iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59401

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

A Novel Approach on Early Prediction of Employee Attrition in an Institution using Machine Learning Techniques

Padmapriya J
Information Technology Department, Vidya Jyothi institute of Technology

Abstract: ***Institutions can only succeed if they have good employees. Retaining a good employee in an institution is a must to its growth. Sometimes, employees face issues in their institution because of overwork, no promotions, no rewards for good work, not seeing eye to eye with their manager, frequent business trips and extreme conditions which lead them to look for new jobs in the market. Employee attrition can be curbed if these causes are found sooner. To predict an employee's resignation, Machine Learning Techniques are utilized. Attrition rates in an organization are predicted by factors such as work-life balance, opportunities, office atmosphere, pay, and other benefits. The Human Resources team will find the attrition rate data to be quite helpful in keeping exceptional employees. Random Forest, K-Nearest, Support Vector Machine and XG Boost are algorithms used to predict the attrition rate in an institution. The Human Resources Management (HRM) dataset is used by the models to detect various data aspects and efficiently estimate employee attrition.***
Keywords: Attrition, Prediction, Institution, Machine learning, Retention

I. INTRODUCTION

In an institution, employee attrition is a key factor for growth. If employees are not satisfied with their work and management there are high chances, they would want to shift their job or move for better opportunities. But if they leave jobs unexpectedly, it may cause huge loss for the institution. Hiring new employees will consume money and time, and also the freshly hired employees take time to make the respective institution profitable. Retention of skilled and hardworking employees is one of the most critical challenges faced by many institutions. Hence, by improving employee satisfaction and providing a desirable working environment an institution can improve attrition rate. The major reasons for the employees to leave their jobs are relocations, disliking management, pursuing higher studies, salary not as per expectation, dissatisfaction in work, lack of opportunities for career growth, poor working environment, unfriendly environment, bad relationship with higher authorities, workload, and overtime. If the employee has recently joined it is difficult to find their interest of leaving their job.

This system is able to predict which employee may leave an institution with what reason, so that they can take several corrective actions in order to ensure that employees stay in the institution and can reduce attrition. Some of the employee retention strategies to control attrition are motivating employees, exposing employees to newer roles and taking constant feedback from employees.

II. LITERATURE SURVEY

Employee Attrition is the normal flow of people out of an institution, due to career or job change, relocation, illness and so on. Employee Attrition is the percentage of employees leaving the institution for whatever reasons. Employees can leave the institution for personal as well as professional reasons. There are two types of turnover, voluntary turnover which is decided by the employee, and the other one is decided by the company and that is why it is called involuntary turnover. Involuntary turnover generally happens when performance of the employee is not up to the expectations of the company.

Retention is necessary for the growth and stability of an institution. The high attrition rate is caused when there are more employment opportunities in the market. Currently the employee attrition is one of the major issues faced by HR managers. There are so many working employees who are not satisfied due to aspects which are not fulfilled by the institution which results in higher attrition rate. IBM HR simulated dataset is a medium sized-dataset provided by IBM and it contains 1470 samples with 34 input features (Age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number of Companies Worked, Over18, Over Time, Percent Salary Hike,

Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager) and its target variable is attrition that is represented as''No'' (employee did not leave) or ''Yes'' (employeeleft).
Kaggle HR dataset is a large sized-dataset supplied by Kaggle that contains 15000 samples where its target variable is'' left'' and its 9 features are satisfaction level; last evaluation; number project; average monthly hours; time spend company; Work accident; promotion last 5 years; sales and Salary. The CSV revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, managers name, and performance score.

III. PROPOSED SYSTEM

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

A. Data Collection

In order to collect employee real data and to tap the factors responsible for attrition in this study, an online questionnaire was prepared and used as a datagathering instrument from respondents.
Features collected through the exploratory method have been divided into three parts. Part 1 comprises demographic variables including: Gender, Age, Education, Marital status, and Tenure. Part 2 is about their overall level of satisfaction, motivation, involvement, and life interest (Job satisfaction, Job involvement, Job performance, Promo ability, Environment satisfaction, Rewards, Relationship satisfaction, Business travel, Grade, Training, Work life/ balance). Finally, part 3 aims to know the most impactful factors according to respondents and to collect their suggestions. From the designed survey we received 450 responses. Respondents were university people from different countries (India, Tunisia, Norway, France, United States, Italy, Pakistan, England, and Germany). The questionnaire was anonymous. 44.5% of respondents were female and 55.5% were male. Age of the respondents varied from 27 to 62. Out of the total participants, 47.3% wanted to leave their jobs and the rest did not have the intention to quit. The HRM dataset used in this research work is distributed by IBM Analytics. This dataset contains 35 features relating to 1500 observations and refers to India data. All features are related to the employees' working life and personal characteristics.

B. Dataset Features

Title must be in 24 pt Regular font. Author name must be in 11 pt Regular font. Author affiliation must be in 10 pt Italic. Email address must be in 9 pt Courier Regular font.

TABLE I
ATTRIBUTES OF HRM DATASET

Sl No	Attribute 1	Attribute 2
1	Age	Monthly income
2	Attrition	Monthly rate
3	Business travel	Number of previous employers
4	Daily rate	Over 18
5	Department	Overtime
6	Distance from home	Percent salary hike
7	Education	Performance rating
8	Education field	Relations satisfaction
9	Employee count	Standard hours
10	Employee number	Stock option level
11	Environmentsatisfaction	Total working years
12	Gender Training times	last year
13	Hourly rate	Work-life balance
14	Job involvement	Years with company
15	Job level	Years in current role
16	Job role	Years since lastpromotion
17	Job satisfaction	Years with current manager
18	Marital status	Yes/No

The dataset contains target feature, identified by the variable Attrition: "No" represents an employee that did not leave the company and "Yes" represents an employee that left the company. This dataset will allow the machine learning system to learn from real data rather than through explicit programming. If this training process is repeated over time and conducted on relevant samples, the predictions generated in the output will be more accurate.

C. Feature Extraction

Feature extraction is a type of dimensionality reduction where a large number of pixels of the image are efficiently represented in such a way that interesting parts of the image are captured effectively. The table 3.1 show features collection which includes different attributes. They are employee number, environment satisfaction and job satisfaction in the institution etc.

Feature selection is done through quantitative method using data collection techniques. One of the common techniques used for data collection is survey method. Survey can be done among the employees by the employer once in a quarter which will help to improve the employee attrition rate.

There are 2 Class labels – Active and Terminated labeled 0 and 1 respectively. Each employee would have a record for every quarter of being active in the institution, until the quarter of turnover, at which time the data point changes class label from active to terminate. The dataset had 73,115 data points with each labeled active or terminated. The data was gathered from the HRM dataset. The HRM dataset is used to provide some key features like demographics features like age and compensation, related features like pay, and team related features like peer attrition etc. The data provided key features like unemployment rate, median household income etc. Overall, there were 33 features of which 27 were numeric while 6 were categorical in nature.

D. Algorithm for Attrition Prediction Model

1) *K-Nearest Neighbor:* K-Nearest Neighbor is considered a lazy learning algorithm that classifies data sets based on their similarity with neighbors. It is one of the most fundamental and simple classification methods and one of the best choices for a classification study of the data. The classification using KNN involve determining neighboring data points and then deciding the class based on the classes of the neighbors.
2) *Decision tree classifier*: As the name implies all decision tree techniques recursively separate observations into branches to construct a tree for the purpose of improving the prediction accuracy. Decision tree is a conventional algorithm used for performing classifications based on the decisions made in one stage. This provides tree structured representation of the decision sets.
3) *Random forest classifier:* Random Forest is used for both Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to enhance the performance of the model. Instead of depending on one decision tree, the random forest takes the prediction from each tree and prediction which have majority of votes will be the final output. As the number of trees increases the accuracy also increases and prevents it from the over fitting problem.[3]
4) *Support Vector Machine (SVM):* Support Vector Machine is kind of classification technique. It is a model used for classification and regression problems. It can solve linear and non- linear problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes. When unknown data is given as input, the SVM predicts which class it belongs to. The margin between the hyper plane and the support vectors are as large as possible to reduce the error in classification.[3]
5) *Long Short-Term Memory Networks:* LSTM are improved version of recurrent neural networks (RNN) that are able to model sequential and temporal data and to predict times series. More specifically, a cell state is added in LSTM to store long-term states and to build more stable RNN for time series prediction by detecting and memorizing the long-term dependencies existing in the time series.
6) *Convolution Neural Networks (CNN):* CNN generally contains four kinds of layers in its structure: an input layer, a convolution layer, a pooling layer, and a fully connected layer. In the convolution layer, which represents the most important CNN part, the input is convoluted with different filters where each filter is considered as a smaller matrix.
   Then, corresponding feature maps are generated after the convolution operation. The pooling operation consists of reducing the size, while preserving the important features. The efficiency of the network is thus improved, and over-fitting is avoided. The main role of the convolution and pooling layers is to extract features, and the main goal of the fully connected layers is to output the information from feature maps together, and then provide them to latter layers.

