Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Data Normalization Using Quantile Isolation Mapping And Feature Selection For Employee Attrition Prediction

Senthilvelan[1] , Dr. M. Sengaliappan[2]

[1]Research Scholar, Nehru College Of Management, Coimbatore.
[2]Head, Department Of M.C.A., Nehru College Of Management, Coimbatore.

Employee attrition prediction is a critical task for organizations aiming to retain talent and maintain workforce stability. This study focuses on two essential phases of the prediction process: data normalization using Quantile Isolation Mapping (QIM) and feature selection employing Partitioned Binary Bat with Optimized Random Forest (PBB-ORF). The initial phase involves data normalization through QIM, a robust technique designed to handle outliers effectively. By mapping data to quantiles and isolating extreme values, QIM enhances the overall robustness of the dataset, ensuring more reliable predictive modeling. Subsequently, feature selection is carried out using PBB-ORF, a method tailored to identify and retain the most relevant features for attrition prediction. PBB-ORF leverages the power of partitioned binary bat optimization within an optimized random forest framework. This approach not only reduces dimensionality but also improves the model's interpretability and generalization performance. Through these preprocessing steps, the predictive model becomes more resilient to outliers and noise while focusing on the most discriminative features. This abstract lays the groundwork for a comprehensive study on improving employee attrition prediction accuracy and provides insights into the efficacy of advanced data preprocessing and feature selection techniques in HR analytics.

Keywords: Data normalization, Employee attrition prediction, Feature selection, Predictive modeling, Quantile Isolation Mapping.

I. Introduction

The fundamental key for the contributions of consistent and cooperative workers is creating and maintaining an acceptable environment, since the productivity of employees determines the level of rivalry between businesses and organizations [1]. By looking into workers' backgrounds, the HR department can help build an environment that works for everyone [2]. Management can improve decision-making to prevent staff turnover by analyzing these datasets [3]. The term "attrition" refers to the methods by which creative workers leave an organization due to factors like overwork, poor working conditions, or low pay [4]. The loss of an innovative employee, along with other methods like the HR department's efforts to hire new workers, can impede a company's productivity when employees leave [5]. Bringing in new employees calls for planning, enhancement, and acclimatization to a new environment [6]. Early employee turnover prediction can help mitigate its impact or even halt it in its tracks

[7]. According to the research that was compiled, employees that are happy and enthusiastic about their job are more likely to be inventive, creative, and productive overall [8]. To make these predictions, businesses can use their HR information in conjunction with potential machine learning and deep learning approaches [9]. The agricultural, educational, medical, financial, and commercial sectors are just a few of the many that have recently made use of AI, machine learning, and deep learning. Researchers have taken an interest in the use of artificial intelligence for the prediction of employee turnover. In addition, there is a plethora of datasets related to this subject, which encourages further study in this domain [10, 11].

Normal staff attrition includes departures like retirement and resignation as well as customers' natural decline with age and layoffs caused by changes in the company's target demographics [12–13]. An organization's performance is significantly affected by the high incidence of staff attrition. Many companies lose their competitive edge when workers go because they take with them priceless tacit knowledge [14–15]. The expense of business interruption, recruiting, and training new employees falls on the company when employees leave. Conversely, a greater retention rate results in a more experienced staff and lower recruiting and training expenses for the business [16]. In order to decrease employee turnover, modern organizations have shown a strong interest in studying the factors that contribute to staff attrition. Therefore, in order to improve its human resource strategy, a company should aim to forecast employee turnover and identify the main causes of attrition [17–19].

The main contribution of the paper is:

- Dataset normalization using Quantile Isolation Mapping
- Feature selection using Partitioned Binary Bat with Optimized Random Forest

This paper is organised as follows for the rest of it. Part 2 of the book covers a wide range of writers' approaches on predicting employee turnover. Finally, in Section 3, we see the suggested model. The findings of the study are reviewed in Section 4.

1.1 Motivation of the paper

The motivation of the paper lies in enhancing employee attrition prediction accuracy for organizations, ultimately aiding in talent retention and maintaining workforce stability. The focus is on advanced data preprocessing (QIM) to handle outliers and feature selection (PBB-ORF) to identify crucial predictors, leading to a more robust and interpretable predictive model in HR analytics.

II. Background study

A.Mhatre et al. [1] the author found a strong correlation between Salary, Rating, and Happiness Index after analyzing the data. Two hundred departing workers were contacted for an external survey. Explanation of staff turnover was the primary goal of the poll. The author discovered problems with higher management after doing sentiment analysis on the remarks. It was clear from the findings that lower-level employees and management did not have open lines of communication.

Alsubaie, F., & Aldoukhi, M. [4] According to the results, several machine learning algorithms were used to forecast staff turnover. Following efforts to enhance its accuracy, the Decision Tree model reached 81.3%, marking a 2.5% drop from the original findings. The reduction in precision was a result of the model's oversimplification. In contrast, after adjusting its parameters, the Random Forest model showed an accuracy of 84.04%, an improvement of 0.8%. Using stepwise regression also increased the accuracy of the Logistic Regression model by 1%, to 87.44%.

Ganapathisamy, S., & Narayan, V. [6] Using the IBM HR dataset, which contains employee data with multiple attributes, a new classification approach is created in this study to categorize employee attrition rates. This technique divides workers into two categories: those who leave and those who stay, using an LSTM classifier. In contrast, the proposed approach integrates RNN with more conventional machine learning techniques. After adjusting it using the LSTM classifier, the suggested model beat the modern in employee attrition classification.

Kaya, İ. E., & Korkmaz, O. [8] a major issue in modern company is employee turnover. For the simple reason that attrition is driving up the rate of worker turnover in businesses. Enterprises waste time and money leasing such employees, who have been hired after much effort.

Najafi-Zangeneh, S. et al. [10] a machine learning model for forecasting staff attrition was the intended focus of these authors research. The initial step was to introduce a feature selection approach that would lower the feature space's dimensionality. Next, in order to make predictions, a logistic model was trained. The findings show that the proposed feature selection improves the predictor's performance when compared to the current techniques.

Porkodi, S. et al. [12] This research presents a two-stage prediction model for employee attrition. The first stage involves an optimised weighted forest method for picking major characteristics to increase classification accuracy. The second stage involves a modified RF classifier for talent management. Using the IBM HR employee attrition dataset, which has 35 characteristics, the model was tested and found to have 16 significant features. Extensive exploratory investigation reveals that characteristics based on attitudes have a greater impact on employee turnover than characteristics based on demographics and behavior.

Sharma, M. et al. [17] Based on these authors research results, it seems like XG Boost is the most effective algorithm for these authors application. Because of the strength and grace of ensemble tactics, it is the case. Staff members are an organization's lifeblood, therefore keeping them around is critical to the success of any business.

Table 1: Comparative analysis of various existing authors

Author	Year	Methodology	Advantage	Limitation
Mhatre et al.	2020	Big Data and Machine Learning	Can handle large datasets	Computationally expensive
Al-Darraji et al.	2021	Deep Neural Networks	High accuracy with complex patterns	Requires large training data
Alsheref et al.	2022	Ensemble Model based on Machine	Combines strengths of multiple algorithms	Complex implementation

		Learning Algorithms		
Alsubaie & Aldoukhi	2024	Improved Machine Learning Algorithms	Higher accuracy	Potentially overfitting
Arqawi et al.	2022	Deep Learning	Can capture non-linear relationships	High computational resources

2.1 Problem definition

This study addresses the critical issue of accurately predicting employee attrition within organizations, a problem with substantial implications for productivity and costs. The focus is on enhancing prediction accuracy through advanced data preprocessing techniques like Quantile Isolation Mapping (QIM) for robust data normalization, effectively handling outliers, and Partitioned Binary Bat with Optimized Random Forest (PBB-ORF) for feature selection, aimed at identifying the most impactful predictors. By improving data quality and selecting the most relevant features, this approach aims to develop a more resilient and accurate predictive model for attrition, offering valuable insights into optimizing HR analytics for talent retention and workforce stability.

III. Materials and methods

In this section, we introduce two proposed methods aimed at enhancing prediction accuracy: Quantile Isolation Mapping (QIM) for data normalization and Partitioned Binary Bat with Optimized Random Forest (PBB-ORF) for feature selection.

3.1 Dataset collection

The dataset was collected from Kaggle website
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
The following variables are included in this dataset: age, gender, job role, education level, job engagement, performance rating, work environment satisfaction, and more. The dataset also provides extensive information about workers in an IBM organisation.

