Comparison Of Support Vector Machine And Naïve Bayes Algorithms For Analyzing Public Interest In Espresso Coffee

Sano Rita[1*], Muhammad Halmi Dar[2], Mila Nirmala Sari Hasibuan[3]

[1,2,3] Faculty of Science and Technology, Universitas Labuhanbatu, Sumatera Utara Indonesia.
[*]Corresponding Author:
Email: sanoritasanorita004@gmail.com

Abstract.

Given its increasing popularity, public interest in buying espresso coffee is an important concern for coffee industry players. To understand and predict this buying interest, the use of classification algorithms in data analysis is crucial. This study was conducted to compare the performance of two popular classification algorithms, namely Support Vector Machine and Naïve Bayes, in analyzing public interest in buying espresso coffee. This research problem is based on the need for an accurate predictive model in the coffee industry to aid in strategic decision-making related to marketing and sales. The proposed solution is to implement two different classification algorithms and assess their performance using a variety of performance evaluation metrics. The purpose of this study is to determine which algorithm is superior in terms of accuracy, precision, recall, and f1-score. The research method entails collecting data on public interest in purchasing espresso coffee, preprocessing data, implementing both algorithms, and evaluating each algorithm's performance. The results show that Naïve Bayes consistently outperforms Support Vector Machine in all performance evaluation metrics. Naïve Bayes achieved 94.00% accuracy, 91.40% precision, 100% recall, and 95.51% F1-Score, compared to Support Vector Machine, which achieved 90.00% accuracy, 88.60% precision, 96.90% recall, and 92.56% F1-Score. The conclusion of this study is that the Naïve Bayes classifier is more effective and efficient in predicting people's purchasing interest in espresso coffee compared to support vector machines. This advantage can be attributed to the ability of Naïve Bayes to handle data that may have non-normal distributions or independent variables.

Keywords: *Coffee, Espresso, Naïve Bayes, Public Interest, and Support Vector Machine.*

I. INTRODUCTION

High pressure and quick brewing produce a rich and concentrated taste in espresso. We carry out the brewing process at a temperature of 88–92 degrees Celsius to produce the ideal espresso. It's also crucial to adjust the brewing time between 22 and 30 seconds. The layers in perfect espresso consist of crema, body, and heart [1]. Usually served in a small cup known as a demitasse, espresso coffee serves as a base for other coffee drinks like cappuccino or latte [2]. Espresso has become an important part of coffee culture around the world, and it is a favorite for many coffee lovers. Espresso also has a higher caffeine content than regular coffee, so it provides a stronger stimulant effect [3]. The espresso brewing process requires a special machine that is able to create high pressure to produce the perfect drink [4], [5]. Espresso coffee has a rich and concentrated taste because of the unique brewing process [6]. Although it has become part of global coffee culture, espresso remains a favorite choice for coffee lovers because of its high caffeine content and strong stimulant effect.Some people enjoy espresso because of the delicious, thick, and complex taste of coffee. Some people don't like espresso because it's bitter. Although some people think espresso is too bitter, that doesn't mean espresso coffee is not delicious. The pleasure of espresso lies in the unique taste and complexity it offers. For many people, espresso is an art form in the world of coffee, where the skill of the barista and the quality of the coffee beans come together to create a drink that is thick, aromatic, and full of character.

The bitterness in espresso, when enjoyed properly, is actually part of its charm, offering layers of deep flavors and a more intense coffee drinking experience. Additionally, one can balance the bitterness of espresso in various ways to enhance its enjoyment for a variety of palates. For instance, many popular coffee drinks like cappuccino, latte, and macchiato often use espresso as a base, adding milk and sugar to enhance the taste and create a smoother, more balanced flavor profile. Despite its strong taste, espresso still offers extraordinary pleasure for those who know how to appreciate it, as demonstrated by the various ways of

serving it. In order to find out how many people are interested and not interested in espresso coffee, machine learning techniques using the Support Vector Machine (SVM) and Naïve Bayes methods were implemented to test it.SVM is one of the machine learning methods used for classification and regression [7], [8]. SVM works by finding the optimal hyperplane that can separate data in feature space with the maximum margin [9]. In the classification context, SVM seeks the hyperplane with the largest margin that divides two classes of data, enabling the model to generate more precise predictions on previously unseen data [10].