7) *XG BOOST:* XG Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weighs data points that are hard to predict. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modeling problems. Machine Learning is a very active research area and already there are several viable alternatives to XG Boost. Microsoft Research recently released Light GBM framework for gradient boosting that shows great potential. Cat Boost developed by Yandex Technology has been delivering impressive bench- marking results. It is a matter of time when we have a better model framework that beats XG Boost in terms of prediction performance, flexibility, expandability, and pragmatism. However, until a time when a strong challenger comes along, XG Boost will continue to reign over the Machine Learning world.

IV. BLOCK DIAGRAM

Fig. 1 shows the employee attrition prediction model with feature collection which collects 16 features from the dataset. The feature selection is done after feature collection through quantitative method. The feature selection is done via a survey method using questionnaires. After that, appropriate data is selected for the model by an algorithm. The available dataset which contains employee details is present within the HRM dataset.

Employee details from the dataset are preprocessed and then the prediction model is introduced for the interpretation of related results.

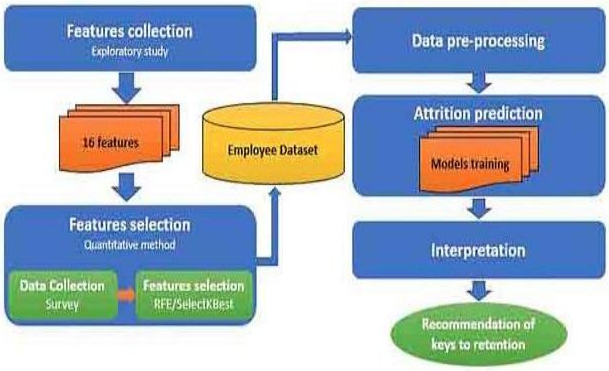


Fig 1: Architecture of Proposed Model

These results are useful for retention of employees inthe organization which is very needed for institutional growth. This will avoid unnecessary time in recruiting new employees to replace the old employees [2].

Hiring and retaining top talent is an extremely challenging task that requires capital, time, and skills. Small business owners spend 40% of their working hours on tasks that do not generate any income such as the hiring process for new employees.

V. RESULTS AND EVALUATION

The population in the dataset is representative of a workforce that is distributed across India, comprising of people at different stages of their careers, different levels of performance and pay, and from different backgrounds. Hence, it is intuitive to assume that a rule-based approach or a tree-based model will most likely perform best, considering the various themes and groups naturally occurring in the data. It is seen that the two tree-based classifiers in Random Forest and XG Boost perform better than the other classifiers during training and that XG Boost is significantly better than Random Forest during testing. The XG Boost classifier outperforms the other classifiers in terms of accuracy and memory utilization.

The XG Boost classifier is also optimized for fast, parallel tree construction, and designed to be fault tolerant under the distributed settings. XG Boost classifier takes data in the form of a DMatrix. DMatrix is an internal data structure used by XG Boost which is optimized for both memory efficiency and training speed. DMatrices were constructed from numpy arrays of the features and classes.

TABLE 2
ACCURACY OF DIFFERENT MODELS WITH DIFFERENT DATASETS

Models	IBM dataset		11 features from IBM dataset		Kaggle dataset	
	Acc.	F1	Acc.	F1	Acc.	F1
DT	0.777	0.318	0.952	0.766	0.972	0.945
LR	0.83	0.368	0.849	0.15	0.782	0.37
SVM	0.85	0.5	0.837	0.42	0.782	0.495
DNN	0.80	0.42	0.84	0.4	0.89	0.18
LSTM	0.71	0.487	0.75	0.57	0.79	0.605
CNN	0.84	0.672	0.89	0.649	0.91	0.7
RF	0.858	0.169	0.953	0.828	0.978	0.967
XGB	0.853	0.434	0.956	0.728	0.976	0.946
VC	**0.93**	0.58	**0.96**	0.62	**0.98**	0.88
Stacked	0.88	0.5	0.9	0.631	0.96	0.67

The Table 2 shows different model accuracies with different datasets. The models SVM, CNN and VC are more accurate when compare to other models. Comparing the IBM and Kaggle datasets, Kaggle shows more efficiency in feature selection.

A. Age vs Attrition Analysis

People of age 29 to 31 years left the company more frequently. Although the number of employees in the age group of 18 to 23 is less, the attrition rate is high in this group too. Also, as Fig 2 illustrates, the likelihood of quitting the company declines with age

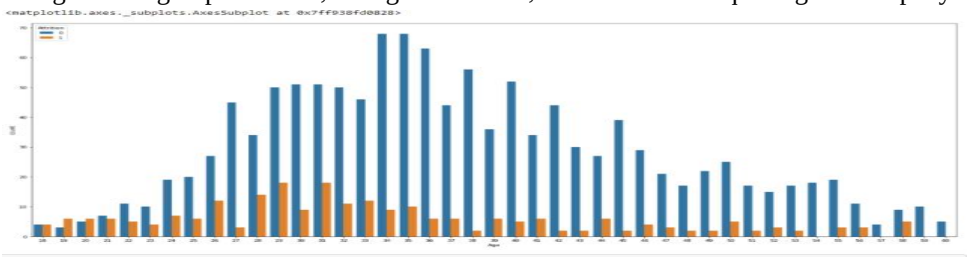


Fig 2 Attrition rate between age and gender

B. Job Role Vs Attrition

The Fig 3 displays the relationship between job role verses attrition for different departments. The sales executive has highest attrition rate and human resources has least attrition rate among these departments. There are very less attrition rate in female employees in manufacturing director, health care representative, manager and research director.

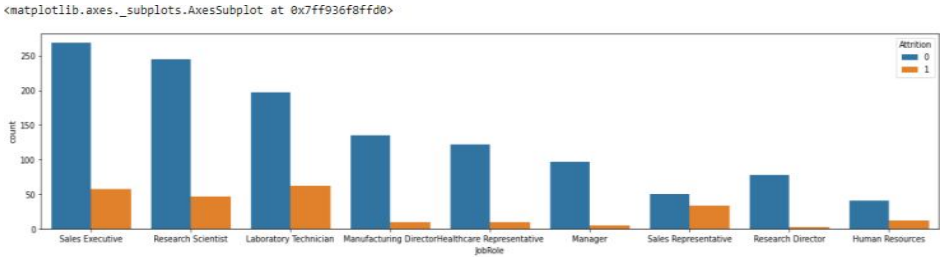


Fig 3 Comparation between different departments

C. Current Manager Vs Attrition

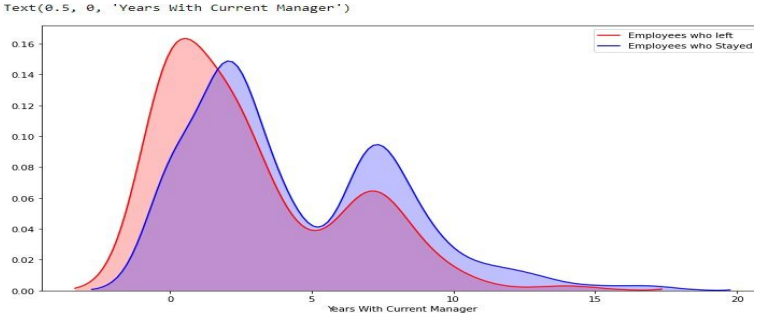


Fig. 4 Attrition rate with respect to current manager

The Fig 4 shows the relationship between number of years working with current manager and attrition rate. The employees more likely to leave their jobs in the first 5 years (0-5) when compared to the next 5 years (5-10). The attrition rate is almost zero for 15 to 20 years