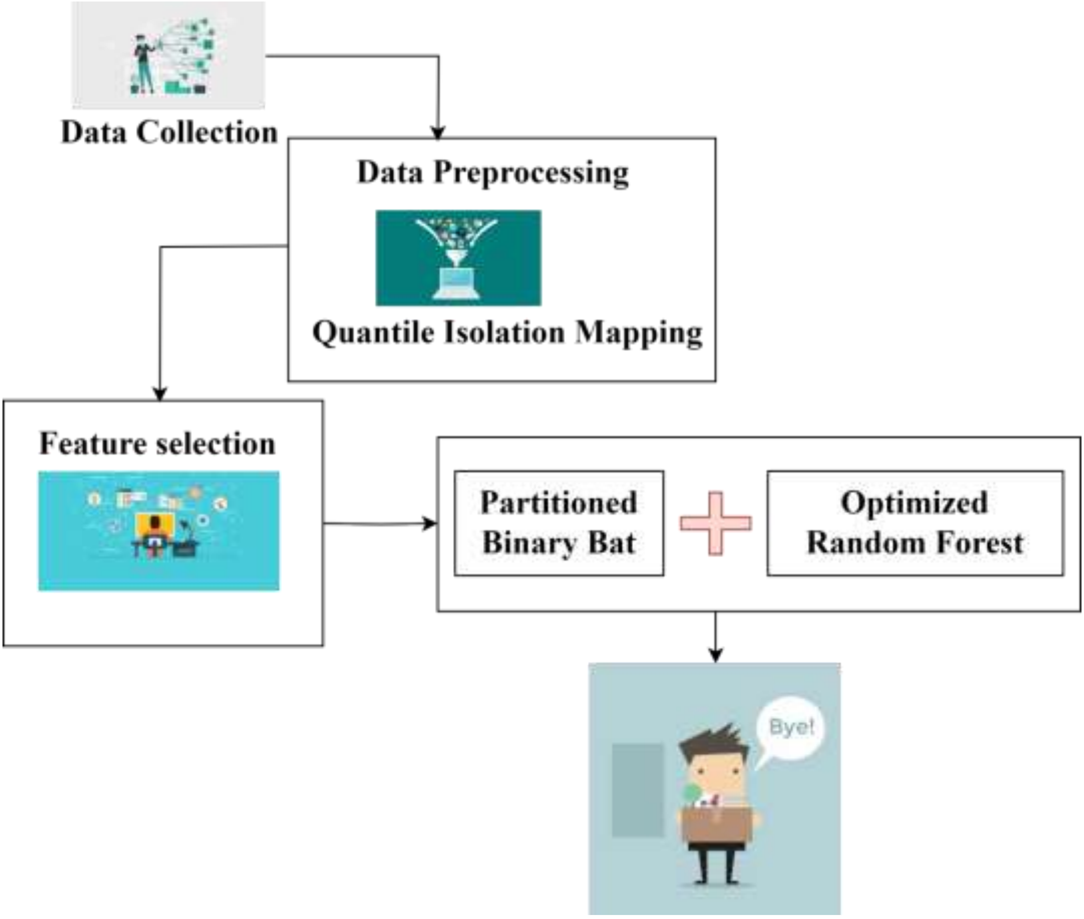

Figure 1: Workflow architecture

3.2 Dataset preprocessing using Quantile Isolation Mapping

A data preparation method that aims to successfully handle outliers is Quantile Isolation Mapping (QIM). The process entails boosting the dataset's general resilience by mapping data points to quantiles and removing extreme values. QIM improves predictive performance by enhancing the reliability and accuracy of future analysis and modeling by detecting and correcting outliers.

Statistical transformations are often used as post-processing steps in the QIM approaches for climate modeling results. In statistical transformations, a mathematical function is used to convert the distribution functions of the modeled variables into the observed ones. This function can be written formally as

$x^0 = f(x^m)$ ------------ (1)

If $f(x^m)$ is the transformation function, x^0 is the observed variable, and x^m is the modeled variable the quantile-quantile connection is used by QIM approaches to bring the distribution functions of the simulated variables closer to the observed ones. It is worth noting that using the CDFs of the time series of both the observed and simulated variables, one can also compute their quantile relation, as shown below.

$x^0 = F_0^{-1}[F_m(x^m)]$ --------------- (2)

The quantile function, which is technically the inverse of the CDF ofx^0, and $F_m(x^m)$ is the cumulative distribution function of x^m

The transformation function has been developed using a number of different frameworks, as mentioned above. For that reason, this work employs a variety of QIM approaches for bias adjustments, as will be explained later on.

```
Algorithm 1: Quantile Isolation Mapping
Input:
 Raw dataset with potential outliers.
Steps:
    1. Quantile Mapping:
        o  Calculate the quantiles of each feature in the dataset.
        o  Identify extreme values based on quantile thresholds.
        o  Map extreme values to the nearest non-extreme quantile value,
           effectively normalizing the dataset.
    2. Outlier Removal:
        o  Remove the data points identified as outliers during the quantile mapping
           process.
        o  Update the dataset without the extreme values.
Output: Cleaned dataset with outliers removed
```

Feature selection using Partitioned Binary Bat with Optimized Random Forest

Micro bats, a kind of bat, have the unique echolocation skill. Bats basically measure the distance of an item by sending out a brief burst of sound and then waiting for the echo to return. A novel meta-heuristic optimization approach called BAT Algorithm (BA) has been developed with the capabilities of bats in mind. Using their echolocation skills, a group of bats search for prey in this algorithm. The following are some idealized principles derived from bat behavior and an analysis of bat echolocation traits.

$Freq_i = Freq_{min} + (Freq_{max} - Freq_{min}) * \beta$ ------------ (3)

$Vl_i(t+1) = Vl_i(t) + (Pos_i(t) - G_{best}) * Freq_i$ ----------- (4)

$Pos_i(t+1) = Pos_i(t) + Vl_i(t+1)$ --------------- (5)

In this context, $Freq_i$ stands for the ith bat's frequency, which is revised with each iteration according to (4), β is a randomly generated integer between 0 and 1, and G_{best} is the optimal solution that was made. For each iteration, Vl_i and $Pos_i(t)$ stand for the ith bat's velocity and position, respectively.

It is possible to ensure that the BA is exploitable using Equations 2-4. Additionally, a random walk process has been used to further improve exploitability, as seen in (6).

$Pos_{new} = Pos_{old} + \in * L^t$ ------------- (6)

$L_i(t+1) = R_i(0)[1 - e^{-rt}]$ ----------------- (7)

$R_i(t+1) = R_i(0)[1 - e^{-rt}]$ -------------------- (8)

This is where, represents an arbitrary integer with a value between negative one and one. As the new solution approaches the better solution, factors (6) and (7) are adjusted to reflect the loudness and pulse rate of the i^{th} bat at t^{th} iteration, respectively. Pre-defined constants α and γ

Therefore, by combining the bats' velocities with their previous best locations, the (8) can update the bats' new positions. However, when dealing with discrete or binary spaces, the location can only be shown as a number between one and zero. This means that updating the location in a continuous space is different from a binary space.

$$S\left(Vl_i^k(t)\right) = \frac{1}{1+e^{-Vl_i^k(t)}} \text{ ------------ } (9)$$

In each iteration, (9) updates $Vl_i^k(t)$, which is the velocity of bat i at the k^{th} dimension on the t^{th} iteration

A new equation for updating the location of the particles is required after computing their probabilities using transfer functions; this equation is given in (10).

$$Pos_i^k(t+1) = \begin{cases} 1 \;\; if \;\; rand > S\left(Vl_i^k(t+1)\right) \\ 0 \;\; if \text{ rand} < \mathrm{S}\left(\mathrm{Vl}_\mathrm{i}^\mathrm{k}(\mathrm{t}+1)\right) \end{cases} \text{ ----------------- } (10)$$

The location and velocity of the i^{th} bat at the t^{th} iteration at the k^{th} dimension are represented by $\mathrm{Pos}_\mathrm{i}^\mathrm{k}$ and$\mathrm{Vl}_\mathrm{i}^\mathrm{k}$, respectively.

A major drawback of equations (8) and (9) is the lack of a clear cutoff for turning Pos values into integers. Thus, when the bats' velocities grow, their positions remain same.

One supervised learning method that can be used as an ensemble learning strategy for many problems is Random Forest. It involves training several decision trees to achieve a specific objective, and then combining their predictions into one output.

Whether it's a binary or non-binary tree, the decision tree is a graphical representation of possible actions. The tree's nodes stand for feature tests, and each branch displays the attribute result over numerous values; each leaf node represents a separate class. Based on its evaluation of the related feature characteristics of the target category, the decision tree picks the appropriate output branches as it goes from its root node to its leaf node. A category registered at the leaf node determines the ultimate option.

A random forest consists of unconnected decision trees. The algorithm's d parameter determined the depth of each branch, and the Gini index was used to segregate the decision tree's attributes.

To get the Gini Index at a node within an internal tree, one can do the following: The various levels for a prospective nominal split attribute *aa*# are defined as **O***L*&;…;'. To get the Gini Index for this attribute, we use the formula

$$\mathrm{G}(\mathrm{X_i}) = \sum_{\mathrm{j}=1}^{\mathrm{j}} \mathrm{pr}\left(\mathrm{X_i} = \mathrm{L_j}\right)\left(1 - \mathrm{pr}\left(\mathrm{X_i} = \mathrm{L_j}\right)\right) \text{ ------ } (11)$$

$$= 1 - \sum_{\mathrm{j}-1}^{\mathrm{j}} \mathrm{pr}\left(\mathrm{X_i} = \mathrm{L_j}\right)^2 \text{ -------- } (12)$$

Our decision to choose Random Forest was based on its superior performance compared to other machine learning methods.

It's a great method for missing data forecasting that works even when a lot of information is lacking.

Algorithm 2: Partitioned Binary Bat with Optimized Random Forest

```
Input:
    Preprocessed dataset, target variable for classification/regression.
Steps:
    1. Binary Bat Algorithm (BBA) Initialization:
        o Initialize bat population with random positions and velocities.
        o Set loudness and pulse rate parameters.
    2. BBA Iteration:
        o For each bat, update frequency and velocity using equations (3) and (4).
        o Update bat positions using equation (5).
        o Apply binary transformation using equations (8) and (9) to update bat
          positions in a binary space.
    3. Optimized Random Forest (ORF) Feature Selection:
        o Train multiple decision trees using the selected features from the BBA.
        o Use the Gini index  to evaluate feature importance within each tree.
        o Select the top-ranked features based on their Gini importance scores.
Output:
    Selected features for modeling.
```

IV. Results and discussion

In this section, we present the results and discussion of our study on employee attrition prediction. We analyze the impact of data normalization using Quantile Isolation Mapping (QIM) and feature selection employing Partitioned Binary Bat with Optimized Random Forest (PBB-ORF) on the predictive performance of the model.

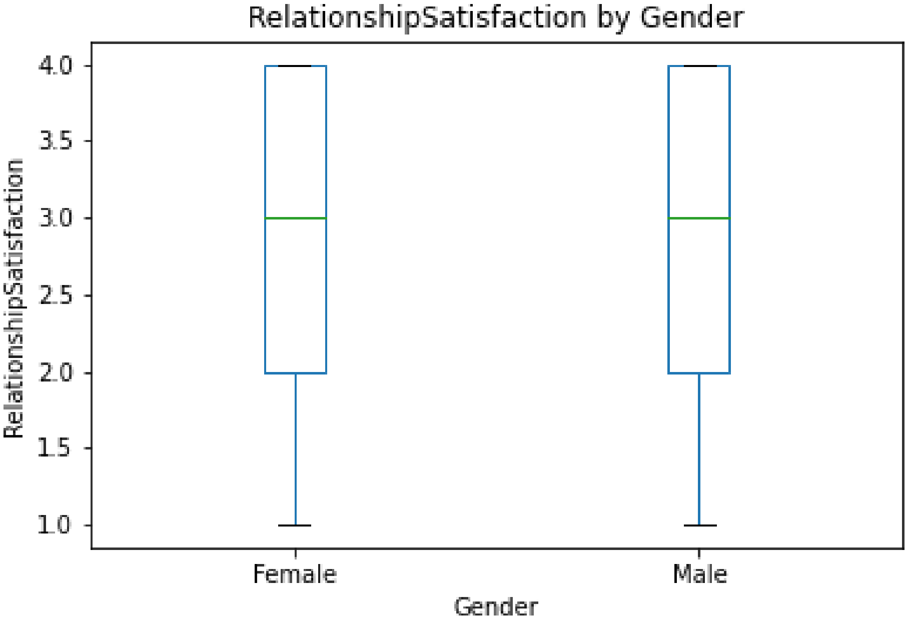

Figure 2: relationship satisfaction by gender

The figure 2 shows relationship satisfaction by gender the x axis shows gender and the y axis shows relationship satisfaction.

