International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 24959-24978

http://www.ripublication.com

Reimagining Employee Connects To Build an Attrition Predictor Model

K. M. Suceendran[1], Saravanan Ramachandran[2], Divya Ananthram[3], Dr. S. Poonkuzhali[4], R. Kishore Kumar[5], Dr. K. Sarukesi[6]

[1]Research Scholar, Hindustan University & Senior Consultant,
Resource Management Group, Tata Consultancy Services (TCS), Chennai.
[2]Consultant, Delivery Excellence Group, TCS, Chennai
[3]Business Analyst, Resource Management Group, TCS, Chennai.
[4]Professor, [5]Assistant Professor, Dept. of IT,
Rajalakshmi Engineering College, Chennai.
[6]Dean, Kamaraj College of Engineering & Technology, Virudhunagar.
TAMIL NADU, INDIA.
sucee.k@tcs.com[1], saravan.r@tcs.com[2], divya.ananthram@tcs.com[3],
poonkuzhali.s@rajalakshmi.edu.in[4], rskishorekumar@gmail.com[5],
profsaru@gmail.com[6]

Abstract

Employee Engagement is a strong factor in driving the growth of a business or an organization. Employees who are committed to their respective organizations and are constantly motivated to work, contribute effectively to the growth of the organization's business. On the other hand, disengaged employees are capable of digressing from organizational goals and influencing the others around too, impacting efficiency in all quarters. In addition to of monetary benefits, there are other psychological and situational factors that influence an employee's level of engagement. Effective communication between employer and employee is one of the strong non-monetary factors impacting employee engagement. This paper discusses how we can build a predictor model, which can profile attrition-likely employees and cluster them based on causative factors. In this paper, various classifier algorithms of different categories are used to create and classify specific employee clusters. Exit interview data are classified into distinct categories based on the reasons cited for attrition. Employees in specific clusters are profiled and specific attributes identified from the organization memory related to these employees. Some of the systems considered for our analysis include leave history, grievance management, and relocation requests.

Trend and pattern from these systems can serve as precursors or indicators to the diminishing level of employee engagement. In this paper, we have analysed the organizational memory of employees to identify attrition-likely employees. We have used various classifier algorithms such as Zero R, One R, Decision Table, IB1, ID3, J48, Naïve Bayes, Bayesian Network, K Star, IBK, Random tree and Random Forest. These algorithms are evaluated on the basis of precision, recall, F-measure and ROC measure. The digital persona thus formed from the clusters can be applied to the entire workforce to identify attrition-likely associates. Additionally, an appropriate engagement strategy can be formulated to suit the possible behaviours based on the classifier algorithm.

Index Terms— Attrition, Classification, Digital Profile, Organizational Memory, Tacit Knowledge

I. INTRODUCTION

Employee connect is a critical attribute of the relationship between an employee and the enterprise. The research on employer-employee relationship is not new and the term ’employee engagement’ was defined in the early 1990s. William Kahn provided the first formal definition of employee engagement as “the harnessing of organization members’ selves to their work roles; in engagement, people employ and express themselves physically, cognitively, and emotionally during role performances”[1]. Over the years, the emphasis and perspective on this relationship have undergone many changes: In the 70’s it started with ’satisfaction’ which shifted to commitment’ in the 80’s. In the 90s, the process of engaging employees became the focus of HR practitioners, particularly in the knowledge enterprises [2]. The imperative of demonstrating the value of HR gave impetus to many studies to correlate employee engagement with tangible business outcomes.

Many organizations with large workforce comprising knowledge workers are confronted with the challenge of effectively leveraging employee engagement. Apart from monetary benefits and incentives, there are other psychological and situational factors that affect an employee’s level of engagement [3]. A recent study covering 142 countries indicates that only 13 percent of employees across the globe are engaged at work. This translates into only about one in eight employees is actually committed to his job [4]. This study from Gallup reveals that most employees, 63 percent, are not engaged and 24 percent are actively disengaged. Furthermore, this study also speaks of the employee engagement levels with respect to the Indian scenario. The study says only 9% of employees are engaged, while 31% are actively disengaged. However there is large variation in engagement levels when comparing Indian employees of different job types and education background. Approximately 17- 18% of people holding supervisorial, managerial, professional or sales related roles are engaged in their work [4]. Effective employee engagement depends on the successful connect between employees and organizational representatives including supervisors, senior leaders, HR personnel [5]. Consistent and effective communication with the employees is imperative for employee engagement and retention [6]. Challenges exist in the form of nature and scale of workforce, type of connects, degree of subjectivity inherent in the process and more importantly, not leveraging the organizational memory about the employee involved in the connect.

Employee Connects refers to various situations where an employee has an opportunity to connect and discuss with an HR personnel or a supervisor. In this paper we have considered the exit interview data, grievance data, relocation data and long leave data as our connects. Exit interviews and discussion with the HR as part of separation process are done so as to understand the reasons of separation of the associate from the company. These connects being the last in an associate's tenure with the company will bring out opinions and feedback from the associate and the reasons will be useful for analysts to study trends of attrition and arrest the same effectively [7]. Similarly relocation, long leave and grievance related discussions with the employee enable tacit knowledge to be generated with respect to the employee.

In this paper, employee connects details from a leading Information Technology organization in India are taken as input data. The data is used for analyzing the various classification techniques using Waikato Environment for Knowledge Analysis (WEKA) data mining tool [8]. In this evaluation process, different features are considered for choosing best algorithm which classifies the employees into different categories of profiles, according to the reason for resignation, leave, relocation or grievance stated.

II. RELATED WORKS

Alao D. & Adeyemo A. B. analyzed the role of Decision tree algorithms in their research on employee attrition. Employees' demographic data and work related data were used to classify employee details into attrition classes [1]. WEKA tool was used to generate decision tree models and rule-sets. Alao et al, used the classifiers C4.5 (J48), REPTree and CART (SimpleCart) decision tree algorithms. Attribute importance analysis was used to determine the significance of attributes. F-measure and the AUC probability tree learning measures were used as evaluation metrics for the classifier models generated [9]. The paper explained that decision trees are more flexible data modeling techniques than neural networks. Missing data values are managed by decision tree algorithms and models can be built with missing values, whereas in neural network and regression models, missing values need to be inserted.

Jayanthi et al (2008) presented the role of data mining in Human Resource Management Systems (HRMS). The paper explains about data mining techniques and the ability of these algorithms to improve decision making processes in HRMS, so as to sustain competitive advantage of the organization among competitors [10]. Hamidah et al suggested that in future works, attribute reduction should be conducted to identify relevant attributes for each of the factor [11].

Chughtai, in his paper studied the mediating role of employee engagement between HR practices and employee turnover intentions in a telecom sector in Pakistan, with sample data from two leading telecom companies in the region. The results highlighted a definite relationship between HR practices, employee engagement and turnover intentions, indicating effective use of basic HR practices leads to low level of turnover intentions, high level of employee engagement, and that relationship between HR practices and turnover intentions gets mediated by employee engagement [12].

Hsin Yun Chang, in the study on Employee Turnover, explains about feature selection, also known as feature subset selection. Feature selection is a common data pre-processing method. Yun Chang uses a mixed feature subset selection method in the study combining Taguchi method and Nearest Neighbour Classification rules to select and analyze factors and find best predictor of employee attrition [13]

S.Poonkuzhali et al. analyzed various classification algorithm for predicting efficient classifier for T53 Mutants [14]. R.Kishore Kumar et al. performed comparative analysis for predicting best classifiers for emails spam [15].

Lillian T Eby et al portrayed in their article the attitudinal changes in employees, post involuntary relocation and voluntary relocation of places of work and its effect on turnover intentions. The predictors used to analyze post-move attitudinal changes and intentions, were, the type of job change (lateral equivalent/ lower level of job, or promotion) and the type of relocation decision (voluntary relocation or involuntary relocation). Employees who relocated to a lateral equivalent job or a lower position felt less organizational support and higher intentions to leave than those who relocated for a promotion. Similarly, those employees taking a downward move or involuntary relocations felt lower organization support was extended towards them and felt stronger intentions to quit than voluntary relocators [16].

Dalton in the article 'Employee Transfer and Employee Turnover: A Theoretical and Practical Disconnect?' discusses the continued interest of researchers in the topic of employee turnover. However the author also notes that there is very little empirical attention to the topic of employee transfers and the organizational or personal consequences it may have. The paper discusses that employee relocation and transfer underline the flexibility in the organizations and also reflect employee development, which is positive [17].

N S Forster in the article Employee Job Mobility and Relocation lends support to the contention that management of mobility and relocation of employees is ad hoc and a kneejerk on the management. This article emphasizes that sudden transfers/relocations are ad hoc and a kneejerk to the management of the organization. The article concludes by outlining future goals of the research project in this area [18]. Thus planning such transfers/relocations and identifying sample digital profiles which are relocation-likely will be useful for workforce management and mitigating randomness/unstructuredness.

Hamidah et al in their paper applied classifier algorithms to existing employee performance data so as to predict employee future performance and their career development. The paper emphasized that performance patterns can be discovered from the existing data records. They have used selected classification techniques to propose the appropriate technique for the dataset. An experiment was conducted to understand the feasibility of the classification techniques using the employee's sample performance data. The result of the experiment showed that all of the classifiers had moderate accuracy of more than 70%. The C4.5/J4.8 classifier was selected as the most efficient classifier for the dataset. [19]

Kalaivani et al, in their paper studied the various classification techniques to classify the employees based on performance and predict their future performance. The objective of their study was human talent prediction. Among various classification algorithms such as Decision trees, neural networks, genetic algorithms, support vector machines, rough set theory, fuzzy set approach etc... this research was conducted by applying decision tree classification algorithms to employee's sample performance data. C4.5, Bagging and Rotation Forest algorithms were used and experiments were conducted with the data collected from an institution and implemented in WEKA tool. As a result the experiment yielded best results from the Rotation forest algorithm with training data set producing 100 % accurate result and the cross validation producing 51.46% accurate result [20].

Tamizharasi et al, in their paper, compared three classification algorithms. They were K- Nearest Neighbour classifier, Decision tree and Bayesian network algorithms. The authors concluded that all decision tree algorithms had lesser error rate than KNN and Bayesian algorithm and also deemed it as an easier algorithm to comprehend. The authors also suggested that efficiency of KNN can be improved by increasing the number of data sets and as for Bayesian algorithm classifier the attributes number can be increased. The authors mentioned that results were validated by a twenty four month data analysis conducted on mock basis [21].

Maria Jose Sousa in her paper analysed the impacts of knowledge sharing processes on organizational practices. The research was qualitative and quantitative involving document analysis on organizational evidences, semi structured group interviews to identify forms of knowledge sharing and collaboration among co-workers, and questionnaires filled in by workers. The paper concluded that knowledge being a crucial element for an organization, to boost innovation and implementation of new practices, communication among employees, and knowledge sharing is highly imperative. In the organization's context, the routines for creating and sharing knowledge were identified as suggestion boxes, openness to communicate and make suggestions to the managers, workshops for employees to participate, and several transversal projects of improvement, quality and maintenance [22].

III. NEED FOR EMBEDDING K-ELEMENTS AND CREATING DIGITAL PERSONA

With many of the non-core HR functions getting progressively digitized or outsourced, the no. of HR personnel actively involved in employee engagement is few in comparison to the size of the workforce. Knowledge-intensive organizations such as IT companies have one HR associate for every 1000 employees. With such a skewed ratio, HR associates meet only a portion of the total associates, the number in accordance with their performance targets. Associates chosen by HR associates is mostly arbitrary and by convenience. There is no definitive process which is derived from the business imperatives or an enabling system which presents HR persons with a complete digital profile.

In our earlier publication "Embedding K-Elements in HR Processes", it was highlighted that data obtained from HR processes is not consolidated or seldom used in real time [23]. Our other paper on identification of mid-level leadership in organizations, discussed on leveraging organizational knowledge and collaborative tacit knowledge to identify and develop leadership pipeline [24]. Though HR increasingly plays the role of a strategic partner and contributes also to operational management, there is often no formal framework for capturing, consolidating, and continually refining tacit data arising from HR touch points or connect sessions. Major connect sessions are mentoring, career counseling, staffing, grievance redressal sessions, relocation or long leave requests and exit interviews. At each of these interactions, pertinent tacit details about the employee are elicited and used in the context of the process. These tacit details are mostly 'soft data' and are subjective and time-specific in nature.

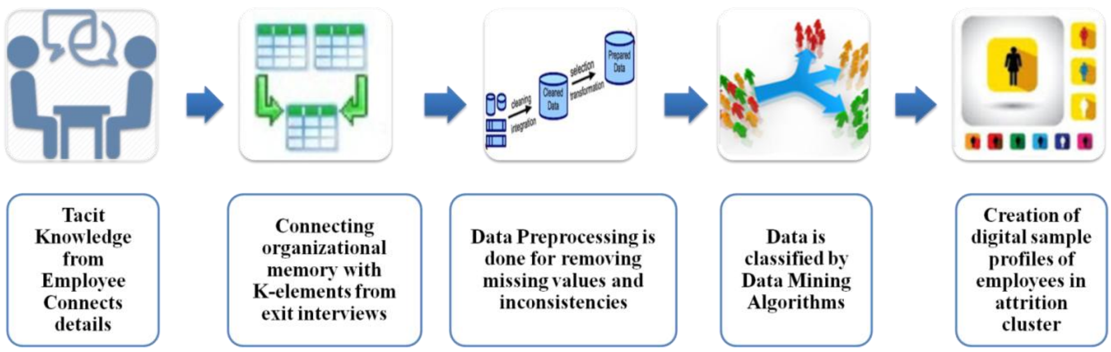

Fig.1. K Elements Framework in Employee Connects

The paper proposes to select set of associates in line with a specific business area in order to identify sample profiles for the business area of predicting associates for attrition. In extension to the exit interview details analyzed in the previous paper [25], this extended paper studies the tacit details obtained from relocation, grievance and long leave details, which are factors of employee turnover, to classify the employee details and derive sample profiles of attrition likely employees.

The pre-processing and classification details are described in proposed framework in subsequent sections.

IV. KNOWLEDGE SOURCES

Employee data related to relocation, long leave, and exit interview and grievance discussion are collected from a leading IT organization in India.

- Exit details consist of 2592 records with 9 attributes after pre-processing and feature selection.
- Relocation details consist of 503 records with 8 attributes after pre-processing and feature selection.

- Long Leave details consist of 2810 records with 10 attributes after pre-processing and feature selection.
- Grievance details consist of 7009 records with 9 attributes after pre-processing and feature selection.

These details are loaded into the WEKA data mining tool for classification after pre-processing and feature selection. The complete set of attributes with more details is given in the appendix. Table 1 provides the source-wise grouping of attributes, which were used for classification after pre-processing and feature selection.

TABLE I KNOWLEDGE SOURCES ATTRIBUTE LIST

Knowledge Sources	**Attribute List after pre-processing and feature selection**
Grievances Data	A2, A3, A4,A5,A11, A12, A16, A17, A31
Long Leave Data	A2, A3, A4, A5, A11, A12, A16, A17, A25, A32
Transfer Data	A2, A3, A5, A11, A12, A16, A17, A33
Exit Interview Data	A2, A3, A5, A9, A11, A12, A16, A17, A34

V. PROPOSED SYSTEM FRAMEWORK

The architecture design of the proposed approach is depicted in figure 2. Initially the details such as employee exit interviews, grievance details, long leave request details, transfer request details are obtained from the HR portal. These data sets are pre-processed and feature selection is applied before classification, to get the subset of input attributes which were relevant and most contributing for classification as well as for increasing the predictive accuracy. Employee records from each data set (Exit / Grievance/ Leave / Transfer) are classified using different classifiers.

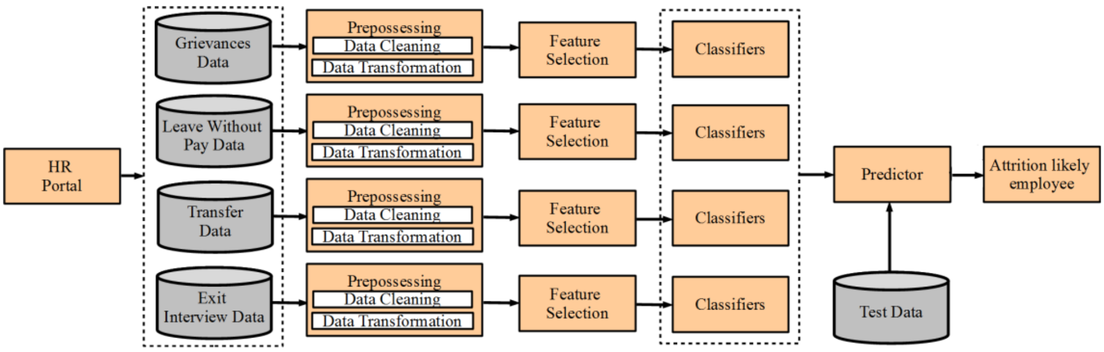

Fig.2. Attrition Predictor Model Framework

The result from each data set is obtained and the best classifiers are determined. The best classifiers are chosen for each of the data set and analysed for determining patterns and sample rules are generated from each classifier output. These sample rules are used to build the predictor engine, which when provided with test data would apply the determined sample rules to identify profiles of employees who are likely to leave the organization due to grievance, denial of leave or relocation requests or other causes of exit.

VI. Results & Discussion

The employee exit details, relocation details, long leave details and grievance details, after preprocessing and feature selection of relevant attributes are loaded into the WEKA data mining tool. Different classifiers are applied to the data sets separately. The classification algorithms applied were namely, Zero R, One R, Decision Table, IB1, ID3, J48, Naïve Bayes, Bayesian Network, K Star, IBK, Random tree and Random Forest.

The results of these classification algorithms are depicted in Table 2, 3, 4, and 5 below for each of the data sets. The results portrayed exhibit accuracy, precision, recall, F-measure and ROC value. The performance of all these classifiers is analyzed based on the accuracy to predict the best classifier for each data type.

For long leave data set, instance based classifiers IBK and ensemble classifiers Random Tree and Random forest with an accuracy of above 92% are identified as best classifiers and results projected in table 2.

Table II Classifier Results for Long Leave Data

Classifiers	Accur	Preci	Reca	F_M	ROC
Zero R	44.64	0.19	0.44	0.27	0.5
One R	60.62	0.557	0.606	0.528	0.677
Naïve Bayes	60.34	0.537	0.603	0.536	0.81
IBK	92.48	0.926	0.925	0.925	0.994
K Star	75.56	0.82	0.756	0.721	0.946
J48	68.61	0.69	0.686	0.626	0.84
Random Forest	92.45	0.926	0.925	0.922	0.99
Random Tree	92.49	0.926	0.925	0.925	0.994

Legend	
Accur	Accuracy
Preci	Precision
Reca	Recall
F_M	F – Measure
ROC	Receiver Operating Characteristic

For relocation details, instance based classifiers IB1, IBK, KSTAR, ensemble classifiers Random tree, Random Forest and tree based classifier ID3 are identified as best classifiers, with an accuracy of above 94%, and results projected in table 3.

TABLE III CLASSIFIER RESULTS FOR RELOCATION DATA

Classifiers	Accur	Preci	Reca	F_M	ROC
One R	42.92	0.551	0.429	0.392	0.612
ID3	94.82	0.95	0.94	0.948	0.999
J48	60.53	0.634	0.605	0.593	0.871
Random Forest	94.82	0.949	0.948	0.948	0.998
Random Tree	94.83	0.95	0.948	0.948	0.999
Decision Table	40.80	0.452	0.408	0.386	0.705
IB1	94.25	0.943	0.943	0.943	0.964
IBK	94.82	0.95	0.948	0.948	0.999
K Star	94.63	0.947	0.946	0.946	0.998
Bayes Net	57.08	0.578	0.571	0.558	0.856

For Grievance Data, instance based classifiers IBK, IB1, K Star provide results with 100% accuracy and ensemble classifiers Random Tree, Random Forest, and tree based J48 classifier, provide results with accuracy above 98% with results depicted in table 4.

TABLE IV CLASSIFIER RESULTS FOR GRIEVANCE DATA

Classifiers	Accur	Preci	Reca	F_M	ROC
Zero R	23.96	0.05	0.2	0.09	0.5
IB1	100	1	1	1	1
IBK	100	1	1	1	1
Random Forest	99.74	0.997	0.997	0.997	0.997
Random Tree	100	1	1	1	1
J48	98.316	0.983	0.983	0.983	1
ID3	100	1	1	1	1
Decision Table	49.22	0.78	0.49	0.49	0.78
K Star	100	1	1	1	1

For Exit details, IBK an instance based classifier and Random tree an ensemble classifiers have been identified as a best classifier for this Employee exit interview dataset as they classified with an accuracy of 97.74%. The results are depicted in table 5.

TABLE V CLASSIFIER RESULTS FOR EXIT INTERVIEW DATA

Classifiers	Accur	Preci	Reca	F_M	ROC
ID3	41.21	0.52	0.41	0.35	0.82
J48	52.10	0.593	0.521	0.47	0.845
Naïve Bayes	40.81	0.406	0.428	0.349	0.808
Bayes Network	45.88	0.456	0.459	0.398	0.83
K STAR	97.43	0.976	0.974	0.974	1
IBK	97.74	0.978	0.977	0.977	1
Random Tree	97.74	0.978	0.977	0.977	1
Random Forest	95.91	0.96	0.959	0.959	0.998

Once the best classifier algorithms are identified, and applied on to the data, the attributes of employees falling under different categories and groups can be studied. The employee exit details are analyzed on the basis of their reason furnished for leaving the organization and the classifier algorithm categorizes the profiles into relevant groups according to their reasons. Similarly the classifiers categorize the employee details in relocation and long leave data according to the reasons and constraints furnished by the employee for obtaining relocation/long-leave. Employee data in grievance data set is classified by the algorithms, with respect to the subject of grievance filed by the employee.

The sample profiles of such employees, along with their different attributes which cause the reasons for exit, leave, relocation, and grievance can be deduced and obtained from the best classifier results. These sample rules are used to build a predictor engine which will collectively consist of all sample profiles, their attributes which affect the reason for leave/relocation/resignation or grievance. Few of the sample rules obtained from the best classifiers are listed below.

Sample Rules obtained from Long Leave Data

```
No of Days Leave <= 371
| No of Days Leave <= 30
| | No of Days Leave <= 27: Marriage
| | No of Days Leave > 27: Maternity
| No of Days Leave > 30: Childcare
No of Days Leave > 371
| No of Days Leave <= 579
| | Marital Status = MARRIED
| | | Gender = Male: Education
| | | Gender = Female: Childcare
| | Marital Status = SINGLE: Education
| No of Days Leave > 579: Education
```

Sample Rules obtained from Grievance Data

```
Total Experience >= 20
| Last Performance Rating = 3 || 2 : Appraisal

Total Experience >4.5
| Age >25
| | Rating on confirmation =4
| | | Marital Status = Married : Overseas Deputation Issues
| | | Marital Status = Single : Compensation Issues
```

Sample Rules from Exit Data

```
Gender = Female
| Role = Junior Developer1
| | Total Experience <= 3
| | | Grade = Grade1
| | | | Last Performance Rating = 4.0: Marriage
| | | | Last Performance Rating = 2.0: Academic pursuit
| | | | Last Performance Rating = 3.0
| | | | | Rating on confirmation = 3.0
| | | | | | Total Experience <= 2
| | | | | | | Age <= 25: Marriage
| | | | | | | Age > 25: Better career opportunities
```

Sample Rules from Relocation Data

```
Gender = Male
| Last Band optional <= 2: Project Confirmation
| Last Band optional > 2
| | Designation = Midlevel Manager1
| | | Total ExperienceYrs<= 4.45: New Project
| | | Total ExperienceYrs> 4.45: Joining Family
| | Designation = Junior Executive1
| | | Age <= 23: Joining Family
| | | Age > 23
| | | | Total ExperienceYrs<= 1.93: Family Health Issues
| | | | Total ExperienceYrs> 1.93: New Project
Gender = Female
| Designation = Junior Executive2
| | Total ExperienceYrs<= 4.48
| | | Age <= 25: Medical
| | | Age > 25: Marriage
| | Total ExperienceYrs> 4.48: New Project
```

| Designation = MidLevel Manager1
| | Age <= 29: Marriage
| | Age > 29: Joining Family
| Designation = JuniorExecutive1: Marriage

Thus, the digital personae of employees most likely to exit the company, due to these reasons can be deduced and formed. These digital personae can be used as a reference to target the next set of employees to conduct employee connect sessions on a priority. Employees whose digital personas match with the attrition-likely profile can be identified and connected with for proactive engagement. Thus randomness and unstructuredness of the Employee Connects session can be mitigated and potential attrition can be arrested considerably.

A. *Performances of the Best Classifiers*

Parameters for performance of classifiers

a) Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method. The accuracy of all the classifiers used for classifying employee details in the results tables.

$$Acuracy = \frac{No. of\ correctly\ classified\ samples}{Total\ no. of\ samples\ in\ the\ class} \tag{1}$$

b) Recall

Recall of the classifier was defined as the percentage of errors correctly predicted out of all the errors that actually occurred.

$$Recall = \frac{True\ Positive}{True\ Positive\ + False\ Negative} \tag{2}$$

c) Precision

Precision of the classifier was defined as the percentage of the actual errors among all the encounters that were classified as errors.

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive} \tag{3}$$

d) F-Measure

F-Measure is defined as the measure of a test's accuracy and can be interpreted as a weighted average of the precision and recall.

$$F\_Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

e) ROC

Receiver Operating Characteristic curve (or ROC curve) provides a graphical plot of the true positive rate against the false positive rate for the different possible cut points of a test. Since the curve follows the left-hand border and then the top border of the ROC space, the classifier is more accurate. The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to classifier's expectation.

1) Instance Based Classifier: IBK (K-Nearest Neighbour):

IBK is a K-Nearest Neighbour classifier that uses that same distance metric for classification. It is an instance based classifier in which the class of the test instance is based on the class of those training instances similar to it as determined by the similarity function based on distance. This algorithm uses normalized distances for all attributes. An object is classified by a majority vote of its neighbour, with the object being assigned to the class most common among its K nearest neighbours.

Pseudo for K-Nearest Neighbour Classification Algorithm

1: Determine the parameter K.
2: Compute the distance between the query instance and all the training attributes using Euclidean distance algorithm.

$$D(q,p) = \sqrt{\sum_{i=0}^{n} (q_i - p_i)^2} \tag{5}$$

3: Sort the training records based on the distance.
4: Find the nearest neighbour based on the Kth minimum distance.
5: Get all the Categories of the training data for the sorted value which fall under K.
6: Predict the measured value using the majority of nearest neighbours.

Figure 3 projects the performance of the IBK classifier in terms of precision, recall, f-measure and ROC value for the different employee connects datasets.

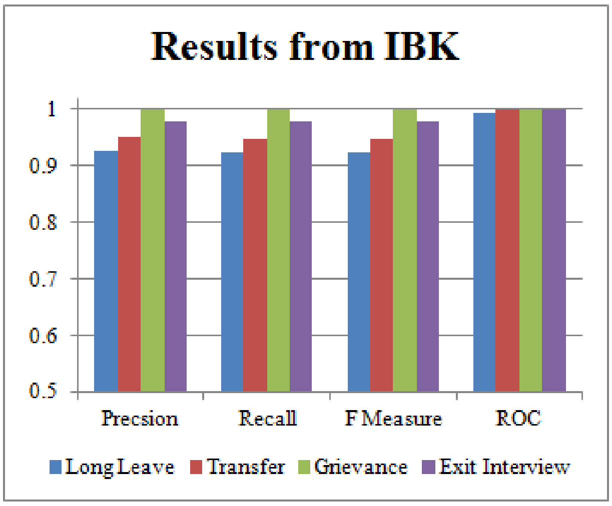

Fig.3. Performance of IBK Classifier

2) Ensemble Classifier: Random Tree

Random tree is an ensemble classifier that consists of many class of many decision trees and outputs the class that is the mode of class output by individual trees. Each tree is fully grown and not pruned. It runs efficiently on large data bases. For predicting a new record is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

Pseudo for Random Tree Algorithm

1: For 1 to N do (N -Number of employee records in HRMS dataset D)
2: Select 'm' input attributes at random from the 'n' total number of attributes in dataset D
3: Find the best spilt point among the 'm' attributes according to a purity measure based on Gini index.

$$G = \frac{2\sum_{i=1}^{n} i y_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n} \tag{6}$$

4: Spilt the node into two different nodes on the basis of split point.
5: Repeat the above steps for different set of records to construct possible decision trees.
6: Ensemble all constructed trees into single forest for classifying stereotyping digital profiles of HRMS data.

Figure 4 projects the performance of the Random tree classifier in terms of precision, recall, f-measure and ROC value for the different employee connects datasets.

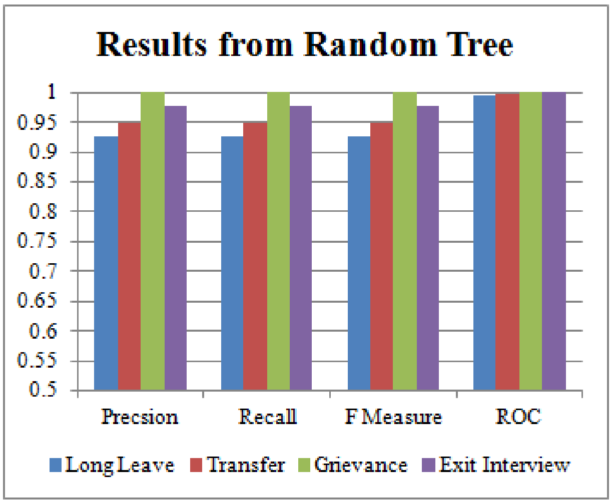

Fig.4. Performance of Random Tree Classifier

3) Tree Based Classifier: ID3

In ID3 decision tree, each node corresponds to splitting attribute. It uses information gain to determine the splitting attribute. The attribute with the highest information gain is taken as the splitting attribute. Information gain is the difference between the amount of information needed to make a correct prediction before and after splitting. Information gain can also be defined as the different between the entropy of the original segment and the accumulated entropies of the resulting split segments. Entropy is the measure of disorder found in the data. ID3 can handle high-cardinality predictor variables. A high-cardinality predictor is a variable which has different possible values thus having different possible ways of performing a split.

Pseudo for ID3 Classification Algorithm - Ref [26]

1: Extract the input training dataset (TD).
2: Let IA be input attributes & TA be the Target Attributes.
3: For a base class construct the tree as follows:
If TD is empty then
Return a single node with the value Failure
If all instance from TD have the same class
value for TA then
Return the single node with that class value.
If IA is empty then
Return a single node with the most frequent
value of TA in TD.
Else
4: Compute the information gain for all IA relative
to TD based on Entropy measure.

$$Info(D) = -\sum_{i=1}^{m}(p_i log_2(p_i)) \tag{7}$$

$$E(A) = -\sum_{j=1}^{v} \left(|D_j| / |D| * Info(D) \right) \tag{8}$$

Where |Dj|/|D| is weight of the j th partition

Gain (A) = Info (D-E (A)) (9)

Where pi is the probability that arbitrary instance in TD

E (A) is the entropy of the attribute A

Gain (A) is the information gain of attribute A.

Info (D) expected information needed to classify an instance in TD.

5: Let X be the split attribute with largest information gain among IA.

6: Partition the TD according to the value of X

7: Return a tree with the root node labelled X and arcs labelled X1,X2…….Xm.

8: Repeat step 4 to 6 until it contains instance belonging to the same category.

Figure 5 projects the performance of the ID3 classifier in terms of precision, recall, f-measure and ROC value for the different employee connects datasets.

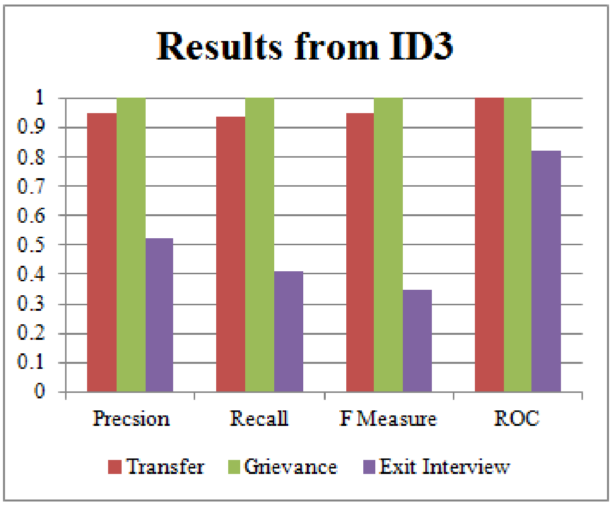

Fig.5. Performance of ID3 Classifier

VII. CONCLUSIONS

Many reasons cited during resignation revolved around relocation, family issues, leave unavailability, better career opportunities, and grievances with respect to work, compensation and appraisal. Thus employee grievance, relocation and long leave details were used in addition with exit interview details to identify sample rules to predict attrition likely profiles.

The employee details have thus been classified by different classifier algorithms successfully. These best classifiers have categorized the employer profiles into relevant groups based on their reason furnished for grievance/ relocation/ leave or resignation, and sample rules have been obtained from the classifications. These rules can be used to build the predictor engine, which when supplied with test data or new employee data, can enable the predictor engine to deduce and identify attrition-likely employees.

The following are the research outcomes:

- Design of the K-Elements framework in Employee Connects process.
- Evolving of the best classifier algorithm for different employee data.
- Performance analysis on various metrics such as Precision, Recall, F-Measure and ROC for the best classifiers.
- Generation of sample rules which create digital personae of attrition likely employees. These sample rules are used for building the predictor engine to identify attrition-likely associates.
- Hence, employee connects sessions can be conducted in a guided fashion and attrition – likely associates can be approached on priority.

Further, rules for profiling or creating digital persona can be applied to the entire workforce to predict attrition-likely associates. Additionally, an appropriate engagement strategy can be formulated to suit the possible reasons based on the classifier algorithm. In future, this approach can be iteratively improved and scaled up as and when better employee connects data is generated and recorded.

REFERENCES

[1] Kahn, William A, "Psychological Conditions of Personal Engagement and Disengagement at Work", Academy of Management Journal, Vol.33, No.4, 1990, pp. 692-724.

[2] Tom O'Byrne, "History of employee engagement - from satisfaction to sustainability", HR Zone, 2013, Available: http://www.hrzone.com/feature/people/historyemployee-engagement-satisfactionsustainability/141048

[3] Sylvia Vorhauser-Smith, "How the Best Places to Work are Nailing Employee Engagement", August 2013, Available:http://www.forbes.com/sites/sylviavorhausersmith/2013/08/14/how-the-best-places-to-work-arenailing-employee-engagement/

[4] Jim Clifton "State of the Global Workplace", Gallup Inc, 2013, Available:http://ihrim.org/Pubonline/Wire/Dec13/GlobalWorkplaceReport_2013.pdf

[5] Kim Harrison, "Good Communication can hugely lift employee engagement", Cutting Edge PR, 2013
Available: http://www.cuttingedgepr.com/articles/goodcommunication-can-hugely-lift-employeeengagement.asp

[6] Thomas R Cutler, "Employee Engagement, Retention and Communication", Business Excellence, August 2014. Available: http://www.bus-ex.com/article/employeeengagement-retention-and-communication

[7] Susan M Heathfield, “4 Tips for Effective Exit Interviews”, Human Resources, Available:http://humanresources.about.com/od/whenemployment-ends/fl/4-Tips-for-Effective-Exit-Interviews.htm
[8] Weka Tool, Available:www.cs.waikato.ac.nz/ml/weka.
[9] Alao D. and Adeyemo A. B., “Analyzing Employee Attrition Using Decision Tree Algorithms”, Computing, Information Systems & Development Informatics, Vol. 4, No. 1, 2013.
[10] Jayanthi, R., Goyal, D.P., Ahson, S.I., “Data Mining Techniques for Better Decisions in Human Resource Management Systems”, International Journal of Business Information Systems, Vol. 3, No. 5, 2008, pp. 464-481.
[11] Hamidah J, AbdulRazak H, and Zulaiha A. O, “Towards Applying Data Mining Techniques for Talent Managements”, International Conference on Computer Engineering and Applications, IPCSIT Vol.2, IACSIT Press, 2009.
[12] Tauseef Ahmad Chughtai, “Role Of HR Practices In Turnover Intentions With The Mediating Effect Of Employee Engagement”, WSEAS Transactions On Business And Economics, Volume 10, Issue 2, April 2013, pp. 97 – 103
[13] Hsin-Yun Chang, “Employee Turnover: A Novel Prediction Solution with Effective Feature Selection”, WSEAS Transactions on Information Science and Applications, Vol. 6, No. 3, 2009, pp. 417-426.
[14] Poonkuzhali. S, GeethaRamani, Kishore Kumar.R, "Efficient Classifier for TP53 Mutants using Feature Relevance Analysis", Proceedings of International Multiconference on Engineers and Computer Scientist, Vol.1, 2012.
[15] Kishore Kumar. R, Poonkuzhali. G, Sudhakar. P, "Comparative Study on Email Spam Classifier using Data Mining Techniques", Proceedings of International Multiconference on Engineers and Computer Scientist, Vol.1, 2012.
[16] Lillian T. Eby and Jacquelyn S. Dematteo, “When the Type of Move Matters: Employee Outcomes Under Various Relocation Situations”, Journal of Organizational Behavior, Volume 21, Issue 6, September 2000, pp. 677–687.
[17] Dalton, Dan R., “Employee Transfer and Employee Turnover: A Theoretical and Practical Disconnect?”, Journal of Organizational Behavior, Volume 18(5), Sep 1997, Pg 411-413.
[18] N.S. Forster, “Employee Job Mobility And Relocation”, Personnel Review, Vol. 19, Issue. 6, 1990, pp. 18–24.
[19] HamidahJantan, MazidahPuteh, Abdul RazakHamdan and Zulaiha Ali Othman, “Applying Data Mining Classification Techniques for Employee’s Performance Prediction”, Knowledge Management International Conference, 2010, pp. 601-607.
[20] V.Kalaivani and M.Elamparithi, “An Efficient Classification Algorithms for Employee Performance Prediction”, International Journal of Research in Advent Technology, Vol.2, No.9, September 2014.
[21] Tamizharasi. K, Dr. UmaRani, “Employee Turnover Analysis with Application of Data Mining Methods”, International Journal of Computer Science and Information Technologies, Vol. 5, No. 1, 2014, pp. 562-566.

[22] Maria José Sousa, "Employees Perceptions About Knowledge Sharing Impacts On Organizational Practices", WSEAS Transactions On Business And Economics, Volume 11, 2014, pp. 718 – 724.

[23] Suceendran. K. M, Saravanan Ramachandran, Sarukesi K, "Embedding K-Elements in HR Processes", Advances in Engineering, Science and Management (ICAESM), 2012.

[24] Suceendran K M, Saravanan Ramachandran, Dhilip Kumar, Divya Ananthram, Dr. K Sarukesi and Sudhakar Pandiarajan, "Leveraging Organizational Memory to Build Talent Pipeline for Mid-level Leadership", World Scientific and Engineering Academy and Society (WSEAS), 9th International Conference on Business Administration (ICBA '15), February 2015.

[25] Suceendran K. M., R. Saravanan, Divya Ananthram, S. Poonkuzhali, R. Kishore Kumar and Dr.K.Sarukesi, "Applying Classifier Algorithms to Organizational Memory to Build an Attrition Predictor Model", World Scientific and Engineering Academy and Society (WSEAS), 9th International Conference on Computer Engineering and Applications (CEA '15), February 2015

[26] Poonkuzhali. S, R. Kishore Kumar and CiddarthViswanathan, "Law Reckoner for Indian Judiciary: An Android Application for Retrieving Law Information Using Data Mining Methods", Advanced Computer and Communication Engineering Technology, Springer International Publishing Switzerland, 2015, Chapter 55, Page 585-593.

APPENDIX

Table VI depicts the total number of attributes of employee connects data before pre-processing and feature selection.

TABLE VI TOTAL NO. OF ATTRIBUTES OF EMPLOYEE CONNECTS DATA.

Attribute No.	Description	Dataset			
A1	Employee No	G	L	R	E
A2	Age	G	L	R	E
A3	Gender	G	L	R	E
A4	Marital Status	G	L		
A5	Designation	G	L	R	E
A6	Qualification				E
A7	Specialisation				E
A8	Date of Joining	G	L	R	E
A9	Role			R	E
A10	Skill			R	E
A11	Grade	G	L	R	E
A12	Total Experience	G	L	R	E
A13	Employment Status		L		
A14	Business Unit		L	R	E
A15	Branch Name		L		E
A16	Last Performance Rating	G	L	R	E
A17	Rating on Confirmation	G	L	R	E
A18	Year of Request	G	L	R	E
A19	Grievance ID	G			
A20	Long Leave Request ID		L		
A21	Date of Request	G	L	R	E
A22	Leave Status		L		
A23	Start Date of Leave		L		
A24	End Date of Leave		L		
A25	No of Days Leave		L		
A26	From Branch			R	
A27	To Branch			R	
A28	Documents verified			R	
A29	HR contact in future location			R	
A30	Resignation Type				E
A31	Reason for Grievance	G			
A32	Reason for Long Leave		L		
A33	Reason for Relocation			R	
A34	Reason for Resignation				E

Legend

G	List of available Grievance data attribute
R	Relocation data attribute
L	Long Leave data attribute
E	Exit Data attribute