D. Working experience with Attrition

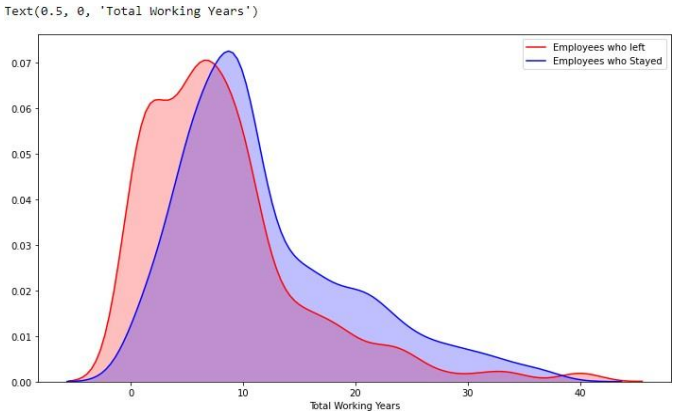


Fig 5 Attrition rate with respect to working hours

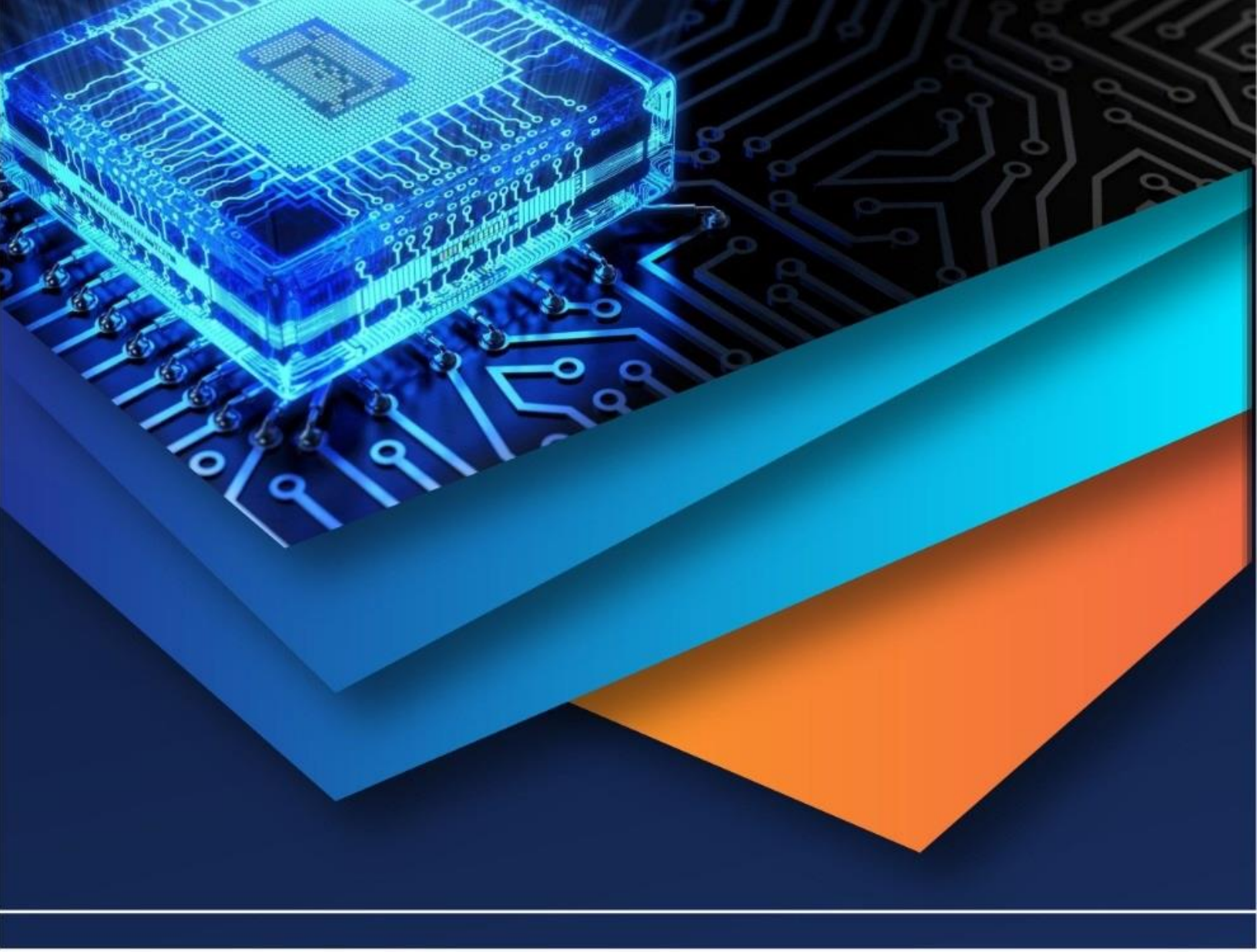
The Fig 5 shows the relationship between work experience of the employees and attrition rate. The attrition rate is highest at ten years of experience. At twenty years of experience the attrition rate low and it is moderate at fifteen years of experience.

VI. CONCLUSION

The importance of predicting employee attrition in institutions and the application of machine learning in building models are presented in this paper. The noise in the data from HRM dataset that compromises the accuracy of these predictive models is also highlighted. Data from the HRM dataset was used to compare the XG Boost classifier against six other supervised classifiers that have been historically used to build prediction models. The results of this research demonstrate that the XG Boost classifier is a superior algorithm in terms of significantly higher accuracy, relatively low runtimes, and efficient memory utilization for predicting attrition. The formulation of its regularization makes it a robust technique capable of handling the noise in the data from HRM dataset, as compared to the other classifiers, thus overcoming the key challenge in this domain. Because of these reasons it is recommended to use XG Boost for accurately predicting employee turnover, thus enabling institutions to take actions for retention or succession of employees.

REFERENCES

[1] Nesrint Ben yahin, Jihen Hlel and Ricardo colomo-palacies, “From Big data to Deep data tosupport people analytics for employee attrition prediction”,2021.
[2] Rohit Punnoose, Pankaj Ajit,” Prediction of Employee Turnover in Institutions using Machine Learning Algorithms A case for Extreme Gradient Boosting”, IJARAI-Vol. 5, No. 9, 2016
[3] R. D. Roscoe and M. T. Chi, “Understanding tutor learning: Knowledge building and knowledge-telling in peer tutors explanations and questions,” Rev. Educ. Res., vol. 77, no. 4, pp. 534–574, 2007.
[4] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly, “Grit: Perseverance and passion for long-term goals,” J. Pers. Soc. Psychol., vol. 92, no. 6, pp. 1087–1101, 2007.
[5] C. Peterson, et al., Character Strengths and Virtues: A Handbook and Classification, vol. 1.Oxford, U.K.: Oxford Univ. Press, 2004.
[6] M. Kapur, “Productive failure,” Cogn. Instruction, vol. 26, no. 3, pp. 379–424, 2008.
[7] J. E. Beck and Y. Gong, “Wheel-spinning: Students who fail to master a skill,” in Proc. Int. Conf. Artif. Intell. Educ., 2013, pp. 431–440. [6] N. Matsuda, S. Chandrasekaran, and J. C. Stamper, “How quickly can wheel spinning be detected?” in
[8] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, “Evaluation of machine learning models for employee churn prediction”, International Conference on Inventive Computing and Informatics(ICICI 2017).
[9] S. Jahan, “Human Resources Information System (HRIS): A Theoretical Perspective”, Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129, 2014.
[10] M. Stoval and N. Bontis, “Voluntary turnover: Knowledge management– Friend or foe?”, Journal ofIntellectual Capital, 3(3), 303- 322,2002.
[11] J. L. Cotton and J. M. Tuttle, “Employeeturnover: A meta-analysis and review with implications for research”, Academy of management Review, 11(1), 55-70, 1986



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY