IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 12, December 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

Analysis of Employee Attrition Rate in Indian IT Industry and Prediction using Machine Learning Approach

Piyush Gangrade [1], Naman Vijayvargiya[2]

Data Engineer, Bangalore, Karnataka, India[1]

Software Engineer, Pune, Maharashtra, India[2]

ABSTRACT:The information technology (IT) industry in India has been facing systemic problems with high churn rates in the past few years, resulting in companies losing money and knowledge. In any company, if a large number of employees leave their jobs with a short notice period, it may lead to a decrease in overall productivity, which in turn will affect turnover. Companies need to put in extra effort in terms of time and cost to fill the vacancy without significant loss to the ongoing business. To overcome these situations, this paper is looking into the possibility of predicting employee attrition on an individual level with machine learning and providing the organization's opportunities to address any issue and improve retention. In this paper, we attempted to predict employee attrition rate using Logistic Regression, Decision tree, Support Vector Machine, Random Forest, Naive Bayes, Machine learning, Data visualization, Feature selection.

KEYWORDS:Logistic Regression, Decision tree, Support Vector Machine, Random Forest, Naive Bayes, Machine learning, Data visualization, Feature selection, Prediction, Employee turnover.

I. INTRODUCTION

Attrition refers to the continuous reduction in the number of employees through the process of retirement, resignation or death. However, turnover may vary from sector to sector, depending on its criteria, conditions, and policies. Attrition rates are a severe problem that can be addressed for any industry or company that pushes for effective talent retention [1]. Despite the efficient operation of the industry, it faces a systematic problem of high labor costs, which in turn affects the performance of the industry. Therefore, companies in the Indian IT industry are focusing on job planning and staff development to retain them, which has become a critical success strategy. As a result of this situation, companies have announced several training and development programs to promote and maintain these programs. Employee turnover is the most significant loss of business as the company has to suffer due to high indirect costs. These costs include hiring to find replacements, costs associated with training and developing new employees, lost productivity due to the departure of experienced employees, and moral damage to the organization. New employees are also recruited and trained over a significant period, which takes up a lot of the organization's time and resources.

With evidence of issues within the IT sector, it's important for the management to require corrective actions to reduce worker turnover. Employee turnover has been found to be affected by several factors, including personal, work-related, and organizational factors. We can use the same factors to predict turnover within the organization. Performance and employee satisfaction also have a significant impact on attritions. For this motive, research should be conducted to understand the satisfaction level of employees and provide a quick solution to the issues for any potential reasons behind the high attrition rate.

The pressure on HR departments to provide value to the organization has led to the introduction of data-driven decisions and machine learning [2]. Using machine learning models can help organizations anticipate attrition. For example, analysts can build and prepare a machine learning model using historical data saved in HR departments. In addition, it can predict the workers leaving the organization. In this study, we implement some of the well-known data classification techniques, namely, Logistic Regression, Decision tree, Support Vector Machine, Random Forest, Naive Bayes, Machine learning, Data visualization, and Feature selection on the Human Resources (HR) Employee Attrition dataset.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

This paper follows the given structure: The second section describes previous relevant studies and the inspiration for the analysis. Section III explains the methodology adopted for conducting this research and then the various machine learning techniques used. The data set description is also presented in this section. Finally, the data visualisation and experimental results using machine learning methods on the given data set are presented in Section IV, which is followed by the conclusion in Section V.

II. RELATED WORK

A large part of human resource management is maintaining employee satisfaction. A study by Ryan, Schmit, and Johnson (1996) [3], showed that employee satisfaction is related to employee turnover. Higher satisfaction tends to lower employee turnover. Some ways in which employee turnover affects organizational effectiveness are: (1) when an employee leaves their tacit and explicit knowledge is lost, which decreases the overall knowledge within that particular organization, the productivity of the organization could be affected negatively, (2) increased workload on other employees, (3) hurt company morale, (4) lower turnover rates reflects that there will be less hiring which will lower costs such as activities for training and employment process (Koys, 2006; James & Mathew, 2012)[4]. In addition, employee turnover has tangible and intangible effects. In order to gain insight into the actual impact and how employee turnover affects competitive advantage, it is essential to consider them at the same time.

Machine learning has taken a long time to enter the HR field and was only used a few years ago. According to Linkedin [5], only 22% of companies have started analysis in the HR department, and it is unclear how well these analyzes are implemented. In areas such as marketing, we have extensive access to big data. For example, there will be data such as the number of products sold in a country, when they were sold, and the number of observations the product received.[6] Machine learning is a very effective tool for analyzing large amounts of data in these situations. That is because HR operations and outcomes affect organizational performance in different ways. These are operations such as onboarding new employees, training new and old employees, identifying good and bad performance, determining who should be promoted, employee retention, and employee benefits.[6]

In [7], they investigated the relationship between exit practices such as delays and absenteeism, job content, place of residence, socioeconomics, and worker turnover rate in rapidly developing market segments such as the Indian IT industry. The extraordinary thing about this study is that five prescient data mining techniques (artificial neural networks, logistic regression, classification and regression trees, classification trees (C5.0), and discriminant analysis) were used on the sample data of 150 employees in a big software organization. The withdrawal behavior and employee turnover are clearly related according to the survey results. This study raises some questions for future studies. Firstly, further research could explicitly collect data on statistical factors from a large sample of organizations to test the relationship between statistical factors and turnover. Second, extensive data can be collected on past academic research variables associated with staff turnover.

A discussion in [8] shows the application of a logistic regression method that relies on employee data to build a risk equation for predicting employee turnover. This equation was later applied to assess the risk of fluctuations in the current group of employees. After estimation, high-risk clusters were identified to discover the reason, and action plans were selected to minimize risk in the future [8]. IBM Watson Team in [9] conducted an analysis of the employee turnover process, clarified the purpose of the turnover, and proposed a structure to identify potential turnover. They calculate the cost of turnover, suggest employee names for the retention process, and try to compare the difference between the expected turnover cost before the retention period (EACB) and the expected turnover cost after the retention period (EACA). [9]

The evaluation in [10] uses the k-nearest neighbors algorithm predictions about whether an employee leaves the company. The features are considered to be characterized by employee performance, average monthly working hours, years spent at the company, and so on. Experimental result demonstrates that k-Nearest Neighbors outperforms well over its peer algorithms like Naïve Bayes, Logistic Regression, and Multi-layer Perceptron (MLP) Classifier. To prevent employee churn, well-known classifiers such as Decision tree, Logistic Regression, SVM, KNN, Random Forest, Naive Bayes methods on the human resource data are implemented in [11]. Feature selection methods are applied to the data, and the results have shown that employee attrition can be prevented.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

Researchers in [12] studied improving the quality of employee attrition prediction and decision support systems. They studied the impact of text data on the churn prediction process. They found that merging unstructured text data into traditional churn prediction algorithms can significantly improve prediction performance. Saradhi and Palshikar determines the employee turnover in [13] by using logistic regression, decision tree, Naive Bayes, and random forest methods. Kane-Sellers' report on the Fortune 500 Industrial Automation Manufacturers database is regarded as the best for professional salesforce. Logistic regression is the main method of Kane-Sellers in the most recent study. [14]

III. DATA AND METHODOLOGY

Apply a wide range of data mining techniques, from simple techniques such as Naive Bayes, linear regression, and nearest neighbors methods to more complex techniques such as SVMs, random forests, and other ensemble techniques. To resolve this issue, use the following steps to analyze employee data analysis and predict churn.

1) Identify and analyze the data set of an employee that comprises current and past records
2) Clean up the dataset, handle the missing data, and derive new features as needed.
3) Select the appropriate features for attrition prediction from employee data.
4) Try several classifications and compare the accuracy, recall, precision and F-measure results of the test data to determine the most appropriate classification.
5) Apply feature selection method, and select the features that are more helpful to foresee representative stir
6) Create a classification model
7) Further the prediction of employees churn when using the model and determine their retention.