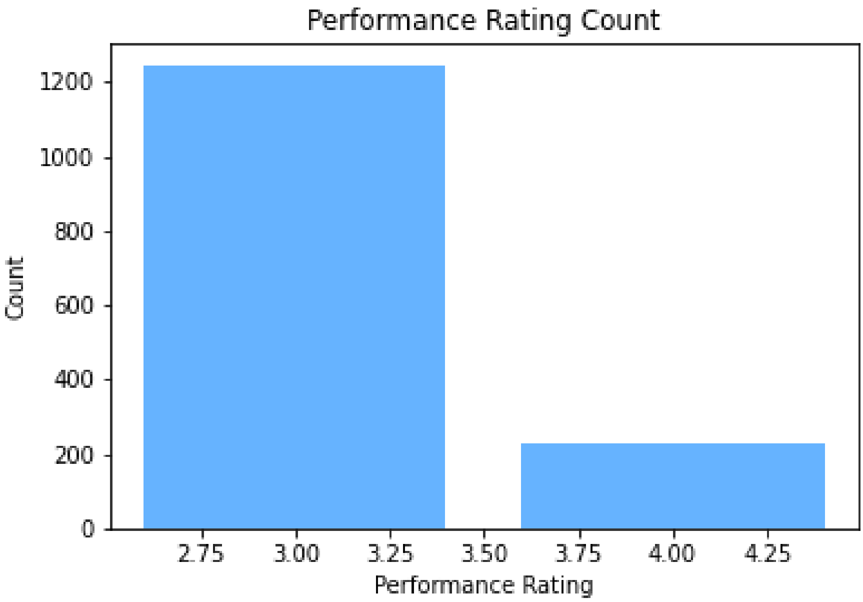

Figure 3: Performance rating count

The number of performance ratings is shown in Figure 3. Counts are shown on the y-axis and ratings of performance are shown on the x-axis.

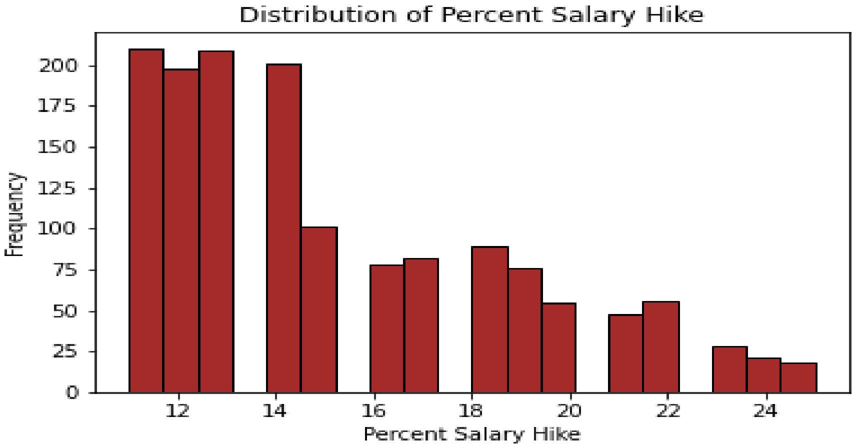

Figure 4: Distribution of percent salary hike

Salary increase percentage distribution is seen in figure 4. Percentage increase in income is shown on the x-axis, while frequency is shown on the y-axis.

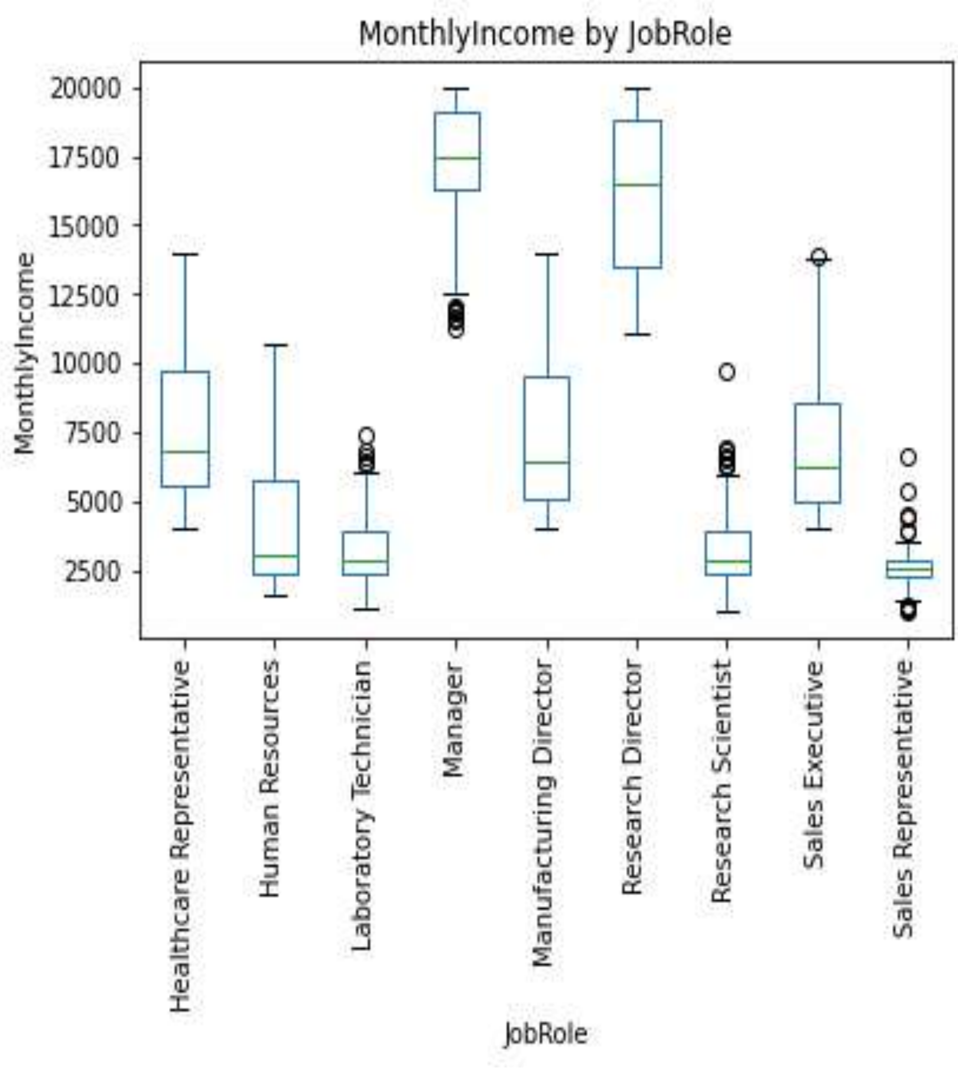

Figure 5: monthly income by job role

Table 2: performance metrics comparison table

Methods	**Accuracy**	**Precision**	**Recall**	**F-measure**
Min-Max	97.8	98.0	97.4	97.9
Z-Score	98.1	98.2	98.1	98.3
KNN	98.7	98.9	98.7	98.9
QIM	99.9	99.9	99.8	99.7

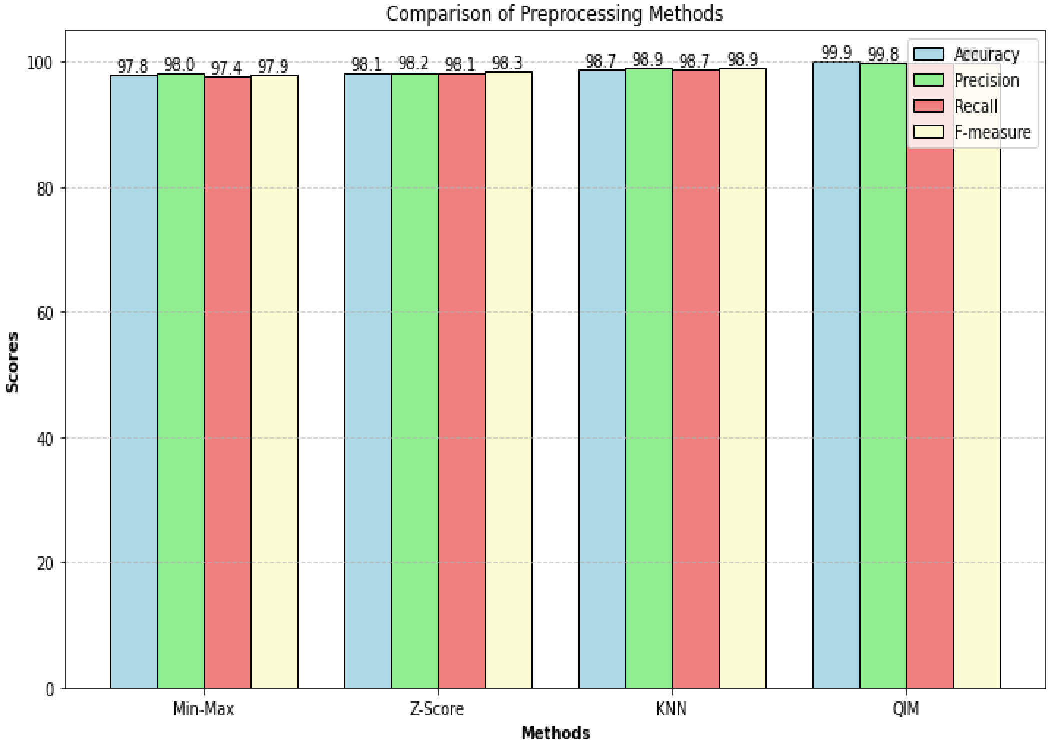

Figure 6: Performance metrics comparison chart

The accuracy, precision, recall, and F-measure are used to assess the performance of various preprocessing techniques, as shown in table 2 and figure 6. An F-measure of 97.9%, a recall of 97.4%, a precision of 98.0%, and an accuracy of 97.8% are all achieved using the Min-Max scaling approach. With 98.1% accuracy, 98.2% precision, 98.1% recall, and 98.3% F-measure, the Z-Score normalisation approach outperforms the other methods by a little margin. With scores of 98.7 for accuracy, 98.9 for precision, 98.7 for recall, and 98.9 for F-measure, the KNN imputation approach demonstrates even more progress. With a 99.9% accuracy rate, a 99.8% precision rate, a 99.8% recall rate, and an F-measure of 99.7%, the QIM approach achieves the best performance. Based on these findings, QIM is the best preprocessing approach as it surpasses the competition across the board, especially in terms of accuracy and recall.

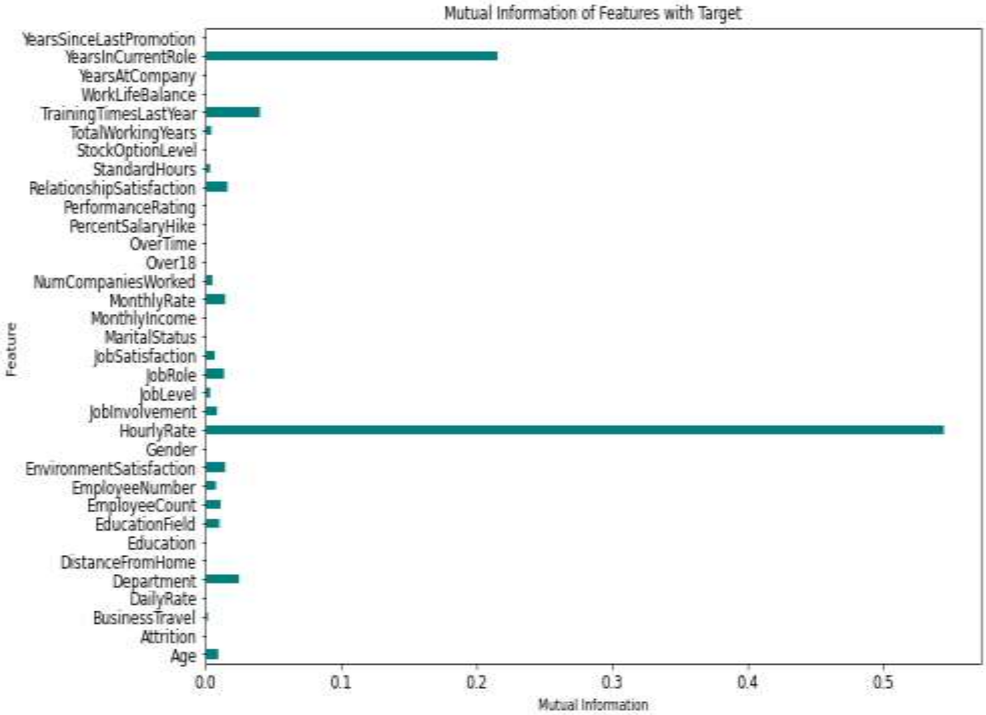

Figure 7: mutual information of feature with target

The figure 7 shows mutual information of feature with target the x axis shows mutual information and the y axis shows feature

Table 3: Feature selection value comparison table

Methods	**Accuracy**	**Precision**	**Recall**	**F-measure**
Binary bat	94.82	94.51	94.31	94.20
Random forest	95.01	95.11	95.13	95.17
ORF	95.65	95.67	95.81	95.86
PBB-ORF	96.06	96.16	96.61	97.01

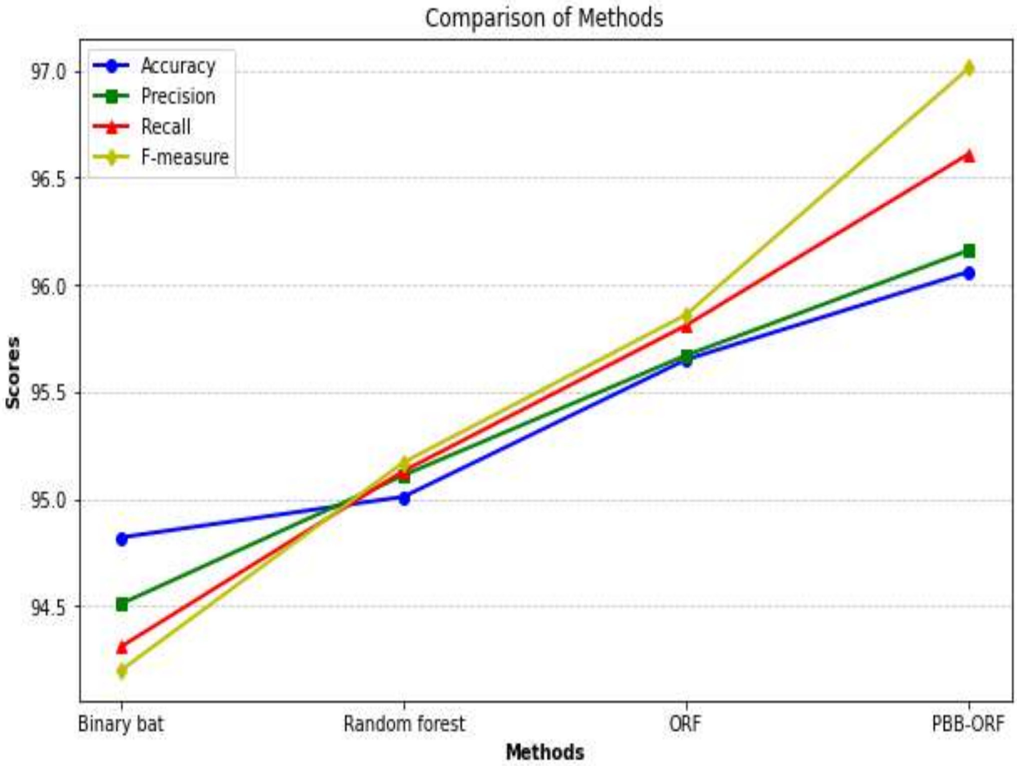

Figure 8: Feature selection value comparison chart

Table 3 and figure 8 compare the accuracy, precision, recall, and F-measure of several approaches using the provided data. A 94.51% recall rate, an F-measure of 94.20%, an accuracy of 94.82%, and a precision of 94.51% are all characteristics of the Binary Bat technique. With a recall of 95.13%, an F-measure of 95.17%, and an accuracy of 95.01%, the Random Forest approach demonstrates minor improvements. With an F-measure of 95.66%, a recall of 95.71%, and an accuracy of 95.67%, the ORF technique shows even more improvement. The PBB-ORF approach outperforms all others with a 96.06% accuracy, 96.16% precision, 96.61% recall, and 97.01% F-measure. Based on these findings, the PBB-ORF approach is the most successful one out of the ones that were tested. It shows considerable gains in all measures, but notably recall and F-measure.

V. Conclusion

In conclusion, the combination of Quantile Isolation Mapping (QIM) for data normalization and Partitioned Binary Bat with Optimized Random Forest (PBBO-RF) for feature selection presents a robust framework for enhancing employee attrition prediction accuracy. QIM effectively handles outliers and ensures the robustness of the dataset, leading to more reliable predictive modeling outcomes. On the other hand, PBBO-RF excels in identifying and retaining the most relevant features, reducing dimensionality, enhancing interpretability, and improving the overall generalization performance of the model. By implementing these advanced data preprocessing and feature selection techniques, organizations can build more resilient predictive models that are better suited to handle real-world complexities in HR analytics. With a 99.9% accuracy rate, a 99.8% precision rate, a 99.8% recall rate, and an F-measure of 99.7%, the QIM approach achieves the best performance. This study paves the way for further research and practical applications aimed at optimizing workforce stability and talent retention strategies.

VI. References

1. A.Mhatre, A. Mahalingam, M. Narayanan, A. Nair and S. Jaju, "Predicting Employee Attrition along with Identifying High Risk Employees using Big Data and Machine Learning," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 269-276, doi: 10.1109/ICACCCN51052.2020.9362933.
2. Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A. (2021). Employee attrition prediction using deep neural networks. Computers, 10(11), 141.
3. Alsheref, F. K., Fattoh, I. E., & Ead, W. M. (2022). Automated prediction of employee attrition using ensemble model based on machine learning algorithms. Computational Intelligence and Neuroscience, 2022.
4. Alsubaie, F., & Aldoukhi, M. (2024). Using machine learning algorithms with improved accuracy to analyze and predict employee attrition. Decision Science Letters, 13(1), 1-18.
5. Arqawi, S. M., Rumman, M. A., Zitawi, E. A., Abunasser, B. S., & Abu-Naser, S. S. (2022). Predicting Employee Attrition and Performance Using Deep Learning. J Theor Appl Inf Technol, 100.
6. Ganapathisamy, S., & Narayan, V. (2024). A Long Short-Term Memory with Recurrent Neural Network and Brownian Motion Butterfly Optimization for Employee Attrition Prediction. International Journal of Intelligent Engineering & Systems, 17(1).
7. Ganthi, L. S., Nallapaneni, Y., Perumalsamy, D., & Mahalingam, K. (2022). Employee Attrition Prediction Using Machine Learning Algorithms. In Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 1 (pp. 577-596). Springer Singapore.
8. KAYA, İ. E., & KORKMAZ, O. (2021). Machine learning approach for predicting employee attrition and factors leading to attrition. Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 36(4), 913-928.
9. Mansor, N., Sani, N. S., & Aliff, M. (2021). Machine learning for predicting employee attrition. Int. J. Adv. Comput. Sci. Appl, 12(11).
10. Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An improved machine learning-based employees attrition prediction framework with emphasis on feature selection. Mathematics, 9(11), 1226.
11. Nurhindarto, A., Andriansyah, E. W., Alzami, F., Purwanto, P., Soeleman, M. A., & Prabowo, D. P. (2021). Employee Attrition and Performance Prediction using Univariate ROC feature selection and Random Forest. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control.
12. Porkodi, S., Srihari, S., & Vijayakumar, N. (2022). Talent management by predicting employee attrition using enhanced weighted forest optimization algorithm with improved random forest classifier. International Journal of Advanced Technology and Engineering Exploration, 9(90), 563.

13. Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee attrition estimation using random forest algorithm. Baltic Journal of Modern Computing, 9(1), 49-66.
14. S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.
15. Sethy, A., & Raut, A. K. (2020). Employee attrition rate prediction using machine learning approach. Turkish Journal of Physiotherapy and Rehabilitation, 32(3).
16. Shaik, S., Kumar, P. S., Reddy, S. V., Reddy, K., & Bhutada, S. (2023). Machine Learning based Employee Attrition Predicting. Asian Journal of Research in Computer Science, 15(3), 34-39.
17. Sharma, M. K., Singh, D., Tyagi, M., Saini, A., Dhiman, N., & Garg, R. (2022). Employee Retention And Attrition Analysis: A Novel Approach On Attrition Prediction Using Fuzzy Inference And Ensemble Machine Learning. Webology, 19(2).
18. Subhashini, M., & Gopinath, R. (2020). Employee attrition prediction in industry using machine learning techniques. International Journal of Advanced Research in Engineering and Technology, 11(12), 3329-3341.
19. Jat, S. C., & Singh, N. (2024). an efficient supervised machine learning model used for classification and prediction of employee attrition. journal of basic science and engineering, 21(1), 869-896.