

# A Survey on Big Data Applicability in Prediction Using Absence Information for Workforce Management

R. Varalakshmi, R.S. Dhivya

*Abstract--- Prediction is a method of finding new insights from large data sets, it's a reliable way to use big data for more accuracy. The overall goal of the process is to predict useful outcomes from dataset and transform into a meaningful insight for decision. Absence data has thousands of leave type information; by analysing the data, new insights can be predicted, since the data is growing exponentially, traditional method of prediction will not be effective. In this paper we discuss about prediction techniques and a survey of works done in the field of absenteeism is performed, This paper concentrates on machine learning algorithms and also presents the applicability of absence data in big data for workforce planning.*

*Keywords--- Big Data, Data Science, Predictive Analytics, Workforce planning, Random Forest.*

## I. INTRODUCTION

Workforce attendance and absenteeism are important to organisations as it affect the productivity cost and may lead to poor quality of goods and service, it also leads to additional work pressure for other employees on duty. Management of absenteeism is the one of the key strategic action in Human Resource function that any organisation wish to pursue in order to assure company performance and success said by Goetzell[1]. In the modern industry, millions of data is generated, each data set is unique and different from other and has valuable information, Absenteeism data is one such data set which should be exploited and utilized, which has valuable information to aid organisation's HR strategic action. Employee Absence are estimated for 3% loss, those absence must be filled either by overtime or with an additional workforce, The ability to predict employee absence would facilitate effective workforce planning. In this paper we use Predictive analytics for Workforce prediction, Human Resource (HR) has a chain of functions starting from recruitment to retirement, however absence data is the large data set in HR, Though numerous machine learning procedures can be used to predict information, Classification is one of the powerful technique will get the prediction need. In this work we attempt to study the different Classification algorithm to predict the absenteeism accuracy, deal briefly on absence data for prediction using Random forest algorithm, glimpse upon Predictive analytics, Absence data as big data, Workforce planning and Random Forest & Decision tree.

## II. LITERATURE SURVEY

Several studies and research has been done in absenteeism management such as [2] As stated by Evans in Nursing home industry, there were 22 categories of absenteeism was classified, 12 of the categories with low absenteeism differed from other nursing homes however the other 10 categories are common across nursing homes, which suggest absenteeism is unique to every firm though the industry is common. [3] Staffs in Wa Municipal Education Office has the habit of absenting from work due to religious and festival occasions which is not related to their health or personal problems as researched by Ghana Education service; which elucidates absenteeism cannot be defined into certain common categories. From the above two studies we infer absenteeism is unique to industry and society, however from the data perspective the absenteeism reasons to be consider as attributes (a1, a2, etc.) of absence data and a common classification algorithm can be used.

Valle, Vara and Ruz, [4] Used Navie Bayes algorithm to predict job performance in call service office to know the degree of individual future performance using the transactional record attribute of average performers. [5] Jantan, Hamdan and Othman used C4.5 classification algorithm based on decision tree algorithm to predict the talent pattern from talent records. [6] Rishi Sai Reddy Sudireddy and Uttam Mande propose improvement measures by using MapReduce in term of the time taken to analyse the data. [7] Jantan propose a Sequential Minimum Optimization (SOM) algorithm in SVM technique, used for employee achievement performance pattern, the accuracy of the model is proposed. [8] Aarya Vardhan states that; apply certain clustering algorithm into a real word data set of absenteeism to understand the mechanism of the algorithm which would help us to identify effectively based on the features. [9] Ferreira used Artificial Neural Network (ANN) these model are inspired by the structure of the brain and aim to simulate human behaviour. [10] Gayathri creates a classification model to predict employee absence for short and long term absence with accuracy, which affects the productivity.

[11] Karakus says Hadoop clusters running on cloud infrastructure can be changed based on the call centre agents call records processed. [12] Aditya Sinha use pruning excessive node and re-evaluate the system accuracy which the ANN classified. [13]

**Revised Version Manuscript Received on 22 February, 2019**

**R. Varalakshmi**, Professor, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai.

**R.S. Dhivya**, Assistant Professor, Department of Computer Applications, J.H.A Agarsen College.  
Research Scholar, VISTAS, Chennai.

Liu presents the improved random forest algorithm to realize the self-adaptation in similar situations and also verified the viability of the new improved method while using the actual data in big data environment. [14] Hanl compares scalable Random Forest algorithm with traditional Random Forest algorithm on MapReduce, which produce equal accuracy and more suitable to classify large datasets. [15] Breiman states random inputs and random features produce good results in classification method using bagging and random features technique. [16] Lin used ensemble random forest algorithm with heuristic bootstrap sampling approach on the large-scale insurance business and suggests parallel computing capability and memory cache mechanism optimized through Spark, Its more suitable for product recommendation or potential customer analysis than traditional strong classifiers like SVM and Logistic Regression etc.

From the above studies we infer data sets differ, however a tailored algorithm can be used based on the attributes of dataset.

### III. PREDICTIVE ANALYTICS

Predictive analytics consists of wide variety of techniques, it uses mining methods, artificial intelligence through machine learning to predict new events and possible events could happen in future. Within the other types of analytics viz. Descriptive, Diagnostic, Predictive and Prescriptive; Predictive analytics uses the inference of descriptive & diagnostic analytics to detects tendencies, clusters and exceptions to predict future trends which aids planning. Though predictive analytics has numerous advantages, the accuracy of prediction depends on the data quality and the flow of consistent data in future period of time, which includes data transformation and data wrangling.

### IV. BIGDATA

Big data is traditionally characterized by three elements, also called the three V's.

- Volume: Terabytes, Records, Transactions, Tables and files.
- Velocity: Batch, Real-time, streams and Near-time data.
- Variety: Unstructured, structured and semi structured data.

Other Vs of Big Data:

- Value: Trends, sentiment, perspective and risk.
- Veracity: Quality and accuracy.

#### Applications of Big Data

Automobile, Tele communication, Retail Industry, Financial services etc., use big data for prediction to cater their business need in the form of:

- Customer analytics
- Compliance analytics
- Fraud analytics
- Operational analytics
- Production targets

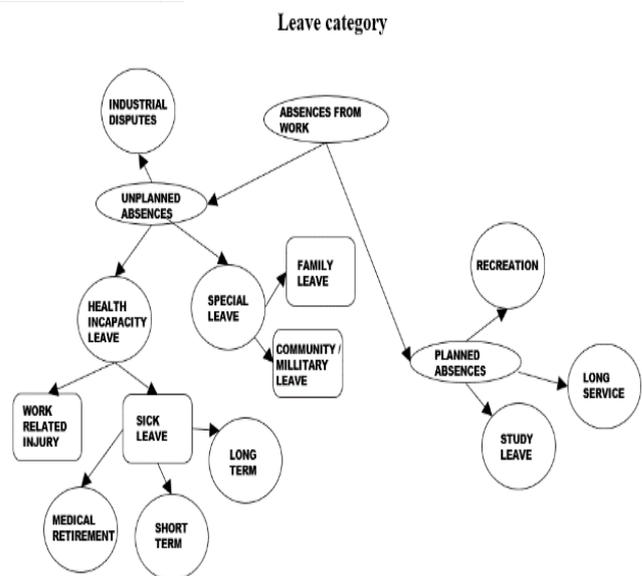
However, Human Resources is one of the business function in any industry can use big data to support business in people needs, here workforce planning.

### V. ABSENTEEISM

Absenteeism is a habit of not attending to duty or work without a proper reason. Generally, absenteeism is unplanned absences. However in this analytics we consider all types of leave of absence either planned or unplanned as in the below picture[17] for our prediction, since absence data includes all type of leave.

#### A. Absence Data

Any organisation with Human Resource function irrespective of its size; absence data tops second in volume of HR data after Payroll data followed by Employees job transaction data, However, in perspective of its usability in predictive analytics absence data gives more flavour the third "V" Variety. Organization with 1,000 employees allowing 21 days of annual leave and 10 days of other leave can approximately generate 9,000 data rows if an average employee use 9 times the online absence module, We also include associated attributes such as age, employment tenure, habit, gender and their past absence data for absence prediction.



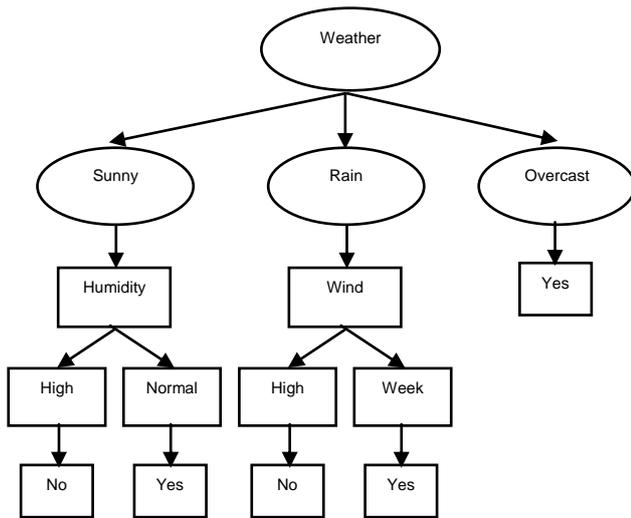
### VI. RANDOM FOREST

Random forest is an ensemble learning method, It is used in regression and classification, this works by combining numerous individual decision trees to form part a Random Forest.

Decision Tree, In decision tree the root of the decision tree is a simple question that has multiple answers, Each answer then leads to a set of multiple questions that help to determine the answers to make the final decision. For example, We use the following decision tree to determine whether to play tennis:

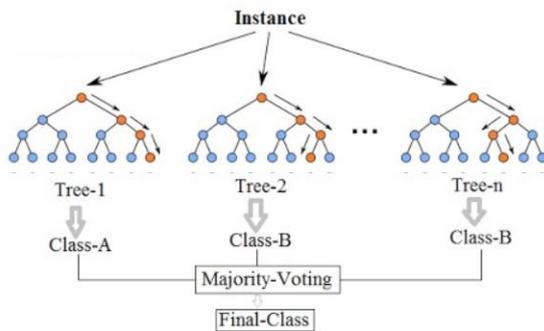


### Decision Tree

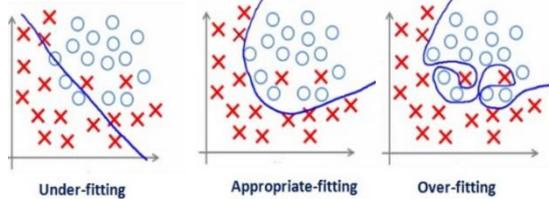


Random Forest is a collection of Decision Trees, its algorithm randomly selects observations to build several decision trees and then averages the results.

### Random Forest



Since its algorithm randomly selects observation, it doesn't use all observations unlike other decision trees, hence Random Forest prevents over fitting.



#### A. Advantage of Random forest

- Efficient on large data set.
- Handle lot of variable without variable deletion.
- Robust on error data.
- Provides an estimate of important variables in classification.
- Applies balance error method in unbalanced data set.

### VII. ABSENTEEISM PREDICTION

Number of predictions human can assume such as below:

- Habitual
- Leave after pay day.
- Vacation Leave on particular season.
- Leave prior or after weekend.
- Relationship Dates.

- Birth date of family members.
- Marriage date.
- Other relationship related dates.
- Entertainment, Sports & Festival dates.
- Cricket / Football events in city.
- New film release.
- Local festival
- Life events other than marriage.
- Maternity.
- Paternity.
- Bereavement.

However we expect the algorithm to provide new insights and the prediction number for the given data set. We expect absence prediction will aid in workforce planning.

### VIII. CONCLUSION

This paper summarise Random forest algorithm method and its advantage, it also covers the importance of absence data and its scope for workforce planning. The future scope of the paper is to choose absence prediction stated in section-VII either by supervised learning.

### REFERENCES

1. Goetzel, R.Z, Long, S.R., Ozminkowski, R.J., Hawkins, K.H., Wang, S. and Lynch, W. Journal of occupational and environmental medicine, on Health, Absence, disability and presenteeism cost estimates of certain physical and mental health condition affecting US employee" 46pp.398-412.
2. Evans - Absenteeism- a complex problem A study on absenteeism in Trondheim's nursing homes Josiane, NTNU Master's thesis in Cultural, Social and Community Psychology
3. Hafiz Bin Salih, Staff Absenteeism: The Case of Wa Municipal Education Office of the Ghana Education Service Coordinator, Ghana Education Service, Wa, Ghana, Open Journal of Social Sciences, 2018, 6, 1-14
4. Valle, M.A., Vara, S., Ruz, G.A, (2012), Job performance prediction in a call centre using a Naïve Bayes classifier, Expert System with Applications, 39(11), pp 9939-9945.
5. Jantan, H., Hamdan, A.R., & Othman, Z.A. (2010). Human talent prediction in HRM using C4.5 classification algorithm, International journal on Computer Science and Engineering, (2008-2010), pp2526-2534.
6. Rishi Sai Reddy Sudireddy and Uttam Mande, GITAM University. Prediction of Road Accident using Correlation based on Map Reducing. 8th International Conference on Computational Intelligence and Communication Networks 2016.
7. Hamidah Jantan, Norazmah Mat Yusoff and Mohamad Rozuan Noh, Universiti Teknologi MARA (UiTM) Terengganu, Malaysia. Proceedings of the International Conference on Data Mining, Internet Computing and Big Data, Kuala Lumpur, Malaysia, 2014. Towards Applying Support Vector Machine Algorithm in Employee Achievement Classification.

8. Aarya Vardhan Reddy Paakaala, Sai Saran Macha & Kumara SakethMudigonda in MSVR Engineering college India. Evaluation of clustering algorithms on absenteeism at work data set. International Journal of Scientific Research & Development Vol.6, 2321-0613.
9. Ricardo Pinto Ferreira., Andréa Martiniano., Domingos Napolitano., Edquel Bueno Prado Farias and Renato José Sassi.Artificial neural network and their application in prediction of absenteeism at work., International journal of recent scientific research Jan'18, vol-9.
10. Gayathri, T. Data mining of Absentee data to increase productivity. International journal of engineering and techniques – vol-4, May' 18, Issue-3.
11. BetülKarakus&Galip Aydin, Computer Engineering Department Firat University Elazig, Turkey. Call Center Performance Evaluation Using Big Data Analytics, 978-1-5090-0284-9/16/\$31.00 ©2016 IEEE.
12. Aditya Sinha, Research School of Computer Science College of Engineering and Computer Science Australian National University ACT 0200 Australia., Predicting Absenteeism At Work using ANN, Effects of Pruning By Badness, and Deep NNs.
13. Yingchun Liu, Random forest algorithm in big data environment , Computer Modelling & New Technologies 2014 18(12a) 147-151.
14. Jiawei Han, YanhengLiul ,Xin Sunl, A Scalable Random Forest Algorithm Based on MapReduce.
15. Breiman, L. Random forests. Machine Learning 45( 1), 5-32 (2001).
16. Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen, AndJin Li, For Fog and Mobile Edge Computing, An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, Special Section On Recent Advance In Computational Intelligence Paradigms for security and privacy.
17. The Australian Faculty of Occupational Medicine: Workforce Attendance and Absenteeism.