ILKOM Jurnal Ilmiah Vol. 18, No. 2, August 2026, pp. 393-403
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

Comparative Analysis of Naïve Bayes Variants for Best-Seller Status Determination in the Air Conditioning Industry

Sasa Ani Arnomo [a,1,*]; **Zada Alzena** [a,2]; **Heri Nuryanto** [b,3]; **Siti Fairuz Nurr Sadikan** [c,4]
[a] *Institut Teknologi Batam, The Vitka City Complex, Tiban, Kota Batam, Kepulauan Riau 29425, Indonesia*
[b] *Politeknik Pariwisata Batam, The Vitka City Complex, Tiban, Kota Batam, Kepulauan Riau 29425, Indonesia*
[c] *Universiti Teknologi MARA, Cawangan Melaka, 77300, Melaka, Malaysia*
[1] *sasa@iteba.ac.id;* [2] *2221025@student.iteba.ac.id;* [3] *Heri@btp.ac.id;* [4] *fairuznurr@uitm.edu.my*
* *Corresponding author*

Article history: Received May 18, 2026; Revised June 04, 2026; Accepted July 13, 2026; Available online August 10, 2026

Abstract

The determination of product status such as the best seller predicate for Air Conditioning products is a fundamental sales promotion strategy. Leveraging machine learning to analyze sales data is essential for maximizing business progress and market positioning within the modern electronics industry. This study aims to evaluate and compare the performance and accuracy of Naïve Bayes (NB) classification models in determining AC product status. The goal is to identify the most effective variant among BernoulliNB, GaussianNB, and MultinomialNB for this specific application. A quantitative comparative analysis was conducted using various data record sizes. The primary features analyzed included brand, inverter type, electrical power cooling capacity, address, price, and sales status. In the first scenario, testing across varying data record sizes revealed that the MultinomialNB variant achieved the highest average accuracy at 85.39 percent. In the second scenario, using the full dataset with an 80% training data split without cross-validation, the Naïve Bayes algorithm as a whole demonstrated robust classification ability, reaching a peak prediction accuracy of 99.312 percent. This peak performance significantly outperformed alternative methods such as SVM and KNeighbors Classifier within those specified parameters. This performance significantly outperformed alternative methods such as SVM and KNeighbors Classifier within the specified parameters. The study concludes that the Naïve Bayes model is highly effective for product status classification in the electronics industry. The MultinomialNB variant is identified as the most consistently reliable model for these specific datasets. Future research should consider incorporating cross-validation techniques to further validate model stability across more diverse data environments.

Keywords: Naïve Bayes Classifiers; Comparative Analysis; Air Conditioning Determination; Probability Distribution; Customer Categorization.

Introduction

Competition in the world of marketing and sales is getting tighter along with the situation. Thus, entrepreneurs compete to improve established businesses by utilizing technological devices so that their companies continue to grow in the face of increasingly tight competition. Efforts to develop technology continue, including how to save air conditioning energy [1]. Managers must be able to identify the status of goods that are cheaper and know what their purchase history is like. Many brands may not match the buyer's desires. There are many brands of Air Conditioner products on the market with various specifications. In this case, an appropriate algorithmic method is needed that can be applied to determine the direction of favourite product choices. In the field of software development, it is one of the elements that causes the creation of a learning method to draw conclusions quickly and effectively. In principle, business intelligence is used to handle large amounts of data such as storing, obtaining, modifying, and displaying statistics [2]. It is capable of providing excellent and reliable information. It is critical for corporate policies to become more resilient. A prediction or estimation approach is one of the strategies needed to solve this challenge accurately [3], [4].

Competition in the global market is intensifying, forcing entrepreneurs to continuously improve their established businesses. This high-pressure environment necessitates the strategic use of technological devices to ensure companies remain competitive and continue to grow. These development efforts cover various critical areas, including energy

http://dx.doi.org/10.33096/ilkom.v18i2.3509.393-403

efficiency, such as how to conserve air conditioning energy. Managers, therefore, need effective tools to identify which products are performing well (or are more cost-effective) and to understand their historical purchase patterns. In a saturated market, consumers face a challenge: many available brands may not meet their specific needs or desires. Especially when dealing with complex items like Air Conditioners that feature numerous specifications, buyers must be able to determine which company's product features best align with their needs. This situation highlights the critical need for an appropriate algorithmic method that can be applied to accurately determine customer product preferences. This is where Business Intelligence (BI) and machine learning methods become essential. BI systems are used to efficiently handle large amounts of data, encompassing storage, retrieval, modification, and statistical analysis, thus providing excellent and reliable information for management. This capability is crucial for making corporate policies more resilient. Consequently, employing an accurate prediction or estimation approach is a vital strategy needed to effectively solve this challenge.

Previous research results state that the most sold AC indicators are determined based on their brands [5], [6]. It predicts the sales of best-selling products using a decision tree method that can be grouped according to the level of certainty. Several studies have also been conducted related to Air Conditioning (AC) in the field of business intelligence monitoring [7], [8], [9]. The methods used include DNN (Deep Neural Network) [10], ANN (Artificial Neural Network)[11], Deep Reinforcement Learning [12], and Meta Ensemble Learning [13]. Furthermore, other research states that the apriori algorithm can be used to find knowledge about the relatedness of products ordered by customers. The resulting learning is how a product is related to another product with 41.67% support and 83.33% certainty. The application of data mining to predict the number of best-selling products can use the Naive Bayes Algorithm. The accuracy of the classification model using confusion matrix obtained 83.3%, precision 84.2% and recall 88.9%. The proposed approach is the Naive Bayes methodology which can classify the product purchasing patterns. The use of Naïve Bayes variations is often used in several cases [14]. It is also considered an effective and superior advanced machine-learning algorithm [15]. However, it is still necessary to improve the accuracy of naïve Bayes classification based on the type of method. There are several methods in the Naive Bayes algorithm including the Bernoullin method, Gaussian method, and Multinominal method. Therefore, it is necessary to compare several of these methods to determine the model's accuracy level. All three Naive Bayes algorithms have their advantages and limitations. By understanding the differences between them, choosing the most appropriate method is necessary.

While several studies have extensively applied the Naïve Bayes algorithm for general product classification and customer pattern analysis [16], [17], most existing literature evaluates the algorithm as a single entity or contrasts it broadly against alternative machine learning models. In the domain of business intelligence and product recommendation, research exploring the granular performance of distinct Naïve Bayes mathematical distributions remains limited [18], [19]. This study addresses this gap by providing a comprehensive comparative analysis of three primary Naïve Bayes variants namely Bernoulli, Gaussian, and Multinomial specifically tailored for electronic product sales data. The scientific contribution of this research lies in its empirical evaluation of how these specific variants handle highly heterogeneous retail features, including continuous numerical data such as price and cooling capacity paired with categorical and geographical attributes like brand, inverter types, and customer addresses. By systematically identifying which probability distribution aligns best with mixed attribute sales datasets, this study establishes a more refined and computationally efficient predictive foundation for managers to optimize inventory positioning and data driven promotion strategies in the modern electronics industry.

Method

A. *Product Specification Variations*

The use of air conditioning is often found in tropical areas which are known for their hot climate. Many people need a suitable temperature and air quality to work well in certain circumstances [20]. In general, it is 20-25 ° C and humidity 40-60%. Heating, humidity, mobility, and breath purity all contribute to the thermal environment of a room. Green products have a significant effect on air conditioning usage [21]. There are even studies controlling the use of air conditioning and heating for room comfort [22]. There are several assessments of quality Air Conditioning such as watt size, specifications, soundproofing, warranty, passing international standards, ACs with machine covers, anti-leakage on the indoor unit, and ACs with LED indicator panels.

B. Operational Optimization and Thermal Comfort

The technical aspects of determining AC operation are heavily influenced by environmental variables and human behavior. Aibin (2020) demonstrated that integrating real-time weather data with smart thermostats can reduce AC operational duration by up to four times without compromising comfort. This aligns with findings by Malik et al. (2022), who explored using fans to raise the temperature threshold for AC activation, significantly reducing energy consumption and greenhouse gas emissions. Regarding automated classification, the use of complex variables such as human facial expressions via *Deep Reinforcement Learning* and thermal comfort prediction using *Artificial Neural Networks* serve as critical foundations for determining when an AC system should be active or adjust its settings.

C. Service Perspective and Customer Satisfaction

Beyond technical aspects, the sustainability of AC usage is determined by service quality and user perception. Andrew Wijaya et al. (2025) highlighted that AC customer retention is strongly influenced by the marketing mix, where store location and accessibility have the most significant impact on satisfaction [23]. Furthermore, Khonglumtan & Srisattayakul (2023) emphasized that in automotive AC service centers, service quality dimensions are the primary determinants of customer satisfaction. Factors such as product quality and price perception consistently prove to influence customer loyalty, mediated by satisfaction [24], [25].

D. Classification Methodology and Data Analysis

To support your Comparative Analysis, the literature indicates that algorithm selection depends heavily on the data type. Widyawati & Sutanto (2020) provide the theoretical basis for comparing Naïve Bayes variants (Bernoulli vs. Multinomial), noting that computational efficiency is a major advantage for large-scale data. Meanwhile, for more complex and non-linear data, *Structural Equation Modeling* (SEM) is frequently used to observe relationships between user behavior variables and AC usage [26], [27].

Table 1. Summary table of Literature

Researcher (Year)	Research Focus	Methods	Key Relevant Findings
Widyawati & Sutanto (2020)	Comparison of Naïve Bayes variations.	Bernoulli & Multinomial NB	Provides the theoretical basis for choosing NB variants based on data type (binary vs. frequency).
Aibin (2020)	Impact of weather on AC operation.	Data Correlation & Smart Thermostat	Weather is a crucial feature (Gaussian) in determining AC active duration.
Malik et al. (2022)	Potential of fans to reduce AC use.	Integrated Framework & GIS	Activation thresholds are influenced by supplementary cooling (fans).
Andrew Wijaya et al. (2025)	AC customer retention.	PLS-SEM & Marketing Mix	Location and price determine customer satisfaction within the AC retail ecosystem.
Wei et al. (2020)	AC control via human expressions.	Deep Reinforcement Learning	Visual inputs (expressions) can be classified to determine AC modes.
Irshad et al. (2020)	Prediction of thermal comfort.	Artificial Neural Network (ANN)	Human comfort variables are the primary labels in smart AC systems.
Khonglumtan (2023)	Service quality in AC centers.	Determinative Analysis	Technical and functional quality influence user perception of the system.
Wahjoedi et al. (2022)	Product quality & loyalty.	Smart PLS 3	Product quality has a direct effect on user loyalty toward the system.
Putri et al. (2025)	Store atmosphere & comfort.	SEM-AMOS	Environmental factors (atmosphere) significantly affect user comfort.
Yunita (2025)	Promotion & brand image on satisfaction.	SEM-PLS	Brand image strengthens user trust in the system/service used.

