ADU Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/adufmbd

Performance of Various Naive Bayes Algorithms on Employee Attrition Detection

Fahriye GEMCİ[1]

[1] Kahramanmaraş Sütçü İmam University, Faculty of Engineering and Architecture, Computer Engineering Department, Kahramanmaras, Türkiye

ARTICLE INFO

Submission *09/12/2024*
Revision *02/01/2025*
Acception *11/01/2025*

Keywords:

Employee Attrition
Naïve Bayes
Gaussion Naïve Bayes
Categorical Naïve Bayes
Supervised Classifier

ABSTRACT

Since companies aim to increase profits, the company's resources must be used correctly. Two of the resources that human resources should consider important for the company are time and the continuation of talented employees in the workplace. In the case of loss of talented employees, training given to new workers requires wages and time. This situation shows that loss of employees is an important problem in company policy.

This study aims to automatically detect employee attrition with machine learning algorithms. Naive Bayes classifiers are one of the successful machine learning algorithms used in many different fields such as text processing. Therefore, in this study, supervised classification has performed on the dataset with 5 different Naive Bayes algorithms in determining the employment loss. Gaussian Naive Bayes and Categorical Naive Bayes show the most successful results in determining the loss of workers.

1. INTRODUCTION

The aim of companies is to maximize profits. Recently, the importance given to human resources has been increasing because qualified and skilled employees constitute an advantage in companies (Vardarlier and Zafer, 2019). Discovering Human Resources (HR) data is very important for the decision-making process and development of companies. Machine learning techniques are often used to discover this HR data.

Companies spend a lot of time on employee recruitment according to their strategic needs. Then, a new employee receives a lot of training to adapt to the company in terms of knowledge. It is a very undesirable situation for an employee who has been oriented, in short, spent time and resources to leave the job. The loss of talented and experienced employees is an important problem that companies must solve.

Classification is the assignment of data to specific classes. Supervised classification means working on a previously known training set of class labels. For supervised classification, naive bayes learn fast and perform great (Dimitoglou et al.; Patil and Sherekar, 2013). Naive Bayes classifiers are one of the most common machine learning algorithms used in many different areas, such as text processing, sentiment analysis (Metsis, 2006; Sabiq et al., 2024). Therefore, in this study, they are compared the performances of 5 different Naive Bayes on a newly published Kaggle dataset in determining employed attrition.

2. RELATED WORKS

In this study, employee attrition has estimated with a data set containing 32 features of 1470 employees

*Corresponding Author: fahriyegemci@ksu.edu.tr

produced by IBM Watson. It has applied the t-test and ADASYN methods to preprocess and 3 different machine learning algorithms such as Support vector machine (SVM), random forest (RF) and K-Nearest Neigbours (KNN) to classify. The KNN algorithm, where k value is 3, shows the highest performance with an F1 score of 0.93 (Alduayj and Rajpoot, 2018)

In this study, employee attrition has estimated with a data set containing 35 features of 1470 employees produced in Kaggle. It has applied Correlation, Information gain Ratio, Gain Ratio, Chi-Square, Fisher's Exact Test based methods to preprocess and 7 different machine learning algorithms such as Artificial Neural Network (ANN), SVM, Gradient Boosting (GB), Bagging, RF, Decision Tree (DT) and KNN to classify. The ANN algorithm and Chi-Square feture selection algorithm shows the highest performance with an accuracy score of 0.89 (Subhashini and Gopinath; 2020).

In this study, employee attrition has estimated with a data set containing 35 features of 1470 employees produced in Kaggle. It has applied k-Fold cross-validation method to preprocess and 7 different machine learning algorithms such as Gaussian Naive Bayes (GNB), Logistic Regression (LR) KNN), Decision Tree (DT), Random forest (RT), Support Vector Machine (SVM) and Linear Support Vector Machines (LSVM) to classify. The GNB algorithm using k-fold cross-validation algorithm shows the highest performance with an F1 score of 0.446 (Fallucchi et al.,2020).

In this study, employee attrition has estimated with a data set containing 35 features of 1470 employees produced in Kaggle. It has applied k-Fold cross-validation method to preprocess and 3 different machine learning algorithms such as DT, RT, and Binary Logistic Regression(BLR) to classify. The BLR algorithm shows the highest performance with an accuracy score of 0.87 (Alsubaie and Aldoukhi, 2024).

Table 1 shows the results of different machine learning algorithms with the highest accuracy or highest F1 score in determining employee attrition. Table 1 also lists the preprocessing steps performed in the studies. Because the success of the preprocessing steps used in improving performance is undeniable.

Table 1. Results of different machine learning algorithms in determining employee attrition

Data Set	Sample Number	Feature Number	Preprocessing Algorithms	Machine Learning Algorithms	The most Success	The most Accuracy Success	The most F1 Score Success
Synthetic data created by IBM data (Alduayj and Rajpoot, 2018)	1470	32	The t-test method, ADASYN	SVM, RF, KNN	KNN (K = 3)	-	0.93
IBM Employee Attrition Dataset (Subhashini and Gopinath; 2020)	1470	35	Correlation, Information gain Ratio, Gain Ratio, Chi-Square, Fisher's Exact Test based methods	ANN, SVM, GB, Bagging, RF, DT, KNN	Chi-Square algorithm, The ANN algorithm	0.8944	-
IBM Employee Attrition Dataset(Alsubaie and Aldoukhi, 2024).	1470	35	K-Fold cross-validation	DT, RF, BLR	BLR	0.8744	-
IBM Employee Attrition Dataset(Fallucchi et al.,2020).	1470	35	K-Fold cross-validation	GNB, LR, KNN, DT, RT, SVM, LSVM	GNB	-	0.446
IBM Employee Attrition Dataset (Krishna and Sidharth, 2024).	1470	35		SVM, RF NB, LR, DT	RF	1	-
Kaggle Data Set in this study	5000	10	One Hot Encoding, ADASYN	BNB, CaNB, CoNB, GNB, MB	CaNB, GNB	0.9513, 0.9567	0.9739, 0.9603

3. MATERIALS AND METHODS

3.1. Data Set

The data set is extracted from Kaggle repository. The data set consists of 5000 samples and 10 features. Features are given in Table 2. Age feature means employee age. Gender feature means employee gender. DistanceFromHome means that it is the distance between where the employee lives and works. Attrition means whether the employee leaves the company or not. JobSatisfaction means that it is an indicator of employee satisfaction. Monthly Income means the monthly income of the employee. YearsAtCompany means that how many years has employee worked in the company. PerformanceRating means employee performance. WorkLifeBalance means how well an employee can balance work with their personal life. Attrition means whether the worker leaves the job or not. They can be used to evaluate workforce diversity, retirement planning, promotions, pay, and job satisfaction and commuting difficulties.

Table 2. Employee Attrition Dataset

Features	Type
Age	Integer
Gender	String
DistanceFromHome	Integer
JobSatisfaction	Integer
MonthlyIncome	Integer
YearsAtCompany	Integer
Overtime	Integer
PerformanceRating	Integer
WorkLifeBalance	Integer
Attrition	Integer

First, the data is prepared by performing preprocessing steps in this study. This data set was checked for missing data. It was observed that there was no missing data. It was observed that there was a categorical feature in the data set. Gender categorical feature values were converted to binary format using one hot encoding. The method is commonly used as preprocessing step. The train data set was held for model learning. The test data set was held for calculating model performance.

3.2. Naive Bayes (NB)

A Bayesian classifier is a statistical and probabilistic classifier based on Bayes' theorem. In terms of statistics, naive bayes can predict the probability that an sample belongs to a particular class (Leung, 2007). In terms of probabilistics, naive bayes is based on approximating a distribution (Wickramasinghe and Kalutarage, 2021). Naïve Bayes uses the information in the sample data to predict the posterior probability (Webb et al., 2010).

Class conditional independence is required for Naive Bayes. It assumes that the features attached to the class label are independent by evaluating them without considering their interdependencies. This independence assumption speeds up the implementation of Naive Bayes (Pajila, 2023).

Bernoulli Naive Bayes (BNB)

Bernoulli Naive Bayes is a Naive Bayes algorithm kind that based on probability of all features. Binary or Bernoulli distribution parameters is used to perform this method (Pajila et al., 2023; Manning, 2008). The Bernoulli NB model is often used for binary or boolean features. Therefore, it is often used in two class classification problems (Sayfullina et al., 2015).

Categorical Naive Bayes(CaNB)

Categorical Naive Bayes is a Naive Bayes algorithm kind that perform on categorically distributed data. Categorical Naive Bayes supposes that each feature in the dataset has itself categorical distribution (Omura et al., 2012).

Complement Naive Bayes(CoNB)

The Complement Naive Bayes classifier has builded to adjust the "intense assumptions" performed with the Multinomial Naive Bayes classifier. So that, classification on imbalanced data sets succesfully performes with CNB.

Gaussian Naive Bayes(GNB)

Gaussian Naive Bayes is a Naive Bayes algorithm kind that perform on continuous data. It is considered that the continuous data features over the dataset maintain a Gaussian distribution (Pajila, 2023).

Multinomial Naive Bayes(MNB)

Multinomial Naive Bayes is a Naive Bayes algorithm kind that based on probability of all features like Bernoulli Naive Bayes. Multinomial distribution parameters is used to perform this method (Pajila, 2023). When the training data is limited in amount, Multinomial Naïve Bayes has been shown to give good results in areas such as sentiment analysis (Sabiq et al., 2024). Multinomial Naive Bayes is often used in feature data with discrete variables. In text processing, multinomial naive bayes has often been

used successfully, since feature data might be word counts.

4. RESULTS AND DISCUSSION

This work is performed based on the "Naïve Bayes" classifier developed in phyton. A dataset with 5000 samples classified in binary categories is used to discover employee attrition. 70% of the samples are selected randomly to create the training dataset for the classifier. The remaining 30% of the samples are used as the test dataset to test the classifier. The employee attrition classification results of different naive bayes types obtained in this study are given in Table 3. Based on Table 3, it is observed that the Complement Naive Bayes algorithm performs lower than other Naive Bayes algorithms in detecting employee attrition. Default hyperparameters of Naive bayes algorithms are used in this study. In determining employee attrition, Gaussian Naive Bayes shows the highest accuracy with an accuracy rate of 0.9567, while Categorical Naive Bayes shows the highest F1 score with an F1 score of 0.9739.

Although the results seem successful, it is observed that there is an imbalanced class problem because one class contains very few examples compared to the other class. Therefore, new examples belonging to the minority class were created to balance the class distributions. Using the ADASYN resampling method, number of minority samples are equalized to the number of majority examples by generating new samples. Naive Bayes algorithms were re-run on samples and the new results are given in Table 4.

In the results obtained in the new data set by adding the samples obtained after the resampling algorithm, Gausssion and Categorical Naive Bayes algorithms are more successful than other naive Bayes algorithms, similar to the previous results.

Table 3. Results of different Naive Bayes algorithms in determining employee attrition

Weighted	Accuracy	Precision	Recall	F1 Score
Bernoulli Naive Bayes	0.9087	0.8257	0.9087	0.9522
Categorical Naive Bayes	0.9513	0.9538	0.9739	0.9739
Complement Naive Bayes	0.7880	0.9043	0.8717	0.8717
Gaussian Naive Bayes	0.9567	0.9706	0.9567	0.9603
Multinomial Naive Bayes	0.8680	0.8859	0.8680	0.8761

Table 4. Results of different Naive Bayes algorithms in employee attrition data set with ADASYN resampling method

Weighted	Accuracy	Precision	Recall	F1 Score
Bernoulli Naive Bayes	0.7751	0.8454	0.7751	0.7638
Categorical Naive Bayes	0.9971	0.9971	0.9971	0.9971
Complement Naive Bayes	0.7769	0.7784	0.7769	0.7764
Gaussian Naive Bayes	0.9427	0.9486	0.9427	0.9425
Multinomial Naive Bayes	0.7772	0.7787	0.7772	0.7768

5. CONCLUSION

The results of the study prove that there is a significant relationship between employee attrition and given characteristics. Such a data set would allow analyzing a newly developed prediction model. The features in the data set were very useful in calculating the employee attrition result.

It is seen that Bayesian classifiers exhibit high accuracy when working with large data sets. These methods are also seen that it shows higher accuracy in low-dimensional data sets by not being affected by the curse of dimensionality.

In employee attrition supervised binary classification, it is seen that Gaussian Naive Bayes competes with 0.9567 accuracy and Categorical Naive Bayes with 0.9513 accuracy. Categorical Naive Bayes comes ahead of Gaussian Naive Bayes with 0.9739 F1 score and 0.9603 F1 score.

Therefore, new samples of minority classes are produced using the ADASYN resampling method for cope with imbalanced data problem. Then the results of testing the algorithms again on the entire sample, it is seen that Gaussian Naive Bayes competes with 0.9427 accuracy and Categorical Naive Bayes with 9971 accuracy. Categorical Naive Bayes with 0.9971 F1 score comes ahead of Gaussian Naive Bayes and 0.9425 F1 score, in employee attrition supervised binary classification.

REFERENCES

Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, 19(3), 62.

Alduayj, S. S., & Rajpoot, K. (2018, November). Predicting employee attrition using machine learning. In *2018 IEEE International Conference on Innovations in Information Technology* (*IIT*) (pp. 93-98). .

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. doi: 10.5267/j.dsl.2023.12.006

Anonim. (2012). US Department of agriculture nutrient database for standard reference, Release 14. URL: http://www.nal.usda.gov/fnic/foodcomp (accessed date: March 23, 2012).

Atalar, M.N. and Türkan, F. (2018). Identification of chemical components from the Rhizomes of Acorus calamus L. with gas chromatography-tandem mass spectrometry (GC-MS\MS). *Journal of the Institute of Science and Technology*, 8(4), 181-187. doi: 10.21597/jist.433743

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. *arXiv preprint* arXiv:1206.1121.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, 9(4), 86. doi: 10.3390/computers9040086

Krishna, S., & Sidharth, S. (2024). HR Analytics: Analysis of Employee Attrition Using Perspectives from Machine Learning. In *Flexibility, Resilience and Sustainability* (pp. 267-286). Singapore: Springer Nature Singapore.

Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

Manning, C. D. (2008). *Introduction to information retrieval*.

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes? In *CEAS* (Vol. 17, pp. 28-69).

Omura, K., Kudo, M., Endo, T., & Murai, T. (2012, November). Weighted naïve Bayes classifier on categorical features. In *2012 IEEE 12th International Conference on Intelligent Systems Design and Applications* (*ISDA*) (pp. 865-870).

Pajila, P. B., Sheena, B. G., Gayathri, A., Aswini, J., & Nalini, M. (2023, September). A comprehensive survey on naive bayes algorithm: Advantages, limitations and applications. In *2023 IEEE 4th International Conference on Smart Electronics and Communication* (*ICOSEC*) (pp. 1228-1234).

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2), 256-261.

Sabiq, F. F., Rahmatulloh, A., Darmawan, I., Rizal, R., Gunawan, R., & Haerani, E. (2024, August). Performance Comparison of Multinomial and Bernoulli Naïve Bayes Algorithms with Laplace Smoothing Optimization in Fake News Classification. In *2024 IEEE International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics* (*ICoABCD*) (pp. 19-24).

Sayfullina, L., Eirola, E., Komashinsky, D., Palumbo, P., Miche, Y., Lendasse, A., & Karhunen, J. (2015, August). Efficient detection of zero-day android malware using normalized Bernoulli naive bayes. In 2015 *IEEE Trustcom/BigDataSE/ISPA,* Helsinki, Finland, Vol. 1, pp. 198-205. doi: 10.1109/Trustcom.2015.375

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. doi: 10.17605/OSF.IO/9XDWE

URL:https://www.kaggle.com/datasets/comrade1234/employee-attrition-using-machine-learning/data (accessed date: September 2, 2022).

Vardarlier, P.; Zafer, C. Use of Artificial Intelligence as Business Strategy in Recruitment Process and Social Perspective. In *Digital Business Strategies in Blockchain Ecosystems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 355–373. doi: 10.1007/978-3-030-29739-8_17

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714. doi: 10.1007/978-0-387-30164-8

Wickramasinghe, I., Kalutarage, H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Comput* 25, 2277–2293 (2021). doi: 10.1007/s00500-020-05297-6