International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 1 Jan. 2026, pp: 206-214 www.ijaem.net ISSN: 2395-5252

An Improved Predictive Analytics Approach to Employee Turnover Prediction Using K-Cross Validation Technique

Abdullahi Jibril Abdullahi [1], Sani Ibrahim[2], Bello Imam Abubakar[3], Abdulhamid Shariff Mahmud[4]

[1234] Department of Computer Science, School of technology, kano state polytechnic

---

Date of Submission: 05-01-2026 Date of Acceptance: 16-01-2026

Abstract:
Employee attrition (employee turnover) refers to the sudden resignation of a staff or working personal from company or business enterprises, this issues had necessitated enterprise's human resource managers to employ an inform datamining techniques in addressing the repercussion that the sudden lost may result into, which in turn became hard to recover and regain. Consequently, identifying a concept applicable in forecasting an intended quitting personal is necessary to minimize it negative impact, in an attempt to solve this issue, past studies uses some influential features like employee's and organization's trait., while in this work, we emphasize on mining a worker's historical profile inform of bipartite graph, aiming at studying the vector pattern of this graph, we achieved this by designing a framework that compute all the vertex sequence for the graph using a random-walk technique and uses the pattern as input to an algorithm called skip gram negative sampling algorithm to get the vector representation for the vertices., add this output to the worker's native data (personal details) and use it as input to the most historically used classifiers to forecast the turnover., specifically, this work proposes a prediction technique that examine the role of a 10k-cross-validation in forecasting worker's attrition called ***Improved Graph-Embedding and Prediction (IGEP).,*** *furthermore, an experimental result prove that the propose technique has a positive impact on the prediction result when compared with others.*

Keyword: *k-cross validation, Graph embedding, Principal component analysis, Bipartite graph,*

I. Introduction

According to [1], Employee attrition (employee turnover) refers to the sudden resignation of a staff or working personal from company or business enterprises.

This problem had necessitated enterprise's human resource unit to employ an inform datamining techniques to model the historical data in a way that can facilitate forecasting such attrition to manage the repercussion that the sudden lost may result into [2].

On the other hand, past studies have demonstrated the role of machine learning in addressing all such unforeseen risk in business activities [3], [4]. However, the utilizing advanced data mining techniques is still underutilized in underdeveloped communities. [5],[33]. These draw aur attention to explore some other areas that we can utilizes data to forecast all worker's attrition behaviour.

This study, consider only an intended (voluntary) attrition from the employee side, as the involuntary attrition behaviours are mostly course by the enterprise' managers themselves.

Many studies have been made to address this problem, some of which include the work of [8], [9] used Machine learning techniques to forecast the employee attrition, while some consider a survey as the effective means of fishing out the reasons for the sudden attrition [10] & [31].

However, previous research focuses mostly on income earn and social wellbeing of the workers, disregarding the workers' working trend and ambition, which contain important details about the workers' attitude. This may in turn result in vital information leakage, that when considered may can help in getting a better output.

Consequently, to address this issue, we design a network-based graph embedding technique by dividing the worker's working history into the employee and the places the worked as depicted in the diagram bellow.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 1 Jan. 2026, pp: 206-214 www.ijaem.net ISSN: 2395-5252

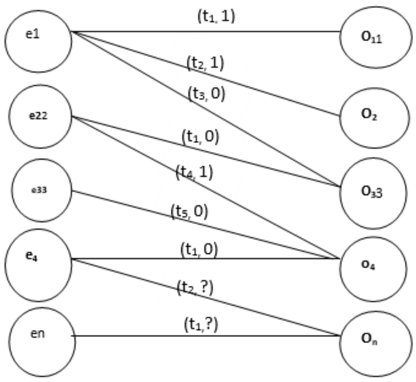

Fig1: employee bipartite graph.

This bipartite graph allows us to effectively study the employee network trend embedding via a horary random walk (HRW), proposed by [33], [34] to produce the sequence for the vertices contain in the bipartite graph.

This output obtained from the HRW in then served as input to a historically best optimization algorithm called Skip Gram Negative Sampling algorithm to transform the vertex walk into a vector representation as observed in the work of [13], [14], [15], [16] & [17].

Finally, we combine the network-based vector information with the employee' demographic data as input to the most historically used prediction algorithms to make the prediction.

In summary, this proposes paper the following important concept:

- ✓ It proposes an improved predication model aimed at enhancing a prediction output of companies' staff attrition.
- ✓ It emphasizes the importance of a Principal Component Analysis (PCA) algorithm, when composing the best part of the data that can fit the model to enhance the prediction result.
- ✓ It evaluates the performance of the improved model in comparison with the other related prediction result.

II. Methodology

This part describes the method use in designing the model and the concept employed and the prediction technique as follow:

2.1.Definition of some term:

2.1.1 The Bipartite graph:

A bipartite graph is a structured graph type with a properties that it's all vertex be partition in to two independent group in a way that no two vertex in same partition are connected directly [3]. Base on the above definition we can state the following.

2.1.2 Bipartite graph:

A bipartite graph is graph of the form $G = (X, Y, E, T)$ partitioning a relation between a staff (s), and company 'c', connected by an edge *(e, c, t)* representing the attrition within a specific time '*T*'. Here, the logic aim is to identify the tendency of 's' to quit services with 'c' at a particular time 't'. [6] [35] & [18].

2.1.3 Horary-random-walk (HRW)

This is a concept used to model a network of dynamic in nature, e.g.., Let $G= (X, Y, E, T)$ be a bipartite graph (dynamic), the HRW select a vertex at random from a graph 'G', and the walker begins to walk from one vertex to another nearby vertex on the other part of 'G. with a condition that on visiting the next vertex, it will walk through the edge with the lowest time frame 't' to walk through or-else it choose any adjacent vertex to walkthrough at random, the walker stop at a last visited vertex if all the edge have been visited l [3].

2.1.4 Skip gram negative sampling (SGNS)

A Skip gram negative sampling is an unsupervised learning algorithm (e.g., Word2Vec.) aimed at learning to convert text into vectors, a process called 'text embedding' [37] & [19].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 1 Jan. 2026, pp: 206-214 www.ijaem.net ISSN: 2395-5252

2.3 IGEP Model:

To have a best solution for the stated problem we, design a model as shown in the figure below:

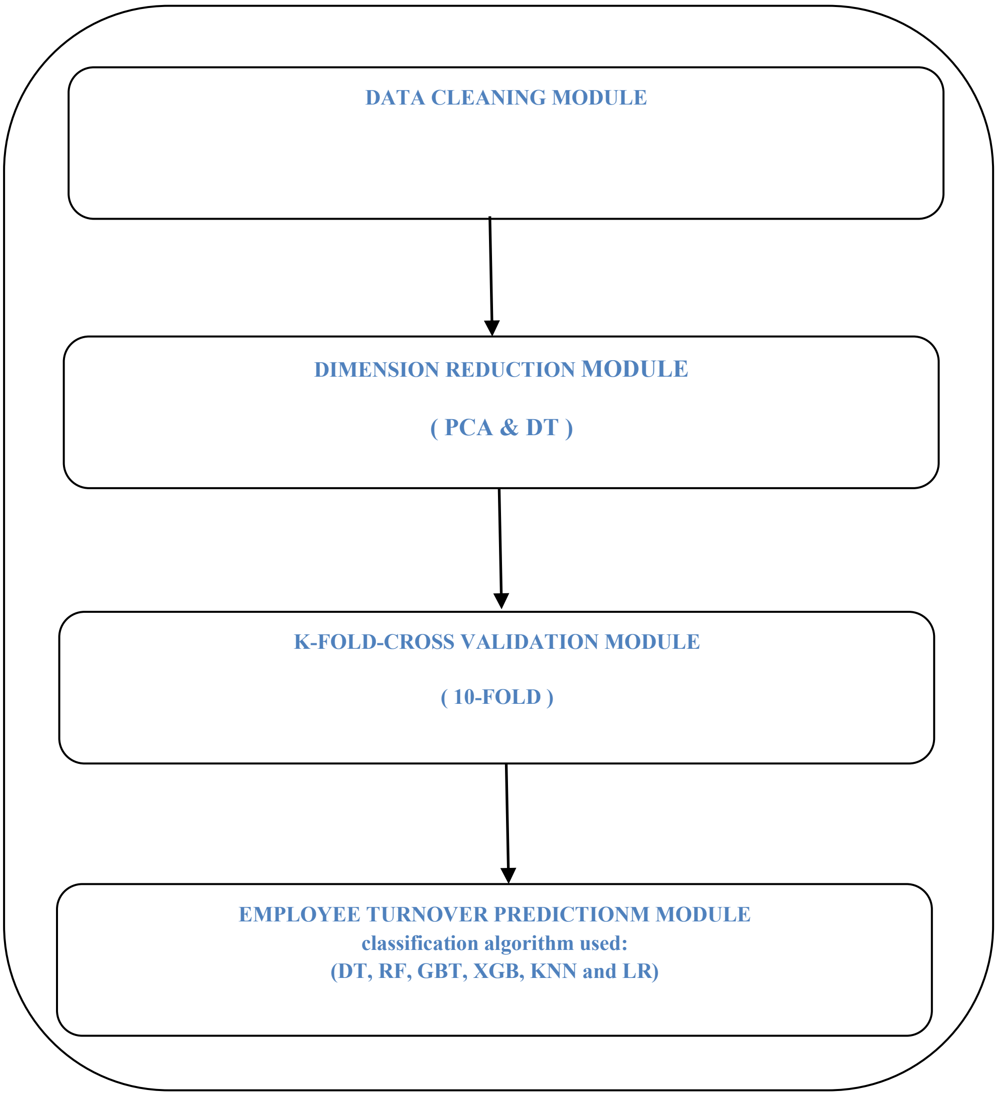

Figure 2.1. A Propose IGEP Model

III. Experiment

3.1 Data collection:

The data used in this work is a secondary data obtain from a public repository. It contains both personal, demographic, turnover status and temporal information of employees within a couple of 15 month starting from 1st January 2018 to 30th march, 2019.

Moreover, as part of the selection criteria, we consider a staff with at least two different working experience and a limit of 18 month as the accepted time of quitting. i.e., any staff that quit before the cutup time will be excluded.

3.2 Data Pre- Processing

As indicated in the model diagram, data cleaning is the first step in data mining after collection. Based on the type of data at hand, we used feed-forward and a decision tree to select the best feature from the demographic information and applied principal component analysis to compose the vectors into a unit of components.

After preparing the data, we form a bipartite graph containing 47,257 staff's vertex and 23,457 companies or enterprises vertices connected by 203,604 edges, respectively.

Also from the staff's biodata, we selected nine relevant basic features namely.

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 3.1 below.

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

Table 3.1: Variable's description tables

S/N	TYPE OF FEATURE	NAME OF FEATURE
1	Staff personal information	Gender, Highest certificate obtained.
2	Present-work information	Date of first Appointment, Unit of work posted, Development levy, Code of present scale, Present status held.
03	Staff Work experience.	Start year of waking career, Turnover status.

3.3 Prediction algorithms used.

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

i. **A Random-forest algorithm (RF):** A random forest algorithm is an ensemble learning that is mostly used in prediction problem, [19], [21] & [32]

ii. **An Extreme gradient boosting Algorithm**: This is a boosting ensemble learning algorithm that is historically known to be good in a prediction problem [5], [25] & [26].

iii. **A Decision-tree algorithm (DT):** Decision tree algorithm is a tree-based structured algorithm having it internal node as a test features while its branches represent the test output. [3], [22] & [30].

iv. **A Gradient-boosting-tree algorithm (GBT)**: this is a boosting learning algorithm that construct many weak decision trees per iteration to reduce the sample loss. [23] & [24].

v. **K-nearest-neighbors algorithm (KNN):** is a classification learning algorithm that work by observing a distance between query, select a number 'k' value that is closer to the query and vote for the most occurring label. [36]

vi. **Logistic-regression algorithm (LR):** This is a gradient descent algorithm that such for an optimized parameter of a linear model having a minimum loss-function.[3].

3.4 Metrics used to evaluate the model performances:

a. F1 score: this is refers to the average of precision & recall

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \ldots\ldots\ldots\ldots\ldots\ldots\ 3.1$$

b. Precision: This is said to be the ratio of all positive cases in all samples identified to be positive.

$$Precision = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.2$$

c. Recall: this is a ratio of apparent positive case in the prediction to all actual positive cases.

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

KEYS**:** True positive [TP], True negative [TN], False negative [FN], and False positive [FP].

4.0 Research result and discussion

1.1 Model evaluation with the staff demographic data only.

To facilited an effective comparison of the new IGEP-Model we simultaneously run and compare the performances of the new model including and then excluding the new embedded features as follows.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 1 Jan. 2026, pp: 206-214 www.ijaem.net ISSN: 2395-5252

Table 4.1 Model performance with the staff demographic data features

S/N	CLASSIFIERS USED	EVALUATION METRICS					
		ACCURACY	PRECISION	RECALL	F1	AUC_ROC	AVERAGE PRECISON
1	DT	0.9154	0.6791	0.7337	0.6109	0.8378	0.5985
2	RF	0.9235	0.7978	0.6996	0.6915	0.8982	0.7844
3	GBT	**0.9572**	**0.8915**	0.6718	0.6930	**0.9088**	**0.8432**
4	XGB	0.9345	0.7416	0.7049	0.6427	0.9321	0.8374
5	LR	0.9001	0.8103	**0.9001**	**0.8528**	0.5621	0.2015
6	KNN	0.9001	0.6593	0.5851	0.5470	0.8184	0.6028

From the above table we can observed the most significance improvement with the GBT algorithm having an accuracy score of 0.9572, precision of 0.8915 AUC-ROC of 0.9088 and an average precision of 0.8432. However, this is because GBT is a boosting ensemble learning algorithms that work with non-linear data effectively.

Moreover, logistic regression is seen to excel in recall and F1-measure with a score of 0.9001 and o.8528, respectively. Again, this might be because a recall metrics aimed to reduce the false positive rate and as well F1 tend to measure the degree to which such false prediction is minimized.

4.2 A Model performance with IGEP features and employee demographic features.

To train and validate the model performance with this features, we composed the embedded features in to six independents component using the stated PCA algorithm, (i.e., PCA-1, PCA-2, … PCA-6), and then combine each with the staff's demographic feature to make the prediction.

The subscript used at each prediction algorithm (Table 4.2) indicate the 'k' value of the PCA Algorithm. (where: 'k' represents the number of components in which the embedded feature was composed)

Table 4.2: A Model performance with IGEP features and employee demographic features.

S/N	CLASSIFIERS USED	EVALUATION METRICS					
		ACCURACY	PRECISION	RECALL	F1	AUC_ROC	AVERAGE PRECISON
1	DT_4	0.9283	0.7546	0.7944	0.7218	0.8752	0.6172
2	RF_6	**0.9566**	**0.9041**	0.7551	0.7676	**0.9453**	0.8888
3	GBT_6	0.9522	0.9003	0.7878	0.7935	0.9386	0.9060
4	XGB_6	0.9393	0.8606	0.8009	0.7820	0.9424	**0.9062**
5	LR_3	0.9004	0.8127	**0.8997**	**0.8526**	0.6428	0.2438
6	KNN_1	0.9318	0.6331	0.6564	0.6153	0.8332	0.6718

Similarly, after the appending the new-IGEP model feature [Table 4.2] it can be observed that random forest exhibits the best result with it corresponding accuracy score of 0.9566, precision score of 0.9041, AUC_ROC score of= 94.53, However, this improvement became possible considering that random forest is a family of ensemble learning that operate a Non-linear that more accurately than other learners.

Table 4.3: Model performance comparison before and after inclusion of embedded feature

S/N	CLASSIFIERS USED	EVALUATION METRICS					
		ACCURACY	PRECISION	RECALL	F1	AUC_ROC	AVERAGE PRECISON
1	DT_4	0.129	0.7550	0.607	0.11.09	0.374	0.187
2	RF_6	**0.331**	0.1063	0.555	0.7610	0.471	**0.1044**
3	GBT_6	-0.500	0.8800	**0.1160**	0.1005	0.298	0.628
4	XGB_6	0.480	**0.1190**	0.9600	**0.1393**	0.103	0.6880

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 1 Jan. 2026, pp: 206-214 www.ijaem.net ISSN: 2395-5252

5	LR_3	0.030	0.2400	-0.040	-0.0200	**0.807**	0.4230
6	KNN_1	0.317	-2.620	0.713	0.6830	0.148	0.6900

Table 4.3 above indicate the performance difference between table 4.1 and that of table 4.2, from this simple comparison we can observed a positive improvement in most of the algorithms which signifies the supremacy the proposed model.

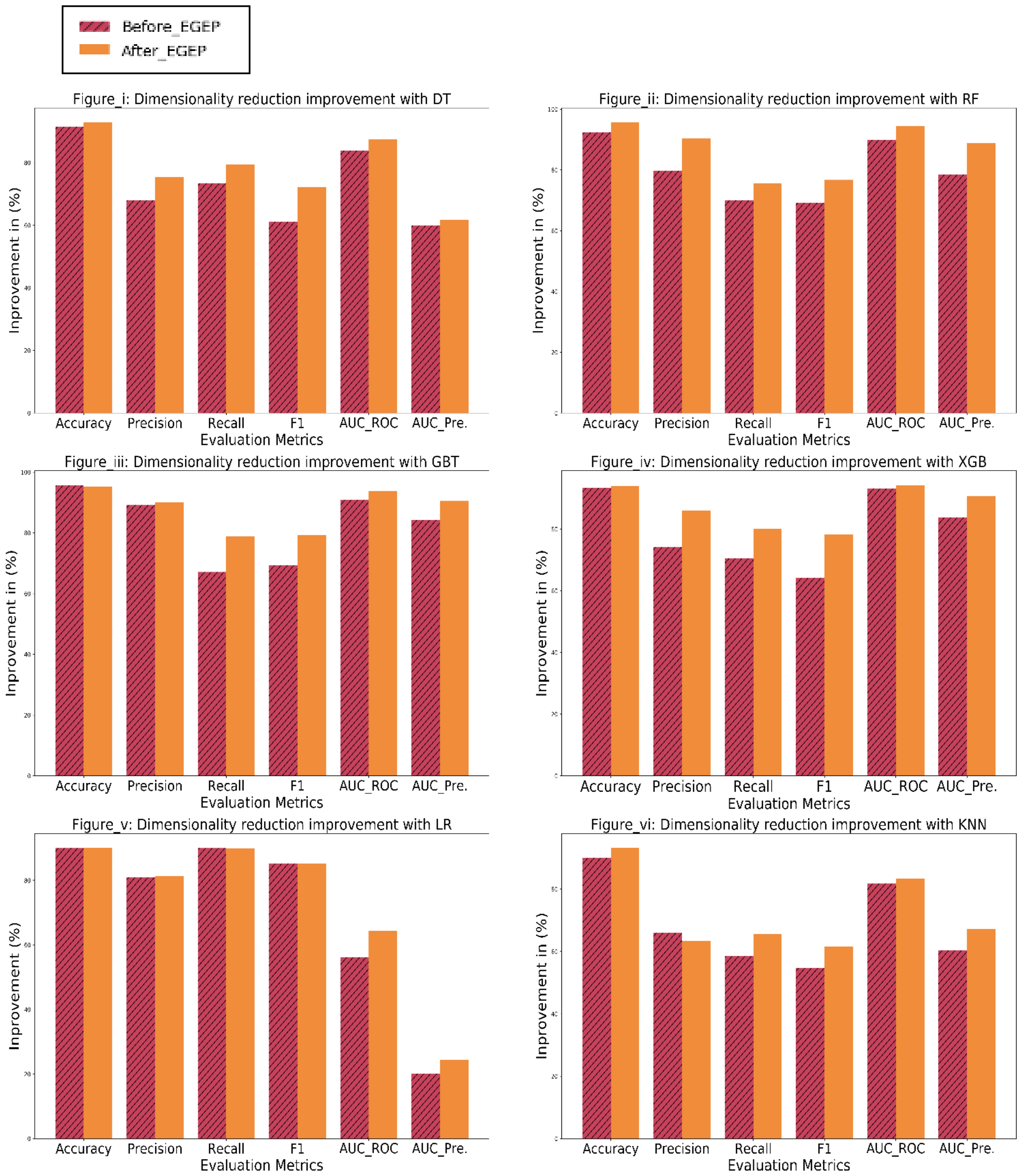

Fig 4.1: model performance attrition prediction with / without IGEP-features

Figure 4.2 below, is a feature importance ranking made with decision tree to further explore the features that contribute immensely to the performance improvement seen in table 4.3 above, from the ranking it is seen that vec-7 and vec-8 emerge 3rd and 4th position in the plot, this tell us that they perform better than most of the demographic features previously used for this type of problem.

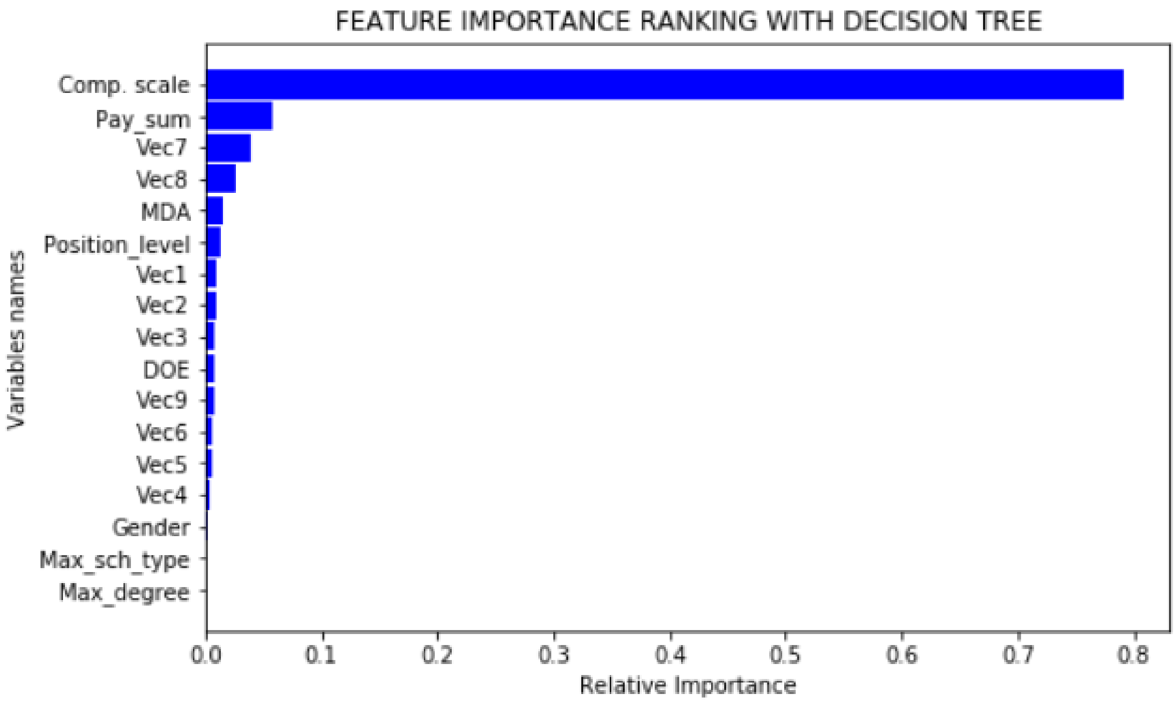

Fig. 4.2: Model feature importance ranking with DT.

V. Conclusion and Future Research Direction.

In this work, we examine the new model call IGEP-model that examine the role of PCA algorithm and cross validation technique in modeling an embedded feature as a set of components and help in solving an unbiased employee turnover prediction performance effectively.

Furthermore future research may focused on using another set of data, possibly from other continent that contain different feature not considered in this context to have a wide coverage of the problem..

References

[1] Vardaman, J. M., Taylor, S. G., Allen, D. G., Gondo, M. B., and Amis, J. M. (2023). Translating intentions to behavior: The interaction of network structure and behavioral intentions in understanding employee turnover. Organization Science, 26(4), 1177-1191.

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

[3] Zhang, Y., Xiong, Y., Kong, X., and Zhu, Y. (2024, November). Learning node embeddings in interaction graphs. In Proceedings of the 2024 ACM on Conference on Information and Knowledge Management (pp. 397 406). ACM.

[4] 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. https://doi.org/10.1109/TKDE.2018.2873341

[5] 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. https://doi.org/10.1109/ACCESS.2020.2965544

[6] Lin, B., Robles, G., amd Serebrenik, A. (2023, May). Developer turnover in global, industrial open source projects: Insights from applying survival analysis. In 2023 IEEE 12th International Conference on Global

Software Engineering (ICGSE) (pp. 66-75). IEEE.

[7] 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. https://doi.org/10.1016/j.tourman.2019.04.030

[8] 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.

[9] Wang, D., Cui, P., and Zhu, W. (2024, August). Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1225-1234). ACM.

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

[11] 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. https://doi.org/10.1109/TKDE.2018.2807452

[12] 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. https://doi.org/10.1145/3269206.32720231997–2005.

[13] Xu, H., Yu, Z., Xiong, H., Guo, B., & Zhu, H. (2015). *Learning Career Mobility and Human Activity Patterns for Job Change Analysis*. 1057–1062. https://doi.org/10.1109/ICDM.2015.122

[14] 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.

[15] 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. https://doi.org/10.1080/00909881003639544

[16] 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*.

[17] 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.

[18] Goldberg, Y., & Levy, O. (2014). *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. *2*, 1–5. http://arxiv.org/abs/1402.3722

[19] 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

[10] 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. http://dx.doi.org/10.1145/2623330.2623732.701–710.

[21] Tang, J., & Qu, M. (n.d.). (2015). LINE: Large-scale Information Network Embedding. International World Wide Web Conference Committee (IW3C2). *P1067-Tang*. 1067–1077. http://dx.doi.org/10.1145/2736277.2741093.

[22] 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. https://doi.org/10.1145/3209978.3209987

[23] Chen, J., Gong, Z., Wang, W., & Liu, W. (2021). HNS: Hierarchical negative sampling for network representation learning. *Information Sciences*, *542*, 343–356. https://doi.org/10.1016/j.ins.2020.07.015

[24] 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.

[25] 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. https://doi.org/10.1109/TKDE.2018.2833443

[26] Zhao, Y., Hryniewicki, M. K., & Cheng, F. (2019). *Employee Turnover Prediction with Machine Learning: A Reliable Approach* (Vol. 1). Springer International Publishing. https://doi.org/10.1007/978-3-030-01057-7

[27] 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. https://doi.org/10.1007/978-3-030-01177-2

[28] 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. https://doi.org/10.1108/IJOA-10-2019-1903

[29] 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. https://doi.org/10.1016/j.eswa.2012.02.005

[30] 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.*

[31] Khera, S. N. (2019). *Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques*. https://doi.org/10.1177/0972262918821221

[32] 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.

[33] 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.

[34] 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.

[35] Saradhi, V. V., & Palshikar, G. K. (2011). Expert Systems with Applications Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006. https://doi.org/10.1016/j.eswa.2010.07.134

[36] 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. https://doi.org/10.1007/978-3-319-63673-3

[37] Yu, W., Cheng, W., Aggarwal, C. C., Zhang, K., Chen, H., and Wang, W. (2022, July). Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2672-2681). ACM.