Dataset Used

After understanding the essential features from various datasets like IBM Watson and HR Analytics available on Kaggle, we created a file containing these fields and sent it to three IT startups. The HR department collected and provided the data at the respective startup. Before aggregating, we redacted the personal details about employees. Sharing the dataset is not allowed as the identity of companies should not be compromised. The dataset contains information relating to HR from 851 employees with 17 highlights. Out of these employee details, there are 562 who stayed and are active, while 284 left the company. As part of data cleaning, we built a dictionary to categorize data columns after encoded values. Finally, they were given integer values instead of non-numerical values so that our machine learning model could interpret them.

	table id	Location	Emp. Group	Function	Gender	Tenure	Tenure Grp.	Experience (YY.MM)	Marital Status	Age in YY.	Hiring Source	Promoted/Non Promoted	Job Role Match	Stay/Left	satisfaction_level	salary
0	1	Pune	B2	IT	Male	0.00	< =1	6.08	Single	27.12	Direct	Non Promoted	Yes	Left	0.38	low
1	2	Noida	B7	Marketing	Male	0.00	< =1	13.00	Marr.	38.08	Direct	Promoted	No	Stay	0.80	medium
2	3	Bangalore	B3	IT	Male	0.01	< =1	16.05	Marr.	36.04	Direct	Promoted	Yes	Stay	0.11	medium
3	4	Noida	B2	Sales	Male	0.01	< =1	6.06	Marr.	32.07	Direct	Promoted	Yes	Stay	0.72	low
4	5	Pune	B2	Support	Male	0.00	< =1	7.00	Marr.	32.05	Direct	Non Promoted	Yes	Stay	0.37	low

Figure 1 : HR data set features and their datatype

Feature Selection

Features can be selected and used to build models with different subsets of the data and determine which features are relevant and which are not. We applied a feature selection method and selected the more convenient features for predicting employee churn. After analyzing the dataset manually, we concluded that these essential features are: Experience (YY.MM), Age in YY., Stay/Left, Satisfaction_level, Location, Promotion, Job Role Match, Hiring Source, Maritial_status, Salary type, Tenure Group, Operations type, Gender. Complex and irrelevant features can adversely affect the performance of the models. Therefore, feature selection and data cleansing should be one of the most important steps in designing a model.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

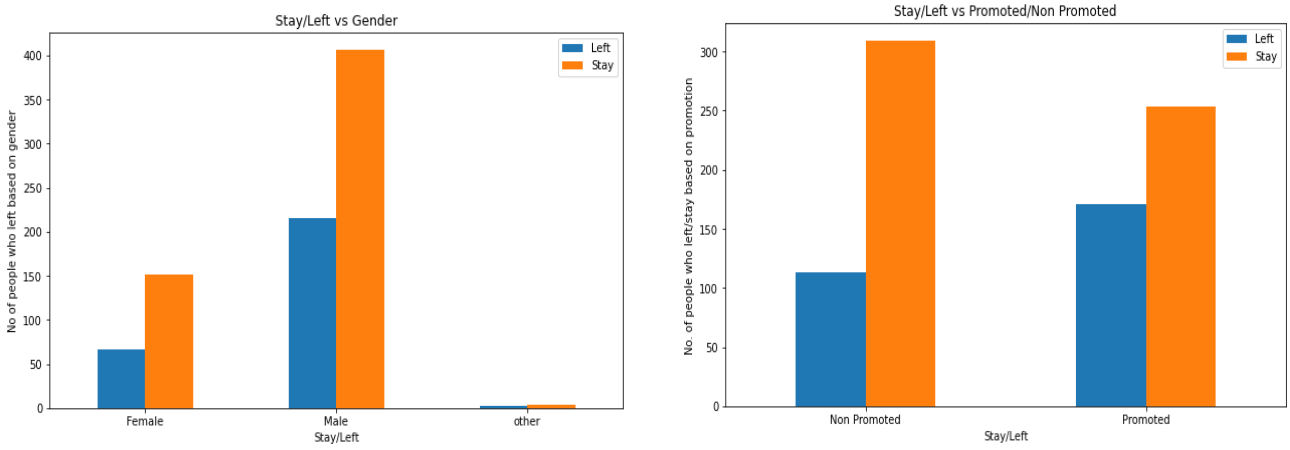
| DOI:10.15680/IJIRSET.2021.1012104 |

Model Building

The entire dataset was split into training data (80%) and test data (20%) to build and train the model for predicting employee turnover. We used the 'train_ test_split' module from the 'sklearn.cross_validation' package, and 0.2 value was assigned to the test_ size parameter. The feature values on which the models are trained are also present in training data. We will employ six predictive models 6-ML algorithms: Logistic Regression, Decision tree, KNN, SVM, Random Forest, and Naive Bayes. Accuracy and predictions are determined using the trained model on the unseen test dataset. Following this, we adapted the training data to hyperparameters. After fitting, class predictions were made on the test dataset, followed by the actual predictions calculated using the precision functions available in the metrics module.

IV. EXPERIMENTAL RESULTS

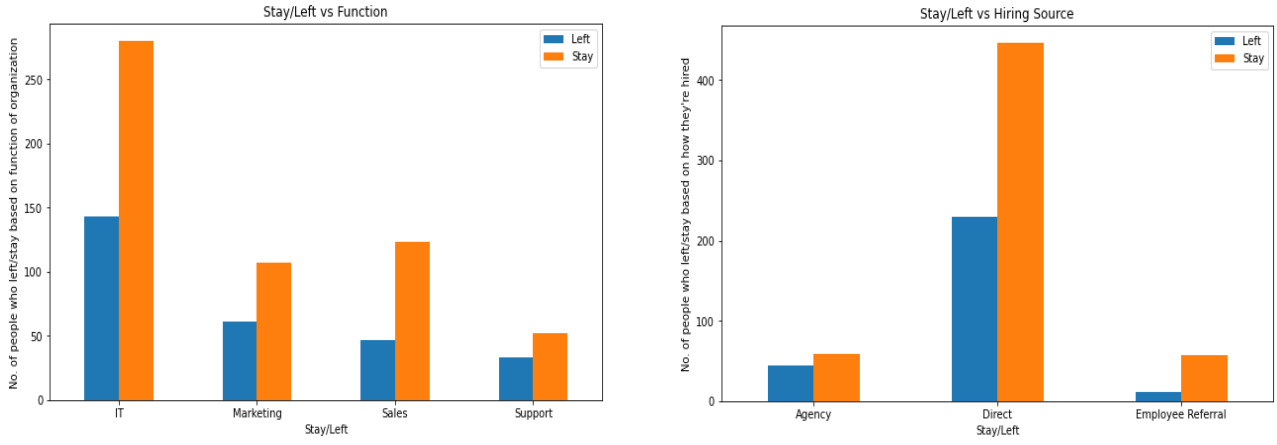
In this part, we perform analysis on a continuous and categorical variable to understand the relationship between the attributes and the output variable. There are many features available, but we cannot show each and every attribute's relation. For simplicity, the important factors are highlighted in the following figures.



(a) (b)

Fig. 2. (a) Shows the attrition based on gender ; (b) Employees who stayed or left on based of promotion

Using the gender and promoted/non promoted field, in Figure 2 for (a) and (b), we determine the relationship between these two features. As we can infer, better and often promotion in the field leads to less attrition. Therefore, it can be analyzed that the attrition rate is not entirely dependent on gender.



(a) (b)

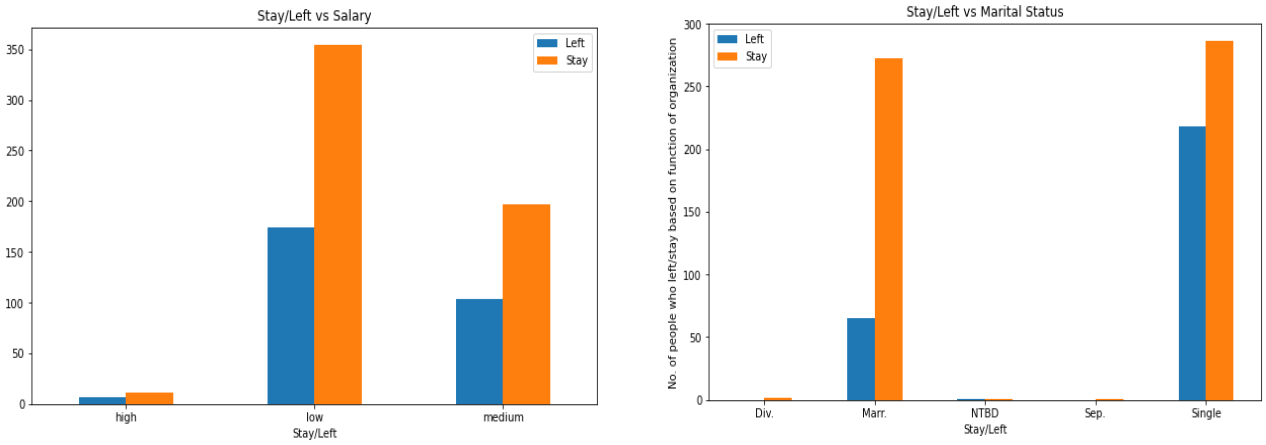
Fig. 3. (a) Attrition for various job types ; (b) People who left/stay based on how they were hired

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

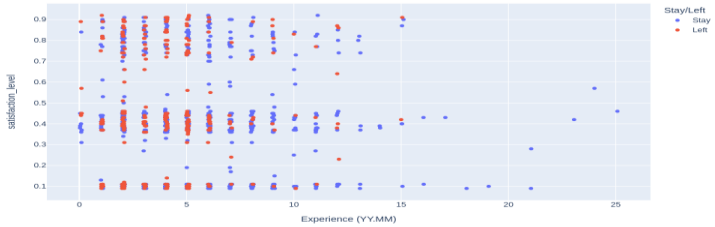
`Figure 3(a) depicts the attrition of employees based on the type of job they had. As data was aggregated from technology-based companies, IT professionals were the main highlight. In figure 3(b), most employees were appointed directly to the job, and interestingly, employees hired through agency had a higher percentage of attrition.



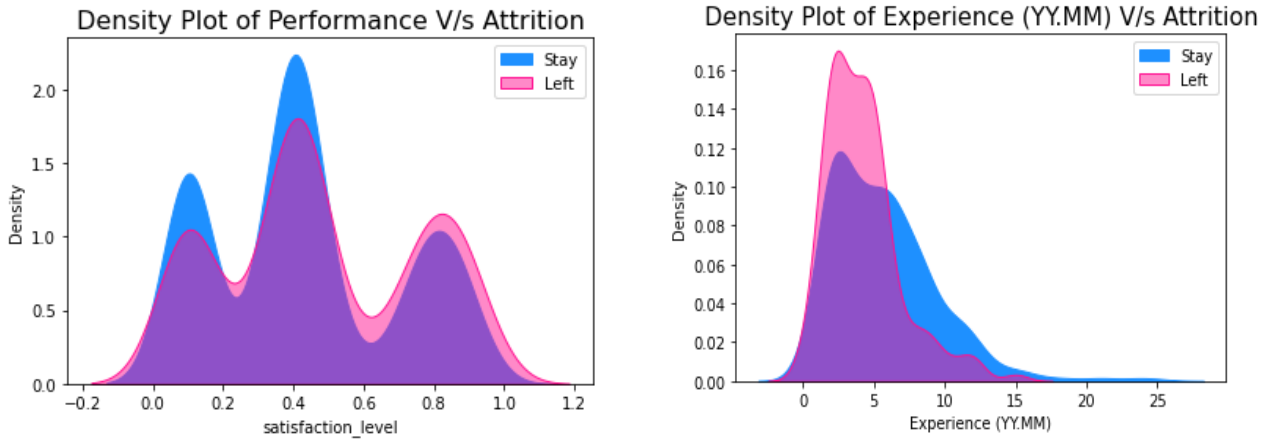
(a) (b)

Fig. 4 (a) Effect of salary on employee turnover; (b) Employee attrition due to their maritial status

As we can infer from Fig 4, the lower the employee's salary, the higher the chances that they will leave the company. Similarly, their marital status is a deciding factor in attrition as a single employee could take the risk by leaving the current job while married employees are risk-averse because they might have a family to support. Nevertheless, as the trends suggest, both substantially impact the attrition rate.



(a)



(b) (c)

Fig. 5 (a) Comparision of satisfaction level and experience of employees for leaving or stay (b) Attrition of employee due to performance (c) People who left/stay based on how they were hired

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

As shown in Figure 5(a), employees who left the company were less satisfied with their job, and their total experience was less than five years. Employees with more than six years of experience stay were inclined to stay. In Figure 5(b), Satisfaction compared to leaving employees can be used to determine that they are more likely to leave the company if satisfaction is 0.1 or less and 0.3 to 0.5. While in Figure 5(c), employees with less experience or overall time in the company are taking chances by leaving the company for better employment or might have been terminated by the employer. All these factors are the most important attributes that influence an employee's decision to leave the company.

COMPARATIVE ANALYSIS

The models are evaluated in terms of key performance measures. The results are summarized along with specified classifiers such as Logistic Regression, Decision tree, KNN, SVM, Random Forest, and Naive Bayes. Table 1 shows the summary results of the baseline classifier. We can see that Random Forest has 1% better accuracy than Logistic regression, but Random Forest is an overfitted model. Therefore, we will select Logistic regression as our final model.

	Logistic Regression	KNN	Support Vector Machine	Random forest	Decision tree	Naive Bayes
Accuracy	0.923529	0.576471	0.911765	0.935294	0.852941	0.841176

Table 1 : Result Analysis over implemented machine learning algorithm

V. CONCLUSION

This paper discussed and shared predictive analyses on employee attrition by effective feature selection. The results of this study also support surveys conducted in predicting employee turnover. It also supports the findings of Nagadevara in [7] that it is possible to predict employee attrition even before they have made their final decision. One of the limitations of this study was that organizations were hesitant to publish data for research purposes. Therefore, organizations wishing to take advantage of such research must provide the data needed to implement the proposed model efficiently.

Although these datasets might differ slightly from other real-world employee profiles and turnover datasets for small, medium, or large organizations, the datasets used in this research provide a framework for performing complete comparative analysis using different machine learning algorithms. This study was limited to a small data set that did not properly train the model, which could give poor results, and it is not easy to get employees' data from an organization because it is confidential.

This project can be expanded as it has much potential for improvement by applying deep learning techniques with a well-designed network of sufficient hidden layers on a large data set that can cover this project's boundaries. There can be time series and trend analysis, which might improve forecast performance if the data is in date format.

REFERENCES

[1] Anne J, Talapatra K, Rungta S, Jagadeesh A. Employee Attrition and Strategic Retention Challenges in Indian Manufacturing Industries : a Case Employee Attrition and Strategic Retention Challenges in Indian Manufacturing Industries : a Case Study. VSRD Int. J Bus. Manag. Res. 2016; 6:8

[2] Tomassen, M. E. 2016 Exploring the Black Box of Machine Learning in Human Resource Management. An HR Perspective on the Consequences for HR professionals. http://purl.utwente.nl/essays/70791

[3] Ryan, A. M.; Schmit, M.J. & Johnson, R. 1996. Attitudes and effectiveness: Examining relations at an organizational level. Personnel psychology 49(4). 853-882. https://doi.org/10.1111/j.1744-6570.1996.tb02452.x

[4] Koys, J. D. 2006. The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness; a unit level, longitudinal study. Personnel Psychology 54(1). 101-114. https://doi.org/10.1111/j.1744-6570.2001.tb00087.x

[5] Linkedin. 2018. The Rise of Analytics in HR - The era of talent intelligence is here.

[6] Cappelli, P.; Tambe, P. &Yakubovic, V. 2018. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. SSRN Electronic Journal. DOI: 10.2139/ssrn.3263878

[7] Nagadevara, V., Srinivasan, V. &Valk, R. (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques, Research and Practice in Human Resource Management, 16(2), 81-99.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012104 |

[8] Khare, Rupesh, et al. "Employee attrition risk assessment using logistic regression analysis." IIMA International Conference on Advanced Data Analytics, Business Analytics. 2015.
[9] Singh, Moninder, et al. "An analytics approach for proactively combating voluntary attrition of employees." 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE, 2012.
[10] Yedida R, Reddy R, Vahi R, Jana R, GV A, Kulkarni D. Employee Attrition Prediction, 2018.
[11] Shankar RS, Rajanikanth J, Sivaramaraju VV, Vssr Murthy K. Prediction of Employee Attrition Using Datamining, IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCA, 2018, 1-8. doi: 10.1109/ICSCAN.2018.8541242.
[12] Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management, 45(3), 164-174
[13] Saradhi, V. V., &Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999-2006.
[14] Kane-Sellers, M. L. (2007). Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis. Texas A&M University.
[15] Abe, S. (2010a). Introduction. In Support vector machines for pattern classification (S. Singh, Ed., 2nd ed., pp. 1–19). London: Springer Science & Business Media. (2010b). Support vector machines for pattern classification. London: Springer Science & Business Media.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details