PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING

R Shiva Shankar,
Department of CSE,
SRKR Engineering College,
Bhimavaram, India

J Rajanikanth,
Department of CSE,
SRKR Engineering College,
Bhimavaram, India

V.V.Sivaramaraju
Department of CSE,
SRKR Engineering College,
Bhimavaram, India

K VSSR Murthy
Department of CSE,
SRKR Engineering College,
Bhimavaram, India

Abstract: **Now a day's Employee Attrition prediction become a major problem in the organizations. Employee Attrition is a big issue for the organizations specially when trained, technical and key employees leave for a better opportunity from the organization. This results in financial loss to replace a trained employee. Therefore, we use the current and past employee data to analyze the common reasons for employee attrition or attrition. For the prevention of employee attrition, we applied a well known classification methods, that is, Decision tree, Logistic Regression, SVM, KNN, Random Forest, Naive bayes methods on the human resource data. For this we implement feature selection method on the data and analysis the results to prevent employee attrition. This is helpful to companies to predict employee attrition, and also helpful to their economic growth by reducing their human resource cost**.

Keywords: Attrition, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, Naïve bayes

I. INTRODUCTION

An employee would choose to join or depart an organization depending on many causes i.e. work environment, work place, gender equity, pay equity and many other. The rest of the employees may think about personal reasons for instance relocation due to family, maternity, health, issues with the managers or co-workers in a team. Employee attrition is a major problem for the organizations particularly when trained, technical and key employees leave for best opportunities from the organizations. This finally results into monetary loss to substitute a trained employee. Consequently, we utilize the present and past employee data to assess the familiar issues for employee attrition. The employee attrition identification helps in predicting and resolving the issues of attrition. We can use this data to stop the attrition rate of the employees.

For this working we use some methodologies of data classification. Those methodologies are Decision Tree (it is tree structure that comprises a branches, root node and leaf nodes. every internal node indicates a test on an attribute, every branch indicates the result of a test, and every leaf node holds a class label), Naive Bayes (it is a classification methodology depending on Bayes Theorem. A Navie Bayes classifier presumes that the existence of a specific in a class is unrelated to the existence of any other feature.

For instance, a fruit may be measured to be an apple if it is red, round, and regarding 3 inches in diameter. Still if these features depend on each other or upon the presence of the rest of the features, all these properties autonomously contribute to the probability that this fruit is an apple) Logistic Regression(it is a statistical approach for assessing a dataset in which there are one or more autonomous variables that establish an outcome.

For instance hours of research enhanced then the probability of passing exams increases, Support Vector Machine (SVM) (it carry outs classification by identifying the hyperplane that entirely differentiates the vector into two non overlapping classes. The vectors that describe the hyperplane are the support vectors), K-Nearest Neighbour (KNN) (compare each value with the neighbour value), and RandomForest (builds the forest with a number of decision trees) methodologies on the Human Resources Employee Attrition dataset given by IBM. The dataset consists of 1470 records with 35 features containing categorical and numeric features. From this dataset we perform a data pre processing (if suppose any attribute value in the record includes any null value or indeterminate value then expel that whole record from the unique dataset and put that record into the training dataset, else if the record include perfect data with all characteristics then put that into the test dataset) technique to select the mainly significant characteristics of the dataset and divide total dataset into two secondary datasets. One is the dataset and the other one is training dataset. Test dataset consists of all the significant characteristics to depict employee attrition or employee attrition and training dataset include not required data. At the end we implemented the above described classification techniques on the test dataset with expelled number of characteristics to get the necessary results which is valuable to stop the employee attrition.

II. LITURATURE SURVEY:

K. Coussement and **D. vanden poel** worked on "**Integrating the voice of customers through call center email into a decision support system for attrition prediction**" [2]. In this research they established that adding unstructured, textual data into a conventional attrition identification. The outcome is raise performance in attrition identification analysis. This study supportive for marketing decision makers to improved recognize customer those have probability to attrition.

