[1,*] Lihe Ma
[2] Kechao Wang
[3] Yan Wang
[4] Lin Liu
[5] Ning Sha
[6] Lin Ma

Employee Turnover Prediction Based on Ensemble Learning DGNK Model

Journal of Electrical Systems

Abstract: - Employee turnover is a problem that can have significant negative impacts on an enterprise. It not only results in the loss of valuable talent and knowledge but also incurs substantial costs in terms of hiring, onboarding, and training new employees. Therefore, predicting employee intent to quit can be crucial for organizations to take proactive measures to prevent it from happening. Early detection of employee turnover intention will help enterprise develop and enhance core competitiveness. This study aims to predict the employee intension to quit. In the present study, more than 1,400 samples containing 31 features of a company's employees were collected from Kaggle website as data sets. A two-layer DGNK model was designed with decision tree, gradient boosting, naive bayes and k-nearest neighbor model as the primary classifier and gradient boosting as the secondary classifier to build the predictive model of employee turnover intention. The experimental outcomes show that DGNK model based on two-layer ensemble learning has the best outcome, while naive bayes model has the worst outcome. In conclusion, this study highlights the importance of predicting employee turnover intention as an effective strategy to enhance organizational performance and competitive advantage. Furthermore, the success achieved in the study suggests that machine learning models like DGNK can play a crucial role in achieving this goal.

Keywords: Ensemble learning, Oversampling, AUC.

I. INTRODUCTION

Machine learning is a highly sought-after research direction in many fields, With extensive applications in classification, regression, and forecasting to meet a variety of needs[1]. In the field of medical and health computer aided system, Qezelbas-Chamak J et al. proposed a machine learning-based model for kidney disease diagnosis, which has been experimentally verified to have good performance. At the same time, it is found that efficiency and accuracy are highly dependent on data set, data preprocessing and model training, the most critical link of which is feature selection method [2]. Big data analysis and machine learning can be used to achieve renewable energy management in smart grid, thereby assisting grid operation planning, monitoring voltage instability, stabilizing boundary prediction and fault detection. Noha Mostafa, Haitham Saad Mohamed Ramadan, et al. suggested a method consisting of five steps involving five distinct machine learning techniques to predict the stability of smart grid [3]. In international economics, recent trade conflicts between prominent economies., as well as shocks to the free trade system brought by the Black Swan event, have been highlighted. Feras A. Batarseh, Munisamy Gopinath et al. used ensemble machine learning to forecast agricultural trade and the impact of abnormal events to provide decision basis for policy makers [4]. However, using a single classifier for scenarios with minor differences in research object features, machine learning has undesirable accuracy [5]. To address the issue at hand, the research incorporates the ensemble method. By merging multiple models together, a highly dependable model, the prediction effect is more ideal. The most common ensemble methods are sequential ensemble method and parallel ensemble method [6]-[9]. The sequential ensemble method includes Bagging and Stacking methods [10], and the parallel ensemble method includes Boosting algorithm [11]. Ensemble methods have shown excellent performance in both regression and classification tasks. They enhance model accuracy by effectively reducing bias and variance.

This study focuses on the issue of employee turnover prediction using an integrated learning approach. Employee turnover is a significant challenge faced by many enterprises, and the loss of key employees can have a substantial impact. Traditional human resource management decisions often rely on intuition or experience, which

[1] School of Information Engineering, Harbin University, Harbin, China; Heilongjiang Provincial Key Laboratory of the Intelligent Perception and Intelligent Software, Harbin, China
[2] School of Information Engineering, Harbin University, Harbin, China; Heilongjiang Provincial Key Laboratory of the Intelligent Perception and Intelligent Software, Harbin, China
[3] Heilongjiang Government Affairs Big Data Center, Harbin, China
[4] School of Information Engineering, Harbin University, Harbin, China; Heilongjiang Provincial Key Laboratory of the Intelligent Perception and Intelligent Software, Harbin, China
[5] Heilongjiang Government Affairs Big Data Center, Harbin, China
[6] Heilongjiang Government Affairs Big Data Center, Harbin, China
*Corresponding author: Lihe Ma

can lead to decision biases. Therefore, accurately predicting employee intentions to quit is a difficult task. In this rapidly evolving technological era, there is a need for improved methods to evaluate the likelihood of employee turnover and effectively prevent brain drain. The objective and scientific prediction of employee turnover can help reduce the subjectivity and uncertainty associated with conventional decision-making processes informed by intuition and experience. By leveraging data-driven analysis, qualitative and subjective judgments can be transformed into measurable and objective assessments. By employing an integrated learning method, this research aims to develop a more reliable approach for predicting employee turnover. This approach will leverage various sources of data and combine different analytical techniques to enhance the accuracy of predictions. The ultimate objective is to deliver enterprises with a valuable tool that enables them to proactively address employee turnover and minimize its negative impact on organizational performance. The primary research contents of this paper are as follows:

(1) By using the ensemble learning method with Stacking, an optimized classifier is designed for predicting the possibility of employee turnover;

(2) The ensemble classifier is compared with four machine learning classifiers, namely naive bayes [12],, k-nearest neighbor [13], gradient boosting [14] and decision tree [15].

(3) The results of the experiment of the ensemble classifier and four learning classifiers were analyzed. The ensemble approach achieves the highest level of accuracy, surpassing 95%.

II. Machine Learning Model Introduction

A. *Naive Bayes*

Naive bayes is a classification algorithm that assumes The probabilities associated each feature dimension within a sample conforms to a Gaussian distribution. By applying the Bayesian formula, It evaluates the likelihood of a new data point belonging to different categories by considering the distribution of features. The category with the highest posterior probability is assigned to the new sample.

B. *K- nearest Neighbor*

The fundamental concept behind the k-nearest neighbor (KNN) model is to estimate the likelihood of a sample belonging to a particular category based on the category of the nearest k samples in the feature space. The calculation of distance plays a crucial role in the KNN model and can be determined using various commonly used techniques. For instance, Euclidean distance can be used to calculate the sample similarity. However, this method is very sensitive to noise features. In order to solve this problem, different weights are assigned to features in the similarity distance calculation formula. The weight selection principle of the features can be set according to the role of each feature in classification.

C. *Gradient Boosting*

Gradient boosting, a member of the boosting algorithm family, draws inspiration from the gradient descent method. It works by leveraging the negative gradient information of the current model's loss function to train new weak classifiers. These trained weak classifiers are then progressively combined with the existing model to improve its overall performance.

D. *Decision Tree*

decision tree is a hierarchical structure used in machine learning and data mining. It consists of internal nodes that symbolize features or attributes, and leaf nodes that represent decisions or outcomes. This type of model is commonly employed in supervised learning to classify and predict unknown data. A full decision tree comprises the following elements: (1) Root node: encompasses the entire sample set; (2) Internal node: signifies a test on a specific feature attribute; (3) Leaf node: indicates the resulting decision.

E. *Ensemble Method*

Ensemble method is to improve the model performance by ensembling multiple models rather than using only one model[16]. The ensemble model greatly improves the result accuracy, making ensemble method well received in machine learning. Ensemble-based solutions can be surprisingly effective when dealing with large amounts of data or small data samples[17]. When the volume of training data is too substantial for a single model to handle, it can be beneficial to partition the data into smaller subsets. Each partition can be used to train different models, and then appropriate algorithms can be selected for ensemble[18].

F. *Confusion Matrix Evaluation*

The model's effectiveness is evaluated using a confusion matrix, which provides information about the accuracy of its predictions. The confusion matrix is a representation of the performance of a classification model. It consists of four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

TP: This category represents the instances where the model correctly predicts a true positive outcome, meaning the predicted value matches the actual value.

FP: In this category, the model incorrectly predicts a positive outcome when the actual value is negative.

TN: This category includes instances where the model accurately predicts a negative outcome, aligning with the actual value.

FN: Here, the model incorrectly predicts a negative outcome when the actual value is positive.

By analyzing these categories in the confusion matrix, we can assess the model's performance and determine its accuracy in making predictions. [19].

Accuracy is a commonly used metric for evaluating the performance of a model, and it represents the percentage of correct predictions out of the total predictions made on a given sample. It is calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision, also referred to as the precision rate, measures the accuracy of a prediction model by determining the likelihood that all predicted positive instances are indeed true positive instances. It calculates the number of correct predictions made among the instances predicted as true positives. The formula for precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also referred to as the recall ratio, represents the likelihood that a true sample will be correctly identified as such in a prediction. This statistical measure is calculated using the following formula:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 score, also known as the F1 measure, is a metric that combines both precision and recall to evaluate the performance of classification models, especially when dealing with unbalanced binary datasets. It provides a more comprehensive measure compared to precision alone. By considering both precision (the ratio of correctly predicted positive instances to the total predicted positive instances) and recall (the ratio of correctly predicted positive instances to the total actual positive instances), the F1 score provides a balanced assessment of a model's effectiveness in capturing both true positives and minimizing false positives and false negatives.

$$F_1 Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

III. Construction Of Dgnk Ensemble Classifier

A. *Idea of DGNK Ensemble Classifier*

DGNK ensemble classifier is designed based on two-layer ensemble classifier with decision tree (DT) , gradient boosting (GB),naive bayes (NB) and k-nearest neighbor (KNN) models as the primary classifier, and gradient boosting (GB) as the secondary classifier. The model is constructed by the five-fold cross validation. The DGNK ensemble classifier framework is shown in Fig 1. The primary classifiers such as DT, GB, NB and KNN models were used for training of the original training set to make prediction on the cross validation set, so that a set of prediction data is obtained. Then, these prediction data were used as a new training set of secondary gradients boosting classifier. For the test set, every model obtained from the original training set should be predicted based on the test set. In the five-fold cross validation, five predictions were made based on the test set. The results of these five predictions were averaged to obtain a new test set, which was used as the new test set of the secondary gradient boosting classifier. Finally, the secondary gradient boosting classifier was used for training on the new training set and test on the new test set to get the final prediction result.

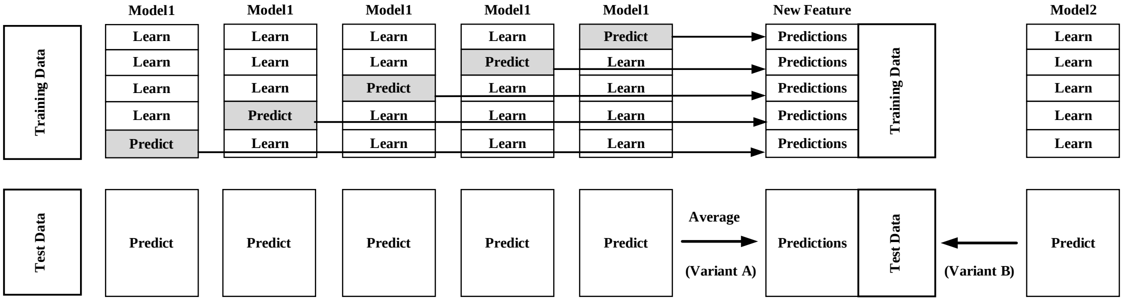

Fig. 1. DGNK ensemble classifier framework

B. Algorithm Flow of DGNK Ensemble Method Model

According to Figure 1, the classification algorithm operation steps of model framework ensemble learning model are described as follows:

(1) Divide the training data set into K-fold to prepare for the training of all primary classifiers. GB, KNN, GB and DT were selected as primary classifiers and GB as secondary classifiers, and K value was selected as 5 for cross validation.

(2) The primary classifiers GB, KNN, GB and DT were trained for K times respectively, and sample of 1/K data set was reserved in each training as the validation data set. After the training of each classifier, the testing data set was used to make predictions. Each classifier will yield corresponding 5 data sets of prediction results. Each part of the 5 prediction result data sets was summed up and averaged. The specific process is as follows:

Firstly, for the first primary classifier GB, the data set is divided into 5 folds, and the 5 data sets are identified as S1, S2, S3, S4 and S5 respectively.

4 data sets such as S2, S3, S4 and S5 are used for GB training. S1 is used as the test data set to save the predicted results of the test data, and then the data set of tests is predicted.

4 data sets such as S1, S3, S4 and S5 are used for GB training. S2 is used as the test data set to save the predicted results of the test data, and then the data set of tests is predicted.

4 data sets such as S1, S2, S4 and S5 are used for GB training. S3 is used as the test data set to save the predicted results of the test data, and then the data set of tests is predicted.

4 data sets such as S1, S2, S3 and S5 are used for GB training. S4 is used as the test data set to save the predicted results of the test data, and then the data set of tests is predicted.

4 data sets such as S1, S2, S3 and S4 are used for GB training. S5 is used as the test data set to save the predicted results of the test data, and then the data set of tests is predicted.

After five rounds of training, five predictive data sets of testing data were obtained. Each part was summed up and averaged, the training data set predicted by each model series was registered.

The three primary classifiers, KNN, GB and DT, were trained by the same method. After all the training, the obtained five prediction result sets were used for prediction of the next layer.

IV. Test And Result Analysis

A. Exploratory Data Analysis

The data used in this study is sourced from a reputable company's employee dataset available on a widely recognized platform. The dataset comprises over 1,400 samples and includes 31 features. These features encompass various aspects such as employee attrition status, age, monthly salary, frequency of business travel, and more. Data processing is the most important step in building an effective machine learning model.

(1) Exploring the relationship between employees' job satisfaction, age, total length of service, and monthly income with turnover

In Fig 2, analyze the characteristics of resigned employees from aspects such as job satisfaction, employee age, total length of service, and employee monthly income:

Resignated employees have lower job satisfaction, below 3.0, while in-service employees have higher job satisfaction, with job satisfaction ranging from 2 to 4, mainly concentrated around 3;

Resignated employees have a lower age range, ranging from 28 to 39 years old, and are mainly concentrated at 32 years old; The working age of employees is relatively high, ranging from 31 to 43 years old, and mainly concentrated around 36 years old;

The total length of service for departing employees is relatively low, ranging from 4 to 10 years, and mainly concentrated in 7 years; The total length of service of in-service employees is relatively high, ranging from 6 to 16 years, and mainly concentrated around 10 years;

Resignated employees have lower monthly income, ranging from $2400 to $6000 annually, mainly concentrated around $3500; The monthly income of in-service employees is relatively high, ranging from $3000 to $9000, and mainly concentrated around $5000;

Therefore, departing employees generally have characteristics such as low job satisfaction, younger age, lower total length of service, and lower monthly income.

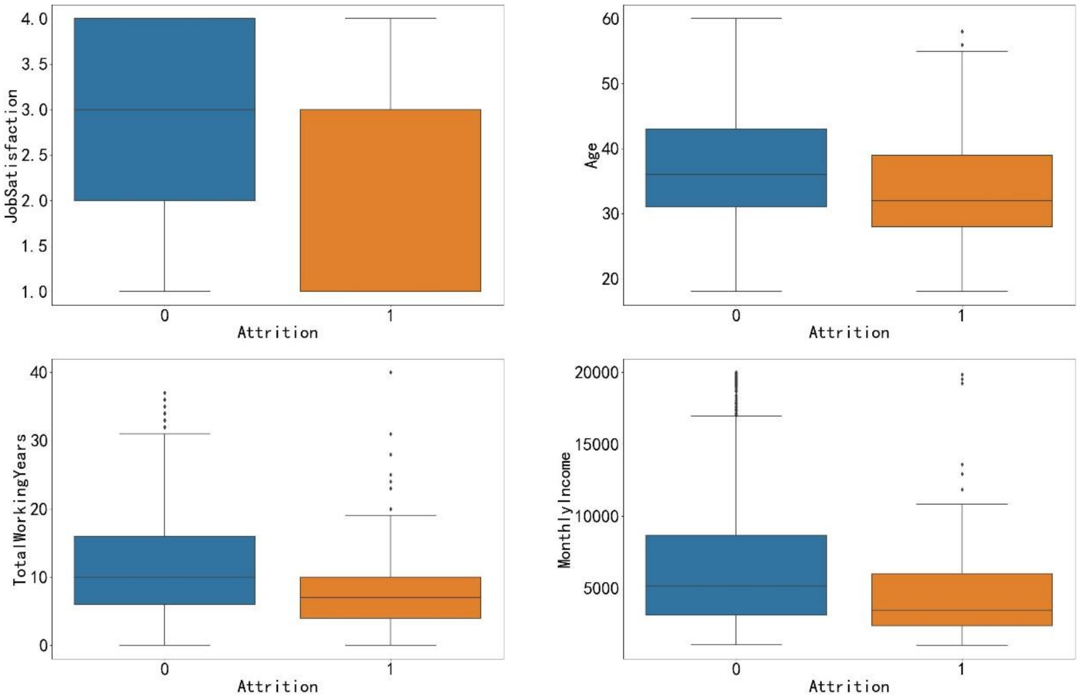

Fig. 2. the relationship between employees' job satisfaction, age, total length of service, and monthly income with turnover

(2) Feature Engineering

The accuracy of a training model is heavily influenced by the choice of feature sets. Feature selection is the process of identifying and removing irrelevant or redundant features from a dataset, reduce data dimensions, mitigate overfitting, decrease training time, and enhance the overall operational efficiency of the model[20], Select a specific set of dependent variables from the dataset that exhibit a significant correlation with the attributes related to resignations and employment. Correlation refers to the extent of association between these variables, which can be either positive or negative. Use a heat map to visually represent the correlation matrix between the selected dependent variables and the characteristics related to resignations and employment. In a positive correlation, as the individual value of a feature increases, the value of the target variable also increases. Conversely, in a negative correlation, an increase in the specific value of a feature leads to a decrease in the value of the target variable. The correlation matrix is shown in Fig 3.

From the heat map, it can be analyzed that there is a negative correlation between resignation and employees' satisfaction with the work environment, monthly income, career level, stock option level, etc. If employees are not satisfied with the work environment and their personal value realization is not high, there is a high possibility of resignation. There is a significant positive correlation between resignation and the number of companies the employee has worked for, as well as the distance between the company and their home address.

