

## **Augmentation of Human Resource Management leveraging Machine Learning Paradigms: An Analytical Study within the Landscape of Information Technology Entities**

**Dipanwita Dey**

Research Scholar, School of Management  
Swami Vivekananda University

&

**Dr. Shantanu Chakraborty**

Professor & Director (Research), School of Management  
Swami Vivekananda University

### **Abstract:**

The information technology (IT) industry heavily depends on the vital function of human resources (HR). With a diverse workforce possessing varying levels of expertise and skill sets, the sector manages its operations efficiently. The integration of machine learning enhances several HR processes, including talent acquisition, candidate matching in recruitment process and employee retention, empowering HR leaders to make more informed, data-centric decisions. This research delves into the potential application of machine learning techniques to enhance various stages of the HR lifecycle. By employing Natural Language Processing based unsupervised machine learning algorithm in candidate matching, a robust and meticulous selection process is established to identify the ideal candidate during recruitment process, aligning with job description. This not only aids in cost-effective and time-efficient recruitment but also mitigates the need for frequent replacements and training efforts. Departing employees often take with them valuable tacit knowledge, a critical source of competitive advantage. To reduce attrition, a machine learning-driven approach is implemented to pinpoint employee attributes contributing to the prediction of attrition

within organizations. The study draws upon human resource data from a service-oriented IT company to derive meaningful insights.

**Keywords:** Employee Attrition, Machine Learning, Human Resources, LDA

### Introduction:

Despite subdued sentiment and extended decision-making cycles, India's IT exports are projected to outpace the country's economic growth in the fiscal year 2022-23. According to The Information Technology industry body NASSCOM, India's export revenue in the IT sector is anticipated to surge by 11.4% in constant currency terms, reaching \$194 billion. In reported terms, the growth stands at 9.4%, with a cross-currency impact of 2%. The overall Indian tech industry is forecasted to reach \$245 billion in the same fiscal year, experiencing an incremental revenue increase of \$19 billion. The country's GDP is expected to grow at a rate of 7% during this period. Proper Recruitment and minimization of attrition is the most challenging task for HR department of IT service company. During recruitment Candidate matching from the bulk of resumes is considered to be one of the laborious, error prone tasks for HR department. Candidates fill up the internal company's profile tools by their experience, skillset, educational qualification. By using unsupervised algorithm Topic Modelling, candidate profiling can be done before the interviews as shown in Fig 1. Candidate profiling discovers two points a) Retrieving data from candidate resumes, b) Assigning the resume to a particular job description. High employee attrition rate poses a significant challenge for IT service company. By using different machine algorithms, the attributes of employee records which are responsible for attrition can be found out.

Fig 1. Standard Recruitment Process



## Literature Review:

Riedl and Biemann in their studies “Text segmentation with topic models” explore a common approach to leverage insights gained from the Latent Dirichlet Allocation (LDA) topic model in text segmentation involves employing topic assignments instead of individual words in established text segmentation algorithms like Text Tiling and C99. This alternative strategy has been demonstrated to yield noteworthy enhancements. Fang (2015) conducted an experiment illustrating that by employing this supervised training procedure, a re-ranking model can be generated, showcasing substantial improvements over conventional search ranking methodologies. The performance metrics for information retrieval, commonly used to assess search outcomes, exhibited significant advancements with this approach.

Shao and Qin in their studies “Text similarity computing based on LDA topic model and word co-occurrence” describe LDA (Latent Dirichlet Allocation) topic model has found extensive use in text clustering due to its effective dimension reduction capabilities. The common approach involves representing a set of texts using the LDA topic model, making inferences through methods like Gibbs sampling, and assessing text similarity using Jensen-Shannon (JS) distance. Nevertheless, JS distance has limitations in distinguishing semantic associations among text topics. To address this drawback, a novel text similarity computation algorithm is introduced. This algorithm is based on a hidden topics model and analysis of word co-occurrences, providing an alternative and potentially more nuanced perspective on the relationships between textual content.

Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, “Employee turnover prediction with machine learning: a reliable approach,” in their studies employed supervised machine learning techniques to assess human resources, specifically focusing on estimating employee turnover

within an organizational setting. Utilizing various methods such as decision tree systems, random forest techniques, gradient boosting trees, extreme gradient boosting, and logistic regression, they conducted computational tests on both real and simulated human resource datasets. These datasets represented organizations with diverse employee populations, ranging from small to medium and large enterprises. The study demonstrates the efficacy of machine learning in identifying key factors within the realm of human resources.

A. Chourey, S. Phulre, and S. Mishra, in their studies “A survey paper on employee attrition prediction using machine learning techniques,”

illustrates the analysis of human resource attrition has become increasingly pivotal. It emerges as a pervasive challenge across all organizational landscapes, marked by a consistent decrease in employee numbers through retirement, resignation, or unfortunate demise. Leveraging machine learning techniques proves remarkably effective in predicting dataset variations. The retention of talent and professional staff poses a critical quandary for HR managers, particularly within the manufacturing sector. This study has pinpointed intricate factors as the primary culprits responsible for employee turnover in specific organizations. Creating an environment conducive to employee retention involves fostering an ethical work culture, cultivating harmonious employee relationships, and diligently implementing organizational policies.

**Objective:** This study aims at finding two most crucial human resource process solution by using machine learning algorithm. Main objectives of the research paper are as follows: -

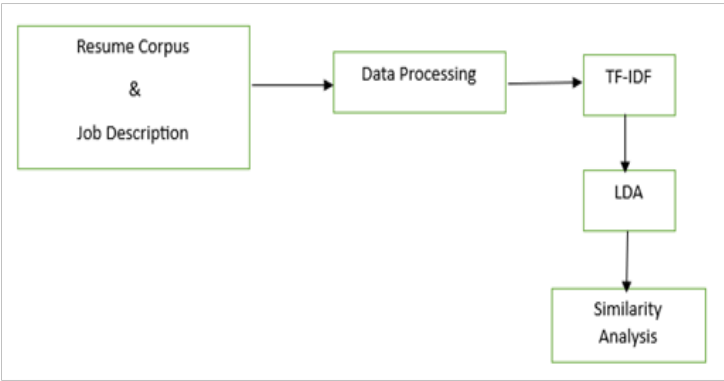
- To match candidates by extracting information from profiles and aligns with their specific job description.
- To determine the factors contributing to employee turnover and precisely identify those employees who are likely to leave in the coming future.

**Methodology:**

**A) Candidate Matching**

During recruitment process, candidate matching can be doneas shown by the Fig 2.

Fig 2. Process Flow for Candidate Matching during recruitment



**Resume Corpus:**

- a. Data Source:The resumes are collected from candidate job portal in free text form.
- b. Data Template: The schema for raw dataset is likeFig 3
- c. Fig 3. Raw Dataset used for Resume Corpus

| ID | Text                                                                                                         |
|----|--------------------------------------------------------------------------------------------------------------|
| 0  | Experience with AWS ecosystem .Proficient with different type of tools such as Terraform,Python,Kubernetes   |
| 1  | Total 3-5 years of experiece in test planning,test case,business analysis.Currently working in service based |

- a. Data Size: Total number of records collected from candidate is 1000

- **Job Description:** There are two job description for recruitment process:

- a) Job Description 1: Designing and implementing cloud architecture solutions based on industry best practices. Implement Infrastructure as Code (IaC) practices using tools such as Terraform or AWS CloudFormation. Demonstrate expertise in at least one major cloud platform (e.g., AWS, Azure, Google Cloud) and the ability to adapt to other platforms as needed. Proficiency in automation tools such as Terraform, Ansible, or CloudFormation. Experience with containerization and orchestration tools (e.g., Docker, Kubernetes). Familiarity with DevOps practices and tools.

- b) Job Description 2: Proven experience as a Testing Manager or in a similar leadership role. Strong understanding of software testing principles, methodologies, and tools. Experience with test automation frameworks and tools. Establish and track key performance indicators (KPIs) to measure the effectiveness of the testing process. Certification in software testing (e.g., ISTQB) is a plus.

- **Data Processing:** The data processing of the resume corpus can be done in three steps as following.

- a) Tokenization: -This method is breaking down a text into smaller units called tokens.

- b) Stop word Removal: - This method will remove all the preposition and punctuation from list of tokens

- c) Lemmatization: -This method will reduce words to their base or root form, known as the lemma. The lemma is the canonical, dictionary form of a word.

As for example, studying to studies, walking to walk

- **TF-IDF:** This is a popular method in language processing and finding information. It helps us figure out how significant a word is in a document compared to a bunch of documents (a collection or corpus). To do this, words in a text document are turned into numbers

that show their importance through a process called text vectorization. TF-IDF is a common way to score this importance, and there are other scoring methods too.

**Term Frequency (TF)** - TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

**Inverse Document Frequency (IDF)** - IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

- **LDA:** LDA stands for Latent Dirichlet Allocation, and it is a probabilistic model used in natural language processing for topic modelling. Topic modelling is a technique that helps identify topics present in a collection of texts. LDA assumes that documents are mixtures of topics and that each topic is a mixture of words.

The following steps are carried out to assign topic (Job Description) to each of the document (Resume)

- 1) Randomly assign a topic from the set of  $K$  pre-defined topics to each word within every document.

- 2) For each document  $d$ :

For each word  $w$  in the document and each topic  $t$ , compute:

- a)  $P(\text{topic } t \mid \text{document } d)$  = Proportion of words in document  $d$  that are assigned to topic  $t$

- b)  $P(\text{word } w \mid \text{topic } t)$ : Proportion of assignments to topic  $t$  across all documents from words that come from  $w$

- 3) Modify the assignment of topic  $t$  to the word  $w$  with a likelihood determined by the product of the conditional probability  $p(t' \mid d)$  and  $p(w \mid t')$  for all other words and their respective topic assignments in the document  $d$ .

Iterate through the final step several times until a stable state is attained, wherein the allocations of topics cease

to undergo further changes. Subsequently, ascertain the distribution of topics for each document based on these finalized topic assignments.

- **Similarity Analysis:**The proportion of topics that have been assigned to the each of document is shown in graph(Details shown in Fig 6.)

B) **Employee Attrition Minimization**

Fig 4. Process Flow for Employee Attrition Minimization



- **Raw Data:**
  - a. Data Source: The existing employee data are collected from Human Resource department.
  - b. Data Template: The schema for raw dataset is like Fig 5.

Fig 4. Raw Dataset used for Employee Attrition

- c) Data Size: Total number of records collected from HR department is 1000000

| Emp ID  | Emp Name     | Emp Location | DoI        | Function | Gender | DoB        | Marital Status | Promoted | Job role | Stay/Left |
|---------|--------------|--------------|------------|----------|--------|------------|----------------|----------|----------|-----------|
| 2010942 | Meenakshi J  | Pune         | 16-04-2002 | Tech     | F      | 16-01-1980 | Married        | YES      | YES      | Stay      |
| 2017561 | Samuel Jhon  | Bangalore    | 17-04-2018 | Tech     | M      | 18-05-1996 | Single         | YES      | NO       | Left      |
| 2117531 | Bose Anindya | Kolkata      | 18-04-2020 | Sales    | M      | 29-06-1990 | Married        | NO       | YES      | Stay      |

- **Explanatory Data Analysis:** This type of analysis identifies the general pattern of data. This will identify the

attributes which can be deciding factor for attrition. Details is shown in analysis section.

- **Data Processing:** The data processing can be done in the following steps:

- a. **Missing Value Treatment:** Replacing the missing record with corresponding column of mean/median. Dropping those columns which has more than 10% of Null value.

- b. **Dummies Variable:** All categorical variables are converted into dummies variable. As for example "Marital Status" has two values "Single", "Married". It can be done by encoding 1 and 0

- **Model:** Three classification algorithms are used to identify those employees who are likely to leave in the coming future. Algorithms are as Logistic Regression, Random Forest, Decision Tree. The input for the table is cleaned data and output is in binary 1 or 0 (Left/Stay)

a) **Logistic Regression:** Logistic regression is a statistical method used for binary classification, which means predicting the outcome of a categorical dependent variable that has two possible values (usually 0 and 1). It is widely used in various fields, including machine learning, statistics, and epidemiology. In logistic regression, the dependent variable is modelled as a function of one or more independent variables using the logistic function (also known as the sigmoid function). The logistic function maps any real-valued number to a value between 0 and 1, making it suitable for representing probabilities. The equation for logistic regression is as follows:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)}}$$

Here:

- $P(Y = 1)$  is the probability that the dependent variable  $Y$  is 1.
- $e$  is the base of the natural logarithm.
- $b_0$  is the intercept term.
- $b_1, b_2, \dots, b_k$  are the coefficients associated with the independent variables

a) **Decision Tree:** Decision tree is a predictive modelling tool used in machine learning and data mining for both classification and regression tasks. It is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

a) **Root Node:** The topmost node in the tree, representing the entire dataset. It is divided into subsets based on the values of a feature.

b) **Internal Nodes:** Nodes in the tree that represent decisions based on the values of features. Each decision node has branches that lead to child nodes based on different feature values.

c) **Leaf Nodes:** Terminal nodes at the ends of the branches that represent the final outcomes or predictions. Leaf nodes contain the predicted class (in classification) or the predicted value (in regression).

d) **Branches:** Connections between nodes, representing the possible outcomes based on the values of features.

e) **Splitting:** The process of dividing a node into two or more child nodes based on a certain criterion. The goal is to create subsets that are more homogeneous with respect to the target variable.

f) **Criterion:** The measure used to decide how to split a node. Common criteria include Gini impurity for classification and mean squared error for regression.

b) **Random Forest:** Random Forest is a popular ensemble learning method used in machine learning for both classification and regression tasks. It is an ensemble of decision trees, where each tree is trained on a random subset of the data and makes independent predictions. The final prediction of the Random Forest is typically the average (for regression) or the majority vote (for classification) of the individual tree predictions.

Here's how Random Forest works:

a) **Bootstrap Sampling (Bagging):** Random Forest builds multiple decision trees by repeatedly sampling (with replacement) from the original dataset. Each sample is used to train a separate decision tree.

b) **Feature Randomization:** In addition to sampling data, Random Forest also introduces randomness in the features used for splitting nodes in each decision tree. Instead of considering all features at each split, it randomly selects a subset of features. This helps in creating diverse trees and prevents overfitting.

c) **Independent Trees:** Each tree in the Random Forest is grown independently of the others. This means that the decision trees are not correlated, and they can make different predictions based on their individual training samples and feature subsets.

d) **Voting or Averaging:** For classification tasks, the Random Forest combines the predictions of individual trees by majority voting. For regression tasks, it averages the predictions of individual trees. This ensemble approach often leads to more robust and accurate predictions compared to individual decision trees.

- **Evaluation of Model:**

The evaluation of classification technique is done by confusion matrix.

**Confusion Matrix:** A confusion matrix serves as a tabular representation designed to assess the efficacy of a classification algorithm. It serves the purpose of visually presenting and condensing the overall performance metrics of a classification algorithm. There are four measurement metrics of classifier.

i) **True Positive (TP):** Instances where the model correctly predicts the positive class.

ii) True Negative (TN): Instances where the model correctly predicts the negative class.

iii) False Positive (FP): Instances where the model incorrectly predicts the positive class (Type I error).

iv) False Negative (FN): Instances where the model incorrectly predicts the negative class (Type II error).

There are four performances of the classifier

i) Accuracy: The ratio of correctly predicted instances (TP + TN) to the total number of instances.

ii) Precision (Positive Predictive Value): The ratio of correctly predicted positive instances (TP) to the total predicted positive instances (TP + FP).

iii) Recall (Sensitivity or True Positive Rate): The ratio of correctly predicted positive instances (TP) to the total actual positive instances (TP + FN).

iv) Specificity (True Negative Rate): The ratio of correctly predicted negative instances (TN) to the total actual negative instances (TN + FP).

From the Fig 5., it is clear that Random Forest algorithm is the best one.

|          | Decision Tree | Random Forest | Logistic Regression |
|----------|---------------|---------------|---------------------|
| Accuracy | 0.84578       | 0.8795        | 0.8598              |

Fig 5. Comparison of Accuracy for three algorithms

### Kolmogorov-Smirnov Table:

The Kolmogorov-Smirnov table is a statistical reference tool used to find critical values for the Kolmogorov-Smirnov test, assessing the goodness-of-fit between observed and expected distributions. It helps determine if a sample comes from a specific distribution. The table provides critical values based on sample size and significance level.

A typical KS table can be constructed with the following measurement:

- i) **Count:** Number of Observation in each group
- Events, non-Events:** Number of events and non-events in each group
- ii) **Expected Event Rate:** Average predicted probability of Event for each Group
- iii) **Actual Event Rate:** Number of Events / Number of Observation
- iv) **Cumulative Event, Cumulative Non-Event:** Cumulative summation of Events and non-Events respectively
- v) **Cumulative Event Rate:** Cumulative Event / (Cumulative Event+ Cumulative Non-Event)
- vi) **Percentage Event Capture:** Cumulative Event/Total Events
- vii) **Percentage non-Event Capture:** Cumulative Non-Event/Total Events
- viii) **KS Measure:** Percentage Event Capture-Percentage Non-Event Capture

**Analysis and Result:**

**A) Candidate Matching**

| Resume   | Similarity With Job Description 1 | Similarity With Job Description 2 |
|----------|-----------------------------------|-----------------------------------|
| Resume 1 | 56%                               | 7%                                |
| Resume 2 | 30%                               | 2%                                |
| Resume 3 | 80%                               | 10%                               |
| Resume 4 | 6%                                | 60%                               |
| Resume 5 | 10%                               | 12%                               |
| Resume 6 | 7%                                | 83%                               |
| Resume 7 | 8%                                | 54%                               |
|          | Shortlisted for Job Description 1 |                                   |
|          | Shortlisted for Job Description 2 |                                   |

Fig 6. Similarity of Resume with Job Description

In Fig 6, similarity of Resume with Job Description 1 and Job Description 2 is calculated using LDA algorithm. Assuming the threshold 50%, Resume 2, Resume 5 can be rejected as similarity with both job description is less than 50%. Resume 1 and Resume 3 are shortlisted for Job Description 1. Resume 4, Resume 6 and Resume 7 are shortlisted for Job Description 2.

## B) Employee Attrition Minimization

- 1) Find out the attributes affecting Employee Attrition

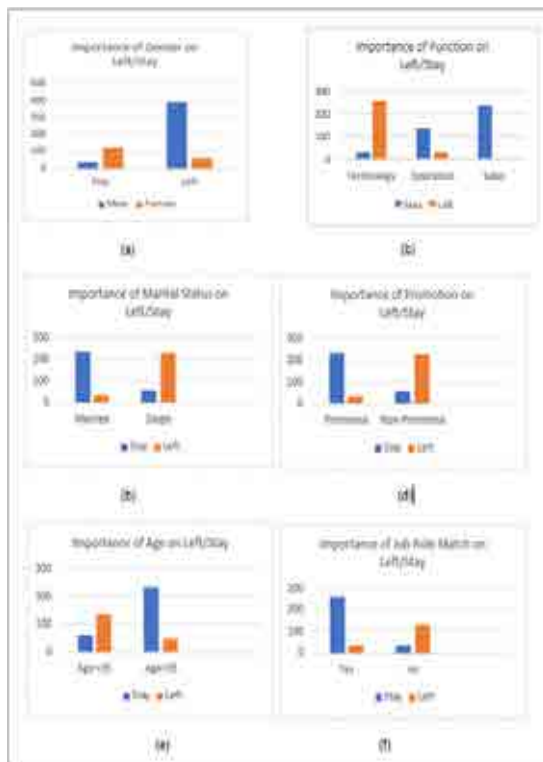


Fig 7. Effects of Different Attributes on Employee Attrition

The attributes effecting the employee attrition are shown in Fig 7

- a) The attrition rate of female employee less then male employee
- b) Employee working in Technology function has high tendency of attiraion than other function
- c) Married people left company less than single people
- d) Employee who get promotion , has less tendency to leave the company
- e) Employee of younger age(less than 35) has higher tendency of attritation
- f) Employee whose job role is matched has less tendency to leave the company

2) Kolmogorov-Smirnov Table Analysis

| Groups | Count   | Event | non-Event | Expected Event Rate | Actual Event Rate | Count Event | Count non-Event | Count Event Rate | % Events | % non-Events | KS-Measure |
|--------|---------|-------|-----------|---------------------|-------------------|-------------|-----------------|------------------|----------|--------------|------------|
| 1      | 100000  | 24910 | 75090     | 17.05%              | 24.91%            | 24910       | 75090           | 24.91%           | 36.03%   | 7.13%        | 28.90%     |
| 2      | 100000  | 23870 | 76130     | 16.09%              | 23.87%            | 48580       | 151420          | 24.28%           | 37.02%   | 14.00%       | 41.02%     |
| 3      | 100000  | 13610 | 86390     | 15%                 | 13.61%            | 62190       | 237810          | 20.72%           | 20%      | 16.96%       | 53.02%     |
| 4      | 100000  | 8180  | 91820     | 14.54%              | 8.18%             | 70330       | 129660          | 17.59%           | 71.34%   | 28.67%       | 45.45%     |
| 5      | 100000  | 5650  | 94350     | 13%                 | 5.65%             | 78000       | 424000          | 15.20%           | 74.67%   | 56.99%       | 17.88%     |
| 6      | 100000  | 4190  | 95810     | 12.40%              | 4.19%             | 80190       | 519810          | 13.37%           | 79.67%   | 66%          | 11.67%     |
| 7      | 100000  | 2590  | 97410     | 11.70%              | 2.59%             | 82780       | 817220          | 11.83%           | 84.23%   | 77.66%       | 6.57%      |
| 8      | 100000  | 2400  | 97600     | 10.30%              | 2.40%             | 83180       | 714820          | 10.85%           | 86.45%   | 79.98%       | 6.47%      |
| 9      | 100000  | 2070  | 97930     | 4.56%               | 1.40%             | 87250       | 812750          | 9.69%            | 89.09%   | 85.99%       | 3.10%      |
| 10     | 100000  | 1234  | 98766     | 0.90%               | 1.23%             | 88484       | 911516          | 8.85%            | 100.00%  | 100%         | 0.00%      |
|        | 1000000 | 88404 | 911596    | 9.15%               | 88.48%            |             |                 |                  |          |              |            |

Fig 8. KS table construction to distinguish between Left and Stay

- Here Events and Non-Events are considered as “Left” and “Stay” respectively and attrition captured in random is 10 %
- KS-Measure determines the discriminating power between Events and non-Events. It attains highest at Group 3 and then decrease to zero (blue Marked). Group 1-3 is considered as the “Left” dataset. (Yellow Marked)
- First three group contain only 30% of data, but it is able to capture 70% of Events (Red Marked). Without using ML model, only 10% of Events is captured

## Conclusion:

Hiring the right candidate and minimizing employee attrition are two most tactical challenges for Human Resource department for any knowledge-based IT company. By using this Machine Learning based approach, following advantages are achieved.

- Interviewers are shared with relevant resumes for evaluation after proper scanning. This helps them to select most appropriate candidates
- After identifying the key factors for attrition, different programs are adopted by the company such as career enhancement series, women leadership program etc
- The employee group identified for probable attrition are offered promotion, better job role etc

## Reference:

‘HR in the IT Industry: Importance, Trends, Best Practices’. (2023, March 16). Darwinbox. <https://blog.darwinbox.com/hr-in-the-it-industry>

Alao, D. and Adeyemo, A. B. (2013). ‘ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS’. Computing, Information Systems, Development Informatics and Allied Research Journal 4(1). <https://iiste.org/Journals/index.php/CIS/article/view/10148>

Phadnis, S. (2023). ‘Indian IT sector to touch \$245 billion in FY23: Nasscom’. The Times of India 2 March. <https://timesofindia.indiatimes.com/business/india-business/indian-it-sector-to-touch-245-bn-in-fy23-nasscom/articleshow/98344616.cms?from=mdr>

Riedl, M. and Biemann, C. (2012) ‘Text segmentation with topic models’, Journal for Language Technology and Computational Linguistics, Vol. 27, No. 1, pp.47–69.

Shao, M. and Qin, L. (2014) ‘Text similarity computing based on LDA topic model and word co-occurrence’, 2nd International Conference on Software Engineering, Knowledge

Engineering and Information Engineering (SEKEIE 2014), Atlantis Press, pp.199-203.

Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu. (2019).“Employee turnover prediction with machine learning: a reliable approach,” in Proc. of SAI Intelligent Systems Conference, 2018, pp. 737-758, doi: 10.1007/978-3-030-01057-7\_56.

A. Chourey, S. Phulre, and S. Mishra, “A survey paper on employee attrition prediction using machine learning techniques,”*Journal of Interdisciplinary Cycle Research*, vol. 11, no. 12, pp. 199-202, 2019. [Online]. Available:<http://www.jicrjournal.com/gallery/24-jicr-december-2247.pdf>

Ujlayan,A. and Sharma,M.(2019).“A study to enhance candidate screening process using similarity analysis”.*Int. J. Business and Data Analytics* , Vol. 27, No. 1

Marakani,S.(2019).“Employee Matching Using Machine Learning Methods”.*Master of Science in Computer Science*. <https://www.diva-portal.org/smash/get/diva2:1338281/FULLTEXT02>

Thomas,W.Edgar and David,O.Manz.(2017).‘Research Methods for Cyber Security’.*ScienceDirect*.

<https://www.sciencedirect.com/topics/computer-science/logistic-regression>

Sarang,N.(2018).‘Understanding Confusion Matrix’.

<https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

Bhalla,D.(2019). ‘CALCULATE KS STATISTIC WITH PYTHON’.<https://www.listendata.com/2019/07/KS-Statistics-Python.html>