C.P. Wei and **I.T. Chiu** worked on **"Turning telecommunications call details to attrition prediction: a data mining approach"[3].** In this study, experimentally assess an attrition identification method that offers attritioning from subscriber contractual data and call pattern modifies mined from call details. This described method is capable of describing potential attritioners for contract level for particular prediction time period.

K. Coussement and **D. Van den Poel** worked on **"Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques" [4].** With the use of preserving customers, academics in addition to practitioners identify it crucial to design an attrition identification model that is as precise as possible. Comparison is prepared between two parameter-selection methods, necessary to perform support vector machines. Both methods are based on grid search and cross-validation.

J. Burez and **D. Van den Poel** worked on "**Handling class imbalance in customer attrition prediction**" [5]. In this research they explored that how to handle class inequity in attrition identification. For this study they use most accurate assessment metrics. That is AUC, lift. AUC and lift prove to be good assessment metrics to calculate accuracy. Result describe that tend to enhance prediction accuracy.

C.-F. Tsai and M.-Y. Chen, worked on **"Variable selection by association rules for customer attrition prediction of multimedia on demand" [6].** Data mining methods have been extensively applied to build customer attrition identification models, like neural networks and decision trees in the domain of mobile telecommunication. This study includes the pre-processing stage for choosing significant variables by association rules. Four assessment measures namely prediction accuracy, precision, recall, and F-measure, all of which have not been taken into account to examine the model performance.

B. huang, M. T. Kechadi and **B. Buckley** worked on "**customer attrition prediction in telecommunications**"[8] to identify employee attrition. During these days customer attrition identification became a main part of telecom sector. This study is helpful to estimate customer attrition behaviour to stop customer attrition. For this research they utilized roughest theory, a rule-based decision making method to mine rules from attrition detection. For this they utilized four rule algorithms those are Exhaustive, Genetic, Covering and LEM2. Rough set classification depends on genetic algorithm. Rule based technique for enhance performance. The outcome of this study can done by **B. huang, M. T. Kechadi** and identify customers from attrition.

V.V. Saradhi and **G.K. Palshikar** worked on "**Employee churn prediction**" [9] Employee attrition directly relates to the customer attrition but not similar to customer attrition. For this they study and compared some major Machine Learning methodologies to stop employee attrition. In this they carried out several methods for creating and comparing predictive employee attrition models. This work is helpful for building best employee prediction model.

R.Khare, D. Kaloya, C. K. Choudhary, and G. Gupta worked on "**Employee attrition risk assessment using logistic regression analysis**"[10]. This study utilized the logistic regression methods to stop employee attrition. In this the researchers gather demographic data of divided as well as present employees. This information made useful to produce risk equation, and make a cluster of the high risk employees. For this to be implemented the organization provide attention on that cluster of employees to stop them from attrition.

M. L. Kane- Sellers, worked on "**Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis"[11].** Diminishing, employee retention and voluntary turnover has been increased to the front position of HRD practitioners. It mainly concentrated on personal features, work characteristics, and human resource development (HRD) intrusion characteristics effecting employee voluntary turnover. The outcome recommended that training and development participation contributes more importantly to employee retention than the salary and job title promotions to the firm's capability to keep sales professionals.

X. Lin, F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang and G. Xu worked

on "**A support vector machine-recursive characteristic elimination feature selection method depending on artificial contrast variables and mutual data**" [12]. Support vector machine-recursive feature elimination (SVM-RFE) is a competent feature chosen method and has displays promising applications in the assessment of the metabolome data. SVM-RFE measures the weights of the characteristics in accordance to the support vectors, noise and uninformative variables in the high dimension data may affect the hyper-plane of the SVM learning model. Therefore we suggested mutual information (MI)-SVM-RFE technique which removes out noise and uninformative variables by means of artificial variables and MI, then performs SVM-RFE to choose the most of the discriminative characteristics.

