Journal of Information Technology applications in Education, Economy, Health and Agriculture
Vol. 01, No. 03, October 2024, pp. 57~63
ISSN: xxx-xxxx, DOI: -

Utilizing Datamining to Predict Sales Trends Based on Historical Data

Alby Afifuddin Junda[1], Maria Rosalina Trisna[2], Yustino Prami Genohon[3], Farrel Muhammad Raihan Akhdan[4], Imam Auwal Salisu[5]

[1,2,3] Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[4] Department of Informatics Magister, Universitas Islam Indonesia, Yogyakarta, Indonesia
[5] Department of Economic Management, Centre for Management Development, Nigeria.

Article Info

Article history:

Received August 05, 2024
Revised September 10, 2024
Received October 20, 2024

Keywords:

Naive Bayes, Support Vector Machine, Apriori Algorithm, Prediction Accuracy.

ABSTRACT(10 PT)

This study aims to compare the performance of the Naïve Bayes and Support Vector Machine (SVM) algorithms in predicting sales trends based on historical data. The results of the study show that SVM is more effective than Naïve Bayes with an accuracy of 34.74% compared to 15.49%. This study helps companies in making strategic decisions and improving operational efficiency. Data Mining is an important tool in predicting sales trends and improving prediction accuracy.

Corresponding Author:

Alby Afifuddin Junda
Department of Informatics Engineering
Faculty of Engineering, University of Widyagama Malang
Jl. Borobudur No. 35 Malang, East Java
Email: dino@widyagama.ac.id

1. INTRODUCTION

In the ever-evolving digital era, Data Mining has become a very important tool in business decision making, especially to predict sales trends based on historical data. With the increasing volume of data generated by companies, data mining techniques enable deeper analysis to understand consumer behavior patterns and market dynamics.[1].

The application of data mining techniques, such as the Apriori algorithm, allows companies to conduct market basket analysis, which is the analysis of consumer purchasing patterns. By using this algorithm, companies can more easily identify product combinations that are often purchased together, which in turn can help in designing more effective promotional strategies and product placement.[2]For example, if data shows that customers who buy bread also tend to buy jam, a store could place the two products close together to encourage sales.

In addition, classification and regression techniques in Data Mining are also used to predict future sales trends.[3]. By continuously analyzing historical data, companies can determine which products are experiencing increased or even decreased demand.

The use of time series methods is also becoming popular in predicting sales. This method uses historical data to identify seasonal patterns or long-term trends.[4], Thus, companies can not only respond quickly to changes in demand but also reduce the risk of losses due to excess or lack of stock.[3].

Overall, the use of Data Mining in predicting sales trends based on historical data is not only about increasing prediction accuracy but can also help companies make better strategic decisions.[5], but can also be used to predict which products the company will increase to increase production, and even which products may decrease to adjust production levels. This actually speeds up the decision-making process so that it can increase sales and reduce losses and can compete with other similar companies.[3].

This study aims to compare the performance of the Naïve Bayes and Support Vector Machine (SVM) algorithms in data mining techniques, in order to explore and process historical sales transaction data in business. The focus of this study is to find sales patterns that can provide useful information for strategic decision making. By exploring these patterns, this study is expected to help overcome obstacles in business operations while increasing sales intensity.[6], so that it can be known which algorithm is more effective in analyzing historical data and predicting sales trends.

2. METHOD

Data mining is the process of extracting relevant information and related knowledge from massive datasets using statistics, mathematics, artificial intelligence, and machine learning. In essence, data mining is a scientific field whose primary objective is to extract knowledge from the data or information we currently possess.

2. 1 Data collection

Historical sales data is collected from the sales of printed products from one of the printing companies in Makassar City, South Sulawesi, Indonesia. This data includes information about daily sales of various printed products. The dataset we take from Kaggle, with previous analysis using the backpropagation algorithm. This analysis will certainly help the company optimize its inventory management and sales strategy, as well as identify opportunities for growth and expansion. The dataset contains the following attributes:

- Transaction Date
- Types of products
- Number of Orders
- Price
- Total Price

The collected dataset will then be processed and analyzed using the Data Mining method, which is implemented on the Google Colab platform. In this study, two leading algorithms, namely Naive Bayes and Support Vector Machine (SVM), will be used to explore and reveal patterns contained in the data. With this approach, it is hoped that useful insights can be obtained in understanding sales trends and consumer behavior in more depth.

2. 2 Data Pre-Processing

Once the data is obtained, it is divided into two subsets: training data (70%) and testing data (30%). The training data is used to train the predictive model, allowing the model to learn from patterns in the data. Meanwhile, the testing data is used to evaluate the performance of the trained model, with the goal of measuring the model's accuracy and effectiveness in predicting outcomes on data it has never seen before.

2. 3 Naïve-bayes implementation

The implementation of the Naïve Bayes algorithm in this study aims to predict the number of sales trends. Naïve Bayes is a classification algorithm based on Bayes' Theorem, assuming that each feature is independent. This algorithm is widely used in classification because of its simplicity and effectiveness, especially when handling large datasets[18]. The data is divided into 80% for training data and 20% for testing data.

The way it works follows Bayes' Theorem, which states that the probability of an event can be calculated based on the available information. The basic formula of Naïve Bayes is as follows:

$$P(X|C) = \frac{P(X|C).P(C)}{P(X)}$$

The probability P(C|X)P(C|X) represents the likelihood of a data point belonging to class CC given the features XX. The probability P(X|C)P(X|C) indicates the likelihood of observing features XX given that the data belongs to class CC. Meanwhile, P(C)P(C) describes the prior probability of class CC, and P(X)P(X) is the overall probability of the features XX in the dataset.

2. 4 Support Vector Machine Implementation

The parameters of the Support Vector Machine algorithm used are C (cost) and Kernel, then find which parameter has the best value. After that, compare which variables get the best prediction results. Although it is a new concept, Support Vector Machines perform better than other methods, particularly when it comes to handwriting recognition and text classification. The concept of SVM begins with the classification problem of two training classes, positive and negative. In addition, this method tries to find the best separator so that it can maximize the boundaries between the two classes. In some cases that have been done, the data cannot be classified using the SVM linear method, so a kernel function is developed to classify data in non-linear form. In non-linear problems, SVM introduces the concept of kernel into a large space. The separator, also known as the hyperplane, seeks to optimize the distance or margin between various data classes in this high-dimensional space. By analyzing the margin and determining its greatest value, the best hyperplane for dividing two classes is found. The crux of the issue is the procedure for identifying this ideal hyperplane as the class separator.

3. RESULTS AND DISCUSSION

3.1. Data collection

Data collection in this study was taken from daily sales data from a printing product sales company between August 2022 and November 2023, with a total of 1,076 transactions. This data collection process was taken from the dataset available on Kaggle as an Open Source, which records various data information from various sources. This data is raw data that has not been processed. The gathered data is displayed.in Table 1.

No	Tanggal	Jenis Produk	Jumlah Order	Harga	Total
1	05/08/2022	Foodpak260	1000	1800	1800000
2	05/08/2022	FoodpakMatte245	1000	1900	1900000
3	05/08/2022	CraftLaminasi290	5000	750	3750000
4	05/08/2022	CraftLaminasi290	1000	1200	1200000
5	07/08/2022	Dupleks310	1000	1550	1550000
...	....	...	...	...	...
1075	15/11/2023	FoodpakMatte	1000	2200	2200000
1076	15/11/2023	GreaseProof	1000	300	300000

Table1 Raw Data

3.2. Data Pre-Processing

In the pre-processing stage, there are three steps taken, namely data cleaning, data transformation, and categorical variable encoding. In the transformation stage, time data is processed into DateTime data to maintain the consistency of existing time data.[8].

a. Data Cleaning

In the cleaning step, irrelevant data and unnecessary attributes were removed, leaving 1,064 data sets without reducing any attributes.

b. Data Transformation

Next, at the data transformation stage, several operations are carried out to make the data more ready for analysis, namely changing the Datetime format and extracting the year, month, and day which have several important purposes in date data analysis to facilitate further analysis.

```
# Preprocessing
df['Tanggal'] = pd.to_datetime(df['Tanggal'], format='%d/%m/%Y', errors='coerce')
df['Tahun'] = df['Tanggal'].dt.year
df['Bulan'] = df['Tanggal'].dt.month
df['Hari'] = df['Tanggal'].dt.day
```

Picture1 Preprocessing for date formats.

c. Encoding Variable

In this step, Label Encoding is used to convert categorical columns such as “Product Type” into a numeric format that can be processed by machine learning algorithms. This technique assigns unique integer values to each category, so that the data can be interpreted by the model without losing the original categorical information. Label Encoding is very useful for target columns (labels) in multi-class classification tasks because it produces an efficient numeric representation. With this approach, each category is assigned an integer label, such as 0, 1, 2, and so on, making the computation process of machine learning algorithms easier.[8].

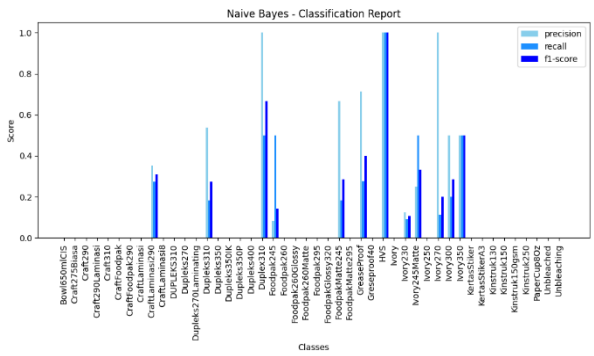
After all categories have been successfully coded into numbers, the data is ready for further modeling or analysis. Here is an example of coding for Label Encoding that has been applied in Figure 2;

```
# Encode target variable yaitu 'Jenis Produk'
le = LabelEncoder()
df['Jenis Produk'] = le.fit_transform(df['Jenis Produk'])
```

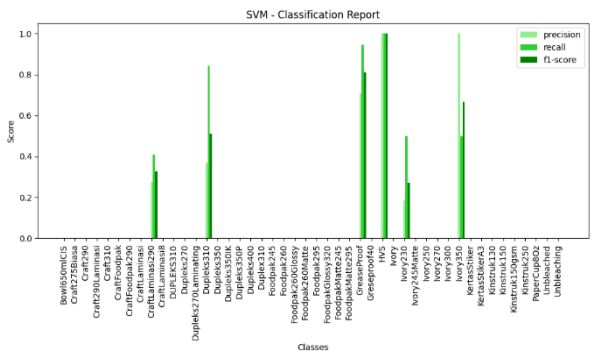
Picture2Encoding Variable 'Product Type'

3.3 Analyzing Data Using Naïve Bayes and SVM

After the data pre-processing stage, the analysis was carried out by applying the Naïve Bayes and SVM algorithms using Google Colab. Both algorithms are applied to predict the realization of printed product sales using the previously processed dataset. Then we will see how big the difference in the Accuracy levels of both is. 20% of the data is used for testing, while the remaining 80% is used for training, to ensure the model can be tested validly. The classification results show that the two models depict different graphs.



Picture 3 Classification Report - Naive Bayes



Picture 4 Classification Report SVM

In the classification there are 3 evaluation matrices for each class.

- Precision: How accurate the model is in predicting for each class.
- Recall: How well the model detects all samples that actually exist in a class.
- F1-Score: The precision and recall harmonic mean, which shows the balance between the two.

Based on the classification evaluation graphs for the Naive Bayes and SVM models, it can be seen that the performance of the two models differs significantly in several classes. Naive Bayes tends to have a lower overall distribution of precision, recall, and F1-score scores than SVM, with many classes showing scores close to zero, indicating that this model is less able to capture patterns in the data. In contrast, SVM shows more consistent performance with higher precision, recall, and F1-score scores in most classes, especially in classes with more regular data distributions. This suggests that SVM is more effective in separating the data in this case. However, there are several classes where both models struggle, as seen from the scores remaining low. This is likely due to data imbalance or complexity of patterns between classes. Overall, SVM appears to be the better choice of model for this dataset as it shows superior performance in most metrics and classes..

3.4 Comparison of Accuracy Results

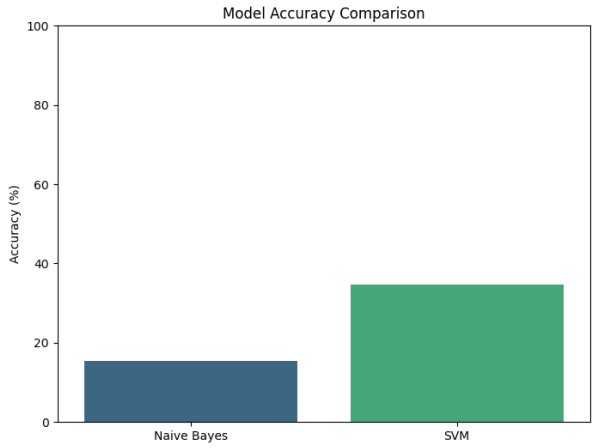


Figure 5 Accuracy graph of NB and SVM

Naive Bayes - Accuracy: 15.49%
SVM - Accuracy: 34.74%

Based on the comparison graph of the accuracy of the Naive Bayes and SVM models, it can be seen that the accuracy of the SVM model is much higher compared to Naive Bayes. SVM achieves an accuracy of 34.74% while Naive Bayes is only around 15.49%. This difference shows that SVM is more effective in capturing patterns and classifying the dataset used. The low accuracy of Naive Bayes can be caused by the assumption of independence between features that is not met in this dataset, thus affecting its performance. On the other hand, SVM, which uses a maximum margin approach to separate classes, is able to handle data complexity better, resulting in higher accuracy. Thus, SVM is a superior model choice for this dataset, both in terms of accuracy and other evaluation metrics.

3.5 Results Analysis

The evaluation results of the Naive Bayes model show very good performance in predicting student graduation based on the dataset used. Based on the confusion matrix, the model successfully predicted all students in the dataset correctly, resulting in a True Positive (TP) evaluation value of 4 students who were predicted to graduate actually graduated based on actual data. This shows that the model has good ability in identifying students who will graduate. True Negative (TN) of 2 students who were predicted not to graduate actually did not graduate. This shows that the model can identify students who are at risk of failure with perfect accuracy. False Positive (FP) there are no students who are predicted to pass but actually do not pass, indicating that the

model does not make mistakes in predicting graduation. False Negative (FN) there are no students who are predicted to fail but actually pass, indicating that the model does not fail to identify students who meet the graduation criteria. With the confusion matrix Accuracy: 100%, indicating that the model predicts all data correctly without error. Precision: 100%, indicating that all students who were predicted to pass actually passed. Recall: 100%, indicating that the model successfully captured all students who actually passed. F1-Score: 100%, indicating a perfect balance between precision and recall.

4. CONCLUSION

This study shows that the use of data mining techniques, especially the Naïve Bayes algorithm and Support Vector Machine (SVM), is effective in predicting sales trends based on historical data. By analyzing daily sales data, companies can identify consumer behavior patterns and optimize sales strategies and inventory management. The evaluation results show that SVM has a much higher accuracy than Naïve Bayes (34.74% vs. 15.49%), indicating its ability to handle the complexity of data patterns better. This study emphasizes the importance of utilizing historical data in strategic decision making to improve sales results and understand consumer preferences. Thus, the application of data mining techniques can be a useful tool for companies in monitoring and improving their sales performance.

REFERENCES

[1] VN Budiyasari, P. Studi, T. Informatics, F. Engineering, U. Nusantara, and P. Kediri, "Implementation of Data Mining in Eyeglass Sales Using the Apriori Algorithm,"*Indonesia. J. Comput. Inf. Technol.*, vol. 2, no. 2, pp. 31–39, 2017.

[2] N. Lestari, "Application of Apriori Algorithm Data Mining in Sales Information System,"*Edik Inform.*, vol. 3, no. 2, pp. 103–114, 2017, doi: 10.22202/ei.2017.v3i2.1540.

[3] MF Haryanti*et al.*, "The Influence of Data Mining, Corporate Strategy on Corporate Performance Reports," J. Manaj. and Business, vol. 3, no. 1, pp. 71–90, 2024.

[4] PA Duran, AV Vitianingsih, MS Riza, AL Maukar, and SFA Wati, "Data Mining for Sales Prediction Using Simple Linear Regression Method,"*Technique*, vol. 13, no. 1, pp. 27–34, 2024, doi: 10.34148/teknika.v13i1.712.

[5] AJP Sibarani, "Implementation of Data Mining Using Apriori Algorithm to Improve Drug Sales Patterns,"*JATISI (Journal of Information Technology and Information Systems)*, vol. 7, no. 2, pp. 262–276, 2020, doi: 10.35957/jatisi.v7i2.195.

[6] H. Setiawan, "Application of Data Mining Using the Support Vector Machine (SVM) Method to Analyze Fashion Retail Products to Determine Trends,"*Acad. Open*, vol. 9, no. 1, pp. 1–12, 2024, doi: 10.21070/acopen.9.2024.8581.

[7] ME Lasulika, "Comparison of Naïve Bayes, Support Vector Machine and K-Nearest Neighbor to Find the Highest Accuracy in Predicting Smoothness of Cable TV Payments,"*Ilk. J. Ilm.*, vol. 11, no. 1, pp. 11–16, 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[8] N. Purnama*et al.*, "Journal of Computer Science and Information Technology (CoSciTech) Prediction Model for the Number of Machine Lubricant Sales at PT. X With the Naïve Bayes Algorithm," vol. 5, no. 3, pp. 609–618, 2024.

BIOGRAPHIES OF AUTHORS

Please attach the close up picture here	Alby Afifuddin Junda description of curriculum vitae in 500 words maximum
Please attach the close up picture here	Maria Rosalina Trisna description of curriculum vitae in 500 words maximum
Please attach the close up picture here	Justin Prami Genohon description of curriculum vitae in 500 words maximum