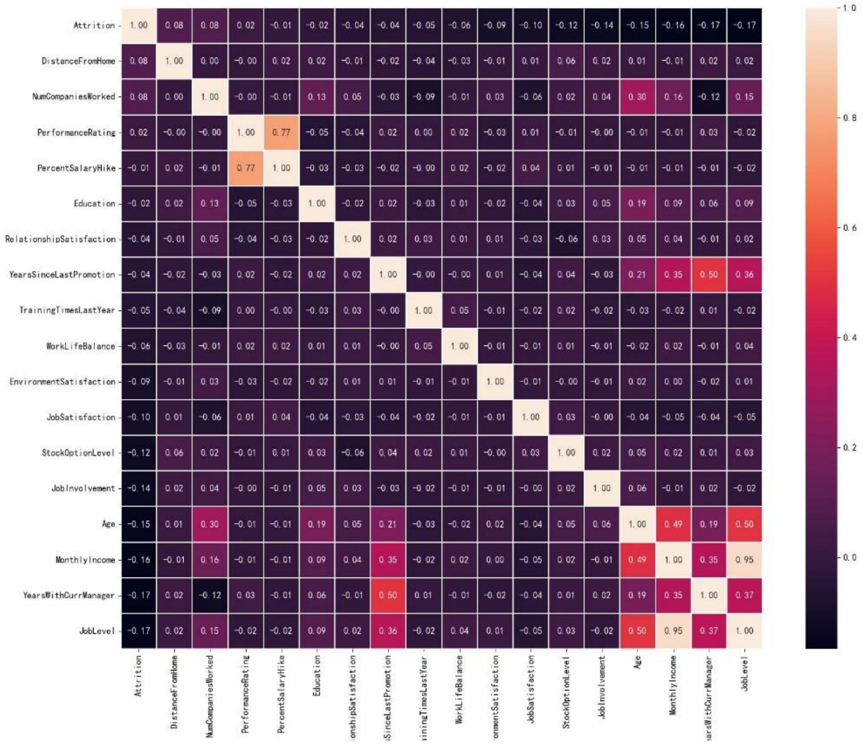

Fig. 3. Feature Engineering

B. Data Preprocessing and Data Set Division

The data processing steps are as follows:

(1) Irrelevant data has been removed from the modeling process. The employee number and age have been excluded as they do not contribute to the modeling. Furthermore, the feature related to standard working hours has also been eliminated from the dataset.

(2) For non-digital attributes, data standardization is achieved through the application of one-hot encoding.

(3) The figure type characteristics are classified into discrete categories, such as age groups, proximity to company, income brackets, and so on.

(4) The new dataset is generated by merging the non-digital dataset, which has undergone data standardization processing, with the discrete processed digital features.

(5) Normalizing feature data involves scaling all the features to a uniform interval, typically [0, 1];

(6) Class imbalance occurs in a classification task when there is a significant difference in the number of training samples belonging to different categories. For example, if the dataset has an approximate ratio of 84:16 (or 5:1) between two classes, it can be identified as a class imbalance issue. This imbalance can lead to the model primarily learning biased information from the more dominant class in the training set. Consequently, most classes may show higher precision, while a few classes may have lower precision. This imbalance can hinder the model's ability to learn crucial features effectively, thereby compromising the overall robustness of the model. Furthermore, since the number of samples in the minority class is relatively small, this study adopts an oversampling technique to address the class imbalance problem.

(7) The data in (6) is divided into a training dataset, which consists of 80% of the data, and a test dataset, which contains the remaining 20%. Four primary classifiers based on machine learning models, namely GB, KNN, GB, and DT, were individually trained on the training set. Subsequently, the models were constructed and evaluated on the test set. The DGNK model is established by using Stacking algorithm with GB model as the secondary classifier to predict the test set.

C. AUC Assessment

To assess the model's performance, a combined assessment using ROC [21] (Receiver Operating Characteristic) and AUC[22] (Area Under the Curve) metrics was conducted. These measures were used to assess the advantages and limitations of the binary classifier. The AUC values of each model can be observed in Fig 4.

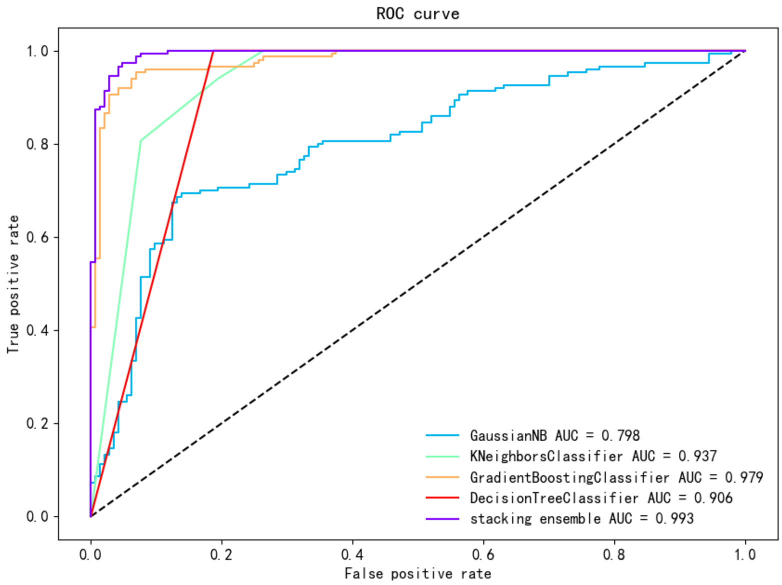

Fig. 4. AUC comparison of each mode

As shown in Fig 4, for the first effect, gradient boosting performs the best among the four models of DT, GB, NB and KNN, with an AUC value of 0.979. naive bayes (NB) has the worst performance at 0.798. For the second effect, the DGNK model built based on the ensemble method performs best, with an AUC value of 0.993, which is superior to the four base models, indicating validity of the DGNK model built based on ensemble method.

D. Evaluation using confusion matrix

In the study, multiple metrics including Precision, Accuracy, Recall, and F1 score were utilized to evaluate the performance of the models. The precision, recall, and F1 score values for each model are available in Table 1. The experimental prediction results showed that the DGNK model built based on ensemble method had better performance than the base model in terms of recall, precision and F1 score. The recall, precision and F1 score of DGNK model were 95.58%, 95.58% and 95.58% respectively, which were higher compared to DT, GB, NB and KNN models. The accuracy, recall, precision and F1 score of naive bayes (NB) model were poor at 67.35%, 67.35% and 67.35% respectively, proving the validity of DGNK model based on ensemble method.

TABLE 1. Comparison of Accuracy, Recall Rate, and F1-Score

Classifier	**Precision**	**Recall**	***F*1-score (%)**
NB	67.35	67.35	67.35
KNN	87.41	87.41	87.41
GB	90.48	90.48	90.48
DT	90.82	90.82	90.82
Ensemble	95.58	95.58	95.58

The accuracy results of each model are presented in Table 2. The experimental prediction results showed that the DGNK model built based on the ensemble method had better performance than the DT, GB, NB and KNN models in terms of accuracy. The accuracy of DGNK model was 95.58%, which was higher compared to DT, GB, NB and KNN models. The accuracy of NB model was poor at 67.35%, and the data also confirmed that the DGNK model built based on the ensemble method had a high accuracy.

TABLE 2. Accuracy of Each Model

Classifier	**Accuracy (%)**
NB	67.35
KNN	87.41
GBDT	90.48
DT	90.82
Ensemble	95.58

V. Conclusion

In this paper, a two-layer DGNK model was designed with DT,NB,GB, KNN as the primary classifier and GB as the secondary classifier to build the model for predicting employee turnover intention. In view of the small size of sample data and the imbalance of sample data, the overmining method was used to build the data set, and the four primary classifiers were optimized to obtain the optimal parameter combination of the model, so that prediction could be made. It solved the shortcomings of the existing methods in prediction of employee turnover intention. The experimental results demonstrated that the accuracy of DT, GB ,NB and KNN models based on rough set classifier was 67.35%, 87.41%, 90.48% and 90.82% respectively, and the accuracy of two-layer ensemble learning model was 95.58%. The ensemble learning model herein has certain practical application value in the prediction of employee turnover intention. The data-driven prediction of employee turnover intention is mainly based on objective data analysis free from subjective and empirical factors, which transforms subjective qualitative analysis to quantifiable and objective analysis based on data. Therefore, the prediction model of employee turnover intention can assist the employee management in human resources department, help the decision makers to obtain valuable information, adjust the talent strategy of the enterprise, thus boosting the sustainable development of the enterprise.

REFERENCES

[1] M. Agrawal and S. Agrawal, “A systematic review on artificial intelligence/deep learning applications and challenges to battle against COVID-19 pandemic,” Disaster Advances, vol. 14, no. 8, pp. 90–99, 2021.

[2] Qezelbash-Chamak J, Badamchizadeh S, Eshghi K, et al. A survey of machine learning in kidney disease diagnosis[J]. Machine Learning with Applications, 2022, 10: 100418.

[3] Mostafa N, Ramadan H S M, Elfarouk O. Renewable energy management in smart grids by using big data analytics and machine learning[J]. Machine Learning with Applications, 2022, 9: 100363.

[4] Batarseh F A, Gopinath M, Monken A, et al. Public policymaking for international agricultural trade using association rules and ensemble machine learning[J]. Machine Learning with Applications, 2021, 5: 100046.

[5] Shin D, Cho W I, Park C H K, et al. Detection of minor and major depression through voice as a biomarker using machine learning[J]. Journal of Clinical Medicine, 2021, 10(14): 3046.

[6] Ribeiro M , Silva R ,Moreno S R , et al. Efficient bootstrap stacking ensemble learning model applied to wind power generation forecasting[J]. International Journal of Electrical Power & Energy Systems, 2022, 136:107712-.

[7] Visser L, AlSkaif T, van Sark W. Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution[J]. Renewable Energy, 2022, 183: 267-282.

[8] Hwangbo L, Kang Y J, Kwon H, et al. Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients[J]. Scientific Reports, 2022, 12(1): 1-9.

[9] Mienye I D, Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects[J]. IEEE Access, 2022, 10: 99129-99149.

[10] Xu S B ,Huang S Y ,Yuan Z G , et al. Prediction of the Dst Index with Bagging Ensemble-learning Algorithm[J]. The Astrophysical Journal Supplement Series, 2020, 248(1):14.

[11] Yen A ,Morgan H E ,Wang K , et al. Interpretable Machine Learning Model Supported by Parallel Ensemble Learning to Predict Local Recurrence for Patients with Cervical Cancer[J]. International Journal of Radiation Oncology, Biology, Physics, 2021(3S):111.

[12] Peña, F., & Ferri, C. (2020). A comparative study of naive Bayes classifiers for imbalanced data sets. Knowledge-Based Systems, 193, 105436.

[13] Celikyilmaz, A., Sezgin, T. M., Inan, H. (2020). K-nearest neighbor graph-based unsupervised dimensionality reduction for hyperspectral image classification. International Journal of Remote Sensing, 41(7), 2636-2655.

[14] Akter, S., Islam, M. H., & Uddin, M. Z. (2020). Gradient boosting machine for predicting bankruptcy: A comparative analysis with logistic regression. Expert Systems with Applications, 152, 113347.

[15] Kumar, A., & Rani, P. (2021). Breast cancer diagnosis using PCA-based feature selection and decision tree classification. Complex & Intelligent Systems, 7(3), 1689-1704.

[16] Keyvanpour, M., & Vahdat, A. (2020). Ensemble models for imbalanced data classification: A review. Journal of Big Data, 7(1), 1-35.

[17] Zhang, Y., & Ma, J. (2021). Ensemble-based deep learning for large-scale image recognition. Neurocomputing, 448, 473-483.

[18] Xu, J., Song, T., Wu, H., & Chen, J. (2022). Multi-model ensemble with hybrid feature selection for energy consumption forecasting in smart building. Applied Energy, 309, 117955.

[19] Chen, J., Lin, J., & Hao, M. (2020). A comparative analysis of machine learning algorithms for credit risk assessment: Evidence from peer-to-peer lending. Emerging Markets Finance and Trade, 56(10), 2313-2327.

[20] Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

[21] Ullah, I., Al-Maadeed, S., Bouridane, A., & Khelifi, F. (2021). A Novel Texture Feature Selection and Classification using Recursive ROC Analysis for Automatic Facial Expression Recognition. Pattern Recognition Letters, 142, 280-287.

[22] Deka, B., & Sarma, N. (2021). A Comparative Study of Machine Learning Models for Predicting the Severity of Dengue Disease. Health Information Science and Systems, 9(1), 1-14.