III. PROPOSED SYSTEM:

Data set:

Data set is a collection of data. Most commonly a data set corresponds to the contents of a single database, where every column of the table represents a particular variable, and each row corresponds to a member of the dataset. For our project we take employee data from IBM which contains 1470 records and 35 fields including categorical and numeric features. Each record in the employee data set represents a single employee information and each field in the record represents a feature of that particular employee.

Data pre-processing:

From the IBM employee dataset we implement a feature selection method to select the most important features of the dataset and divide total dataset into two sub datasets. One is test dataset another one is training dataset. That is if suppose any feature value in the record contain any null value or undefined or irrelevant value then separate that entire record from the original dataset and place that record into training dataset, else if the record contain perfect data with all features then place that into test dataset. Test dataset contain all important features to predict employee attrition or employee attrition and training dataset contain irrelevant data.

Test dataset and training dataset:

Separating data into test datasets and training datasets is an important part of evaluating data mining models. By this separation of total data set into two data sets we can minimize the effects of data inconsistency and better understand the characteristics of the model. The test data set contains all the required data for data prediction and training data set contains all irrelevant data. Here we have 788 records in test dataset and 682 records in training dataset. We apply data classification and data prediction on the test dataset of 788 records.

Data classification techniques:

Data classification is the process of organizing data into categories for its most effective and efficient use. Data classification techniques are Decision tree, K nearest neighbour (KNN), Support vector machine (SVM), Logistic regression, Naive bayes.

Decision Tree: It is tree structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.

Naive Bayes: It is a classification technique based on Bayes Theorem. A Navie Bayes classifier assumes that the presence of a particular in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all these properties independently contribute to the probability that this fruit is an apple.

Logistic Regression: It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. For example hours of studying increases then the probability of passing exams increases.

Support Vector Machine (SVM): it performs classification by finding the hyperplane that completely separates the vector into two non overlapping classes. The vectors that define the hyperplane are the support vectors.

K-Nearest Neighbour (KNN): compare each value with the neighbour values or Nearest values. It is a non parametric method used for classification.
Here we apply classification techniques on the test data set and categorize the data into different departments. That is sales department, human resource department, research and development department. After that we apply performance analysis, education analysis, and salary analysis on all categorized departments data to predict the employee attrition by finding the best employees.

Performance analyzer: It is used to analyze the average performance of employees in each department. Here we have average performance in

sales department is 3.13677, average performance in research and development is 3.16233, and average performance in human resource department is 3.14285. From the average performances we find out the employees who have the highest performance than average performance and predict those employees from employee attrition.

Salary analyzer: It is used to analyze the employees who have high salary and who have low salary. If any of the employees getting low salary even though their performance is high then we identify those employees and prevent them from employee attrition by incrementing their salaries. The employees who have their salary lower than 6000 they belongs to low salary category and who have more than 6000 salary those employees belongs to high salary categories. Here we have 914 employees getting low salary and 556 employees getting high salary. In those 556 employees those getting high salary 231 are females and 325 are males.

Education analyzer: It is used to analyze the employees who have higher qualification and who have lower qualification. Here we divide the total employees in five categories according to their educational qualification. Those five categories are employees with single degree, employees with double degree, triple degree, employees with four degrees, and finally employees with five degree qualification. Finally we find out the average performance, average job satisfaction and average monthly pay to the employees of each category.

Predicted data: By this total analysis we find out the best employees and we prevent those employees from employee attrition by providing the all requirements.

IV. IMPLEMENTATION

ALGORITHM FOR SVM:

Input: Taking some of the records as a sample data and also we take averages of Performance Rating, Monthly Income and Training Times Last Year.

