ANALYSING EMPLOYEE ATTRITION USING MACHINE LEARNING

Usha.P.M. [1] *Dr. N.V.Balaji* [2]

ABSTRACT

Technology has brought in revolutionary changes in the business processes of organizations with technologies like Big Data Analytics, Artificial Intelligence and Robotics. Predictive analytics is one such innovative technology which is widely used by organizations. With the advent of computers in all areas of work, organizations possess massive amounts of data in structured and unstructured forms. An analysis of these data by using technology like data mining paves way for predicting future trends and behavior, which in turn results in more data driven decision making. These predictions can be made in various functional areas of an organization.

The rationale of this study is to use data mining techniques to understand the factors influencing attrition of human resources using Weka. Weka is a data science tool that can be used for predictive analytics. In knowledge-based organizations, attrition is a severe apprehension because it affects the competitive strength of business. Weka can be employed to cluster data with techniques like k-means algorithm to explain the factors leading to attrition. A comparison of algorithms in Weka can also be made to understand the effectiveness. The dataset provided by IBM is used for this study.

Key words: Attrition, Machine Learning, Clustering, Classification.

I INTRODUCTION

Human resource are the most imperative of all the resources of an organization since it decides how to optimally utilize other resources. And it remains a reality that a human resource cannot be replaced by another. High attrition will lead to low productivity. Hence, attrition has to be dealt with utmost importance and measures have to be taken by organizations to prevent this. If an organization can foresee the risk of an employee parting ways with the organization, necessary measures can be taken to prevent it. In the current era of fourth industrial revolution, with technologies like predictive analytics where future is predicted by means of statistical modeling techniques and machine learning, it is not impossible to predict the likelihood of an employee departing from the organization. This paper discusses the use of classification and clustering for attrition analysis. A comparison is done to check the accuracy of various data mining algorithms with Weka. It is an assortment of machine learning algorithms which are used for performing data mining. It has tools which can perform clustering, classification, association, visualization and the like.

II RELATED WORK

In the study "Analyzing Employee Attrition Using Decision Tree Algorithms" Alao D & Adeyemo A. B (2013) is using Weka to classify employees. A decision tree is generated from the dataset containing 309 records of employees. Various classifiers like J48, REPTree, CART, JRip and SeeTree were compared, and SeeTree was found to be the best in terms of accuracy[1].

Qasem A, Al-Radaideh and Eman Al Nagi (2012) in their work "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", use data mining to make classification models which can be instrumental in predicting the performance of employees. The accuracy percentages were found to be low in C4.5 and Naïve Bayes [2].

Hamidah J, AbdulRazak H and Zulaiha A. O (2011), in their

[1]Research Scholar, Department of CA, CS & IT,
Karpagam Academy of Higher Education, Coimbatore, India.

[2]Research Supervisor, Department of CA, CS & IT,
Karpagam Academy of Higher Education, Coimbatore, India.

study "Towards Applying Data Mining Techniques for Talent Managements" attempt to find the accuracy of classifier algorithms for talent management. They have considered data mining techniques such as decision trees and neural networks. The classifier algorithms used under decision tree are C4.5 and Random forest. For neural network the algorithms used by researcher are Multilayer Perception and Radial Basic Function Network. It was found that Radial Basic Function and C4.5 shows very good performance [3]. Jayanthi et al. (2008) in her paper has explained the contributions of data mining in handling and managing human resources. Human resource data are of utmost importance for all organizations for gaining a competitive advantage in business. The paper explains how data mining can help in an informed and effective decision making [4]

III METHODOLOGY

The tool used for carrying out the experiments is Weka, which is an assortment of machine learning algorithms. The dataset used for the process is a fictional data set created by IBM. It contains 1470 rows and 35 columns.

Four attributes EmployeeCount, EmployeeNumber, Over18 and StandardHours having same value irrespective of values of other attributes are removed. The dataset after removal of irrelevant attributes is saved in a comma-separated file format. Weka contains a tool as arff viewer which can be used for converting a csv file to arff (attribute relation file format). It is an ASCII format developed for using in Weka machine learning software.

Using preprocess button in Weka, the arff file is selected and opened and the attributes are displayed. There are different classification and clustering algorithms in Weka. An attempt is made in this study to compare the accuracy of certain models.

A. Classification

Classification is a machine learning algorithm which comes under supervised learning. From the data set the classifications are created, and based on this learning new observations are classified.

J48

Decision tree is a graphical tool used for decision making in various situations where alternative choices are represented. Under a particular circumstance, the best option can be selected using a decision tree. The figure given below depicts the use of decision tree to make a choice on whether to buy a particular car or not by evaluating the alternatives.

TABLE I: Attributes in the data set

Field Name	Possible inputs
Age	Number
Attrition	"No", "Yes"
BusinessTravel	"Non-Travel", "Travel_Frequently", "Travel Rarely"
DailyRate	Number
Department	"Human Resources", "Research and development", "Sales"
DistanceFromHome	Number
Education	1,2,3,4,5
EducationField	"Human Resources", "Life Sciences", "Medical", "Other", "Marketing"
EmployeeCount	1
EmployeeNumber	Number
EnvironmentSatisfaction	1 2 3 4 (Satisfaction with The Environment)
Gender	"Female", "Male"
HourlyRate	Number
JobInvolvement	1 2 3 4
JobLevel	1 2 3 4 5
JobRole	Laboratory Technician, Healthcare Representative, Manufacturing Director, Human resources, Manager, Research Director, Research Scientist, Sales Executive, Sales Representative
JobSatisfaction	1,2,3,4
MaritalStatus	"Divorced", "Married", "Single"
MonthlyIncome	Number
MonthlyRate	Number
NumCompaniesWorked	Number
Over18	Y
OverTime	"No", "Yes"
PercentSalaryHike	Percentage increase in salary
PerformanceRating	Number
RelationshipSatisfaction	1 2 3 4
StandardHours	Number
StockOptionLevel	0 1 2 3 (The stocks of the company owned by the employee)
TotalWorkingYears	Number
TrainingTimesLastYear	Number
WorkLifeBalance	1 2 3 4
YearsAtCompany	Number
YearsInCurrentRole	Number
YearsSinceLastPromotion	Number
YearsWithCurrM ger	Number

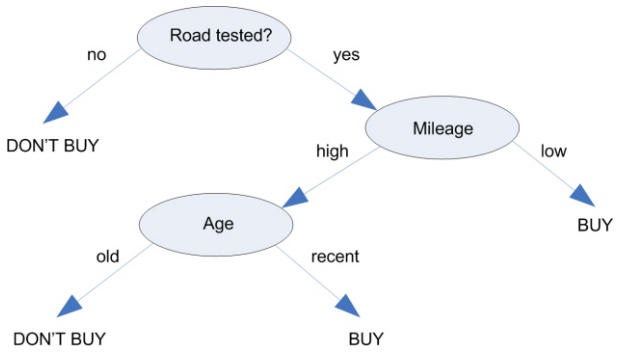

Fig. 1: Example of a decision tree for decision of buying a car [14]

Decision trees can be used for classifying instances based on their features. Decision trees use a top down approach. Each node is considered as a representation of an instance, and branch represents the values taken by the instance.

J48 is a decision tree implementation using Java in Weka. It uses C4.5 algorithm which is a successor of the algorithm developed by Ross Quinlan which is named ID3. C4.5 can be implemented to create a classifier in the structure of a decision tree.

The dataset which is converted to arff format is classified using the classify option in Weka. The algorithm is chosen as J48. The accuracy of the classification attained by the algorithm using 10-fold cross validation is given below

TABLE II: The measures of accuracy of J48 using tenfold cross validation

Algorithm	TP Rate	FP Rate	Precision	Recall	F measure	MCC	ROC Area	PRC Area	Class
J48 (10 fold cross validation)	0.266	0.068	0.429	0.266	0.328	0.242	0.608	0.29	Yes
	0.932	0.734	0.868	0.932	0.899	0.242	0.608	0.854	No

TABLE III: The correct and incorrect classification percentage and time

Algorithm	Correct classification	Incorrect classification	Time
J48 (10 fold Cross validation)	82.449%	17.551%	0.09s

The confusion matrix reveals the following result::

TABLE IV: Classification shown by confusion matrix

	Classified as Yes	Classified as No
Yes	63	174
No	84	1149

J48 using percentage split as 70

The accuracy of the classification attained by the algorithm using percentage split as 70 is given below:

TABLE V: The measures of accuracy of J48 using percentage split 70

Algorithm	TP Rate	FP Rate	Precision	Recall	F measure	ROC Area	Class
J48 (Percentage spilt 70)	0.321	0.058	0.553	0.321	0.406	0.697	Yes
	0.942	0.676	0.86	0.942	0.899	0.697	No

TABLE VI: The correct and incorrect classification percentage and time

Algorithm	Correct classification	Incorrect classification	Time
J48(Percentage split as 70%)	82.7664%	17.2336%	0.14s

The confusion matrix reveals the following result:

TABLE VII: Classification shown by confusion matrix.

	Classified as Yes	Classified as No
Yes	26	55
No	21	339

Naïve Bayes

Naïve Bayes is yet another classification algorithm which works based on Bayes' Theorem. Bayes theorem assumes that each feature of the data set makes independent and equal contribution to the output.

The dataset opened in the previous example is classified using the Naïve Bayes algorithm in Weka. The following result is achieved:

TABLE VIII: The measures of accuracy of Naïve Bayes

Algorithm	TP Rate	FP Rate	Precision	Recall	F measure	ROC Area	Class
Naïve Bayes	0.641	0.183	0.402	0.641	0.494	0.774	Yes
	0.817	0.359	0.922	0.817	0.866	0.774	No

TABLE IX: The correct and incorrect classification percentage and time.

Algorithm	Correct classification	Incorrect classification	Time
NaiveBayes(10 fold Cross validation)	78.8435 %	21.1565 %	0.03s

TABLE X: Classification shown by confusion matrix.

	Classified as Yes	Classified as No
Yes	152	85
No	226	1007

Naïve Bayes with percentage split 70

The result obtained after using percentage spilt as 70 is given below:

Table XI: The measures of accuracy of Naïve Bayes with percentage spilt 70

Algorithm	TP Rate	FP Rate	Precision	Recall	Fmeasure	ROC Area	Class
Naïve Bayes(percentage spilt 70)	0.691	0.164	0.487	0.691	0.571	0.819	Yes
	0.836	0.309	0.923	0.836	0.878	0.819	No

Table XII: The correct and incorrect classification percentage and time

Algorithm	Correct classification	Incorrect classification	Time
NaiveBayes(Percentage split as 70%)	80.9524%	19.0476%	0.03s

The confusion matrix displays the following result:

TABLE XIII: Classification shown by confusion matrix.

	Classified as Yes	Classified as No
Yes	56	25
No	59	301

B Clustering

Cluster stands for a group of similar objects and so clustering stands for the process of making clusters. Clustering in data mining is an unsupervised method for creating clusters. Known outcomes are not readily available in unsupervised learning. Instead clusters are formed from the provided data set based on the characteristics of the dataset under analysis.

K_means

K-means is one of the r unsupervised machine learning algorithm which is popular and is used for clustering. K means algorithm learns underlying patterns. It creates k clusters from the dataset. K centroids are fixed and the data points are analyzed and allotted to the nearest cluster. Again, centroids are calculated and the process is continued until all data points are allotted to clusters.

The given data set of employees is given as input to the k-means algorithm in Weka. The number of clusters required is given as 2. The option **Classes to clusters evaluation** is selected and the attribute Attrition is mentioned as the class for clustering. Here Weka ignores the attribute mentioned as Classes to cluster, and it generates the clusters. In the test phase, the maximum occurrence of the class attribute in each cluster is found, and these classes are assigned to the clusters. The classification errors are mentioned in the confusion matrix generated.

The performance of the algorithm is as follows:

TABLE XIV: The correct and incorrect classification percentage and time

Algorithm	Correct classification	Incorrect classification	Time
K-Means	57.2789%	42.7211%	0.17s

The confusion matrix displays the following result:

TABLE XV: Classification shown by confusion matrix.

	Assigned to 0	Assigned to 1
0	720	513
1	115	122

The confusion matrix shows that 115 instances are incorrectly clustered as "Yes" and 122 instances are correctly clustered as Yes. 720 instances are correctly clustered as No and 513 instances are wrongly clustered as No.

EM (Expectation Maximization)

EM is a clustering algorithm where the memberships to clusters are computed with probability. Classification probabilities are calculated in this algorithm. In Weka, for this algorithm, the Classes to cluster option is selected, and the attribute Attrition which assumes two values "Yes" or "No" is given. The confusion matrix, which is generated, can specify the correct and incorrect assignments.

The performance of the algorithm is as follows:

TABLE XVI: The correct and incorrect classification percentage and time

Algorithm	Correct classification	Incorrect classification	Time
EM	55.102%	44.898%	1.69s

TABLE XVII: Classification shown by confusion matrixa

	Assigned to 0	Assigned to 1
0	647	586
1	74	163

The confusion matrix shows that 74 instances are incorrectly clustered as "Yes" and 163 instances are correctly classified as Yes. 647 instances are correctly classified as No and 586 instances are wrongly classified as No.

IV RESUTS AND DISCUSSION

The classification and clustering methods used in Weka will help you to understand the characteristics of the employee, which will lead to the employee leaving the organization. The output of the J48 algorithm, which is a decision tree, will help you differentiate the characterics of employees who may leave the organization and who may not leave. Similarly, clustering algorithm creates clusters with similar characteristics leading to attrition or retention of employees. The paper has tried to understand the performance of two classification and two clustering algorithms in performing the attrition prediction of employees. The data set used is selected from Kaggle. The following observations are made after the analysis using Weka.

TABLE XVIII Comparison of two classification algorithms

Algorithm	Correct classification	Incorrect classification	Time
J48(10 fold Cross validation)	82.449%	17.551%	0.09s
J48(Percentage split as 70%)	82.7664%	17.2336%	0.14s
NaiveBayes(10 fold Cross validation)	78.8435%	21.1565%	0.03s
NaiveBayes(Percentage split as 70%)	80.9524%	19.0476%	0.03s

Chart 2 showing time taken by different algorithms

As far as accuracy is concerned J48 has shown a better performance. But it has taken more time than Naïve Bayes. NaiveBayes is showing a slightly better performance than cross validation when a percentage split is used.

TABLE XIX Comparison of two clustering algorithms

Algorithm	Correct classification	Incorrect classification	Time
K-Means	57.2789%	42.7211%	0.17s
EM	55.102%	44.898%	1.69s

Considering both time and accuracy K-Means has exhibited a better performance. The Classes to cluster option is selected in both cases, and attrition parameter is given as the class for clustering.

V CONCLUSION

Employee attrition is a major concern of many organizations now as qualified hands increase as do opportunites. The machine learning algorithms can be used for predicting the chances of attrition. This can be used for taking preventive measures. A comparison of various algorithms is made on a selected dataset to understand the accuracy of each. Further research can be conducted with the same data set using python, and the result can be compared.

REFERENCES

[1] D.Alao, A.B. Adeyemo, Analysing Employee Attrition using decision tree algorithms*, Computing, Information Systems & Development Informatics Vol. 4 No. 1 March, 2013*

[2] Al-Radaideh, Q.A, Al-Nagi, E (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, International Journal of Advanced Computer Science and Applications

[3] Hamidah J, AbdulRazak H, and Zulaiha A. O (2011). Towards Applying Data Mining Techniques for Talent Managements, *2009*
International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, *IACSIT Press.*

[4] Jayanthi, R., Goyal, D.P., Ahson, S.I. (2008) Data Mining Techniques for Better Decisions in Human Resource Management Systems.
International Journal of Business Information Systems, 3(5) 464 – 48

[5]Allen D. G (2008), Retaining Talent: A Guide to Analyzing and Managing Employee Turnover, *SHRM Foundation's Effective*
Practice Guidelines Series, SHRM Foundation.

[6]Nagadevara, V, Srinivasan, V & Valk, R (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours:
Application of Data Mining Techniques, *Research and Practice in Human Resource Management*, 16(2), 81-99.

[7] Allen, M.W, Armstrong, D.J, Reid, M.F, Riemenschneider, C.K (2009). "IT Employee Retention: Employee Expectations and Workplace Environments", SIGMIS-CPR'09, May 2009, Limerick, Ireland.

[8] Jantan, H., Hamdan, A.R. and Othman, Z.A. (2010b). "Human Talent Prediction in HRM using C4.5 Classification Algorithm", International Journal on Computer Science and Engineering, 2(08-2010), pp. 2526-2534.

[9] Witten I. Frank E., and Hall M. (2011). Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition, Morgan Kaufmann Publishers.

[10] Valle, M.A., Varas, S, Ruz, G.A, (2012). Job performance prediction in a call center using a Naive Bayes classifier, Expert Systems with Applications

[11]Lavrac, N (1999). "Selected Techniques for Data Mining in Medicine", Artificial Intelligence in Medicine.

[12]https://cs.ccsu.edu/~markov/ccsu_courses/datamining-ex3.html

[13]https://hackernoon.com/a-machine-learning-approach-to-ibm-employee-attrition-and-performance-b5d87c5e2415

[14]https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm