Employee Behaviour and Attention of Career Building

T.V.Suneetha
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
takkellapati9@gmail.com

Bhanda Shivani
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
shivanibhanda@gmail.com

Chidrapu Shreshta
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
shreshtachidrapu@gmail.com

Mehrajunnisa Begum
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
muskanbegum1416@gmail.com

Nargees
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
nahednargis@gmail.com

***Abstract** -- More study into machine learning's potential uses in corporate settings is warranted in light of rising interest among business executives and stakeholders. The loss of skilled workers would result in less productivity of the organization. In this study, we apply ML approaches to investigate the root reasons for employee dissatisfaction. There were three major studies that employed the IBM Watson-generated dataset to foretell employee turnover. Many different machine-learning models, mostly built from scratch but some using the original lesson dataset, were trained in our initial experiment. Two such models were a small-sample size support vector machine (SVM) and a stochastic K-nearest neighbour (KNN) approach. We have updated the dataset using the aforementioned machine learning methods and employed the ADASYN synthetic methodology to reduce class disparity in our experiment. In the third trial, data types were manually standardised through an iterative and gradual process. Regarding this, we found that KNN training on an ADASYN-balanced dataset with a K-size of 3 yielded the best results (0.93 F1-score). Using feature selection and random forest, we were able to get the F1-score up to 0.909 while using just 12 of the possible 29 characteristics. The accuracy achieved is 85% at value K=5.In conclusion, we can say that this model would help a lot of organizations to figure out the reasons behind the low performance of the employees. Also, they can try to improve whatever is causing the employees to leave the company.*

Keywords -- :(ML) Machine learning, Supervised, KNN, Random forest, Employee attrition, Behaviour analysis.

I. INTRODUCTION

Attrition is the gradual loss of employees over time as a result of a decrease in the total number of working days. There are a variety of categories for the types of harm that companies can inflict on their staff: Disbanding or dismissal of key personnel. One type of attrition is "involuntary attrition" which means the company decides to dismiss the employee based on some reasons such as poor performance, stealing and not meeting the organizational needs. The organisation made more of an effort to retain workers with lower levels of accuracy, while those with higher accuracy were more likely to quit freely. Attrition rates may be low if workers leave voluntarily. Companies that truly value their employees' contributions invest in them by providing opportunities for professional growth and a positive work environment. There are additional costs associated with replacing employees, such as forming a selection committee, advertising the open position, and training and orienting the new hires. When important employees prepare to leave the firm, upper management may opt to institute stricter internal regulations and procedures. To keep talented workers from leaving for better opportunities elsewhere, the company may provide them incentives like pay hikes and more professional development opportunities. Calculations made by machine learning algorithms may foresee workers leaving their positions. Human resources data may be used to train a new technology model that can be used to anticipate which workers will be departing the organisation. These models are educated by comparing the characteristics of the employed with those of the unemployed. Employees leaving their jobs voluntarily may be devastating to businesses. Finding a competent worker's replacement is time-consuming and costly [1]. Research has been conducted on the causes and dynamics of employee turnover. Based on the analysed research, many variables were identified as key drivers of employee turnover. For example, compensation packages have been shown to have a significant impact on employee retention and productivity [2, 3]. The low turnover rate necessitates a higher salary. The high turnover rate in retail is due in part to factors other than pay, as discussed by [1], which include heavy workloads, bonuses based on performance, and few chances to advance in one's career.

979-8-3503-9725-3/23/$31.00 ©2023 IEEE

II. RELATED WORK

In terms of management, anything that can be done manually is managed by command. Once an employee leaves an organisation, their file is updated by hand. It requires time, effort and errors are possible. It's possible that an employee can be thinking of quitting their job for a number of different reasons. Employees may opt to quit their current position due to low compensation, a lack of amenities, or other similar issues. It's just not feasible for humans to foresee these kinds of events. To address these concerns, we apply a range of machine-learning techniques to create a model of employee turnover. On feeding it with enough information, the model learns to make reliable forecasts. The random-forest technique is used to create models for making predictions. In this article, we examine the use of dependent variables and grammar structure vectors in the estimation of employee turnover. The effectiveness of a model may be gauged, in part, by selecting an appropriate accuracy score. Just two are taken into account by the standard method for assessing attrition loss. The current confusion matrix serves its purpose admirably. The processing time for the Random Forest approach is higher than that of the K-nearest neighbour approach. Although turnover and downtime are taken into consideration by current methods, other characteristics, such as extensive travel and relationship status, may be just as crucial. Many studies have examined machine learning's potential to predict how employees would react in a variety of settings. In [4], the authors use a Naive Bayes model and decision trees to forecast worker output. Their research showed that job title was the most influential factor, whereas age, gender, and education level had no bearing. This study used a dataset of 1575 records and 25 attributes to investigate several data mining techniques for predicting employee turnover (or attrition). Neural networks, support vector classes, logistic regression, and ensemble methods are all examples of the kinds of supervised and unsupervised techniques that are used. With an accuracy of 84.12%, support vector classification (SVC) was shown to be the most successful classification technique.C4.5, C5, REP Tree, and classification and regression trees are only some of the decision tree methods investigated in [6]. (CART) In order to test and train the decision trees, the researchers used a dataset consisting of 309 employee records out of 4326 records and six variables in total. C5's 74% accuracy was the highest of any judgement tree we examined.

They also found that income and tenure were major drivers in the test dataset, which indicates that these factors are crucial to the organisation under review. The authors of [7] made an annual prediction of the growth rate of neural networks in use by small West Coast industrial businesses. Using a neural network concurrent optimization method (NNSOA) and 10-fold cross-validation, they accurately predicted employee turnover 94% of the time. Moreover, they used a genetic algorithm to zero in on "tenure of an employee on January 1" as the most crucial, controllable factor. Using a web-based database that includes profiles of 6,909,746 present and past employees, the authors of [8] make predictions about employee turnover. The usual employee profile would include the person's past and current jobs, as well as their education and training, providing details about the business. SVM model training and testing were both successful in this study. Around 55% of projections were correct, so the model did a good job overall. In order to effectively train the system, the researcher suggested including employee demographic information in the dataset. It was projected that a multinational store established in the United States would have a high turnover rate of employees [9]. A total of 33 features and 73,115 observations were included in the dataset. The area under the curve (AUC) for XGBoost was 0.88, making it the best performing of the seven data mining methods studied. It also had a lower memory requirement than its rivals. In [10], the author presents a model she used to predict when employees at Swedbank could leave the company. A random forest model achieves higher accuracy (78.5%) than competing models like the support vector machine (SVM) and the monolayer perceptron (MLP).

Many accuracy measurements and techniques of machine learning have been reported by researchers in the past. Because of this, settling on a single preferred version is a formidable challenge. Also, actual attrition data shows a class imbalance that has not been addressed in any of the prior research. As a result, we investigated a wide range of strategies for lowering social distinctions, all of which improved educational settings.

III. PROPOSED APPROACH

Important factors like quarterly revenue, last promotional year, wage raise, and others are collected during the first pre-processing of the Kaggle data. Those traits are seen as being fairly acceptable in terms of employee absenteeism. The degree to which a given element is reliant on a certain component of the workforce may be readily determined by using dependent or expected variables. Turnover rates are little affected by factors like workforce size or employee demographics. Exploratory data analysis may assist you in identifying who could be at risk of quitting their job and when by providing a summary of the data's numerous qualities. It would suggest that a positive work environment and an optimistic mindset among employees might go a long way towards reducing the severity of this issue. Predictions may be made with the use of the system's KNN model. Random forest was the model-building process employed in the prior approach. An example of ambient learning, this technique uses many decision trees to make classifications. Classification is performed on every tree in the randomly created forest, as well as on an unlabeled sample. The data that has not yet been classified is filed under the most popular classification system. Raising morale and encouraging more collaboration among workers might be one approach to solving this issue. a measure of how often employees leave a companyThe present method's main benefit is that it employs the KNN (K closest neighbour) model, which is advocated for usage in the proposed system. In contrast to the new method's use of 25 parameters for attrition prediction, the old one only used two. Some of the factors that contribute to an organization's staff turnover rate include the location of workers, the number of projects they are working on, the amount of hours they work each day, and the marital status of those workers.

IV. MACHINE LEARNING

Training Data involves analysing a massive, already-classified dataset with the help of a classifier and a regression algorithm. It is expected that this retraining process will need to be done many times before the required level of improvement is reached. By contrast, the data-driven method known as unsupervised learning does not rely on labels to make progress. It may be used in conjunction with other methods, such as factor and cluster analysis, to learn more about the data.

The reason behind choosing ML is it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products.

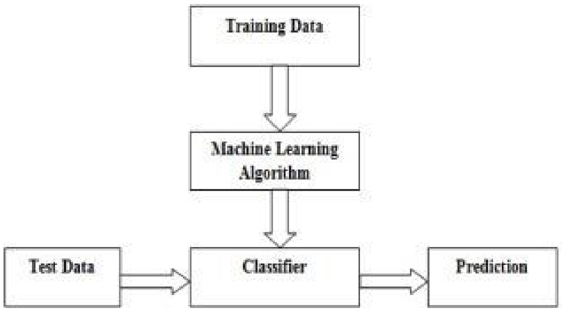


Fig.1 Machine Learning Model

V. ALGORITHM IMPLEMENTATION

One of the most used ML algorithms is logistic regression, which is part of the larger category of machine learning techniques that also includes supervised learning. The method may now be used to choose a category variable from a group of unrelated variables.

Logistic regression is used to make predictions about the outcome of a classification regression. That's why the ultimate tally has to be something that can be readily partitioned into smaller sums. As the range of the provided statistics is from 0 to 1, we may offer answers like "0" or "1," "Yes" or "No," "true" or “false” etc.

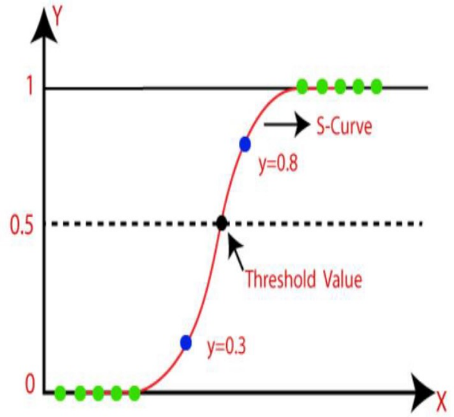


Fig.2 Logistic Regression

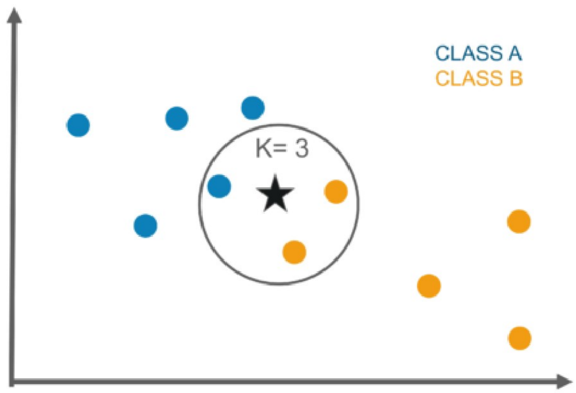


Fig. 3 Classifying using KNN

K-Nearest Neighbors (KNN) is a popular supervised learning technique used in many machine learning algorithms. The use of machine learning (ML) to problems of regression analysis and classification is strongly recommended. It uses ensemble learning, which combines many classifiers, to take on difficult problems and boost detection rates.

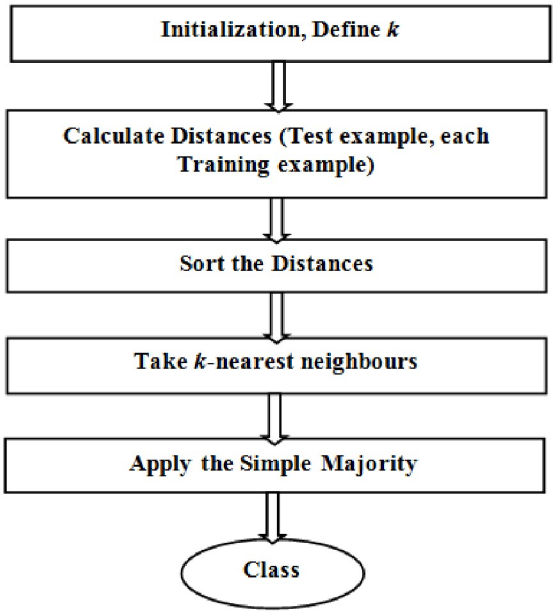

Initialization, Define *k*
Calculate Distances (Test example, each Training example)
Sort the Distances
Take *k*-nearest neighbours
Apply the Simple Majority
Class

Fig.**4** KNN Work Flow

Step 1: Pick K training variables randomly.

Step 2: construct the trees of choice that correspond to the specified data (Subsets).

Step 3: Decide how many judgment trees, N, you'd want to construct.

Step 4: Repeat Stages 1 and 2.

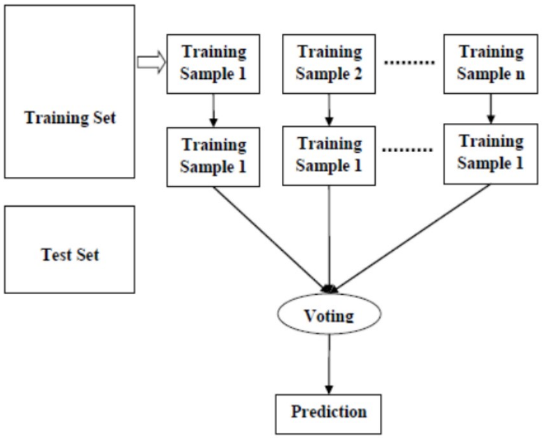


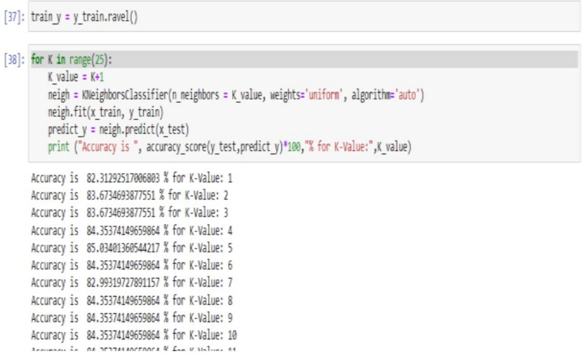
Fig. 5 Splitting Training and Testing data

k-Nearest Neighbors Algorithm:

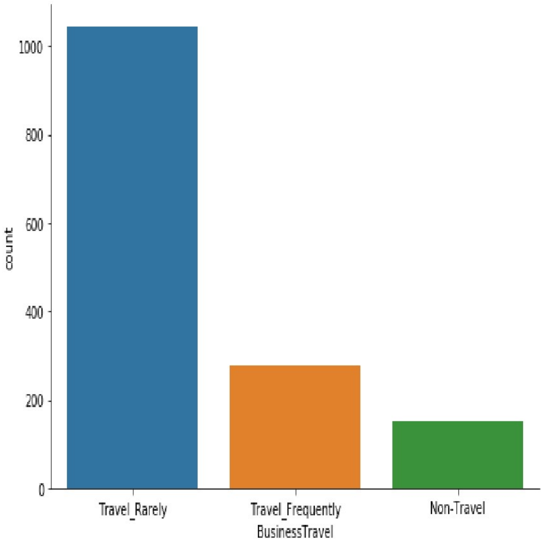
$$\left(\boldsymbol{d}(\boldsymbol{x},\boldsymbol{y})\sqrt{\textstyle\sum_{i=1}^{n}(\boldsymbol{yi}-\boldsymbol{xi})}\right)$$

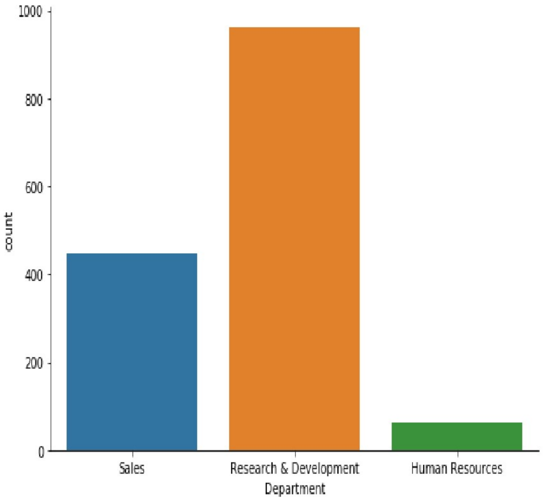
VI. RESULTS

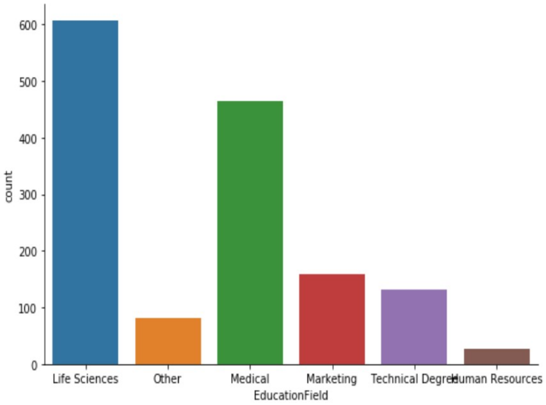
The output imports all required libraries (OS, Numpy, Pandas, Matplotlib, and Sklearn) for my project. We use the pandas package to import a CSV file containing data on IBM's employee turnover rates into our data frames. The function read csv is used in that library. This output is used for cleaning the dataset before analysis, which improves the quality of the findings.This is why we relied on the Pandas' drop function. The drop operation is used in conjunction with other algorithmic procedures to rid a dataset of unwanted data records.The K-Nearest Neighbors classifier was used after applying the leaner regression, which yielded an accuracy score of 82%. Our algorithm achieved an accuracy of 85%.

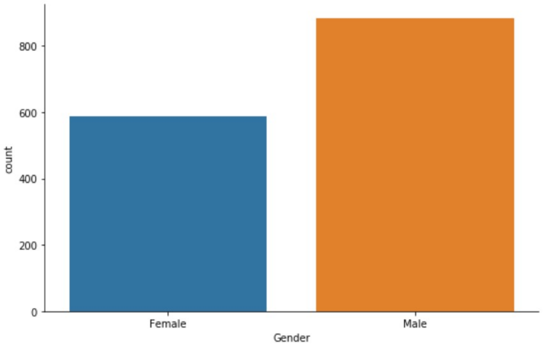


```
[37]: train_y = y_train.ravel()

[38]: for K in range(25):
          K_value = K+1
          neigh = KNeighborsClassifier(n_neighbors = K_value, weights='uniform', algorithm='auto')
          neigh.fit(x_train, y_train)
          predict_y = neigh.predict(x_test)
          print ("Accuracy is ", accuracy_score(y_test,predict_y)*100,"% for K-Value:",K_value)

      Accuracy is  82.31292517006803 % for K-Value: 1
      Accuracy is  83.6734693877551 % for K-Value: 2
      Accuracy is  83.6734693877551 % for K-Value: 3
      Accuracy is  84.35374149659864 % for K-Value: 4
      Accuracy is  85.03401360544217 % for K-Value: 5
      Accuracy is  84.35374149659864 % for K-Value: 6
      Accuracy is  82.99319727891157 % for K-Value: 7
      Accuracy is  84.35374149659864 % for K-Value: 8
      Accuracy is  84.35374149659864 % for K-Value: 9
      Accuracy is  84.35374149659864 % for K-Value: 10
      Accuracy is  84.35374149659864 % for K-Value: 11
```









VII. CONCLUSION

Modern data-driven decision-making processes often include predictive modelling. The goal of this study is to use a K-Nearest Neighbors classifier to forecast employee turnover. The model's efficacy is analysed once it has been put through its paces. With an experimental accuracy of 85%, the suggested model seems trustworthy and optimises employee retention. Considering that the presented model may be utilised to efficiently increase staff retention rates, it is a valuable tool for decision-making.

VIII. REFERENCES

[1]. Abhiroop Nandi Ray, Judhajit Sanyal, Machine Learning Based Attrition Prediction, Global Conference for Advancement in Technology, IEEE, (2019).

[2]. Sandeep Yadav, Aman Jain, Deepti Singh, Early Prediction of Employee Attrition using Data Mining Techniques, IEEE, (2018).

[3]. Shawni Dutta, Samir Kumar Bandyopadhyay, Employee attrition prediction using neural network cross validation method, International Journal of Commerce and Management Research, (2020).

[4]. Rachna Jain, Anand Nayyar, Predicting Employee Attrition using XG-Boost Machine Learning Approach, International Conference on System Modeling & Advancement in Research Trends, IEEE, (2018).

[5]. Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, Xiaoyu Zhu, Employee Turnover Prediction with Machine Learning: A Reliable Approach, Springer Nature Switzerland AG 2019.

[6]. Sarah S. Alduayj, Kashif Rajpoot, Predicting Employee Attrition using Machine Learning, 13th International Conference on Innovations in Information Technology (IIT), IEEE, (2018).

[7]. Rohit Hebbar A, Rajeshwari S.B, Sanath H Patil, S S M Saqquaf, Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees, International Conference on Recent Trends in Electronics, Information & Communication Technology, IEEE, (2018).

[8]. Usha.P.M, N.V.Balaji, An Analysis of the Use of Machine Learning for Employee Attrition Prediction – A Literature Review, Journal of Information and Computational Science, (2020).

[9]. Rohit Punnoose, Pankaj Aji, Prediction of Employee Turnover in Organizations using Machine Learning Algorithms, International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, (2016)

[10]. Sandeep Yadav, Aman Jain, Deepti Singh, “Early Prediction of Employee Attrition using Data Mining Techniques” in IEEE 2018.

[11]. R Shiva Shankar, J Rajanikanth, V.V.Sivaramaraju, K VSSR Murthy, ” PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING”, in IEEE 2018.

[12]. Rachna Jain, Anand Nayyar,” Predicting Employee Attrition using XGBoost Machine Learning Approach”, in IEEE 2018.

[13]. Nagadevara, Vishnuprasad. (2018). Early Prediction of Employee Attrition in Software CompaniesApplication of Data Mining Techniques.

[14]. Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38, 1999–2006. [7] Prashant Gupta (18th May 2017) Decision Trees in Machine Learning [Online]. [Accessed: 10-Oct2018].