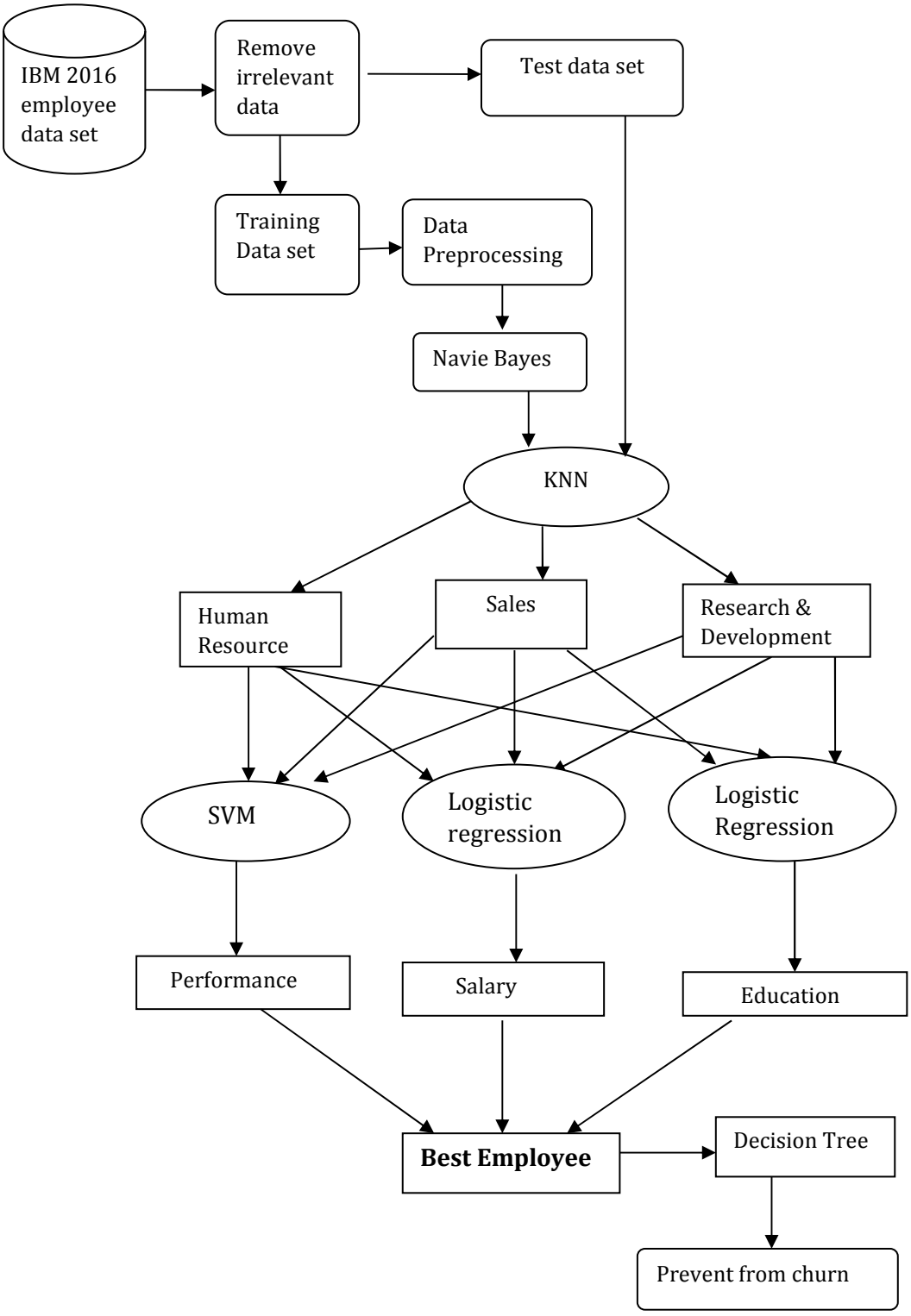
Output: Classified records.

Step1. We calculate the average rating of performance of an employee in the sample data.
Step2. Repeat Step1 for Monthly income.
Step3. Repeat Step1 for Training Times Last Year.
Step4. Compare each employee's performance, monthly income and training times last year with the average values of performance, monthly income and training times last year.
Step5. Now take each record under that particular column and add them to the output.
Step6. Remaining columns are also calculated according to step4 and step5.
Step7. Now all these employee records are taken into a single class.
Step8. We display these records to know which employee is under attrition.
Step9. Take all those employee records and preventing them from leaving the Organization.

