Prediction of Employee Attrition using Random Forest Classifier Technique

M. Ravi, Assistant professor, Department of Information Technology, JB Institute of engineering and technology(JBIET), Hyderabad, Telangana

A. Nirmai (Student), A. Krishitha (Student) , CH. Madhan Mohan Reddy (student), JB Institute of engineering and technology(JBIET), Hyderabad, Telangana.

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**.**

1. 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.

2. Literature 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.

3. OVERVIEW OF THESYSTEM

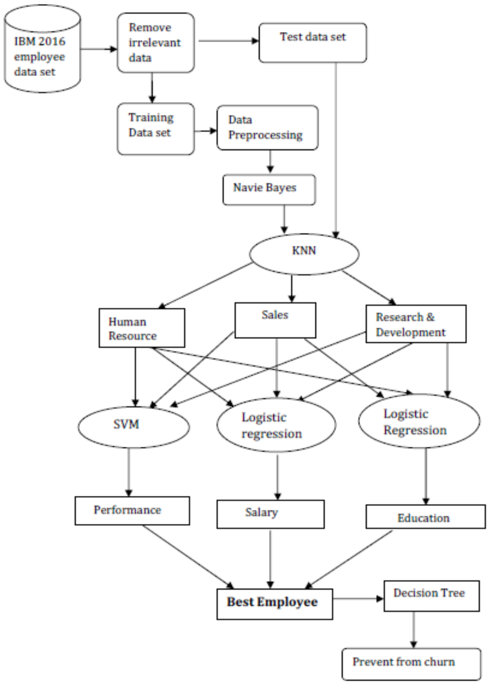


Fig 3.1 System Architecture

3.1 Existing system:

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.

3.2 Disadvantages:

In the existing systems they used only few of data mining techniques for data prediction.

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.

3.3 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

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 branched notes the outcome of a test, and each leaf node holds a class label.

3.4 Advantages:

We try to find to analyze the past and existing employee information to estimate the future attritions 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.

3.5 Algorithm For Decision 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 testattribute=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 generatedecision-tree (si, attribute-list, test-attribute)

4. SYSTEM DESIGN

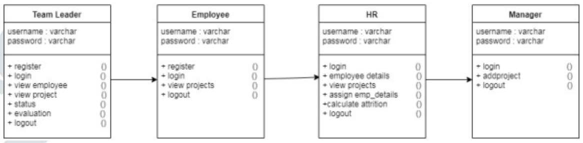


Fig 4.1 Class Diagram

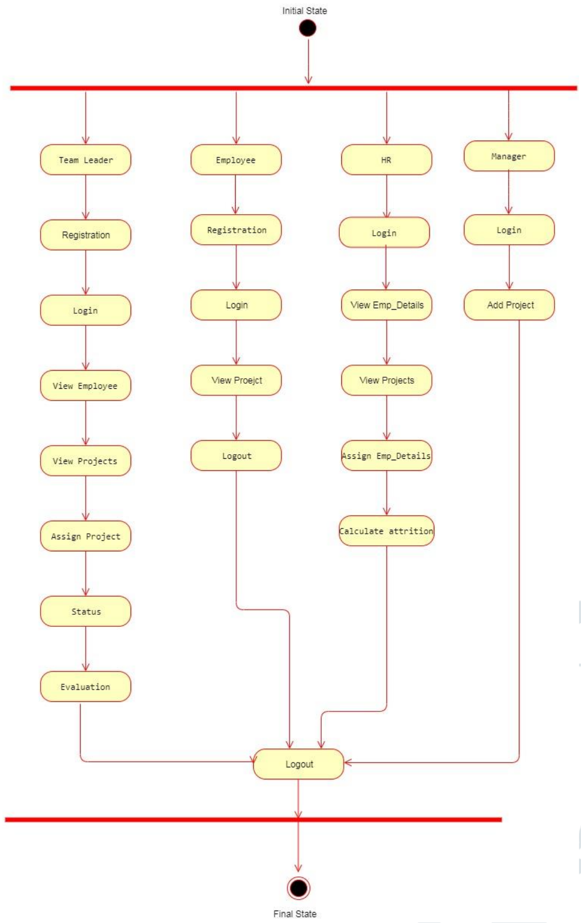


Fig 4.2 Activity Diagram

5. CODE & OUTPUT RESULTS

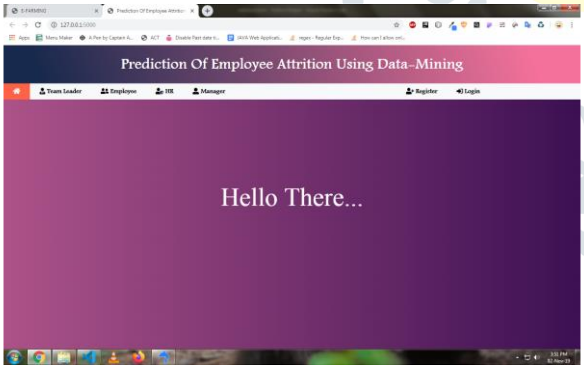
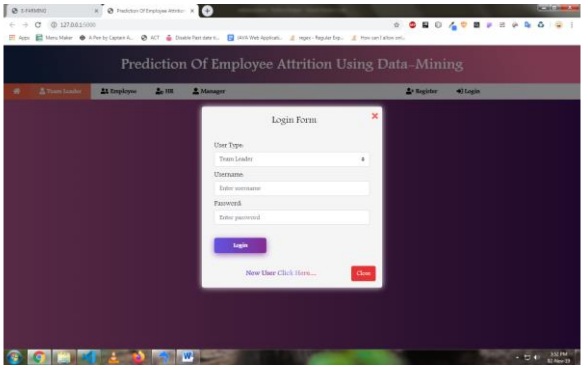


Fig 5.1: Home Page


Prediction Of Employee Attrition Using Data-Mining

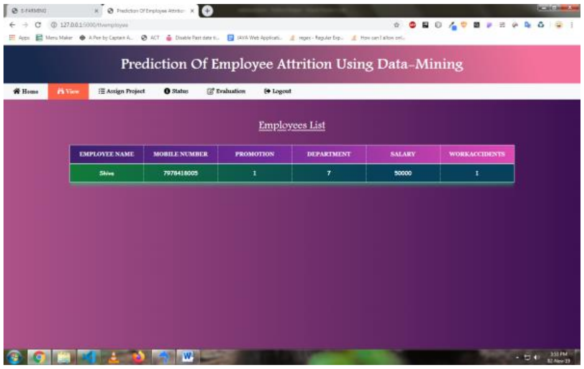
Login Form

User Type: Team Leader

Username: Enter username

Password: Enter password

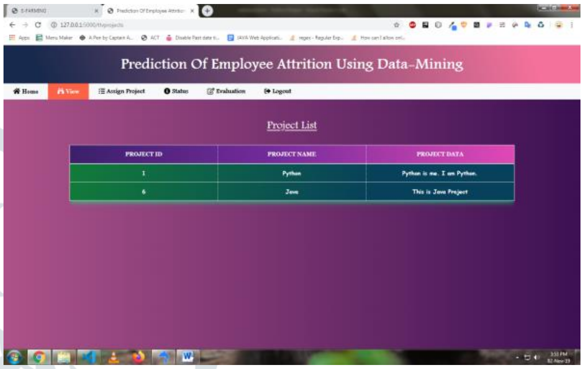
Fig 5.2: Login Page


Prediction Of Employee Attrition Using Data-Mining

Employees List

EMPLOYEE NAME	MOBILE NUMBER	PROMOTION	DEPARTMENT	SALARY	WORKACCIDENTS
Shiva	7978418005	1	7	50000	1

Fig 5.3: Employee List Page


Prediction Of Employee Attrition Using Data-Mining

Project List

PROJECT ID	PROJECT NAME	PROJECT DATA
1	Python	Python is me. I am Python.
6	Java	This is Java Project

Fig 5.3: View Project List Page

5. 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.

7. 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.