Employees Performance Prediction

B. Prasanthi

Assistant Professor, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

ABSTRACT

Employee is an important element of the organization. The success or failure of an organization depends on the employee performance. Human resource plays an important role for the success of the organization. Human Resource Management, HRM is a set of functions to maintain and develop a proficient human resource. A performance prediction procedure helps the human resource management in identifying the strengths and weaknesses of an employee. The performance evaluation of employee is depended on several different parameters based on work domain and objectives of an organization. This process of employee performance evaluation has a vital role in making strong and strategic decisions of manpower planning instead of salary reviews. The main objective of this prediction model constructed in this paper is to assist HR personnel in decision making by predicting the performance of an employee.

Keywords : Machine learning, Human Resources, HRM

I. INTRODUCTION

Human Resource Management has an important role in taking the decisions for the competitiveness, effectiveness and growth of the organization. Organizations consider HRM as a people practice. Therefore, it is the responsibility of the human resource management to recruit the employees to the suitable job at the correct time, train them and qualify them to improve the organization standards [1].

In the existing system HR employee need to handle huge amount of the data manually. This large amount of the information is important for them and it is challenging to mine the data and get the useful amount of data from it. Several errors may occur during the prediction of the performance manually.

With the growth and advancement of technology in all fields such as business organizations, education etc., Human Resource Managers will not have to handle the huge amount of data manually. It is a challenging task to mine and extract the useful and best data from this large amount of data. From here, the role of Machine Learning comes. Machine Learning is currently gaining great attention and mostly used by many people. It is considered as an emerging technology.

Machine Learning techniques like classification, association, clustering are used to get the knowledge from large amount of the data.

Classification is one of the ML technique, which performs prediction on values of the data using the results found from the various data. Classification is a supervised machine learning technique where the class label is already known. The classification techniques which are used to build the models can also be used to predict the future data.

The main objective of this research is to help the decision makers to discover potential talents of employees as follows:

- Collecting a dataset with corresponding predictive values
- Identify various factors which show impact on the employees' performance.
- Using proposed ML techniques for building a model and observe the relationships between the various features which effect the overall efficiency.

There are several data classification techniques such as Decision Tress, SVM, Linear Regression, Naïve Bayes classifier, Random Forest, KNN etc. In this paper, Logistic Regression is used. Many other techniques can also be used for classification such as Random Forest, Neural Networks etc.

II. LITERATURE SURVEY

Several studies have researched the use of the machine learning to predict the employee performance and behaviour.

Mona Mohamed Nasr, Essam Shaaban , Ahmed Samir published "A Proposed Model for Predicting Employees' Performance Using Data Mining Techniques "[1] was published in the journal International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 1 in January 2019. This paper presents various Data Mining algorithms used in the performance prediction of the employees. The algorithms used in this paper are Decision Tree, Support Vector Machine, Naive Bayes. An accuracy of 79.31% , 82.07% and 86.90% is achieved using Decision Tree, Naïve Bayes, SVM algorithms respectively.

Riyanto Jayadi published "Employee Performance Prediction using Naïve Bayes",[2] was published in the journal International Journal of Advanced Trends in Computer Science and Engineering, 8(6), in November - December 2019, 3031- 3035 . The proposed technique uses Naïve Bayes algorithm to predict the employee performance. It proposed that number of instances is related to level of accuracy and also as number of instances getting increased, the higher the accuracy became.

Sarah S, Kashif Rajpoot published "Predicting Employee Attrition using Machine Learning"[3] in 13th International Conference on Innovations in Information Technology (IIT), January 2019. In this paper SVM, random forest, and KNN classification models were evaluated This paper proposed predicting employee attrition using the original class-imbalanced dataset which resulted in the decrease in accuracy. The ADASYN algorithm solves the class imbalance problem by creating new synthetic instances based on the density distribution of the minority class. ADASYN will accomplish this by using adaptive learning to change the weights for the minority class instances. As a result, it will shift the decision boundary, which will make it easier to learn from difficult instances.

Namrata Bhartiya, Sheetal Jannu. Purvika Shukla , Radhika Chapaneri published "Employee Attrition Prediction Using Classification Models"[4] in 5th International Conference for Convergence in Technology (I2CT) Pune, India, 2019 Mar 29-31, 201.

This paper proposed that accuracy of the model has increased with the feature reduction. For Attrition evaluation Decision Tree, SVM, KNN, Random Forest and Naïve Bayes are used. An accuracy of 81.0%, 76.5%, 76.8%, 83.3% and 71.3%is obtained from Decision Tree, SVM, KNN, Random Forest, Naïve Bayes respectively.

John M. Kirimi, Christopher A. Moturi, published "Application of Data Mining Classification in Employee Performance Prediction"[5] in International Journal of Computer Applications (0975 – 8887) Volume 146 – No.7, July 2016. This paper proposed that various attributes were grouped and used for testing the model so that accuracy can be determine accurately. Decision Tree, ID3 and Naïve Bayes algorithms have been used to predict the performance. C4.5 algorithm had the highest accuracy of 92.69% and best suited for the training and development of the classification model.

Previous studies displayed different accuracy where they used different machine learning methods and various datasets.

III. Proposed Methodology

The proposed methodology was performed to help the human resources which is to build the prediction model by examining various features that would affect and predict the employees' performance.For this objective to become successful, it is required to develop a model that includes various steps that involve problem statement and definition, collection of data, understanding the data , preprocessing and modeling the data , testing and evaluating the results. If the classifier's accuracy was appropriate , the model or classifier can be used to apply to new data to give prediction about specific unknown label class and this is considered the third step as shown in Figure 1.
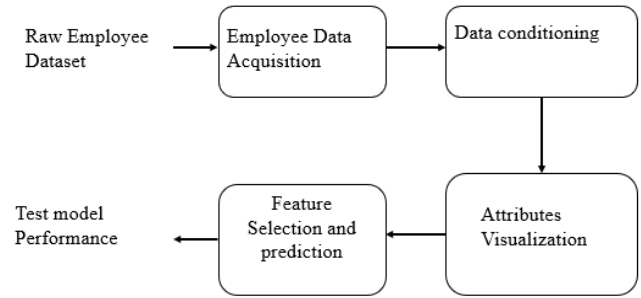

Figure 1: Process flow of the prediction

A. Data and Tools Acquisition

Initially, the data required for performing the prediction of the employee performance are acquired and also the tools required are also acquired. Initially, the required libraries such as numpy, pandas, matplotlib, scikit-learn are imported. The dataset is imported and stored as a data frame for further analysis and predictions. The dataset is stored into data frame such that any modifications done to the data will affect the data frame, and the original data remains unchanged.The dataset used for the training and testing of this INX Future Inc. The dataset of employees has the 1200 data. The features present in the data are 28 in total. The Shape of the data is 1200x28. The 28 features are classified into quantitative and qualitative where 16 features are qualitative and 11 features are quantitative.

B. Data Conditioning

Data Conditioning is cleaning the data, normalizing the datasets and performing the transformations. In this the following steps were performed.

- Exploring dataset properties: The objective of this is to understand the relationship among the variables and also to analyze the problem domain. It is for detecting the frequent issues such as outliers, null values, redundancies etc. The Figure2 shows the count of the attributes and also whether the attributes are null or not.

- Data Preparation: In this step the data is explored before the modeling and analysis. It involves feature reduction and data cleaning. In the feature reduction decisions about the feature in the dataset are made on which feature are to be retained and which are to be removed for prediction.

C. Data Visualization

The step is performed to gain overview of data and examine the data quality. Visualizations helps in understanding the data characteristics including outliers, consistencies and inconsistencies, skewness and relation among the attributes.

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   EmpNumber                    1200 non-null   object
 1   Age                          1200 non-null   int64
 2   Gender                       1200 non-null   object
 3   EducationBackground          1200 non-null   object
 4   MaritalStatus                1200 non-null   object
 5   EmpDepartment                1200 non-null   object
 6   EmpJobRole                   1200 non-null   object
 7   BusinessTravelFrequency      1200 non-null   object
 8   DistanceFromHome             1200 non-null   int64
 9   EmpEducationLevel            1200 non-null   int64
 10  EmpEnvironmentSatisfaction   1200 non-null   int64
 11  EmpHourlyRate                1200 non-null   int64
 12  EmpJobInvolvement            1200 non-null   int64
 13  EmpJobLevel                  1200 non-null   int64
 14  EmpJobSatisfaction           1200 non-null   int64
 15  NumCompaniesWorked           1200 non-null   int64
 16  OverTime                     1200 non-null   object
 17  EmpLastSalaryHikePercent     1200 non-null   int64
 18  EmpRelationshipSatisfaction  1200 non-null   int64
```

Figure 2: A part of the dataset while preprocessing

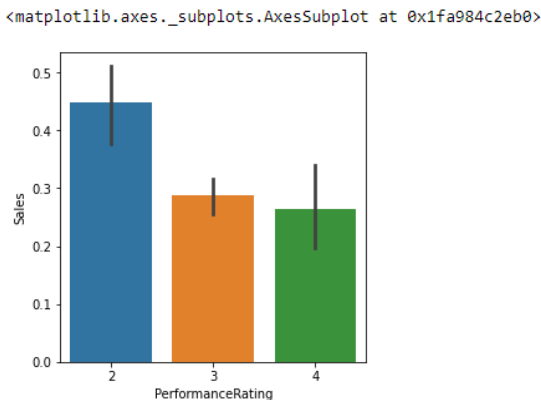The Figure 3 shows the bar plot show the bar plot of the sales department vs performance.

Figure 3 : Bar plot of Sales Vs Performance

D. Feature Selection and Prediction

Feature selection is an important concept of Machine Learning. It is a process where features which are necessary are selected to improve the accuracy and the performance. During this the dataset is split into training and testing set. The logistic Regression algorithm is used for the performance prediction.

The figure 4 shows the part of the training dataset sample.

```
[[-0.70098033 -1.04482374 -1.5407879  ... -0.06272953 -0.36255543
  -0.29563402]
 [ 1.07615291  0.8521185   0.27063524 ... -0.34450925 -0.36255543
  -0.57847295]
 [ 0.48377516 -0.41250966  1.17634682 ... -0.06272953 -0.36255543
  -1.14415081]
 ...
 [ 1.07615291  0.8521185  -1.5407879  ... -0.06272953 -0.66863237
  -0.29563402]
 [-0.70098033 -1.04482374  1.17634682 ... -0.06272953 -0.66863237
  -0.01279509]
 [-1.29335808 -1.2555951   0.27063524 ... -1.18984842 -0.66863237
  -1.14415081]]
```

Figure 4 : Training dataset sample

The figure 5shows the part of the testing dataset sample

IV. Results

It was observed that when complete dataset attributes are considered then accuracy obtained is 79.4% whereas when attributes with correlation greater than 0.1 then accuracy increased to 83.3%.

```
[[-1.29335808  1.27366121  1.17634682 ...  2.19150824  0.86175235
   1.96707741]
 [-1.29335808 -1.2555951   0.27063524 ...  1.34616907  3.31036791
  -0.29563402]
 [-1.29335808  1.90597529 -1.5407879  ...  0.78260963 -0.66863237
   0.83572169]
 ...
 [-0.10860259 -0.62328102 -1.5407879  ... -0.62628898 -0.05647848
  -0.57847295]
 [-1.29335808 -1.2555951  -1.5407879  ... -0.62628898 -0.36255543
  -0.01279509]
 [ 1.07615291  0.8521185  -0.63507633 ... -1.18984842 -0.66863237
  -1.14415081]]
```

Figure 5 : Testing dataset sample

The Figure 6 shows the accuracy when complete dataset is being considered.

```
0.7944444444444444
            precision    recall   f1-score   support

        2       0.70      0.51      0.59        63
        3       0.84      0.91      0.87       264
        4       0.54      0.45      0.49        33

 accuracy                          0.79       360
macro avg       0.69      0.62      0.65       360
weighted avg    0.78      0.79      0.79       360
```

Figure 6: Accuracy when complete dataset is considered

The Figure 7 shows the accuracy when features with co-relation greater then 0.1 are considered

```
0.8333333333333334
            precision    recall   f1-score   support

        2       0.65      0.48      0.55        63
        3       0.86      0.94      0.90       264
        4       0.81      0.67      0.73        33

 accuracy                          0.83       360
macro avg       0.78      0.69      0.73       360
weighted avg    0.82      0.83      0.82       360
```

Figure 7 : Accuracy when dataset with features with correlation greater than 0.1 are considered

The important features that are positively correlated are Environment Satisfaction, Last Salary Hike Percent & Work life Balance. This means that if these factors increase, Performance Rating will increase. On the other hand, the features that are negatively correlated are Years Since Last Promotion, Experience Years at this Company, Experience years in current role & Years with Current Manager. This means that if these factors increase, Performance Rating will go down.

V. REFERENCES

[1]. Mona Mohamed Nasr, Essam Shaaban, Ahmed Samir, "A Proposed Model for Predicting Employees' Performance Using Data Mining Techniques ", International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 1, January 2019

[2]. Riyanto Jayadi, "Employee Performance Prediction using Naïve Bayes ", International Journal of Advanced Trends in Computer Science and Engineering, 8(6), November - December 2019, 3031- 3035

[3]. Sarah S, Kashif Rajpoot, "Predicting Employee Attrition using Machine Learning" 13th International Conference on Innovations in Information Technology (IIT), January 2019.

[4]. Namrata Bhartiya, Sheetal Jannu, Purvika Shukla, Radhika Chapaneri, "Employee Attrition Prediction Using Classification Models", 5th International Conference for Convergence in Technology (I2CT) Pune, India, 2019 Mar 29-31, 201

[5]. John M. Kirimi, Christopher A. Moturi, " Application of Data Mining Classification in Employee Performance Prediction", International Journal of Computer Applications (0975 – 8887) Volume 146 – No.7, July 2016

Cite this article as :