Bima Journal of Science and Technology, Vol. 7 (1) Mar, 2023 ISSN: 2536-6041

DOI: 10.56892/bima.v7i01.411

# APPLICATION OF ADAPTIVE CROSS VALIDATION AND PRINCIPAL COMPONENT ANALYSIS OPTIMIZATION FOR EMPLOYEE TURNOVER PREDICTION USING ENHANCED GRAPH EMBEDDING

ABDULLAHI JIBRIL ABDULLAHI, NASIMA IBRAHIM and AISHA FARIDA AHMAD

Department of Computer Science, Kano state polytechnic, Kano, P.M.B.3401, Kano, Nigeria.

email: ajabdullahi.kanopoly.cs@gmail.com

## ABSTRACT

A rapid disclosure of an experience employee from organization (known as employee turnover), had necessitate organizations to make use of the recent development of information technology, to model their data collected over a period of time to enhance decision making, considering that, employee turnover continue to degrade performance of their organizations., Sometime, it requires much time and money for organization to get the equivalent replacements and get them train., consequently, predicting the likely hood of an employee to resign will allow the organization to take proactive measures to control the losses and costs., To address this problem, previous researches focus on exermining some impact factors., In this study, we consider modelling employee's job historical data to form a dynamic bipartite graph between employees and organizations and learnt a vector representation of this graph, we achieved this by developing a model that generates a sequence for each vertex in the graph using a horary random walk (HRW) method and input the sequence to a skip-gram-negative-sampling (SGNS) to obtain the vector representation for each vertex., and add this vectors to the employee's basic information and use it as input to machine learning classifiers to predict the employee's turnover., specifically , we proposed a graph embedding and prediction model that investigate the role of cross validation in predicting employee turnover called *Enhanced Graph Embedding and Prediction (EGEP)., Moreover*, an experimental result indicated that our method had significantly enhanced the prediction performance of the employee turnover with 11.09%, 11.06% and 13.93% in precision, recall and F1 respectively.

**Keywords:** Cross validation, Graph embedding, Bipartite graph, Horary random walk**.**

## INTRODUCTION

According to Yadav, (2018), employee turnover is defined as a disclosure or leaving of an intelligent skill personal from industry or organization. With the development of information technology, particularly with the implementation of human resource information systems (HRIS) in organizations, has triggered human resource managers and professionals to further make use of the data collected over time to enhance decision making (Xu et al. 2018).

Researchers have proven the significance of machine learning in some areas like commerce, finance, and health to reach a better group of customers or reduce the risk of investment for better performance (X. Cai et al., 2020; Stamolampros et al., 2019). But the use of advanced data analytics in the human resources (HR) area is still lacking (Jain & Nayyar, 2018).

This drives our motivation to further explore this area by analyzing the data to predict employee's turnover behaviour, as it became one of the most problematic aspect of organizational HR efficiency encountered over some time and as well difficult to introduce a noticeable measure to avoid in the organization's skilled workforce (Alao, 2013).

DOI: 10.56892/bima.v7i01.411

Additionally, the cost of employee turnover is ranging from 1.5 to 5 times the employee's annual pay sum, depending on how difficult it is to make his/her replacement(H. Cai et al., 2018) . Sometimes, it demanded a couple of times and money for the organization to get the equivalent replacements and get them trained. As such, ability to forecast the likely hood of an employee to resign will allow the organization to take appropriate action to minimize the cost.

In this work, we emphasize on voluntary turnover of the employee, because the involuntary turnover is mainly triggered by the organization human resources section.

There exist lots of research concerning employee turnover behaviour, to mention a few; Many researchers like Dave et al., (2018) and Xu et al., (2015, 2018), used Machine learning techniques to classify and predict the employee turnover at different organizational sectors, while some used survey to figure out the factors that lead to an employee's turnover decision(Chourey, 2019; Punnoose, 2016).

However, previous researches focus mostly on the impacting factors like employee's demographic information while ignoring the employee's working experience network structure that contain vital information about the employee behaviour., This in turn, may lead to the loss of some vital information that when incorporated may aid the prediction performance.

On the other hand, the network-based researches (Feeley et al., 2010; Vardaman et al., 2015..), are mainly restricted to the use of static network features, and because employee's working historical data is mainly in sequence, static network-based features cannot maintain the time-related information contained in the employee's work experience (X. Cai et al., 2020), while others are computationally expensive.

In consideration of the above limitation, we propose a graph embedding technique that lower the computational complexity and enhance performance. This decision become possible considering that, employee's working experience record can be depicted using a bipartite graph through partitioning the vertices in-to employee's and organization's vertices respectively as shown in the Figure 1 bellow

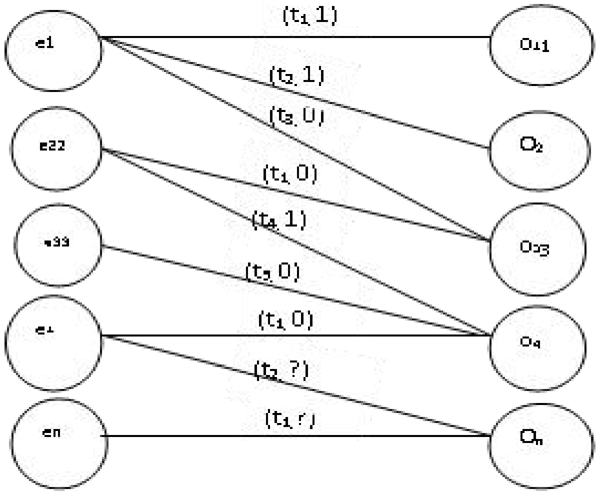


**Figure 1:** Dynamic Bipartite Graph

To address this problem, we proposed an Enhanced graph embedding and prediction (EGEP) model. Furthermore, for us to dealt with the graph embedding problem, we employed a Horary Random Walk (HRW), proposed by Cai et al., (2020) (which consider the time dependence information when carrying out the random walk) to generate a sequence for each vertex in the bipartite graph.

This sequence (the output of the HRW), is then inputted in to a Skip-gram with negative sampling (SGNS) model to get a low-dimensional vector representation for each vertex sequence. SGNS is proven to be the model that can produce a high-quality vector representation for a given sequence (Chen et al., 2018; Goldberg & Levy, 2014; Huang et al., 2016; Perozzi & Skiena, n.d. 2014; Tang & Qu, n.d., 2015).

DOI: 10.56892/bima.v7i01.411

Ultimately, we append these network-based features with the employee's basic information, apply a ten-k cross validation technique and use them as input to various classic machine learning classification algorithms to solve the employee turnover prediction problem more effectively.

In short, this study improves the existing body of knowledge as follows:

1. This study proposes an enhanced graph embedding and prediction (EGEP) model that improve the prediction performance of employee turnover.
2. This paper indicates the role of a principal component analysis (PCA), in extracting a good component that, when added to the employee's basic information can dynamically improve the prediction performance of employee turnover.
3. This paper investigates the role of cross validation techniques in producing an averagely stable result in each prediction execution as opposed to a one-held-out train test split method, employed by the state-of-the-art employee turnover prediction.
4. The paper evaluates and selects an appropriate classification technique with high performance in comparison with other states-of-the-art turnover prediction accuracy.

The remaining sections of this paper are organized as follows: Section 2 presents the adopted methodology for the research., Section 3 present the experimental setup., Section 4 present Result and Discussion., and Section 5 present Conclusion and Future Research Direction.

## MATERIALS AND METHODS

In this section, we explain the model of our EGEP by describing some important concept used in the model and the method used to predict employee turnover.

**Bipartite graph:** A bipartite graph is a structured network type with the properties that its vertices be partition in to two independent groups in such a way that no two vertices in the same partition be connected directly (X. Cai et al., 2020). Base on the above definition we can state the following;

**Definition 1: Turnover Prediction Problem**

Let $G = (X, Y, E, T)$ be a dynamic bipartite graph, describing an interaction between Employee (e), Organizations *(o)*, and an edge *(e, o, t)* connecting *"e"* to *"o"* within timeframe "*t*". The central objective of the turnover prediction problem is to predict whether *"e"* will terminate his/her appointment with *"o"* after a certain timeframe of length *"*$\Delta t$*"* based on the basic and interconnected-based features, , (Alao, 2013; Gao et al., 2018) .

**Definition 2: Dynamic Bipartite Graph**

A dynamic bipartite graph is a bipartite graph $G = (X, Y, E, T)$, $X \cup Y = V$, where *X and Y* represent the sets of two types of vertices, *V* is the set of all the vertices in the graph *"G"*, $E \in (X \times Y)$ defines the inter-set edges, and *T* is the timeframe of *E*. i.e. for each edge $e = (x,y,t) \in E$ connecting a vertex $x \in X$ and a vertex $y \in Y$, *"e"* is associated with a unique timeframe ' t', (H. Cai et al., 2018; Gao et al., 2018).

**Definition 3: Horary random walk (HRW)**

Let $G= (X, Y, E, T)$ be a dynamic bipartite graph, the HRW choose a vertex (v) randomly from G, and the visitor begins to traverse from "v" to the neighbouring vertex on the opposite partition of G, if the visitor is to visit $x \in X$, it will check all the edges in $E_x$ that has not been visited before and chooses to visit the edge with minimum timeframe, else it will randomly choose an adjacent vertex to visit irrespective of whether the edge has been visited or not, the visitor terminate it visitation at a vertex $x \in X$ when there is no unvisited edge or it reaches the maximum length l (X. Cai *et al*., 2020) .

**Skip-Gram Negative Sampling (SGNS)** A SGNS is an unsupervised learning technique used in natural language processing (Word2Vec) to learn and represent text/node in form of vectors (known as Text/node Embedding)

SGNS algorithm is designed to alleviate the deficiency of computational complexity (i.e., computational cost and fairly qualitative vector generation) as it is in Skip-Gram (SG) and Skip-gram Hierarchical SoftMax (SGHS) (J. Chen et al., 2021). .

**EGEP Model**

To overcome the existing employee turnover problem more effectively, we propose an EGEP model that operate in the following phases;

- First, after all the data pre-processing, we divide the clean data in to Network-based features and basic information features,
- we learn the network-based features as low dimensional vectors for both employee and organization through enhanced graph embedding EGE module,
- perform all the necessary data extraction and data selection for both the network-based and basic information features respectively using appropriate techniques,
- Appended the obtained extracted features with the selected features (in 3 above) in a single data-frame to obtained a more Combine Features,
- Then, we feed the combine features (in 4 above) to the historically classic used machine learning cross-validated classification algorithms to predict the employee turnover separately.
- Finally, we evaluate the prediction output of the selected used classifiers to obtain the best classification.

The above phases are diagrammatically summarized in the Figure 2 below;

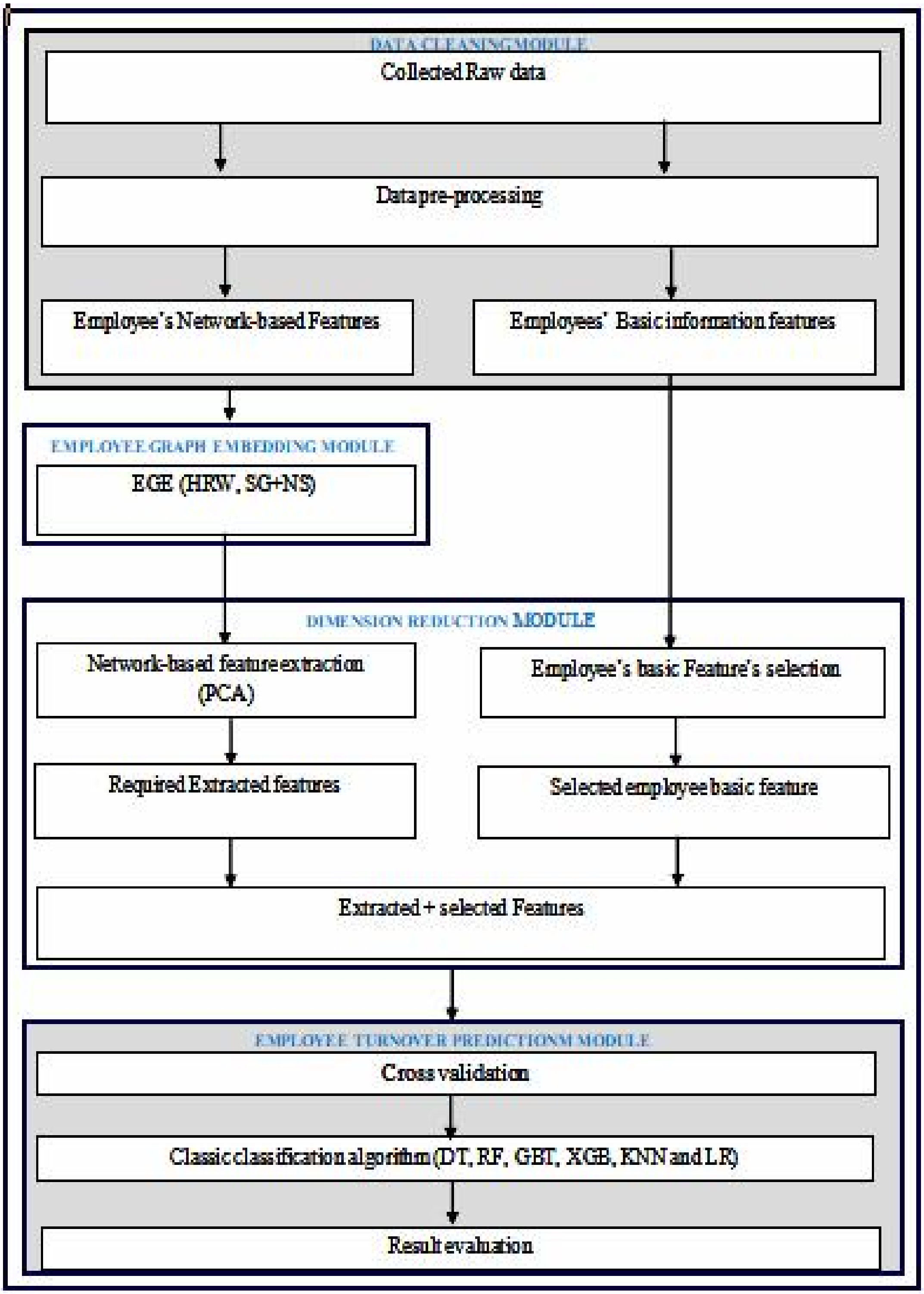
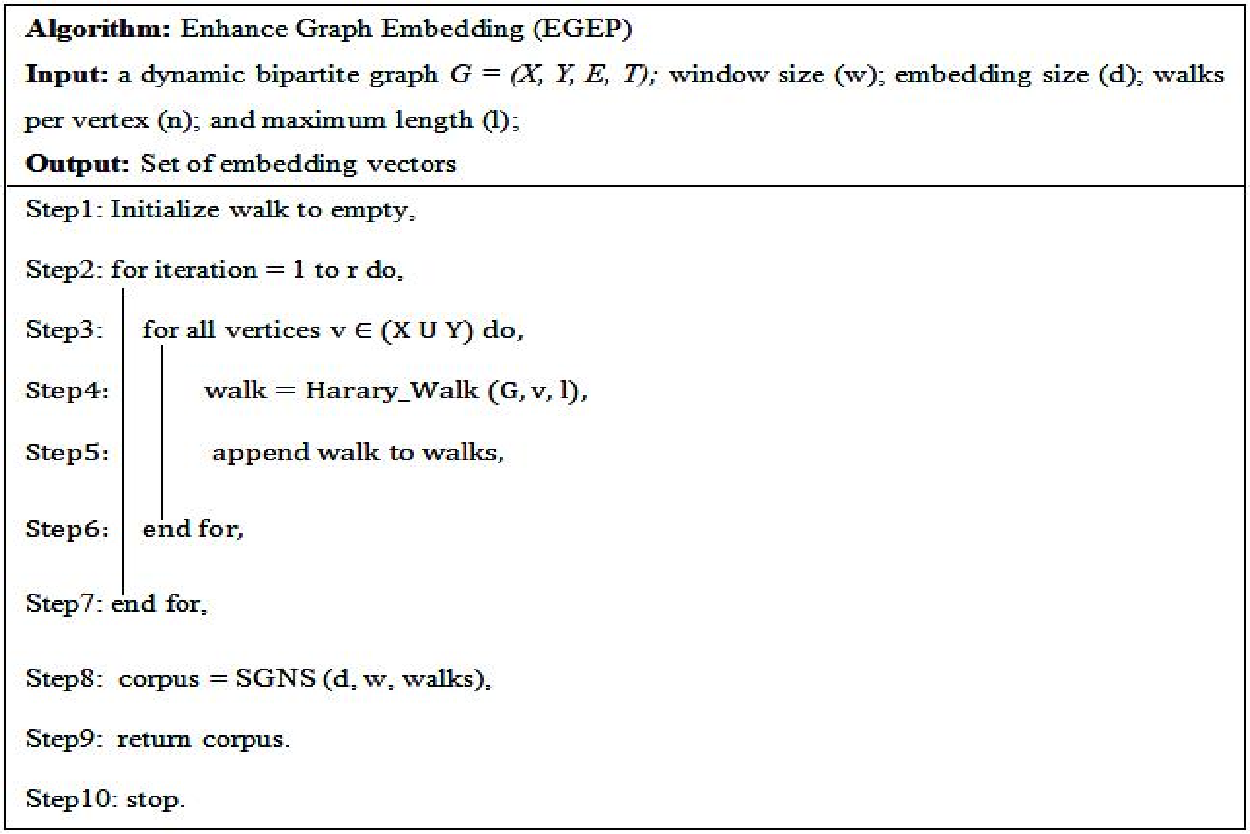

DATA CLEANING MODULE
Collected Raw data
Data pre-processing
Employee's Network-based Features
Employees' Basic information features
EMPLOYEE GRAPH EMBEDDING MODULE
EGE (HRW, SG+NS)
DIMENSION REDUCTION MODULE
Network-based feature extraction (PCA)
Employee's basic Feature's selection
Required Extracted features
Selected employee basic feature
Extracted + selected Features
EMPLOYEE TURNOVER PREDICTIONM MODULE
Cross validation
Classic classification algorithm (DT, RF, GBT, XGB, KNN and LR)
Result evaluation

**Figure 2:** A Propose EGEP Model

DOI: 10.56892/bima.v7i01.411

## Pseudocode of EGEP algorithm



**Algorithm:** Enhance Graph Embedding (EGEP)

**Input:** a dynamic bipartite graph *G = (X, Y, E, T);* window size (w); embedding size (d); walks per vertex (n); and maximum length (l);

**Output:** Set of embedding vectors

```
Step1: Initialize walk to empty,
Step2: for iteration = 1 to r do,
Step3:     for all vertices v ∈ (X U Y) do,
Step4:         walk = Harary_Walk (G, v, l),
Step5:         append walk to walks,
Step6:     end for,
Step7: end for,
Step8:  corpus = SGNS (d, w, walks),
Step9:  return corpus.
Step10: stop.
```

**Figure 2:** Pseudocode of EGEP algorithm:

### Learning Computational Complexity of EGEP

To learn the computational complexity of our proposed EGEP algorithm, we use the concept of big O-notation to analyze the time complexity by the number of nested loops that the model iterates upon as; O *(l * **E**$_{max}$ * n ** $V$+ *w (d +d(k+1))).*

*Where: l = maximum walk length, E$_{max}$ = the maximum degree of employee vertex, n = number of vertex's walk,* $V$ = *$E$ + $O$ is the total of all vertices, and K = Number of negative samples drawn, w = window size and d= the dimension of the embedding.*

**<u>Proof:</u>**

Considering that our proposed model depends on the HRW algorithm [3], which involve n-stages of a random walk with condition that, at each stage, if the current vertex ($c_v$) is an employee vertex, then it will traverse all of it adjacent edges to get the next un-visited edges having the least timeframe, which spends *O (E$_{c)}$ ≤ O (E$_{max}$) time*, where *E$_c$* is the degree of the current vertex,

Otherwise (i.e., if the current vertex is an organization vertex), it will select at random, the nearest vertex to visit, which take complexity of O (1) time, as such, at worst case, the time complexity of HRW is: *O (**l** * **E**$_{max}$ + 1) ≈ O (**l** * **E**$_{max)}$* (X. Cai et al., 2020) .

Moreover, we can easily deduce that the proposed EGEP algorithm call HRW *(n * v)* times through both the inner and the outer loop, appended with SGNS having it

DOI: 10.56892/bima.v7i01.411

complexity equals to $O\ (w\ (d + d(k+1)))$ (Mikolov *et al.*, 2013).

Base on the above scenario we can evaluate that the overall time complexity of the proposed EGEP to be equals to;

$O\ (l * \boldsymbol{E}_{max} * n * \$V\$ + w\ (d + d(k+1)))$ □

### Comparative cost analysis with state-of-the-art's complexity

Thus, when comparing the complexity of the proposed model; $O(l * \boldsymbol{E}_{max} * n * \$V\$ + w\ (d + d(k+1)))$ with its counterpart (the baseline work); $O(l * K_{max} * r * \$V\$ + w\ (d + d\log_2(\ \$V\$\ )))$, we can notice that, when updating the output neurons weight matrix (ONW),

the baseline embedding model is computationally very expensive, as it requires scanning through the entire output embedding matrix to compute the probability distribution up to $\log_2(\$V\$)$ vertices for each training sample, where \$V\$ can be millions or more.

Thus, in contrast with the proposed embedding model that makes used of Negative sampling to update only k+1 weights out of the output neurons weight matrix (ONW) for each training sample, (where \$K\$ is a hyper-parameter that can be empirically fine-tuned, with a typical range of \$ [5-20] \$ vertices for a small dataset and \$ [3-5] \$ for a very large dataset (Mikolov et al., 2013).

For example, with the proposed dataset having a total number of vertices V=70,714 and a number of the hidden (input) layers N = 128, has V*N = 9,051,392 ONW in the neuron's output weight matrix that requires an update for each training sample.

Therefore, the baseline model reduces the above update to (Log2 (V)) * N = 2,048 ONW from the output weight matrix, which is still considered much. But with the proposed model, we reduce this update drastically to only (k+1) * N = 768 ONW (where k = 5) from the output weight matrix., i.e., the proposed model has saved about 1,280 Units of time and space compared to the baseline model.

Moreover, the space complexity of the propose embedding model is independent of the training data size and in the worse-case, it requires an O(n) space to store the vertex embedding and its frequency count (where n is the unknown variable).

## EXPERIMENT

### Data Collection

This study makes use of data collected from one of the China's largest online Employees recruitment platform in its secondary data type form, covering 15 months from 1st January 2018.

This data set is made up of a composition of an employee's historical Job record (which is modeled as a dynamic bipartite graph) and his/her basic information containing demographics and cognitive features such as age, gender, Highest qualification, Type of School awarding institution, organization type, Organizational Employee strength, Date of First employment, place of primary assignment, the department posted, number of promotions, salary structure, salary-grade, and salary-level, Maximum turnover, etc. were recorded for the data mining study.

Random sampling is used to select employee so that we can get accurate results without been biased.

Moreover, to limit our approach, we only take in to account an employee with minimum of two working experience records, also, as part of inclusion and exclusion criteria, we set a time limit of 18-month as the time of turnover, i.e., if an employee leaves an organization after 18 months from the time of employment, then we consider it as a voluntary turnover and will be included in the graph else it will be considered as involuntary turnover and will be excluded.

**Data Pre- Processing**

This stage involves dataset preparation before applying the data mining techniques. However, in this context of employee turnover, the steps involved in data preprocessing are data cleaning which involves handling missing data and noisy data,

Data transformation where the data is transformed into a form suitable for mining, data scaling to scale the data within a common appropriate scale for all the features, then dimensionality reduction which involve data selection (Using forward feature selection and decision tree) and data extraction process (using PCA) which aims to optimize storage effective utilization as well as to retaining data information and reduce analysis costs.

After cleaning the data, we established a dynamic bipartite graph with 47,257 employee's vertices and 23,457 organization's vertices connected by 203,604 edges with their associated time frame.

We also select the nine most relevant basic variables in order of their variation ranking, from the clean data set, these variables can be divided in to three categories as shown in the Table 1 below.

Moreover, we apply a Ten-k cross validation to perform the train-test split.

Classifiers.

In this research, we have adopted six classifiers as follows:

a. Random Forest (RF): This approach is a classic bagging method in ensemble learning, , (X. Cai et al., 2020 ; Shi et al., 2019).

b. Extreme Gradient Boosting (XGBoost): This approach is a type of boosting method in ensemble learning , (Jain & Nayyar, 2018; X. Cai et al., 2020).

c. Decision Tree (DT): This method is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category(Zhao et al., 2019).

d. Gradient boosting tree (GBT): GBT is a classic machine learning algorithm that iteratively constructs an ensemble (CART) of weak decision tree learners through boosting and reduce the sample loss. (Alam & Mohiuddin, n.d. 2018; El-Rayes et al., 2020).

e. K-nearest neighbors (KNN): is a supervised machine learning classifier that work operate by observing a distance between query and all example in the dataset, choosing a number of example (k-value) that is closer to the query and vote for the frequent label(Alam & Mohiuddin, n.d. 2018).

f. Logistic Regression (LR): This approach uses the gradient descent method to iteratively find the optimal parameters of the linear model that minimize the loss function, and then outputs the probability value of the classification through the sigmoid function. Finally, the classification result is obtained by comparison with the threshold (X. Cai et al., 2020).

**Table 1:** Variable's description tables

| S/N | Variable Type | Variable Name |
|---|---|---|
| 01 | Demographics Variable | Gender,<br>Highest certificate obtained. |
| 02 | Present work Variables | Date of employment,<br>Ministry, Department or Agency (MDA) posted,<br>Development levy,<br>Pay Scale code,<br>Position held. |
| 03 | Work experience Variables | Start year of waking career,<br>Turnover status. |

**Evaluation metrics:**
The following are the three-evaluation metrics in evaluating our proposed model;

a. F1 score: this is the harmonic average of precision and recall,

i.e., $F1\ score = 2 * \frac{Precision * Recall}{Precision+Recall}$ (3.1)

b. Precision: A precision is defined as the proportion of positive cases in all samples classified as positive.

i.e., $Precision = \frac{TP}{TP+FP}$ (3.2)

c. Recall: is the proportion of samples that are predicted to be positive in all positive samples.

i.e.,$Recall = \frac{TP}{TP+FN}$ (3.3)

*where**:** TP= True positive, TN= True Negative, FN= False Negative, FP= False Positive.*

**Parameter Setting**

As recommended in the work of [16], we use the hyperparameter setting as follows:

*Window Size (w) =5 Vertices, embedding size (d) = 128 dimension, Walk per vertex (n)= 80, Maximum walk length (l) =15.*

To efficiently learn the low-dimensional vectors for both employees and organizations vertices respectively, at the end of the EGE module's execution, we produced a 128-dimensional vector representation for each vertex. (64 vectors for the organization nodes and 64 for the employee nodes).

However, to append the basic variable of employees with these vectors representation, we apply Principal Component Analysis (PCA) extraction techniques on the obtained 128 low-dimensional vectors, and composed them in to one, two, three, four, five and six, seven and eight separate components respectively, and append each with the basic employee features explain in table 1 above to study the prediction performance of each component as reported in table in table 3 below.

As for the Classic machine learning classification algorithms, we apply a hyperparameter tunning technique to obtain the hyperparameter values used as the most effective parameter setting for each classifier.

## RESULTS

### Result Evaluation Metric for the Classifiers Using Employee Basic Information Only

To verify if the proposed EGEP model can address the employee turnover prediction problem more effectively, we simultaneously run and compare the prediction performance of the classifiers, with and without the EGE variables respectively as presented in the various result tables bellow:

**Table 2:** Classifier's evaluation metric's 10-fold cross validated performance with only employee basic features

| | | EVALUATION METRICS | | | | | |
|---|---|---|---|---|---|---|---|
| **S/N** | **CLASSIFIERS USED** | **ACCURACY** | **PRECISION** | **RECALL** | **F1** | **AUC_ROC** | **AVERAGE PRECISON** |
| 1 | DT | 91.54 | 67.91 | 73.37 | 61.09 | 83.78 | 59.85 |
| 2 | RF | 92.35 | 79.78 | 69.96 | 69.15 | 89.82 | 78.44 |
| 3 | GBT | 95.72 | 89.15 | 67.18 | 69.3 | 90.88 | 84.32 |
| 4 | XGB | 93.45 | 74.16 | 70.49 | 64.27 | 93.21 | 83.74 |
| 5 | LR | 90.01 | 81.03 | 90.01 | 85.28 | 56.21 | 20.15 |
| 6 | KNN | 90.01 | 65.93 | 58.51 | 54.7 | 81.84 | 60.28 |

DOI: 10.56892/bima.v7i01.411

As can be seen from table 2 above, is 10-k cross validation prediction performances before inclusion of the EGE variable, as can be seen, GBT has the highest prediction performance in Accuracy, Precision, AUC_ROC and Average Precision, having their corresponding prediction values equals to 95.72, 89.15, 90.88 and 84.32 respectively.

However, this has probably been achieved because a GBT is among the ensemble learning algorithms that can handle non-linear data more effectively. However, LR is seen to excel in recall = 90.01 and F1-measures = 85.28 and this might be because Recall tend to minimizes the false negative rate in the prediction process and F1 measure the extent to which such False negative is reduced.

### Result Evaluation Metric for the Classifiers with EGEP Variables and Employee Basic Information

To used EGEP variable, we drastically composed the dimension to six-independent component (pca1, pca2, … pca6) and observe the influence of each component on the employee turnover prediction problem using same classifiers with same evaluation metrics as shown in the figure 3 bellow.

These components are then added to the basic employee variables individually and fed to the six historically used machine learning classifiers to predict the employee turnover problem more effectively.

For each classifier, we chose and report the PCA component that exhibit the best result to maintain a good performance. The indicated subscript attached with each classifier in table 3 shows that we use the k-dimensionality of the PCA reduced component with the employee basic variable to train the learning classifier corresponding to the subscript.

**Table 3:** 10-fold cross validated Classifier's result evaluation table with ' EGEP ' variables reduce by PCA algorithm;

| | | EVALUATION METRICS | | | | | |
|---|---|---|---|---|---|---|---|
| **S/N** | **CLASSIFIERS USED** | **ACCURACY** | **PRECISION** | **RECALL** | **F1** | **AUC_ROC** | **AVERAGE PRECISON** |
| 1 | $DT_4$ | 92.83 | 75.46 | 79.44 | 72.18 | 87.52 | 61.72 |
| 2 | $RF_6$ | **95.66** | **90.41** | 75.51 | 76.76 | **94.53** | 88.88 |
| 3 | $GBT_6$ | 95.22 | 90.03 | 78.78 | 79.35 | 93.86 | 90.6 |
| 4 | $XGB_6$ | 93.93 | 86.06 | 80.09 | 78.2 | 94.24 | **90.62** |
| 5 | $LR_3$ | 90.04 | 81.27 | **89.97** | **85.26** | 64.28 | 24.38 |
| 6 | $KNN_1$ | 93.18 | 63.31 | 65.64 | 61.53 | 83.32 | 67.18 |

Similarly, for both adopted classifiers, in table 3 above, we observed that their prediction performances have further improved with the help of the propose graph embedding variable, learnt through our EGEP approach as compared to their respective performances presented in table 2 that lack the EGEP variables, and this performance differences are tabulated in table 4 bellow.

After the inclusion of the propose EGEP variable (table 3) RF algorithm has the best performance in the proposed evaluation metrics., having it prediction values as; Accuracy = 95.66, Precision = 90.41, AUC_ROC = 94.53, However, this has probably been achieved because RF is also a family of ensemble learning algorithms that can handle non-linear data more effectively.

Moreover, its ability to randomize it state at different execution has assist in reducing the effect of overfitting and improve their general performance.

**Table 4:** 10-fold cross validated Classifier's improvement (in %) with EGEP variables reduced by PCA algorithm.

| S/N | CLASSIFIERS USED | EVALUATION METRICS | | | | | |
|---|---|---|---|---|---|---|---|
| | | ACCURACY | PRECISION | RECALL | F1 | AUC_ROC | AVERAGE PRECISON |
| 1 | $DT_4$ | 1.29 | 7.55 | 6.07 | 11.09 | 3.74 | 1.87 |
| 2 | $RF_6$ | **3.31** | 10.63 | 5.55 | 7.61 | 4.71 | **10.44** |
| 3 | $GBT_6$ | -0.5 | 0.88 | **11.6** | 10.05 | 2.98 | 6.28 |
| 4 | $XGB_6$ | 0.48 | **11.9** | 9.6 | **13.93** | 1.03 | 6.88 |
| 5 | $LR_3$ | 0.03 | 0.24 | -0.04 | -0.02 | **8.07** | 4.23 |
| 6 | $KNN_1$ | 3.17 | -2.62 | 7.13 | 6.83 | 1.48 | 6.9 |

To measure the effectiveness of the reduced network-based feature (generated by the proposed model) on each adopted classifiers, in Table 4 above, we further compute the difference in the prediction performances between the network-based embedding prediction (Table 3), and the basic employee variable-based prediction (Table 2) for each classifier with respect to the proposed evaluation metrics.,

As we can observed in Table 3, we cannot precisely state the power of any classifier over another, because, we obtained the classifiers performances base on the component that give us the best output over all the evaluation metrics.

This improvement is achieved by the use of different data component as indicated in the subscript of each classifier; however, the performance has improved significantly with the help of the proposed EGEP embedding algorithm.

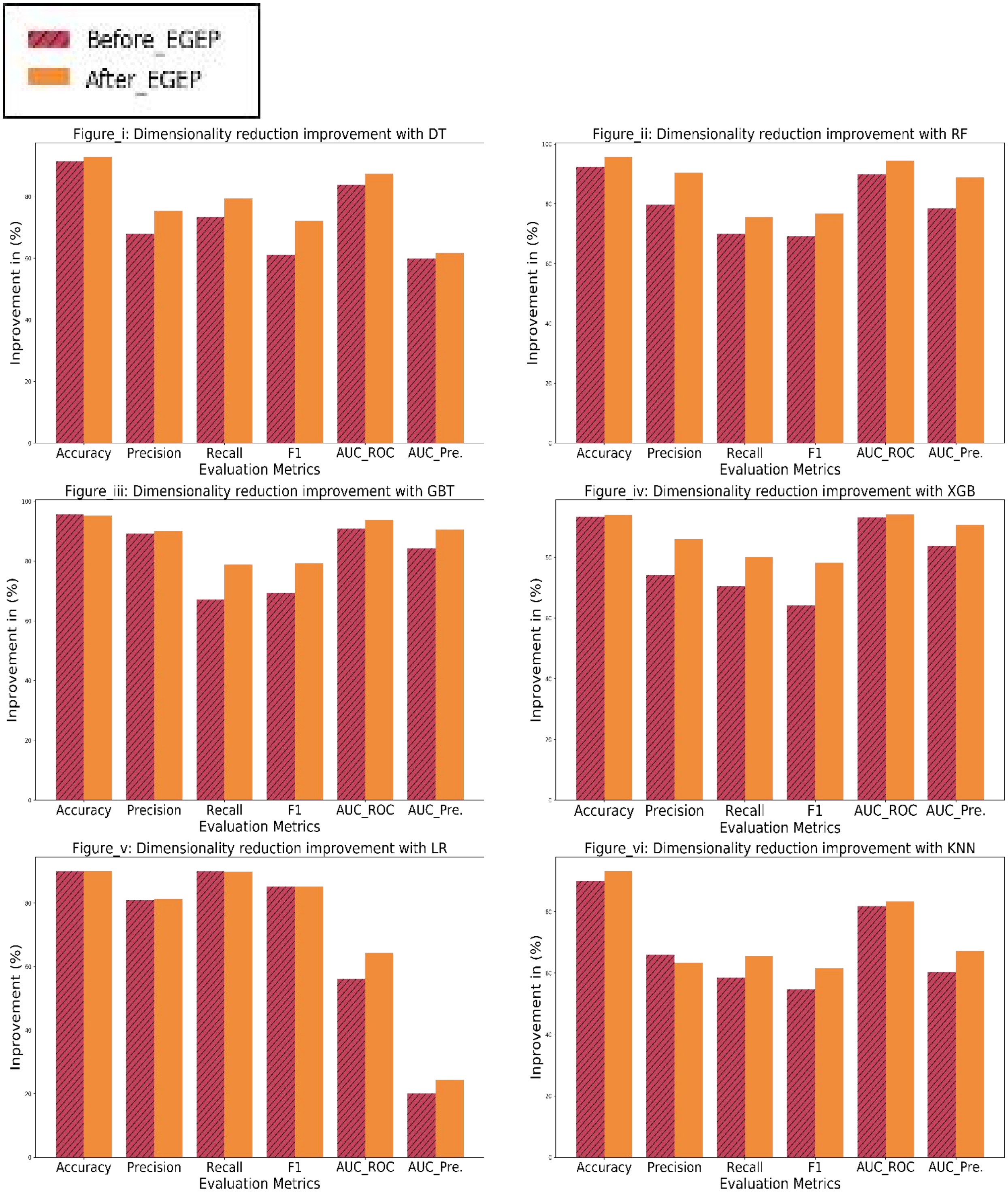


**Figure 3:** Classifier's performance on the employee turnover prediction before and after EGEP

DOI: 10.56892/bima.v7i01.411

To further prove the prediction power of our EGEP model, we computed, ranked and graphed the importance of all the (both the basic employee variable-based and network-based) variables using decision tree classifier.

From figure 4 below, we observed that Vec7 and vec8 has the third and fourth rank in the plot, moreover lots of the PCA-Vector-component have high rank compared to the rank of most of the employee-basic variable counterpart, and this has clearly indicated the prediction power of the proposed EGEP variable in predicting employee turnover more effectively as it positively impacted the employee turnover prediction in all the adopted classifiers.

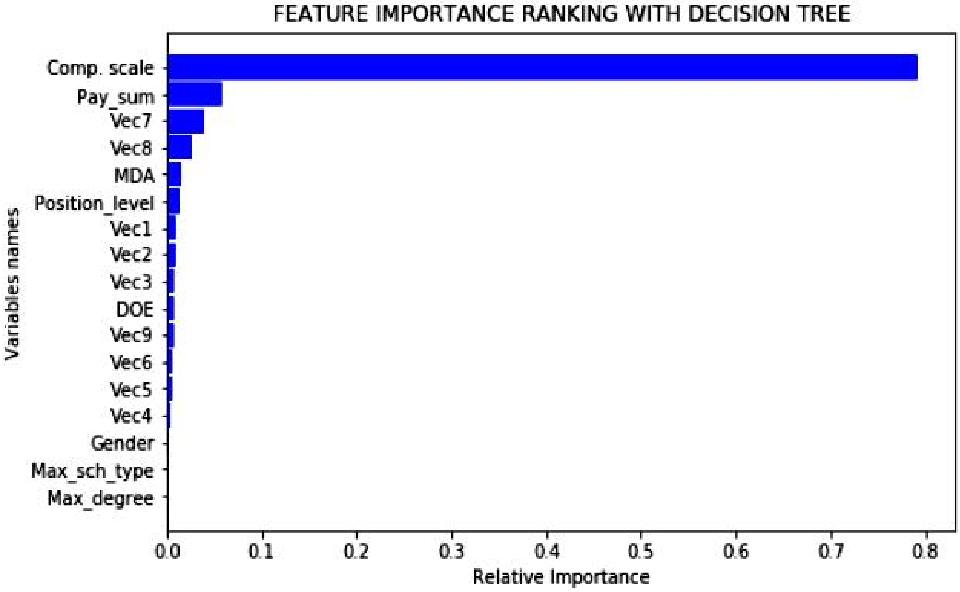


**Figure 4:** Decision Tree Variable importance ranking for the PCA component.

## CONCLUSION

In this paper, we proposed an enhanced graph embedding and prediction (EGEP) model that investigate the impact of PCA extraction algorithm and cross validation train-test-split technique in dealing with the network-based representation learned features as a set of components and assist in addressing an unbiased employee turnover prediction performance respectively.

After performing all the necessary data extraction and data selection for both the network-based and basic information features respectively, we appended the obtained extracted network-based features with the selected employee basic features in a single data-frame to obtained more Comprehensive Features, and feed them to the most historically used classic machine learning classification algorithms to predict the employee turnover separately. Finally, we evaluate the prediction output of the selected classifiers to obtain the best classification.

We conduct an extensive experiment using a real-world data set and the result indicated that our proposed approach was effective in obtaining a best result for the employee turnover prediction problem.

In the future, researchers might:

✓ Consider incorporating other features like changes in the services situation or impact of market strategies.

DOI: 10.56892/bima.v7i01.411

✓ Researchers might also investigate and addressed the issues of negative improvement (seen in table 4) we experience when evaluating the Recall and F1-measure of Logistic regression or the precision of k-nearest neighbor algorithm respectively.

## REFERENCES

Yadav, S. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. *2018 IEEE 8th International Advance Computing Conference (IACC)*, 349–354.

Xu, H., Yu, Z., Member, S., Yang, J., Xiong, H., & Member, S. (2018). Dynamic Talent Flow Analysis with Deep Sequence Prediction Modeling. *IEEE Transactions on Knowledge and Data Engineering*, *PP*(c), 1.

Cai, X., Shang, J., Liu, F., Jin, Z., Qiang, B., & Xie, W. U. (2020). DBGE: Employee Turnover Prediction based on Dynamic Bipartite Graph Embedding. *IEEE Access*, *PP*, 1.

Stamolampros, P., Korfiatis, N., Chalvatzis, K., & Buhalis, D. (2019). Job satisfaction and employee turnover determinants in high contact services: Insights from Employees' Online reviews. *Tourism Management*, *75*(May), 130–147.

Jain, R., & Nayyar, A. (2018). Predicting Employee Attrition using XGBoost Machine Learning Approach. *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, 113–120.

Alao, D. (2013). *ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS*. Computing, Information Systems & Development Informatics Vol. 4 No.

Cai, H., Zheng, V. W., & Chang, K. C. C. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1616–1637.

Dave, V. S., Aljadda, K., & Korayem, M. (2018). *A Combined Representation Learning Approach for Better Job and Skill Recommendation*. Association for Computing Machinery. ACM ISBN 978-1-4503-6014-2/18/10. . . $15.00.

Xu, H., Yu, Z., Xiong, H., Guo, B., & Zhu, H. (2015). *Learning Career Mobility and Human Activity Patterns for Job Change Analysis*. 1057–1062.

Chourey, A. (2019). *A SURVEY PAPER ON EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING TECHNIQUES*. *XI*(199), Journal of Interdisciplinary Cycle Research Volume XI, Issue XII, December/2019. 199–202.

Feeley, T. H., Moon, S. Il, Kozey, R. S., & Slowe, A. S. (2010). An erosion model of employee turnover based on network centrality. *Journal of Applied Communication Research*, *38*(2), 167–188.

Vardaman, J. M., Taylor, S. G., Allen, D. G., Gondo, M. B., Amis, J. M., Vardaman, J. M., Taylor, S. G., Allen, D. G., Gondo, M. B., Amis, J. M., Intentions, T., Taylor, S. G., & Allen, D. G. (2015). *Translating Intentions to Behavior: The Interaction of Network Structure and Behavioral Intentions in Understanding Employee Turnover Translating Intentions to Behavior: The Interaction of Network Structure and Behavioral Intentions in Understanding Emp*. *May*.

Chen, H., Hu, Y., Perozzi, B., & Skiena, S. (2018). HARP: Hierarchical representation learning for networks. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2127–2134.

DOI: 10.56892/bima.v7i01.411

Goldberg, Y., & Levy, O. (2014). *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. *2*, 1–5.

Huang, Y., Yu, L., Wang, X., Cui, B., Vergara, P., Qian, Y., Tang, J. A., Yang, Z., Huang, B., Wei, W., Zhang, Y., Wallace, B., Kumar, S., Carley, K. M. K. M., Hirshman, B. R., Jones, L. A., Carroll, K. T., Tang, J. A., Proudfoot, J. A., … Frankowski, D. (2016). node2vec Real-time Video Recommendation Exploration Categories and Subject Descriptors. *World Neurosurgery*, *95*(1), 41–50. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.1745%5Cnhttp://link.springer.com/10.1007/s41019-015-0002-9%5Cnhttp://dx.doi.org/10.1016/j.ejso.2016.07.137%5Cnhttp://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379%5C.

Perozzi, B., & Skiena, S. (n.d.). (2014). *Deep Walk: Online Learning of Social Representations Categories and Subject Descriptors*. ACM 978-1-4503-2956-9/14/08.

Tang, J., & Qu, M. (n.d.). (2015). LINE: Large-scale Information Network Embedding**.** International World Wide Web Conference Committee (IW3C2). *P1067-Tang*. 1067–1077.

Gao, M., Chen, L., He, X., & Zhou, A. (2018). BiNE: Bipartite network embedding. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, *1*, 715–724.

Chen, J., Gong, Z., Wang, W., & Liu, W. (2021). HNS: Hierarchical negative sampling for network representation learning. *Information Sciences*, *542*, 343–356.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (n.d.) (2013). *Distributed Representations of Words and Phrases and their Compositionality arXiv : 1310 . 4546v1 [ cs. CL] 16 Oct 2013*. 1–9.

Shi, C., Hu, B., Zhao, W. X., & Yu, P. S. (2019). Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *31*(2), 357–370.

Zhao, Y., Hryniewicki, M. K., & Cheng, F. (2019). *Employee Turnover Prediction with Machine Learning: A Reliable Approach* (Vol. 1). Springer International Publishing.

Alam, M. M., & Mohiuddin, K. (n.d.) (2019). *A Machine Learning Approach to Analyze and Reduce Features to a Significant Number for Employee' s Turn Over Prediction Model* (Vol. 1). Springer International Publishing.

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *International Journal of Organizational Analysis*, *28*(6), 1273–1291.

Fan, C., Fan, P., Chan, T., & Chang, S. (2012). Expert Systems with Applications Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications*, *39*(10), 8844–8851.

Jaffar, Z., Noor, W., & Kanwal, Z. (n.d.). *Predictive Human Resource Analytics Using Data Mining Classification Techniques*. *International Journal of Computer (IJC) (2019) Volume 32, No 1, pp 9-20.*

Khera, S. N. (2019). *Predictive Modelling of Employee Turnover in Indian IT*

DOI: 10.56892/bima.v7i01.411

*Industry Using Machine Learning Techniques*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Peng, H., Li, J., Yan, H., Gong, Q., Wang, S., Liu, L., Wang, L., & Ren, X. (2019). Dynamic network embedding via incremental skip-gram with negative sampling. arXiv:1906.03586v1 [cs.LG] 9 Jun 2019.

Punnoose, R. (2016). *Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016 5*(9), 22–26.

Saradhi, V. V., & Palshikar, G. K. (2011). Expert Systems with Applications Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006.

Srivastava, D. K., & Nair, P. (2018). *Employee Attrition Analysis Using Predictive Techniques*. Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1, Smart Innovation Systems and Technologies 83.

.