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems.
In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

SYSTEM ARCHITECTURE

The below figure shows the architectural diagram representing the overall system framework. In this system architecture consists of employee dataset and different data mining techniques and data pre-processing techniques. Here we maintain two data sets in our data base. They are Test datasets and Training datasets, In Training datasets we can use pre-processing technique form given chosen data set in organization.
Navie Bayes, KNN, SVM, Logistic Regression and decision algorithm are used to generate a Best employee in a organization. By using this we clean and classify the employee data set into different departments like human resource, sales and research and development.



4.1 System Archietecture

ALGORITHMS:

ALGORITHM FOR KNN:

Input: W set of records
Output: Classified records
Step1. Take W= {z1, z2,…,zn}be the set records representing a class where z1,z2…zn are the employee records.
Step2. Taken input as x which represent a record to be classified.
Step3. Now we initialized i --> 1.
Step4. DOUNTIL distance from each record to x computed.
Step5: Compute distance from zi to x.
Step6: Now increment the value of i until all records are compared.
Step7: Determine minimum distance to any class.
Step8. If the class exist then classify x into class found of minimum distance,otherwise goto step9.
Step9. Classify x as a class of closest prototype.
In k-NN classification, the output is a class membership. 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 neighbour (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

ALGORITHEM FOR NAVIE BAYES:

INPUT: D set of records.
OUTPUT: Predicted records.
Step1. Take each record is an 'n' dimensional attribute vector X contains x1, x2, x3,….xn.
Step2. where xi is the value of attribute Ai.
Step3. Consider 'm' classes i.e. c1, c2, c3,… cm.
Step4. Bayesian classifier predicts X belongs to class ci
if(p(ci/x)>P(cj/x) for i<=j<=m, j!=i
Step5. Maximum Hypothesis
P(ci/x)=(p(x/ci)p(ci))/p(x)
Step6. Maximize p(x/ci)p(ci) as p(x) is costant.
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

Look at the equation below:

$$p(c/x) = \frac{p(x/c)p(c)}{p(x)}$$

$$p(c/x) = P\left(\frac{x1}{c}\right) P\left(\frac{x2}{c}\right) \ldots\ldots p\left(\frac{xn}{c}\right) p(c)$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

ALGORITHM FOR DECISSION TREE:

Input: Employee data samples, samples (always contains the value of either YES or NO), records with attributes a1, a2,…, an as attribute-list.
Output: Implemented Tree Structure.
Step1. Create an empty node N.
Step2. If samples are all of the same class c, then return N as a leaf node labeled with the class c.
Step3. If attribute-list in the record is empty then return N as a leaf node labeled with the most common class in samples.
Step4. Select test attribute, the attribute among attribute-list in the record.
Step5. Label node N with test attribute.
Step6. For each known value ai of test attribute.
Step7. Grow a branch from node N for the condition test-attribute=ai.
Step8. Let **si** be the set of samples for which test-attribute=ai;
Step9. If **si** is empty then attach a leaf node labeled with the most common class in samples;
Step10. Else attach the node returned by generate-decision-tree (si, attribute-list, test-attribute);

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.
Decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

V. EXPERMENTAL ANALYSIS:

In the existing systems they used only few of data mining techniques for data prediction. In the proposed systems we use five algorithms that is KNN, SVM, logistic regression, decision tree, and navie bayes. Here we also used feature selection method on employee data set. Generally the employee data set contains employee information like skills, nature of work, salary, performance ratting etc.

By using feature selection we can select some required features from employee data set for our analysis. In the below table 5.1 we shows the features of datasets and their type.

Features	**Data type**
Age	Number[10]
BusinessTravel	Varchar[20]
DailyRate	Number[10]
Education	Number[10]
DistanceFromHome	Number[10]
Department	Varchar[20]
EducationField	Varchar[20]
EmployeeNumber	Number[10]
Gender	Varchar[20]
EnvironmentSatisfaction	Number[10]
Hourly rate	Number[10]
Job level	Number[10]
Jobinvolvement	Number[10]
JobRole	Varchar[20]
Mirtalstatus	Varchar[20]
MonthlyRate	Number[10]
Monthly income	Number[10]
Jobsatisfaction	Number[10]
OverTime	Varchar[20]
NumCompaniesworked	Number[10]
PercentSalaryhike	Number[10]
PerfomanceRating	Number[10]
RelationshipSatisfaction	Number[10]
Stock option level	Number[10]
Totalworkingyears	Number[10]
TraningTimeslastYear	Number[10]
WorkLifeBalance	Number[10]
YearsAtCompany	Number[10]
Years In current Role	Number[10]
YearsSinceLastPromotion	Number[10]

5.1 Table Name: Employee Dataset

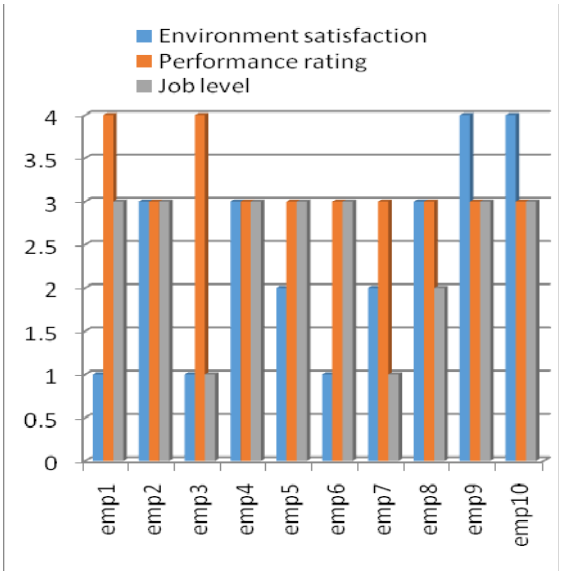
We take only 32 features that is required for our analysis. This is very helpful to increase accuracy of the system. In the below graphs we shows the difference between the accuracy of existing and proposed system.

The employee data is collected from different departments of an organization is stored in a database. Here we have considered a sample data of ten employees on the basis of Environment Satisfaction, Performance rating and their Job levels. These record values in chosen data sets.

emp Id	Environment satisfaction	Performance rating	Job level
emp1	1	4	3
emp2	3	3	3
emp3	1	4	1
emp4	3	3	3
emp5	2	3	3
emp6	1	3	3
emp7	2	3	1
emp8	3	3	2
emp9	4	3	3
emp10	4	3	3

5.2 A sample dataset of ten employees on the basis of Environment Satisfaction, Performance rating and their Job levels

The below graph 5.3 will be displayed by using the sample data set of 10 employees in an organization. From the table we chosen the performance rating based on that we have generated the graph drawn below.



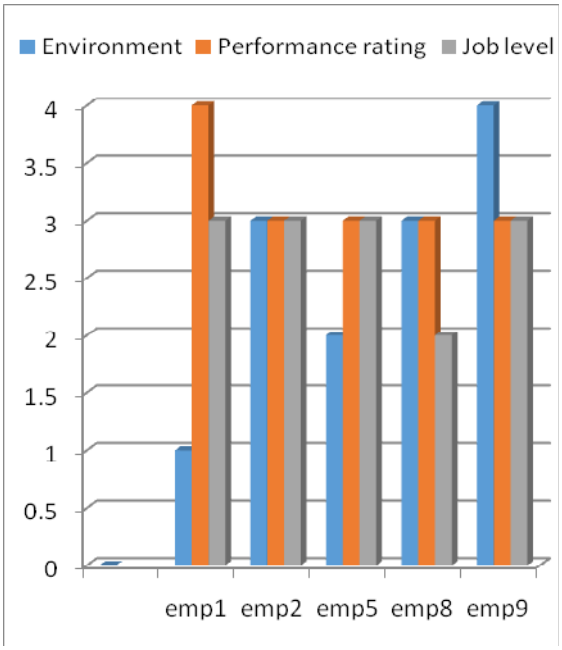
5.3 Graph for sample data set for 10 employees

Emp Id	Environment Satisfaction	Performance rating	Job level
emp1	1	4	3
emp2	3	3	3
emp5	2	3	3
emp8	3	3	2
emp9	4	3	3

5.4 The sample data set of 10 employees chosen the best employees using the performance rating

The above table 5.4 will be displayed by using the sample data set of 10 employees in an organization from that we have chosen the best employees using

the performance rating. With the data set of best employees we have generated the graph 5.5 drawn below.



5.5 Generated the best employees

The above graph 5.5 will be displayed by using the sample data set of 5 employees in an organization. From the above graph we chosen the performance rating based on that we have generated the best employees listed in the table 5.6 below.

Emp Id	Environment satisfaction	Performance Rating	Job level
emp1	1	4	3
emp9	4	3	3

5.6 Generated the best employees

5.7 Graph Generated the best employees

From the updated table we generated a new graph 5.7 and observed that employee1 is the best employee in our organization based on performance rating, job level and environment satisfaction. Since in both employees employee1 have a chance to leave the organization, so we have to prevent him from leaving by incrementing the salary or by any bonus.

VI. Conclusion

Employee attrition effects in financial, time and effort loss for organizations. It is a big issue since a trained and experienced employee is difficult to substitute and it is cost effective. We try to find to analyze the past and existing employee information to estimate the future attritioners and study the reasons of employee turnover. The results of this learning describe that data extraction algorithms can be utilized to construct reliable and accurate predictive methods for employee attrition. The issue of attrition identification is not just to depict attritioners from no attritioners. By using tentative data study and data extraction methids, we can depict the attrition probability for each one employee and provide them score to build the retention techniques.

REFERENCES

[1]. W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Vaesens,"New insights into a churn prediction in the telecommunication sector. An profit driven datamining approach," European journal of operational research, vol. 218, no. 1, pp. 211-229, 2012.

[2]. K.Coussement and D. VandenPoel, "Integrating the voice of customers through call center emails into a decision support system for attrition prediction," Information & Management, vol. 45, no. 3, pp. 164–174, 2008.

[3]. C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to attrition prediction: a data mining approach," Expert systems with applications, vol. 23, no. 2, pp. 103–112, 2002.

[4]. K. Coussement and D. Van den Poel, "Attrition prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," Expert systems with applications, vol. 34, no. 1, pp. 313–327, 2008.

[5]. J. Burez and D. Van den Poel, "Handling class imbalance in customer attrition prediction," Expert Systems with Applications, vol. 36,no. 3, pp. 4626–4636, 2009.

[6]. C.-F. Tsai and M. Y. Chen, "Variable selection by association rules for customer attrition prediction of multimedia on demand," Expert Systems with Applications, vol. 37, no. 3, pp. 2006–2015, 2010.

[7]. K. Coussement, D. F. Benoit, and D. Van den Poel, "Improved marketing decision making in a customer attrition prediction context using generalized additive models," Expert Systems with Applications, vol. 37, no. 3, pp. 2132–2143, 2010

[8]. B. Huang, M. T. Kechadi, and B. Buckley, "Customer attrition prediction in telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414–1425, 2012

[9]. V. V. Saradhi and G. K. Palshikar, "Employee attrition prediction," Expert Systems with Applications, vol. 38, no. 3, pp. 1999–2006, 2011.

[10]. R. Khare, D. Kaloya, C. K. Choudhary, and G. Gupta, "Employee attrition risk assessment using logistic regression analysis,"

[11]. M. L. Kane-Sellers, Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis.

[12]. X. Lin, F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang, and G. Xu, "A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables.