Comparison of NB and NB-PSO to determine level of vehicles sales

I Romli\textsuperscript{1*}, E Pusnawati\textsuperscript{1}, M Jaenal\textsuperscript{1} and A Siswandi\textsuperscript{1}

\textsuperscript{1}Universitas Pelita Bangsa, Bekasi 17530, Indonesia

Abstract. Nowadays, there are various types and brands of vehicles in Indonesia, especially kinds of motor vehicles. In this case, motor vehicles are cars, trucks, and buses (exclude the motorcycles). This research classified the various brands of motor vehicles in the form of classes that are Well Selling (Laris) and Not Selling (Tidak Laris), so that consumers and producers can find out which motor vehicles brands are well selling based on their category and output. This study analyzed 3908 data into 3126 training data and 782 testing data. The data was obtained from the GAIKINDO (Indonesian Automotive Industries Association) site. There were 19 attributes but to ease the research, the attributes used are 8 (including 1 Class attribute to facilitate the search for the best-selling motor vehicles). This research compared the accuracy value among Naive Bayes method and NB-PSO (Naive Bayes- Particle Swarm Optimization) by using such dataset. NB-PSO is adjusted parameters of 1 inertia weight and 5 population size. The results of classification accuracy with the Naive Bayes method produces accuracy values of 92.11\%, Precision values: 86.57\% and Recall values: 97.12\%. Meanwhile, the solutions of NB-PSO have accuracy values of 92.44\%, Precision values: 87.07\% and Recall values: 97.18\%, so PSO method was able to improve the accuracy of classification of NB as many as 0.33\%.

1. Introduction
Motor Vehicle is one of vehicles which is needed in our life. Along with the times, there are many kinds of motor vehicle brands so that consumers need to make decision fast and precisely to choose which motor vehicle is well selling (laris) and not selling (tidak laris), of which these two labels need to be classified in this research. Based on its specification, motor vehicle has various type or variation. Data attributes that will be used are 8 cars specification from that data which are Category, Brand, Type, Model, CC, Transmission, Original Country, Output, and Class. Those specifications are taken from a data which obtained from GAIKINDO (Gabungan Industri Kendaraan Bermotor Indonesia) site. The obtained data is data from January 2015 until March 2019 with the total 3908 record data. Those various motor vehicle brand will be classified into Classes which are well selling and not selling so that consumers, producers, and researchers can figure out which motor vehicle brand is best-selling. Therefore a distinctive classification is needed, so that the costumers know the best-selling motor vehicles brands in terms of the most demand category. Besides that, it encourages the motor vehicles producers to produce kinds of motor vehicles which are most demanded by the consumers.

There are some algorithms that can be used to handle data classification problems, including K-NN (K-Nearest Neighbor), C4.5, SVM (Support Vector Machine) but the algorithm that is frequently used in handling data classification problems is Naive Bayes. In this case, some advantages of naive bayes
are simple, short in time consuming, and high in accuracy [1]. In 2016, the Naive Bayes algorithm was used by Thomas Wiga to classify the types of Diabetes Mellitus by classifying Diabetes Mellitus with Hyperglycemia, Diabetes Mellitus with Hypertension, Diabetes Mellitus with Ulcers, Diabetes Mellitus with Neuropathy or Diabetes Mellitus only [2]. This algorithm has also been used by Deden Rustiana and Nina Rahayu in analyzing the Automotive Market Sentiment on Twitter where the implementation of naïve bayes in this study resulted in a value of 93%, positive sentiment of 90%, negative sentiment of 90%, and negative sentiment of 100% [3]. Then Astrid in 2017 used the algorithm in activities at the University of Semarang's ICT Faculty [4]. Naive Bayes algorithm is also applied in DSS (Decision Support System) based on Web and Mobile to determine the choice of the Hijab model can be used by Muslims correctly and accurately in choosing the hijab model [5]. After that, Ghulam Asrofi conducted a research related to sentiment analysis of the East Java 2018 governor candidates using the Naive Bayes Method which could produce a classification of sentiments (positive, neutral and negative) Indonesian tweets for East Java Governor Candidates 2018 [6].

At first, this research only propose to test naïve Bayes method with those dataset [7], but recently, researchers continue to make improvements to Naive Bayes that has weakness in choicing attribute by adding method which is comparing accuracy value between Naïve Bayes method and NB-PSO (Naïve Bayes–Particle Swarm Optimization) using those data set, since it is based on previous research conducted by Marlina et al that made a comparison between performance of classification methods Naive Bayes and C4.5 algorithms to find the best accuracy of those methods [8], and then Jie Lin and Jiankun Yu have succeed by using NB-PSO to improve the accuracy and more effective [9]. This is also proven by Widiastuti et al that PSO method can improve the accuracy of Naïve Bayes as many as 10.72% [10]. One of the optimization methods which can be used to improve Naive Bayes performance is PSO. This optimization method is adopted from the behavior of flock of bird. Based on Carlisle and Dozier, something that influence in increasing the performance of Naïve Bayes is the size of population [11], so that in this research, it is needed to adjust the parameter of which inertia weight is 1 and population size is 5.

2. Research Method

Based on the figure 1, the first step starts from collecting dataset as many as 3908 data, then preprocessing is done which are cleaning, replacing, reduction, and transformation. After preprocessing has done, it results new dataset. This research compared naïve Bayes algorithm and naïve Bayes PSO optimization (once cross validation) and also Naïve Bayes PSO optimization (twice cross validation). The following is cross validation with k=10 folds so dataset is divided into 10 to 90 which 10% is used as data testing and 90% is used as data training. Validation is conducted to test algorithm method used. K-folds cross validation is a method used to find out the level of success of an algorithm model by re-testing input attributes randomly.
Figure 1. Steps of Research Method

The next is evaluation process of each algorithm performance that has been determined by using confusion matrix technique and ROC Curve. The function of confusion matrix is to measure the level of accuracy, precision, and recall value from the evaluated algorithm model. The accuracy value is the level of accuracy of the percentage between the predicted value and the actual value, then precision value is the accuracy value with the predicted class. While, recall value is the percentage of the algorithm succeed value that is used. The following is the table of confusion matrix [2].

| Confusion Matrix | Actual Values | Positive | Negative |
|------------------|--------------|----------|----------|
| Predicted Values | Positive     | TP       | FP       |
|                   | Negative     | FN       | TN       |

Classification of the accuracy value of AUC itself is divided into 5 categories, which are:
1) \(0.90 - 1.00\) = Excellent Classification
2) \(0.80 - 0.90\) = Good Classification
3) \(0.70 - 0.80\) = Fair Classification
4) \(0.60 - 0.70\) = Poor Classification
5) \(0.50 - 0.60\) = Failure

Then, at the last step, the results calculation of accuracy, precision, and recall value are compared.
3. **Result and Analysis**
The research conducted is divided into 3 model, which are:

a. Model by using Naïve Bayes algorithm. The following is the 1st model.

![Figure 2. First Model](image)

b. Model by using Naïve Bayes and optimized using Particle Swarm Optimization with twice cross validation inside PSO and outside PSO. The following is the 2nd model.

![Figure 3. Second Model](image)

c. Model by using Naïve Bayes algorithm and optimization with once cross validation outside Optimized Weight (PSO). The following is the 3rd model.

![Figure 4. Third Model](image)
The first model is done using accuracy value 92.11%, precision value 86.57%, and recall value 97.12% [1], while the second testing model is done by Jaenal et al [7]. Yet, the dataset used is different; Jaenal et al did not use GAIKINDO dataset. The result of the comparison of the three model can be seen in histogram as follow.

![Figure 5. Comparison of Three Model](image_url)

Based on the test results, it was obtained that if we compare the first model and the second model, there is a decrease on accuracy value 0.35%, recall value 0.3%, precision value 0.43%, while the AUC value is stable.

Then, if we compare between the first testing model and the third testing model, there is an increase on accuracy value 0.33%, recall value 0.06%, precision value 0.5%, and also the AUC value 0.002. It can be concluded that the third testing model has the highest accuracy level than the first and the second testing model. Those results are in accordance with the previous study by Jaenal et al that PSO can increase the accuracy of NB with the average 0.33% from 7 dataset tested [12].

The following is a table of classification results of vehicles sales rate by using dataset from GAIKINDO of which the data representative can performed 16 from the total 3908 dataset.

**Table 2. Result of Classification to Determine Level of Vehicle Sales**
4. Conclusion

From this research, it can be concluded that:

a. Testing model using Naïve Bayes algorithm and optimization using Particle Swarm Optimization with once cross validation outside the Optimize Weight (PSO) has the highest accuracy level than testing model and also testing model using Naïve Bayes algorithm and optimization using Swarm Optimization with twice cross validation inside and outside PSO.

b. This model can be used to be implemented in Decision Support System that can be used to make decisions precisely and quickly in term of choosing a vehicle.

Acknowledgement

The authors would like to thank to DRPM of The Ministry of Higher Education Indonesia for the financial support under Penelitian Dosen Pemula (PDP) Research Grant Scheme No. 7/E/KPT/2019, and also to LPPM of Pelita Bangsa University for supporting and providing facilities.

References

[1] Cholissodin I and Riyandani E 2016 Swarm Intelligence (Teori & Case Study) (Malang: Universitas Brawijaya)
[2] Wiga T 2016 Naive Bayesian Classifier Method Implementation for Diabetes Melitus Diagnose Classification (Yogyakarta: Universitas Sanata Dharma)
[3] Rustiana D and Rahayu N 2017 Analisis Sentimen Pasar Otomotif Mobil: Teewt Twitter Menggunakan Naive Bayes Jurnal SIMETRIS 8(1) 113-120
[4] Putri A 2017 Penerapan Naive Bayesian Untuk Perangkkingan Kegiatan di Fakultas TIK Universitas Semarang Jurnal SIMETRIS 8(2) 603-610
[5] Rahmawati N, Nugroho A and Romli I 2018 Penerapan Metode Naive Bayes dalam Sistem Pendukung Keputusan untuk Menentukan Model Hijab Seminar Nasional Sekolah Pascasarjana UGM
[6] Ghulam A 2019 Analisis Sentimen Calon Gubernur Jawa Timur 2018 Dengan Metode Naive Bayes Classifier Jurnal JIPN 4(1) 27-32
[7] Romli I, Pusnawati E and Siswandi A 2019 Penentuan Tingkat Penjualan Mobil di Indonesia SNasTekS. Sidoarjo
[8] Marlina L, Muslim and Siahaan A 2016 Data Mining Classification Comparison (Naive Bayes and C4.5 Algorithms) International Journal of Engineering Trends and Technology (IJETT) 38(7) 380 - 383
[9] Lin J and Yu J 2011 Weighted Naive Bayes Classification Algorithm Based on Particle Swarm Optimization Proceedings of Communication Software and Networks (ICCSN)
[10] Widiastuti N A, Santosa S and Supriyanto C 2014 Algoritma Klasifikasi Data Mining Naive Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung Jurnal Pseudocode 1(1) 2355-5920
[11] Carlisle A and Dozier G 2001 An off the Shelf PSO Proceedings of the Workshop on Particle Swarm Optimization Indianapolis
[12] Jaenal M, Nugroho A and Romli I 2018 Analysis of Effectiveness Particle Swarm Optimization In Improving The Preformance of Naive Bayes Algorithm ICID Yogyakarta