to absenteeism of employees and thus work delay. Even though there various laws been enacted for safeguarding the interest of both Employers and Employees but they too have various constraints.

Background Study:

G. Bergstrom, M.Lohela-Karlsson, L. Bodin, I. Jensen, L. Kwak., in BMC Public Health, 2017 The strategy says that the seasonal changes are also the reasons for health issues. This research gives the accuracy of 97% for women's and 60% for men. For employees with mental health problems or stress related symptoms, failure to take the work environment into account may lead to reduced work ability and repeated and/or prolonged spells of sick leave. The current intervention looks at both the individual and the workplace context. If the intervention proves successful and is implemented at large within the OHS sector, it may result in increased work ability, reduced rates of sick leave and improved quality of life among employees with CMDs or occupational stress. This may also reduce costs, both for the employer and for society at large.

Zaman Wahid, 2018 done his research based on tree-based machine learning

classifiers. Decision Tree, Gradient Boosted Tree, and Random Forest to predict absenteeism time of employees and to find out the insights that cause employees to perform higher absenteeism at work. The dataset contains 21 categories of the reason for absence which are attested by the International Classification of Disease (ICD) and 7 other categories without the ICD that have proved to be effective in detecting the absenteeism at work. Based on the seven evaluation metrics such as True Positive, True Negative, False Positive, False Negative, Sensitivity, Specificity, and Accuracy we have evaluated the model performance in predicting absenteeism at work. His strategy analysis found that Gradient Boosted Tree produces the best result with an accuracy rate of 84.46% whereas Decision Tree performed the lowest with the accuracy rate of 80.41%. The Random Forest classifier performs in between with an accuracy rate of 82.43%. Using the tree model, he discovered that the reason for absence class as diseases that are attested by International Code of Diseases (ICD).

M. Vijaya Bhaskar Reddy, S. K. Avez and P. Chakradhar, Golden Research Thoughts, June 2014, describes that the absenteeism occurs mainly on some stages like Maladjustment with the working, Social and religious ceremonies, Unsatisfactory Housing, Industrial housing, Unhealthy Working Conditions, Poor Welfare Facilities, Alcoholism, Indebt ness and Maladjustment with job demands. This research give the accuracy of 97%. The study also says that the age is also a major reason for the absenteeism i.e., that the aged people will be low in their immunity level and thus in many cases the absenteeism is occurred.

Wouter Langenhoff, Dr. S.J.A. Hessels, May 11 2011, outcomes with a conclusion that the absenteeism is mainly occurred due to many reasons like, gender, age, education, climate, health status, life style, relationship status, children. The study also says that the absenteeism of an employee is also based on the family issues and the author also used a formula in detecting the absenteeism. The mental stress is also a major affect of employee absenteeism.

Biosci. Biotech. Res. Comm. Special Issue Vol 12 No.1 January 2019, brings the accuracy of 91% of accuracy. The study takes the seasons, transportation expense, Distance from Residence to Work, service time, Disciplinary failure, hit target and children. The author uses a chart representation of explanation to explain the strategy of absenteeism. Weight is also a major defect of absenteeism. The employee weight is also important. Because it also places a major role in health. There are three models used in this research. Naïve Bayes model The Reduced gives 91% of accuracy, Decision Tree model The Reduced gives 90% accuracy and Random Forest model The Base gives 92% of accuracy.

Harshit Trivedi, in July 2010 had found an Artificial Neural Networks algorithm to predict this employee absenteeism. In his research he found 58% of accuracy. This study has been performed to train a back propagation trained feed-forward neural network to predict the employees' absenteeism at workplace at a courier company during the time between July 2007 to July 2010. a better neural network model can be built using this technique to explain employees' absenteeism at workplace.

Likhitha, in 2017 published a research paper in employee absenteeism on the basis of data analytics method. In his research she explained many reasons for absent like, bullying and harassment, burnout and stress, childcare and eldercare, depression, illness, injuries and job hunting. In his research she gives an accuracy of 65%. She used data analytics model to predict the employee absenteeism.

Machine learning

A specific task without using explicit instructions, relying on patterns and inference instead is known as machine learning.It is also said as a subset of artificial intelligence. The training data inmachine learning algorithms build a mathematical model based on sample data, in order to make predictions or decisions without being explicitly programmed to perform the task.wide variety of applications use machine learning technique, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning in computational statistics, which focuses on making predictions. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machinelearning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Some Machine Learning methods:

1.supervised machine learning

2.Unsupervised machine learning

3.Semi supervised machine learning

4.Reinforcement machine learning algorithm.

Significance of Machine Learning:

1. Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

Supervised Machine Learning

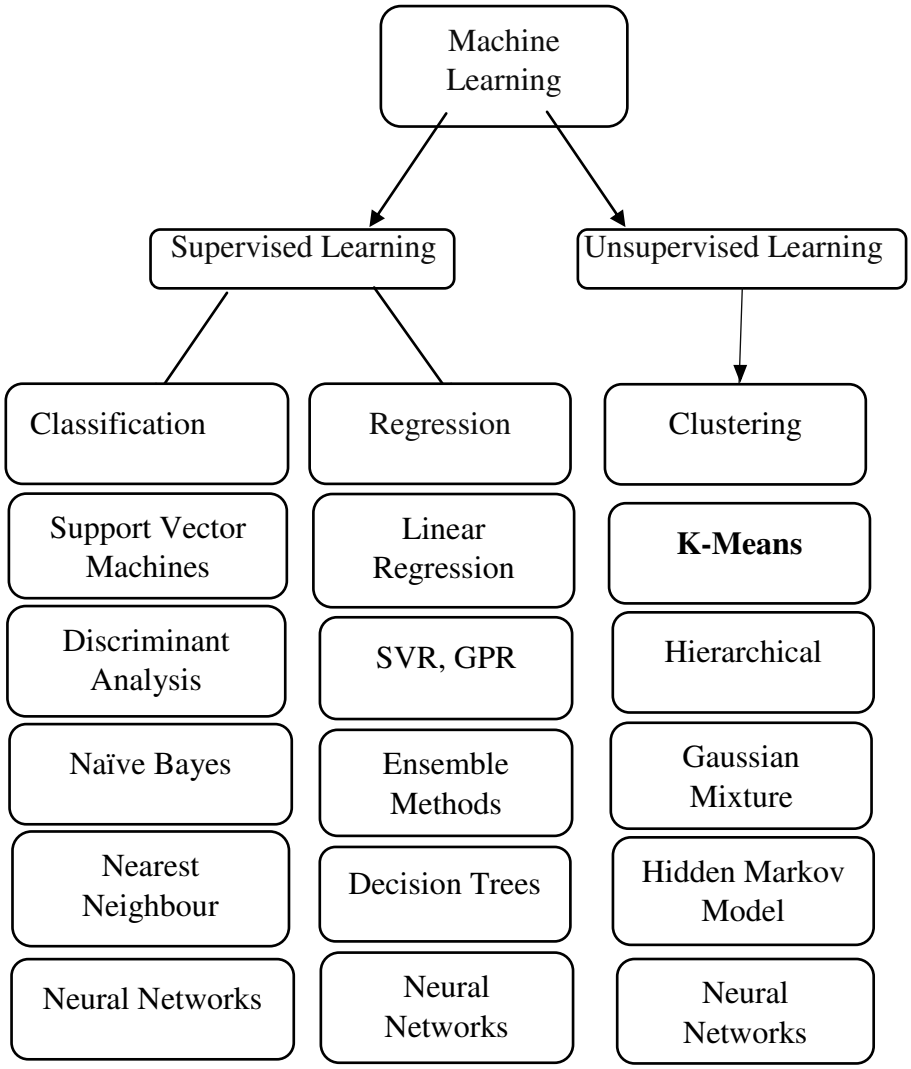


Figure 4.1

Machine Learning in Education:

Machine learning in education is a form of personalized learning that could be used to give each student an individualized educational experience. Here, the students are guided for their own learning, can follow the pace they want and make their own decisions about what to learn. Machine learning, it is a concept which allows the machine to learn from examples and experience. The machines don't write the codes and instead of that data is fed in the generic algorithm. The algorithm that machine builds is the logic based on the given data.

In simple terms, it can be defined as a field of computer science which uses statistical techniques to give computer systems the ability to "learn". This helps to progressively improve performance on a specific task with data. So, without being explicitly programmed, this can be done. For example – in education, we see machine learning in learning analytics

and artificial intelligence working successfully.

Essential machine learning applications in educational field are:

1. Adaptive Learning
2. Increasing Efficiency
3. Learning Analytics
4. Predictive Analytics
5. Personalized Learning

Significance of Employee Absenteeism:

Absenteeism in work place affects the development and growth of the company and leads to affect the financial growth rate. Employees get absent due to different reasons like distance, children, alcohol, health, family, poor welfare facilities, low level wages and so on. In some cases there will be a huge rate of loss in company due to employee absenteeism.

Significance of Machine Learning in Employee Absenteeism:

In machine learning there are many clustering methods to create employee absenteeism algorithms. k-means clustering method is used to create accurate result for predict absenteeism.

Machine Learning Approach:

k-means clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbour classifier, a popular machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbour classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

Training Dataset

There are two kinds of data sets – Training and Test that are used at various stage of development. Training dataset is that the largest of two of them, while test data functions as seal of approval and you don't have to use till the tip of the event.

Test Dataset

This is the data typically used to provide an unbiased evaluation of the ultimate that are completed and fit on the training dataset. Actually, such data is employed for testing the model whether it's responding or working appropriately or not.

Feature Extraction

A data in the dataset are imported as data frames, for that machine learning numpy package is used. The data can be understood by the pipeline. A pipeline consists of a chain of processing elements arranged so that the output of each element is the input of the next; the name is by analogy to a physical pipeline.

Reading Data

We need to read the data-sets into data frames which can be understood by the pipeline. We will use pandas module in python for that task.

Splitting Data

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. In order to use the splitting method we have to import pandas library

- **Training set** – a subset to train a model (80%)
- **Testing set** – a subset to test the training model (20%)

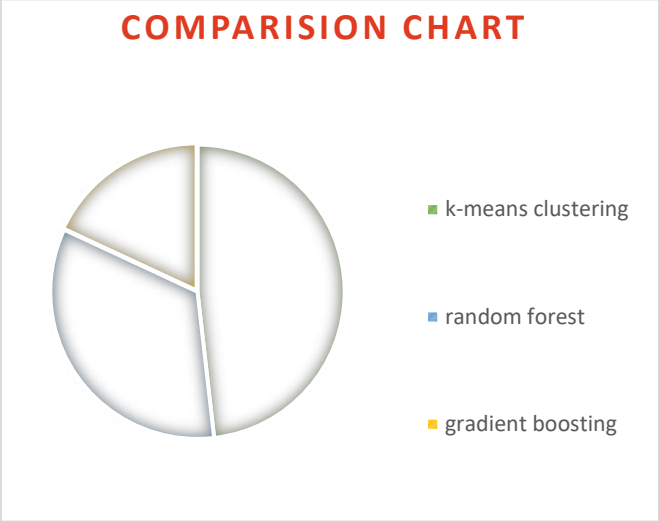
Training Model

To train the model Stratified ShuffleSplit - cross validator that is imported using sklearn model from python sci-kit library. (From sklearn.model_selection import StratifiedShuffleSplit). It uses train data set for learning. After learning it prints score of trained model.

Testing Model

Pass the students information as inputs in to trained model. It provides whether the student will continue or dropout their studies and returns the output as Yes/No.

Comparison Chart:



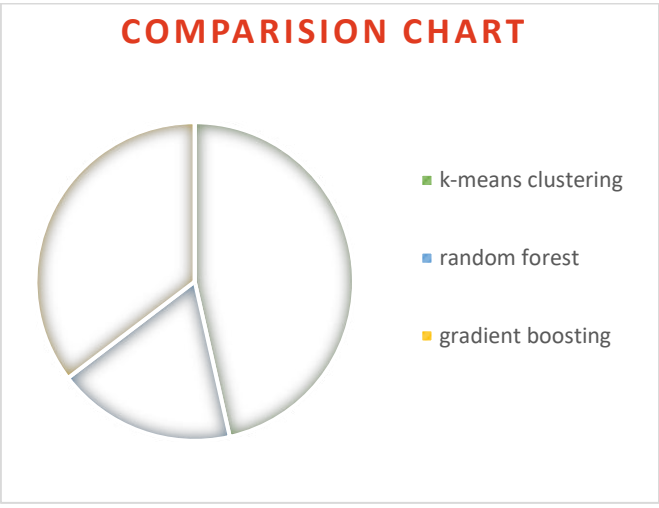
The chart shows that the comparision of three methods results, k-means clustering is more accurate than random forest and gradient boosting.

The k-means clustering method shows that the employees where absent mainly due to health issues and because of personal problems.

The gradient boosting technique shows that low pecentage of output than the other two.

From the 1:1 ratio, the random forest, k-means clustering and the gradient boosting, the k-means clustering shows 80% of result.

Whereas the random forest and the gradient boosting shows 52% and 30% of result.



This comparision chart figures that k-means clustering alogorithm is more efficient than the other two booting methods.

In this comparision the random forest is the least resultant one.

From the 4:1 ratio, the random forest, k-means clustering and the gradient boosting, the k-means clustering shows 92% of result.

Whereas the random forest and the gradient boosting shows 36% and 70% of result.

Output:

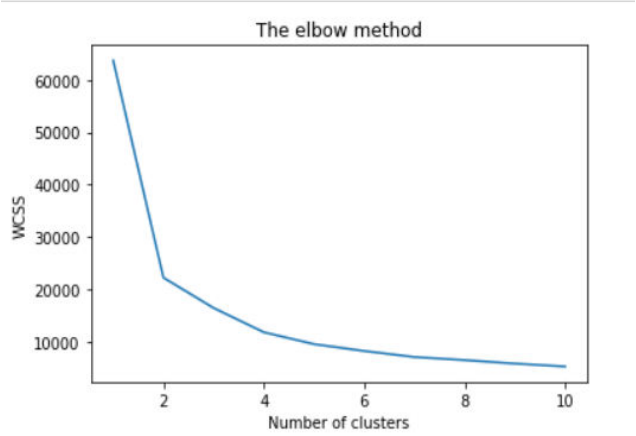


Figure 1.2

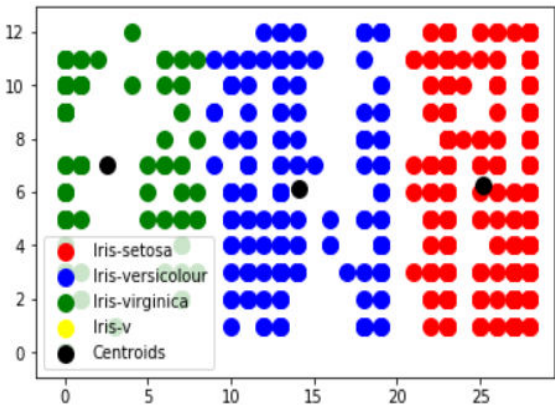


Figure 1.3

Research Challenges

Absenteeism at work acts as a bottom line in an organization. Employers around the world believe that the absenteeism of employees can have a major effect on company finances and other factors.They do not expect those employees who perform excessive absenteeism at work which cause reducing productivity and thus cost the company

Conclusion

After analyzing the results it was determined that the best algorithm for predicting employee absenteeism is the k-means clustering method. The predictive capacity of this algorithm was the best of the alternatives evaluated. In addition to yielding the best sensitivity metrices and true positives, the k-means clustering method with the prespective discussed previously shows a smaller gap between these metrics, and more appropriate behaviour through time. It was found that both the ability to correctly detect absenteeism and sensitivity increases over time.

For future work, plan to explore different datasets so as to show and compare results of different train, test and validation and evaluate several imbalance techniques for student dropout prediction using more measures for results comparison.

REFERANCE:

1. G.Bergström, M. Lohela-Karlsson, L. Kwak, L. Bodin, I. Jensen.IEEE, Golden research thoughts, *Preventing sickness absenteeism among employees with common mental disorders or stress-related symptoms at work: Design of a cluster randomized controlled trial of a problem-solving based intervention versus care-as-usual conducted at the Occupational Health Services*, Centre for Occupational and Environmental Medicine, Stockholm County Council, SE-113 65 Stockholm, Sweden.

2. Zaman Wahid, IEEE, 2018, International Classification of Disease (ICD),*Predicting absenteeism of employees at workplace using tree-based algorithms*, Dhaka

University of Engineering & Technology, Gazipur, 151-35-953.

3. M. Vijaya Bhaskar Reddy, S. K. Avez and P. Chakradhar, SPRINGER, Golden Research Thoughts, June 2014, *Employee absenteeism: a case study of hindustan coco cola beverages private limited*, Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India, .2231-5063

4. Wouter Langenhoff, Dr. S.J.A. Hessels, May 11 2011, SPRINGER, Netspar , *Employee Absenteeism: Construction of a Model for International Comparison of Influential Determinants erasmus university rotterdam Erasmus School of Economics, Master Thesis*, pp.2011-011.

5. Biosci. Biotech. Res. Comm. Special Issue Vol 12 No.1, IEEE, Communication Special Issue Vol 12 No.1 January 2019,*Employees Absenteeism Factors Based on Data Analysis and Classifi cation,* Faculty of Computing and Information Technology, King Abdul Aziz University, Jeddah, Saudi Arabia, pp.2019-201

6. Harshit Trivedi,July 2010, IEEE Research School of Computer Science, The Australian National University, Canberra ACT - 2600, Australia, *Explaining Absenteeism at Workplace Predicted by a Neural Network,* Australlia, pp.77-235.

7. Likhitha, Mr. Praveen, 2017, IEEE, employee absenteeism, 2017*Data analytics and interruption, by using formula*, Mangalore, pp.1110-1212.