E. Classifier Model

The method for analyzing numbers is to create automation to ensure categorization findings from the data. Classification methods aim to partition knowledge into groups while categorizing them based on the importance of parts of the dataset. Bayesian analysis is a numerical methodology for performing inductive reasoning on multiclass classification. Summarized by IBM, various types of Naive Bayes algorithms are distinguished based on the distribution of values. Three popular types of Naive Bayes algorithms include Gaussian Naive Bayes (GaussianNB), Multinomial Naive Bayes (MultinomialNB), and Bernoulli Naive Bayes (BernoulliNB)[28], [29], [30]. Gaussian Naive Bayes (GaussianNB) is a variant of the Naive Bayes algorithm that uses a Gaussian distribution (normal distribution) and continuous variables. This model is used to find the mean and standard deviation. The formula for Gaussian Naive Bayes is:

$$P(Y = y \mid X = x) = P(Y = y) * \prod P(Xi = xi \mid Y = y) \quad (1)$$

P(Y = y | X = x) is the posterior probability, the probability of class Sales_Status given the feature vector x. feature x meliputi Brand, Inverter, Type, Eletrical_Power, Month, Cooling_Capacity, Address, Price. P(Y = y) is the prior probability, the probability of class Sales_Status occurring. P(Xi = xi | Y = y) is the likelihood, the probability of feature Xi having value xi given the class y. In Gaussian Naive Bayes, we assume that the features Xi are conditionally independent given the class Y. This means that the probability of a feature value given a class is not influenced by the values of other features. The likelihood term P(Xi = xi | Y = y) is modeled as a Gaussian distribution:

$$P(Xi = x_i \mid Y = y) = (1 / \sqrt{(2\pi\sigma^2 y)}) * exp(-(x_i - \mu y)^2 / (2\sigma^2 y)) \quad (2)$$

µy is the mean of feature X_i for class y. σ²y is the variance of feature X_i for class y. The Gaussian Naive Bayes classifier predicts the class y with the highest posterior probability. Multinomial Naive Bayes (MultinomialNB) is an algorithm that assumes that features come from a multinomial distribution. This variant is useful when using discrete data, such as frequency counts. It is commonly applied in natural language processing (NLP) use cases, such as email spam classification. In Multinomial Naive Bayes, we assume that the features X_i are categorical and follow a multinomial distribution. This is often used for text classification where the features represent the frequency of words occurring in a document. The likelihood term $P(X_i = x_i \mid Y = y)$ is modeled as a multinomial distribution:

$$P(Xi = xi \mid Y = y) = (n! / (x1! * x2! * \ldots * xk!)) * \prod (pyi)xi \quad (3)$$

n is the total number of occurrences of the feature X_i for class Sales_Status. xi is the number of occurrences of the value x_i for feature X_i for class Sales_Status. pyi is the probability of the value x_i for feature X_i for class y. The Multinomial Naive Bayes classifier predicts the class y with the highest posterior probability. In Bernoulli Naive Bayes, we assume that the features $X_i$are binary (either 0 or 1) and follow a Bernoulli distribution. This is often used for text classification where the features represent the presence or absence of a word in a document. The likelihood term $P(X_i = x_i \mid Y = y)$ is modeled as a Bernoulli distribution [31]:

$$P(Xi = xi \mid Y = y) = pyixi * (1 - pyi)(1 - xi) \quad (4)$$

pyi is the probability of feature X_i being 1 for class y. The Bernoulli Naive Bayes classifier predicts the class y with the highest posterior probability. The research framework of this study is shown:

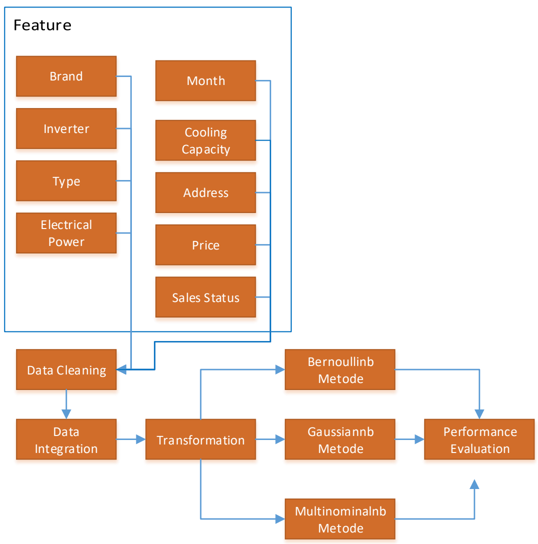

Figure 1. Research Framework

The research framework emphasizes the use of various features. These features include Brand, Inverter, Type, Electrical Power, Month, Cooling Capacity, Address, Price, and Sales Status. Preprocessing is required before applying a classification model such as data cleaning, integration, and transformation. The comparison of methods in Naive Bayes classification is Bernoullinb, Gaussiannb, and Multinomialnb. The final step is to measure the performance of the classification method.

The Naïve Bayes algorithm sometimes has the lowest performance in both accuracy and F1 score [32]. To ensure a comprehensive performance evaluation, the classification models are validated through two distinct testing scenarios. The first scenario focuses on evaluating the stability of the three Naïve Bayes variants across different data scales using 100, 200, and 300 data records. This scenario applies the K Fold Cross Validation technique with configurations of 5 fold, 10 fold, and 15 fold to measure the average accuracy of Bernoulli, Gaussian, and Multinomial variants. The second scenario evaluates the peak predictive performance of the models utilizing the entire dataset of 366 records. This second approach employs the holdout validation method with a conventional split of 80 percent for training data and 20 percent for testing data without applying cross validation. The peak results from this 80:20 split are then cross evaluated against alternative algorithms namely Support Vector Machine and KNeighbors Classifier to determine the ultimate superior classifier for air conditioning product status.

Results and Discussion

Based on the data obtained, a visualization report is produced with various graphic variations depending on the attributes used. The graph reviewed from the attribute differences is shown:

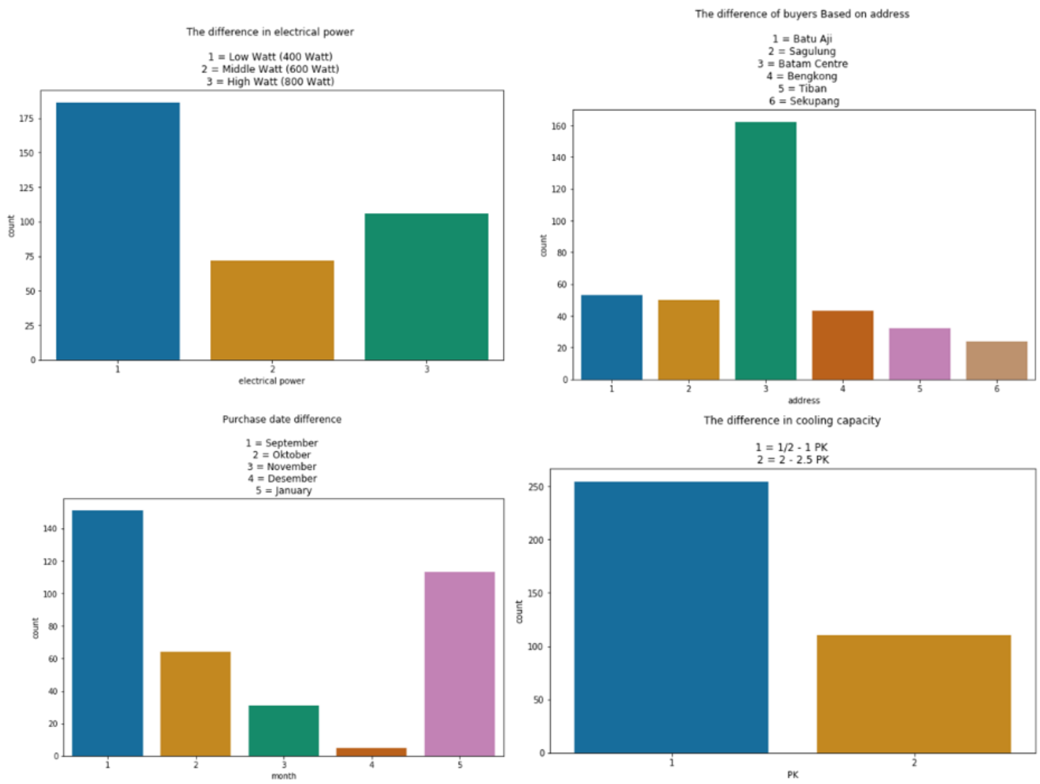

Figure 1. Visualization Report

Figure 2 describes the difference in user interest based on the selected electrical power of Air Conditioning. The number of air conditioning users with electrical power of Low-Watt is the most popular type. That is 186 customers. Air conditioning with Middle-Watt power has 72 users. Meanwhile, the number of air conditioning users with High-Watt power is 106. If seen from the differences based on customer location, the largest number of buyers came from Batam Center which reached 162 customers. The number of buyers from Sekupang was the smallest, reaching only 24 units. The largest number of transactions occurred in September, which was 151. Meanwhile, in December the number of buyers was only 5. That makes it the month with the lowest number of transactions. When viewed from the difference in cooling capacity, the number of buyers who chose 1/2 - 1 PK was 254 unit. Meanwhile, buyers who chose a cooling capacity above 1 PK reached 110 units. The pattern of product selection by customers is shown in the following figure:

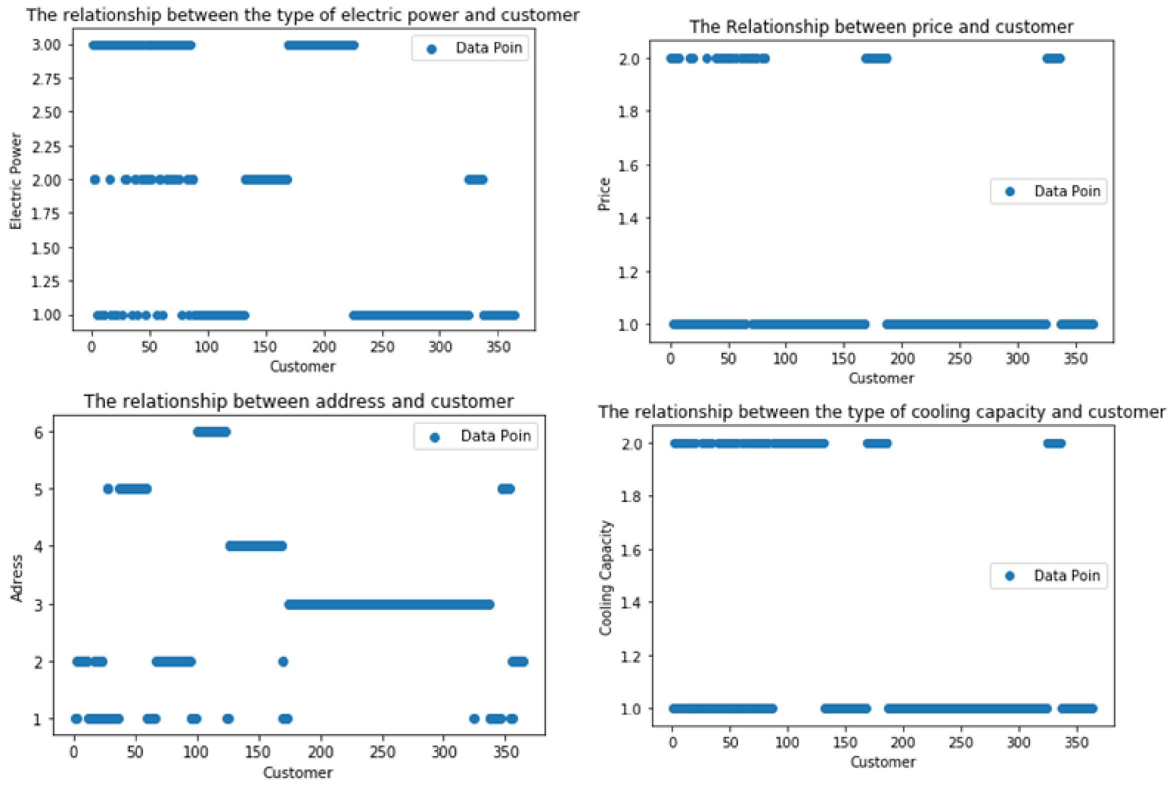

Figure 3. Relationship Between Features

The **Figure 3** explains how the pattern of product purchases by customers. Based on the visualization, it is evident that lower-level categories (value 1.0) are the most popular and dominant across almost all variables, particularly in Price and Cooling Capacity, where the majority of customers are concentrated. Conversely, the highest values (such as level 3.0 in Electric Power or level 6 in Address) show the lowest frequency, appearing only in specific customer clusters, which suggests that high-specification AC units or certain locations are not the majority within this sample. In the Address graph, the highest concentration of customers is found in category 3, spanning IDs from approximately 175 to 340, while categories 5 and 6 are the lowest in terms of data points. Overall, the dataset is dominated by "standard" or low-level choices (level 1.0), which provides a strong indication to the Naïve Bayes model that the prior probability for lower-specification classes will be significantly higher than for high-specification ones. One of the techniques used in machine learning for data preprocessing is MinMaxScaler. It functions to normalize numeric data into a certain range, usually between 0 and 1. The results of MinMaxScaler's work are shown in the **Table 2**.

Table 2. Min Max Scaler

Brand	Type	Electrical Power	Cooling Capacity	Price	Inverter	Address	Month
0.0	0.333	1.0	0.0	0.0	1.0	0.4	0.25
0.0	0.000	1.0	0.0	0.0	0.0	0.0	0.00
1.0	0.000	0.5	0.0	0.0	0.0	0.0	0.00
0.0	1.000	0.0	0.0	0.0	0.0	0.8	1.00
0.0	0.333	0.5	0.0	0.0	0.0	0.8	0.00

The mean of a feature is the average value of all data on a particular feature in a dataset. Mean provides a midpoint or center of the data distribution on a feature. While Variance of feature is a measure of how spread out the data is on a particular feature. The greater the variance, the more spread out the data is from the average. The **Table 2** below describes the Mean and variance that can be used to compare the average of a feature between Most_seller and Least_seller:

Table 3. The Mean and Variance

Status	Mean of feature (μ)	Variance of feature(σ^2)	Status	Mean of feature (μ)	Variance of feature(σ^2)
Most_seller	μ1: 1.047337 μ2: 2.352071 μ3: 1.748521 μ4: 1.254438 μ5: 1.094675 μ6: 1.215976 μ7: 3.088757 μ8: 2.754438	σ1²: 0.045230 σ2²: 1.878646 σ3²: 0.758529 σ4²: 0.190262 σ5²: 0.085966 σ6²: 0.169833 σ7²: 1.725036 σ8²: 2.969194	Least_seller	μ1: 1.653846 μ2: 1.653846 μ3: 2.192308 μ4: 1.923077 μ5: 2.000000 μ6: 1.500000 μ7: 2.730769 μ8: 1.000000	σ1²: 0.235385 σ2²: 0.635385 σ3²: 0.561538 σ4²: 0.073846 σ5²: 0.000000 σ6²: 0.260000 σ7²: 2.524615 σ8²: 0.000000

Frequency in statistics is the number of occurrences of a value or category in a dataset. Frequency indicates how often something occurs in a data set. Frequency can be used to calculate the probability of an event. A table of the frequency and probability of each feature is shown in the **Table 4**.

Table 4. The Frequency and Probability

No	Feature	Frequency	Probability	No	Feature	Frequency	Probability
1	Brand_1 Brand _2	331 33	0. 909341 0. 0.090659	6	Inverter Non_Inverter	278 86	0.763736 0.236264
2	Type_1 Type_2 Type_3 Type_4	162 60 12 130	0.445055 0.164835 0.032967 0.357143	7	Address_1 Address_2 Address_3 Address_4 Address_5 Address_6	53 50 162 43 32 24	0.145604 0.137363 0.445055 0.118132 0.087912 0.065934
3	Low Watt Middle Watt High Watt	186 72 106	0.510989 0.197802 0.291209	8	1st_month 2nd_month 3rd_month 4th_month 5th_month	151 64 31 5 113	0.414835 0.175824 0.085165 0.013736 0.310440
4	PK_1 PK_2	254 110	0.697802 0.302198	9	Most_seller Least_seller	338 26	0.928571 0.071429
5	Low_Price High_Price	306 58	0.840659 0.159341				

Table 4 provides a basic overview of the data distribution of each feature. The table also includes information about probability. It is a measure of the likelihood that an event will occur. The probability value ranges from 0 to 1. Judging from the product features of air conditioning, the highest probability covers low-watt (electrical power), low PK (cooling capacity), inverter type, and low price. The results of the comparison of the three methods were obtained in the research conducted. The results of the comparison accuracy of the classification model obtained:

Table 5. Accuracy Results Based on n-Fold

Naive Bayes Classifier	Cross Validation			Average
	5-Fold	*10-Fold*	*15-Fold*	
GaussianNB	95%	96%	96.667%	95.89%
BernoulliNB	97.5%	98%	98.333%	97.94%
MultinomialNB	97.5%	98%	98.333%	97.94%

Table 5 above describes the accuracy results of the classification model using 366 records. It shows that GaussianNB is the method with the lowest value compared to the others. It obtained an accuracy of 95.89%.

BernoulliNB and MultinomialNB obtained the same accuracy of 97.94%. The best number of fold cross-validation is 15-fold. The accuracy data of all methods is adjusted to the data tested, which is 20% of the training data.

Table 6. Accuracy Results Based on Data Amount

Naive Bayes Classifier	Number of Record			Average
	100	*200*	*300*	
GaussianNB	70.83%	85.354%	96.875%	84.35%
BernoulliNB	62.5%	89.899%	100%	84.13%
MultinomialNB	70.83%	85.354%	100%	85.39%

Table 6 explains the accuracy results of the classification model based on varying amounts of data. Changes in the amount of data tested affect accuracy. On average, MultinomialNB has the highest accuracy when compared to the others. It reaches 85.39%. The classification model is also carried out by comparing others. A comparison of accuracy through several classification methods can be seen in the following table:

Table 7. Comparison of Performance Evaluation Accuracy Results

Classifier Model	Class	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
NaiveBayes	Most Seller	99.45	99.68	99.56	99.3127%
	Least Seller	98.60	97.58	98.08	
	Weighted Avg	99.31	99.31	99.31	
SVM	Most Seller	98.72	99.52	99.12	98.6254%
	Least Seller	98.15	95.06	96.58	
	Weighted Avg	98.62	98.62	98.62	
KNeighborsClassifier	Most Seller	98.40	99.40	98.90	98.2817%
	Least Seller	97.74	94.02	95.84	
	Weighted Avg	98.28	98.28	98.28	
LogisticRegression	Most Seller	98.38	99.42	98.90	98.2817%
	Least Seller	97.80	93.94	95.83	
	Weighted Avg	98.28	98.28	98.28	

Based on the integration of additional evaluation metrics in the **Table 7**, it is clear how each algorithm performs when faced with class imbalance between Most Sellers and Least Sellers. The NaiveBayes algorithm proved to be the most robust model, not only achieving a peak accuracy of 99.3127 percent but also maintaining a very high and balanced F1-score across both classes: 99.56 percent for the Most Seller class and 98.08 percent for the Least Seller class. This indicates that NaiveBayes has excellent recall of 97.58 percent in detecting low-selling products (Least Sellers) without sacrificing precision. Conversely, although competing algorithms such as SVM, KNeighborsClassifier, and LogisticRegression recorded high global accuracy above 98 percent, their recall values for the minority Least Seller class decreased to between 93 and 95 percent. This decrease in recall values for the minority class reflects a tendency for the model to bias toward the majority Most Seller class. Thus, this comprehensive metric testing successfully strengthens the argument that NaiveBayes is the best and most stable model to be implemented in this AC product promotion business intelligence system.

Based on the analysis results in **Tables 5** and **6**, determining the best algorithm must be viewed from two different evaluation perspectives. In **Table 6**, when the model was tested using a limited number of recorded data variations through the K-Fold Cross Validation scheme, the MultinomialNB variant showed the best performance consistency with an average accuracy of 85.39 percent. However, when testing was conducted using the entire dataset with an 80:20 data split as shown in **Table 5**, the Naïve Bayes algorithm achieved overall peak performance with an accuracy of 99.312 percent, significantly outperforming competing algorithms such as the Support Vector Machine and the K-Neighbors Classifier. Thus, there is no contradiction between the two results but rather shows that MultinomialNB is very robust to dynamic data standardization while Naïve Bayes generally achieves the highest optimization at maximum data volume.

Discussion

The results of this study demonstrate that the Naïve Bayes (NB) model is highly effective for classifying the sales status of Air Conditioning (AC) products, achieving a peak prediction accuracy of 99.312%. This performance significantly outperforms alternative classification methods, including SVM (98.6254%) and KNeighbors Classifier (98.2817%), within the specified parameters. Such a high level of accuracy indicates that the Bayesian approach is robust for handling the multiclass classification of sales data in the electronics industry.Regarding the comparison of Naïve Bayes variants, the MultinomialNB model was identified as the most consistently reliable variant for these datasets, achieving the highest average accuracy of 85.39% across various data record sizes. While BernoulliNB and MultinomialNB showed identical performance in n-fold cross-validation testing both reaching 98.333% at 15-fold the GaussianNB variant generally yielded lower results, with an average accuracy of 95.89%. This suggests that while GaussianNB is suitable for continuous variables, the frequency-based nature of the sales features in this study aligns better with the Multinomial distribution.The analysis of customer behavior reveals a clear preference for specific product features, which serve as critical indicators for "Most Seller" status. Customers show a strong preference for Low-Watt electrical power (186 units) compared to middle or high-wattage options. Furthermore, the highest probability for a successful sale is associated with products featuring low prices, low cooling capacity (1/2 - 1 PK), and inverter types. Geographically, the Batam Center area emerged as the primary market with 162 customers, while transaction volume showed significant temporal fluctuations, peaking in September and dropping to its lowest point in December.The implementation of MinMaxScaler was vital in the preprocessing stage to normalize numeric data, ensuring that features like brand, price, and cooling capacity were scaled effectively for the classification models. These findings align with previous literature emphasizing that algorithm selection depends heavily on data type and that computational efficiency is a major advantage of Naïve Bayes for large-scale data. Ultimately, this research provides a strategic foundation for managers to utilize business intelligence and machine learning to maximize market positioning by aligning product offerings with identified consumer preferences.

Conclusion

The distribution of sales data indicates that customers significantly prefer low watt air conditioning units over middle watt or high watt options. The product features with the highest purchase probability include low electrical power, low cooling capacity, inverter types, and lower pricing structures. Regarding the machine learning evaluation, the results confirm that the Naïve Bayes model is highly effective for determining product sales status. In the first validation scenario across varying data record sizes, the MultinomialNB variant demonstrated the highest consistency with an average accuracy of 85.39 percent. Furthermore in the second evaluation scenario using the complete dataset under an 80:20 training and testing split, the overall Naïve Bayes algorithm achieved a peak prediction accuracy of 99.3127 percent. This optimal performance significantly outperformed alternative classification methods namely Support Vector Machine and KNeighbors Classifier. In conclusion MultinomialNB is established as the most reliable variant for handling this mixed attribute retail dataset, thereby providing a solid business intelligence framework for product promotion strategies in the electronics industry.

Acknowledgement

The authors are grateful to the Rector and the Research and Community Service Unit (LPPM) of Institut Teknologi Batam (ITEBA) for the support and for providing the necessary laboratory facilities to conduct this research. Special thanks are also due to our colleagues at the Faculty of Information Technology for their invaluable help and encouragement.

References

[1] Z. Jin, Y. Zheng, and Y. Zhang, “A novel method for building air conditioning energy saving potential pre-estimation based on thermodynamic perfection index for space cooling,” *J. Asian Archit. Build. Eng.*, vol. 22, no. 4, pp. 2348–2364, 2023, doi: 10.1080/13467581.2022.2109645.

[2] A. K. Shofa and W. Yustanti, “Implementation of Xception Algorithm with Convolutional Block Attention Module (CBAM) for Waste Type Detection in Visual Images,” *J. Emerg. Inf. Syst. Bus. Intelligen*, vol. 7, no. 1, pp. 82–90, 2026.

[3] X. Zhang, C. Zhou, Q. Sun, Y. Liu, Y. Yang, and Z. Zhuang, “Prediction of Solid Soluble Content of Green

Plum Based on Improved CatBoost," *Agric.*, vol. 13, no. 6, 2023, doi: 10.3390/agriculture13061122.

[4] S. A. Arnomo and N. B. Ibrahim, "MT-Forward Algorithm for Prediction of Project Categories Based on Selection of Mutants," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 5, pp. 1938–1943, 2022, doi: 10.18517/ijaseit.12.5.14296.

[5] Adriansyah, Desmarita Leni, and R. Sumiati, "Comparative analysis of energy-efficient air conditioner based on brand," *Polimesin*, vol. 20, no. 2, pp. 121–127, 2023.

[6] K. Devi, A. A. Juliet, and M. Rafic, "A contemporaneous sample on the consumer brand preferences for air conditioners in west Chennai," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11 Special Issue, pp. 698–702, 2019, doi: 10.35940/ijitee.K1119.09811S19.

[7] X. Kong, B. Sun, B. Sun, J. Zhang, S. Li, and Q. Yang, "Power Retailer Air-Conditioning Load Aggregation Operation Control Method and Demand Response," *IEEE Access*, vol. 8, pp. 112041–112056, 2020, doi: 10.1109/ACCESS.2020.3003278.

[8] L. Kang, G. Wang, Y. Wang, and Q. An, "The Power Simulation of Water-Cooled Central Air-Conditioning System Based on Demand Response," *IEEE Access*, vol. 8, pp. 67396–67407, 2020, doi: 10.1109/ACCESS.2020.2986309.

[9] T. Jiang, P. Ju, C. Wang, H. Li, and J. Liu, "Coordinated Control of Air-Conditioning Loads for System Frequency Regulation," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 548–560, 2021, doi: 10.1109/TSG.2020.3022010.

[10] S. Maurya, V. Singh, N. K. Verma, and C. K. Mechefske, "Condition-Based Monitoring in Variable Machine Running Conditions Using Low-Level Knowledge Transfer with DNN," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 1983–1997, 2021, doi: 10.1109/TASE.2020.3028151.

[11] K. Irshad, A. I. Khan, S. A. Irfan, M. M. Alam, A. Almalawi, and M. H. Zahir, "Utilizing Artificial Neural Network for Prediction of Occupants Thermal Comfort: A Case Study of a Test Room Fitted with a Thermoelectric Air-Conditioning System," *IEEE Access*, vol. 8, pp. 99709–99728, 2020, doi: 10.1109/ACCESS.2020.2985036.

[12] Q. Wei, T. Li, and D. Liu, "Learning Control for Air Conditioning Systems via Human Expressions," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7662–7671, 2021, doi: 10.1109/TIE.2020.3001849.

[13] Y. Chen, G. Fu, and X. Liu, "Air-Conditioning Load Forecasting for Prosumer Based on Meta Ensemble Learning," *IEEE Access*, vol. 8, pp. 123673–123682, 2020, doi: 10.1109/ACCESS.2020.2994119.

[14] A. Muis, Z. Razilu, U. R. Senga, H. Wulandari, and R. Andriyani, "Comparative Analysis of Naïve Bayes Variants for Predicting Stunting-Risk Families," *J. Inform.*, vol. 13, no. 1, pp. 70–78, 2026.

[15] M. Alenazi and S. Mishra, "Cyberatttack Detection and Classification in IIoT systems using XGBoost and Gaussian Naïve Bayes: A Comparative Study," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, pp. 15074–15082, 2024, doi: 10.48084/etasr.7664.

[16] W. Gunawan, R. A. Wiradiputra, A. P. Sari, D. Prayama, and E. R. Nainggolan, "Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers Using C 4.5 and Naïve Bayes Algorithms," *Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2145–2153, 2023, doi: 10.30630/joiv.7.4.1514.

[17] A. Lia, A. Hananto, B. Huda, A. Yuniar, and E. Novalia, "Determination of Training Participants in Community Work Training Centers Using the Naïve Bayes Classifier Algorithm," vol. 8, no. September, pp. 1162–1167, 2024.

[18] K. A. Ogudo and D. M. J. Nestor, "Modeling of an Efficient Low Cost, Tree Based Data Service Quality Management for Mobile Operators Using in-Memory Big Data Processing and Business Intelligence use Cases," *2018 Int. Conf. Adv. Big Data, Comput. Data Commun. Syst. icABCD 2018*, pp. 1–8, 2018, doi: 10.1109/ICABCD.2018.8465410.

[19] M. Saleem, M. Azam, Z. Mubeen, and G. Mumtaz, "Machine Learning for Improved Threat Detection:

LightGBM vs. CatBoost," *J. Comput. Biomed. Informatics*, vol. 7, no. 1, pp. 571–580, 2024, [Online]. Available: https://doi.org/10.56979/701/2024

[20] M. Mannan and S. G. Al-Ghamdi, "Indoor air quality in buildings: A comprehensive review on the factors influencing air pollution in residential and commercial structure," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, pp. 1–24, 2021, doi: 10.3390/ijerph18063276.

[21] F. R. Amalia, "Green Purchase Behavior Review Of The Role Of Price, Green Product, And Lifestyle," *J. Mark. Bus. Res.*, vol. 2, no. 1, pp. 11–22, 2022, doi: 10.56348/mark.v1i2.35.

[22] M. Aibin, "The Weather Impact on Heating and Air Conditioning with Smart Thermostats," *Can. J. Electr. Comput. Eng.*, vol. 43, no. 3, pp. 190–194, 2020, doi: 10.1109/CJECE.2020.2978459.

[23] A. Wijaya, H. A. Tan, R. M. Tanoro, B. Suwarno, and S. Noor, "Air Conditioner Customer Retention Based on Satisfaction and Loyalty with A Marketing Mix Analysis," *Int. J. Sci. Environ. Air*, vol. 6, no. 1, pp. 821–834, 2026.

[24] Tri Wahjoedi, Wulandari Harjanti, and Sri Rahayu, "Product quality and price perception on customer loyalty mediated by customer satisfaction," *World J. Adv. Res. Rev.*, vol. 14, no. 2, pp. 14–2, 2022, doi: 10.30574/wjarr.2022.14.2.0486.

[25] T. Khonglumtan and P. Srisattayakul, "Analysis of Perceived Determinants in Customer Service Quality: Focused on Automotive Air Conditioning Service Centers," *Glob. Bus. Financ. Rev.*, vol. 28, no. 3, pp. 102–116, 2023, doi: 10.17549/gbfr.2023.28.3.102.

[26] S. C. Putri, D. Lusianti, and D. Wismar'ein, "Enhancing Customer Satisfaction: The Impact of Service Quality and Store Atmosphere on Shopping Convenience," *J. Appl. Bus. Technol.*, vol. 6, no. 1, pp. 1–10, 2025, doi: 10.35145/jabt.v5i3.164.

[27] F. Yunita and E. Puspitaloka Mahadewi, "The Influence of Promotion, Service Quality, Brand Image On Indomaret Customer Satisfaction and Loyalty," *Int. J. Sci. Technol. Manag.*, vol. 6, no. 3, pp. 491–496, 2025, doi: 10.46729/ijstm.v6i3.1299.

[28] B. Ankrah, L. Brew, and J. Acquah, "Multi-Class Classification of Genetic Mutation Using Machine Learning Models," *Comput. J. Math. Stat. Sci.*, vol. 3, no. 2, pp. 280–315, 2024, doi: 10.21608/cjmss.2024.267064.1040.

[29] M. Ismail, N. Hassan, and S. S. Bafjaish, "Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task," *J. Soft Comput. Data Min.*, vol. 1, no. 2, pp. 1–10, 2020, doi: 10.30880/jscdm.2020.01.02.001.

[30] B. Spoorthi, S. S. Kumar, A. P. Rodrigues, R. Fernandes, and N. Balaji, "Comparative Analysis of Bank Loan Defaulter Prediction Using Machine Learning Techniques," *2021 IEEE Int. Conf. Distrib. Comput. VLSI, Electr. Circuits Robot. Discov. 2021 - Proc.*, no. November, pp. 24–29, 2021, doi: 10.1109/DISCOVER52564.2021.9663662.

[31] A. S. Dawood, "Performance Evaluation of Machine Learning Nave Bayes Algorithms for Network Traffic Classification," *Tech. Rom. J. Appl. Sci. Technol.*, vol. 13, pp. 12–26, 2023, doi: 10.47577/technium.v13i.9473.

[32] C. Brooks, V. Kovanović, and Q. Nguyen, "Predictive modeling of student success Using Machine Learning," *Handb. Artif. Intell. Educ.*, vol. 17, no. 1, pp. 350–369, 2023, doi: 10.4337/9781800375413.00027.