The SVM algorithm has the ability to handle the problem of overfitting, which can occur if the model is too complex. SVM also performs well on high-dimensional data [11]. Naïve Bayes is a classification algorithm that uses probability and statistical methods discovered by Thomas Bayes [12]. The Naïve Bayes method predicts the probability of a document category by combining previous experience with new knowledge with the basic idea of unifying word and category probabilities [13]. The Naïve Bayes method for classification consists of two stages. The first is training to gain past experience; the second is classification by calculating the probability value of all data [14]. The Naïve Bayes algorithm is very easy to use and performs well in data classification [15]. The Naïve Bayes method is the best method for carrying out classification with high accuracy results [13].This study aims to compare the performance of the SVM and Naïve Bayes algorithms in analyzing public interest in espresso coffee. We measure each algorithm's performance using an evaluation matrix that includes accuracy, precision, recall, and f1-score. By comparing these two methods, the author hopes to obtain accurate and reliable results regarding public preferences for espresso coffee. This study will not only provide in-depth insights into public views on espresso, but it will also apply machine learning techniques to social data analysis.

II. METHODS

In a study to determine public interest in espresso coffee, the SVM and Naïve Bayes methods were used to classify based on the data obtained. The research process begins with the data collection stage, which involves a direct survey of respondents. Once we gather the data, we proceed to the preprocessing phase. Preprocessing is the process of selecting and selecting data that is suitable for use and will later be used in this study. Therefore, we need to reselect the acquired data to ensure its appropriateness. This is due to the fact that each study has its own data requirements and needs. The preprocessing stage involves cleaning the data to remove errors and handle missing data. Moreover, we convert the data into a format better suited for analysis, like transforming text data into numeric data. This stage involves transforming the data into a different form and format.

Each study requires the data to be prepared for analysis. Therefore, the author will transform the data into an xlsx format for this study. After the data is ready, the implementation stage is carried out using the SVM and Naïve Bayes methods. SVM works by finding the optimal hyperplane to separate data that shows interest and disinterest in espresso coffee, while Naive Bayes calculates probabilities based on existing features to determine the class of each data sample. We evaluate the model results by comparing the performance of these two methods using metrics such as accuracy, precision, recall, and f1-score obtained from the confusion matrix results. The conclusions of this study provide insight into people's preferences for espresso coffee, as well as showing which method is more effective in classifying people's interest data. With this information, coffee businesses can better understand the market and develop more targeted marketing strategies.

III. RESULT AND DISCUSSION

We describe the study's results sequentially, following the research steps outlined in the previous methodology section. The selection stage involves the collection of data for this study.

Table 1. A part of Dataset

Customer	Gender	Taste	Type	Price	Category
Customer 1	Male	Delicious	Good	Cheap	Interest
Customer 2	Female	Delicious	Good	Cheap	Interest

Customer 3	Female	Delicious	Good	Cheap	Interest
Customer 4	Female	Delicious	Good	Expensive	Interest
Customer 5	Male	Delicious	Not Good	Cheap	Interest
Customer 6	Male	Not Delicious	Good	Cheap	Interest
Customer 7	Male	Delicious	Good	Expensive	Interest
Customer 8	Female	Delicious	Not Good	Cheap	Interest
Customer 9	Male	Not Delicious	Good	Cheap	Interest
Customer 10	Male	Not Delicious	Not Good	Expensive	Not Interest

Table 1 shows some of the datasets used. The total data set consists of 70 rows. This study employs two distinct data sets: the training data and the testing data. Training data is training information used to help the data classification process in data mining. Therefore, the training data can effectively assist in the classification process. We used 20 respondent data as the training data, and classified the remaining 50 data as testing data. Next, we create a model capable of analyzing the data. So the model that will be designed will later be used to classify data using the SVM method and the Naïve Bayes method. Figure 1 displays the model design.

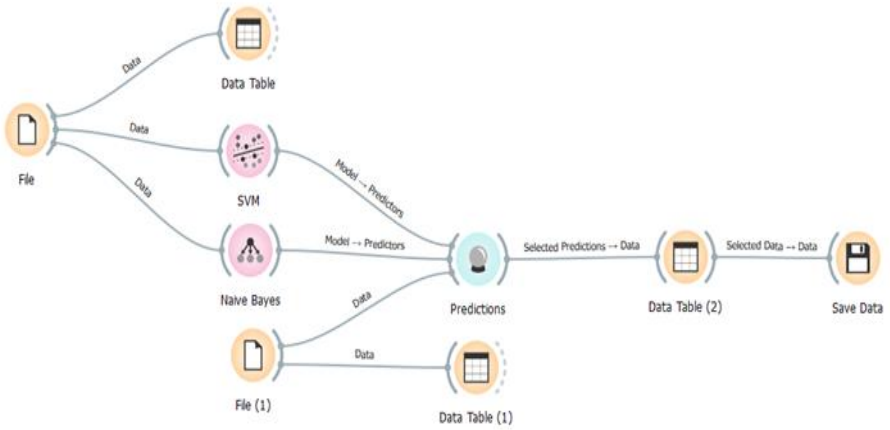

Fig 1. Model Design

Figure 1 shows a workflow diagram of a data analysis project that aims to compare the performance of two classification algorithms, namely SVM and Naïve Bayes, in predicting people's purchase intention for espresso coffee. This component indicates that the dataset, or raw data, originates from an external file. This dataset contains information relevant to the analysis, such as demographic attributes, preferences, or historical data on espresso coffee purchases. The classification model requires all the data. Once we import the data from the file, we may load it into a data table for further manipulation. This may include data cleaning, removing irrelevant data, or filling in missing values. We will enter the cleaned and ready-to-use data into this data table, serving as the prediction model's input. We use the SVM algorithm to build a classification model based on the training data. SVM works by finding a hyperplane that separates the classes with the largest margin. SVM is generally effective in classification problems with small to medium-sized data, especially if the data has clear class boundaries.

However, if there is a lot of data with irrelevant features or the data is too large, SVM may be less effective. The Naïve Bayes algorithm is used as an alternative classification method. This algorithm works based on probability and the assumption of independence between features. Naïve Bayes is usually very fast and efficient, especially for data with many features. However, its main drawback is that the assumption of feature independence, which is rarely true in practice, can reduce the accuracy of the model. This component combines the predictions generated by both models (SVM and Naïve Bayes). This prediction can be used to classify people's purchase intentions for espresso coffee. We compare the output of both models to assess their accuracy and performance. We can use these results to determine which model is more effective in predicting purchase intention. The models may store their predictions in this data table. You can use this data for further analysis or to validate the results.

Table 2. Test Results

Model	AUC	MCC
SVM	0.993	0.782
Naïve Bayes	1.000	1.000

Table 2 shows the results of the evaluation of the comparative model of the SVM and Naïve Bayes algorithms in analyzing people's buying interest in espresso coffee. This table presents the performance metrics of the two models, namely Support Vector Machine (SVM) and Naïve Bayes, which are compared based on several evaluation metrics. AUC (area under the curve): measures the model's ability to distinguish between classes. Values range from 0 to 1, with 1 indicating perfect classification ability. Both models have very high AUCs, indicating almost perfect ability to distinguish classes. However, Naïve Bayes shows perfect results with an AUC of 1,000, which is better than SVM (0.993). MCC (Matthews Correlation Coefficient): Measures the quality of a binary classification, with values ranging from -1 to 1. A value of 1 indicates perfect classification, 0 indicates random classification, and -1 indicates completely wrong classification. A high MCC value for Naïve Bayes indicates excellent and balanced classification, with an MCC of 1.000 indicating perfect classification. SVM also has a good MCC (0.782), but it is lower than Naïve Bayes.

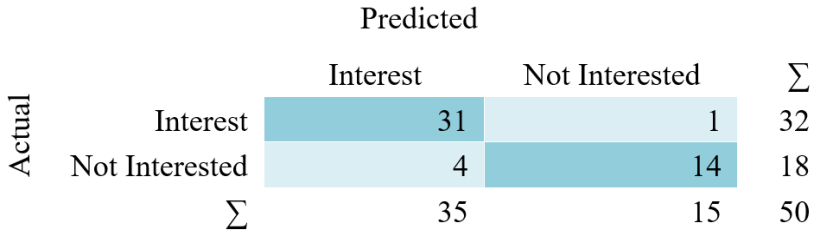

Fig 2. Confusion Matrix of SVM

Figure 2 presents a confusion matrix that the SVM algorithm uses to analyze people's interest in buying espresso coffee. This matrix displays the number of predictions made by the model and classifies them into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). A true positive occurs when the model predicts "interest" and the actual result is "interest." In this case, the model correctly identified 31 out of 32 individuals who are interested in espresso coffee. A True Negative occurs when the model predicts "not interested" but the actual response is "not interested." In this case, the model correctly identified 14 out of 18 individuals who are not interested. False positive refers to situations where the model predicts "interest" when the actual state is "not interested." In this case, there are 4 errors where the model incorrectly classifies individuals who are not interested as interested. These errors indicate that the model has some false positive predictions. A false negative is when the model predicts "not interested" when the actual is "interested." In this case, there was 1 error where the model incorrectly classified an interested individual as uninterested. This indicates that the model has a very low rate of false negative predictions.

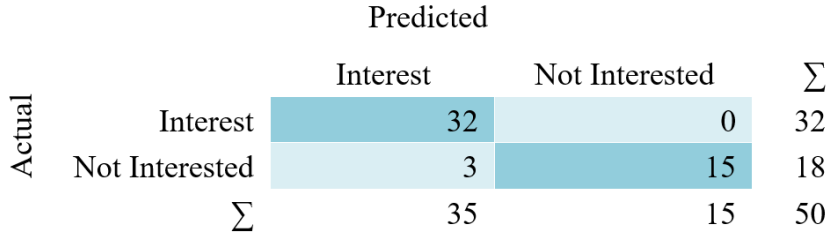

Fig 3. Confusion Matrix of Naïve Bayes

Figure 3 shows the confusion matrix of the prediction results using the Naïve Bayes model in analyzing people's interest in buying espresso coffee. This matrix provides information about the number of correct and incorrect predictions made by the model based on the categories "Interest" and "Not Interested." In the True Positive (TP) case, the model successfully identified all 32 individuals who were interested in espresso coffee. There were no errors in this category's prediction. In the True Negative (TN) case, the model correctly identified 15 out of 18 individuals who were not interested in espresso coffee. In the False Positive (FP) case, there were three errors where the model misclassified people who were not interested as

interested. This indicates some false positive predictions. While in the false negative (FN) case, there were no errors where the model failed to identify individuals who were interested. This shows that the model has excellent recall for the "Interest" class.The confusion matrix shows that the SVM model has a fairly excellent performance in classifying people's purchase interest in espresso coffee. This model has high precision, but it is slightly lower than the recall model.

This means that the model is a little more careful in predicting "interest," with several false positive errors and one false negative. Although there are several errors in classification (especially in the "not interested" class), this model still has a high overall accuracy. The almost perfect recall rate also shows that the model is very good at identifying individuals who are interested in espresso coffee, although there are several false positives indicating that some individuals who are not actually interested are classified as interested. Overall, although SVM does not have perfect performance, this model is still quite effective for this classification task. From this confusion matrix, we can conclude that the Naïve Bayes model has a very good performance in classifying people's purchase interest in espresso coffee. This model has high precision and recall, especially in detecting interested individuals. There are no errors in the prediction of "interest" (FN = 0), which means that the model has perfect recall for that category. Although there are some false positive predictions (FP = 3), this number is relatively small compared to the total predictions, so the precision remains high. The model also shows high overall accuracy. Overall, the Naïve Bayes model appears to be very effective in this classification task, with only a few small errors that can be considered as margins of error. Several factors, such as variations in the data or features not completely separated between the "Interest" and "Not Interested" categories, can cause these errors. However, when compared to the SVM model, Naïve Bayes has almost perfect results. The SVM appears to be slightly more prone to misclassifications in the "Not Interested" category.

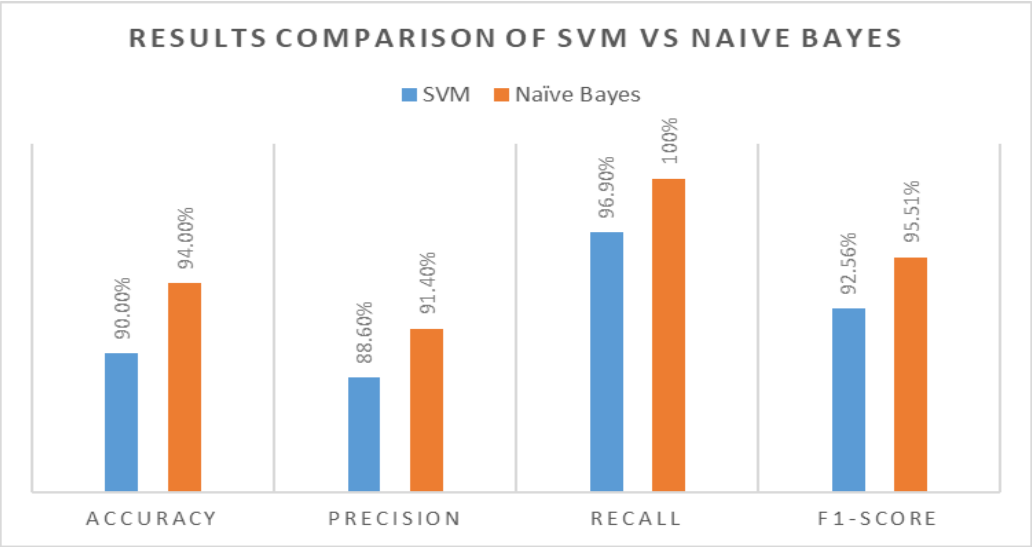

Fig 4. Results Comparison

Figure 4 shows a comparison of the performance results of two classification algorithms, namely Support Vector Machine (SVM) and Naïve Bayes, in analyzing people's buying interest in espresso coffee. The performance evaluation matrices used include accuracy, precision, recall, and F1-score. Naïve Bayes has a higher accuracy (94.00%) compared to SVM (90.00%), indicating that the Naïve Bayes model predicts correctly more often than SVM. Naïve Bayes is again superior (91.40%) compared to SVM (88.60%) in the precision metric, which means Naïve Bayes is better at minimizing false positives than SVM. Naïve Bayes achieves a perfect recall value (100%), indicating that this model is able to find all existing positive cases without missing any. Meanwhile, SVM also performs well, with a near-perfect recall (96.90%). Naïve Bayes has a higher F1-Score (95.51%) than SVM (92.56%), indicating that this model not only has high precision and recall but is also balanced between the two.

IV. CONCLUSION

This study successfully applied the SVM and Naïve Bayes algorithms in determining people's buying interest in espresso coffee. From the results of the study, it can be concluded that Naïve Bayes has a better performance compared to SVM in analyzing people's buying interest in espresso coffee based on all evaluation metrics used. Naïve Bayes excels in accuracy, precision, recall, and F1-Score, indicating that this model is more reliable in predicting buying interest with fewer errors in both false positive and false negative predictions. In real applications, the advantages of Naïve Bayes can be translated into a more effective and efficient model for use in predicting people's buying interest in espresso coffee products, so that it can help in making better decisions related to marketing and sales strategies. Suggestions for further research include exploring other algorithms that may be more effective or combining several algorithms to improve prediction accuracy. In addition, further research can include additional variables that may affect people's buying interest, such as demographic factors or personal preferences, to obtain a more comprehensive and accurate picture. Thus, the results of this study provide valuable insights for the coffee industry in an effort to understand and predict consumer behavior, which can ultimately improve marketing and sales strategies for espresso coffee.

REFERENCES

[1] R. Danutirta and R. Setiawati, "Teknik Pembuatan Perfect Espresso pada Operasional Lobby Lounge, Redtop Hotel Jakarta," ***J. Indones. Tour. Policy Stud.***, vol. 2, no. 1, 2019, doi: 10.7454/jitps.v2i1.114.

[2] D. Hamdan and A. A. Sastra, A to Z Memulai dan Mengelola Usaha Kedai Kopi, Cetakan Pe. PT AgroMedia Pustaka, 2020.

[3] G. E. S., "Pengaruh Kopi terhadap Kelelahan Otot pada Sprint 100 Meter (Studi pada Mahasiswa Universitas Diponegoro)," Universitas Diponegoro, 2017.

[4] S. Rais, Dailami, T. Pratama, N. O. Sipayung, and A. Saputra, "Pelatihan Barista Untuk Siswa-Siswi Di SMK Al-Azhar Batam," ***J. Keker Wisata***, vol. 2, no. 2, pp. 261–271, 2024, doi: 10.59193/jkw.v2i2.259.

[5] A. R. Yusuf and S. Rais, "Peranan Barista dalam Menyajikan Minuman Kopi Berkualitas di Cafe Excelso Vitka Point Tiban Kota Batam," Menata, vol. 2, no. 1, pp. 44–49, 2023, doi: 10.59193/jmt.v2i1.182.

[6] D. Gumulya and I. S. Helmi, "Kajian Budaya Minum Kopi Indonesia," ***J. Dimens.***, vol. 13, no. 2, pp. 153–172, 2017, doi: 10.25105/dim.v13i2.1785.

[7] H. Hendriyana, I. M. Karo Karo, and S. Dewi, "Analisis perbandingan Algoritma Support Vector Machine, Naive Bayes dan Regresi Logistik untuk Memprediksi Donor Darah," ***J. Teknol. Terpadu***, vol. 8, no. 2, pp. 121–126, 2022, doi: 10.54914/jtt.v8i2.581.

[8] F. Handayani *et al.*, "Komparasi Support Vector Machine, Logistic Regression dan Artificial Neural Network dalam Prediksi Penyakit Jantung," ***J. Edukasi dan Penelit. Inform.***, vol. 7, no. 3, p. Vol. 7 No. 3, 2021, doi: 10.26418/jp.v7i3.48053.

[9] J. Rusman, B. Z. Haryati, and A. Michael, "Optimisasi Hiperparameter Tuning pada Metode Support Vector Machine untuk Klasifikasi Tingkat Kematangan Buah Kopi," ***J. Komput. dan Inform.***, vol. 11, no. 2, pp. 195–202, 2023, doi: 10.35508/jicon.v11i2.12571.

[10] F. Qardhawi, "**Penerapan Algoritma Support Vector Machine (SVM) untuk Mendiagnosa Diabetes Melitus dan Hipertensi (Studi Kasus: Puskesmas Bnugoro)**," Politeknik Negeri Ujung Pandang, 2023.

[11] V. K. S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization," ***J. Nas. Tek. Elektro dan Teknol. Inf.***, vol. 9, no. 2, pp. 162–170, 2020, doi: 10.22146/jnteti.v9i2.102.

[12] Afdhaluzzikri, H. Mawengkang, and O. S. Sitompul, "Perfomance of Naive Bayes method with data weighting," **SinkrOn**, vol. 7, no. 3, pp. 817–821, 2022, doi: 10.33395/sinkron.v7i3.11516.

[13] N. Azhar, P. P. Adikara, and S. Adinugroho, "Analisis Sentimen Ulasan Kedai Kopi Menggunakan Metode Naive Bayes dengan Seleksi Fitur Algoritme Genetika," ***J. Teknol. Inf. dan Ilmu Komput.***, vol. 8, no. 3, p. 609, 2021, doi: 10.25126/jtiik.2021834436.

[14] A. B. P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain," ***J. Teknol. Inf. dan Ilmu Komput.***, vol. 7, no. 3, p. 599, 2020.

[15] T. Astuti and Y. Astuti, "Analisis Sentimen Review Produk Skincare Dengan Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO)," ***J. Media Inform. Budidarma***, vol. 6, no. 4, p. 1806, 2022.