JSCE (Journal of System and Computer Engineering)
Vol.6, No.2, April 2025, pp. 180-187
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v6i2.1860

Classification of Chocolate Consumption Using Support Vector Machine Algorithm

Firman Aziz [1,a]; Jeffry [2,b]; Nur Ayu Asrhi [3,c,*]; Supriyadi La Wungo [4,d]
[1,3] Universitas Pancasakti, Andi Mangerangi Street 73, Makassar 90121, Indonesia
[2] Institut Teknologi Bacharuddin Jusuf Habibie, Jl. Balaikota No.1, Parepare 91122, Indonesia
[4] STMIK Kreatindo, Jl. Kali Bambu, Manokwari 98312, Indonesia
[a] firman.aziz@unpacti.ac.id; [b] jeffry@ith.ac.id; [c] asrynurayu@gmail.com; [d] supriyadi.la.wungo@gmail.com
** Corresponding author*

Abstract

Chocolate, derived from the processing of cocoa beans (Theobroma cacao), is a widely consumed product with potential health risks when consumed excessively. This study investigates the classification of chocolate consumption behaviors using the Support Vector Machine (SVM) algorithm and evaluates its classification performance. A benchmark dataset on chocolate consumption was employed, partitioned into nine folds for training and testing purposes. To mitigate issues related to data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. The experimental findings indicate that SVM, enhanced by SMOTE, demonstrates a reliable capacity for classifying chocolate consumption categories. Performance evaluation across multiple experiments revealed variations in Accuracy, Precision, Recall, and F1-Score, with overall accuracies ranging from 50% to 60%, suggesting moderate but consistent classification performance.

Keywords : *Classification, Chocolate, Support Vector Machine Algorithm, SMOTE, Imbalanced Data*

1. Introduction

Chocolate, derived from cocoa beans (*Theobroma cacao*), is a significant agricultural commodity that has been cultivated in Indonesia since 1560, although it only became an important economic product around 1951. Today, Indonesia ranks as the third-largest producer of cocoa beans in the world, following Côte d'Ivoire and Ghana (Fauzi et al., 2022; Kusmawanto, 2022). The cocoa industry holds considerable potential within Indonesia's agricultural sector, contributing to both domestic consumption and international trade.

Beyond its economic value, chocolate offers various health benefits. It is known to stimulate the release of neurotransmitters that enhance mood and to provide high levels of antioxidants (Kesehatan, 2020). Chocolate also contains vitamins and minerals and promotes the release of endorphins in the brain, contributing to pain relief and overall well-being (Febriansyah, Nuha, & Kamal, 2021; Ikawati & Studi IV Kebidanan, n.d.). Endorphins act as natural analgesics, reducing pain intensity, particularly in cases such as menstrual pain (Jain et al., 2019; Mannem et al., 2022).

Additionally, chocolate contains various alkaloids, such as theobromine and phenylethylamine, which produce psychological effects (Cova et al., 2019; Fusar-Poli et al., 2022; Gopalakrishnan et al., 2021). The presence of tryptophan, an amino acid that

serves as a precursor to serotonin, further supports chocolate's impact on mood regulation (Dala-Paula et al., n.d.; Kanova & Kohout, 2021).

Despite these benefits, chocolate also contains compounds that can have addictive effects. Substances like tryptophan, phenylethylamine, enkephalin, and anandamide may encourage frequent consumption and generate psychoactive effects that reduce anxiety (Ross, 2021; Swidan & Bennett, 2020). Excessive consumption, however, may result in negative health outcomes such as weight gain, obesity, anxiety, irregular heartbeat, dental and bone health problems, gastrointestinal disorders, and even increased cancer risk.

Given these concerns, it is important to understand consumption patterns to anticipate potential risks. Therefore, this study aims to classify chocolate consumption using the Support Vector Machine (SVM) algorithm to predict the likelihood of individuals developing chocolate addiction. The consumption levels will be categorized into seven classes: never consumed, more than a decade, a decade, a year, a month, a week, and daily consumption (Murphy, 2022; Raschka & Mirjalili, 2019).

2. Method

2.1 Research Design

This research employs a quantitative approach with a classification model to predict chocolate consumption behaviors using the Support Vector Machine (SVM) algorithm. The study utilizes a benchmark dataset on chocolate consumption patterns, categorized based on frequency. The research is designed to determine the effectiveness of the SVM algorithm in classifying different levels of chocolate consumption, with an emphasis on addressing data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE).

2.2 Data Collection

The data used in this study was obtained from a benchmark dataset available in the UCI Repository. The dataset used is the Drug Consumption Quantified dataset, which can be accessed via UCI Repository. This dataset contains 1885 rows of data and 13 columns, with 12 variables representing respondent characteristics and 7 classes representing different levels of chocolate consumption.

Chocolate consumption is categorized into seven classes, namely:

- Never consumed
- More than a decade
- A decade
- A year
- A month
- A week
- A day

These classes represent the frequency of chocolate consumption reported by individuals in the dataset. Prior to being used in the research, the data has been normalized to ensure all variables are on the same scale, which is crucial for improving the performance of machine learning algorithms, particularly the Support Vector Machine (SVM) model.

The variables in the dataset include factors such as age, gender, socio-economic status, as well as responses to questions regarding other chocolate consumption behaviors. Data collection was conducted by studying reports related to chocolate consumption, which have been categorized and grouped into classes based on different consumption frequencies. This data is expected to reflect varying chocolate consumption patterns across a broad population.

After the data collection phase, the dataset is then processed and split into training data and test data for the purpose of classification modeling using SVM.

2.3 Data Preprocessing

Before analysis, the dataset undergoes several preprocessing steps to prepare it for the classification process. The following steps are performed:

- Data Cleaning: Missing or inconsistent data is handled by imputation techniques or removal of affected data points.
- Normalization: Numerical features are normalized to ensure that the SVM algorithm can process them effectively.
- Data Balancing: Given the imbalanced nature of the dataset (with some categories having fewer samples), the SMOTE technique is applied to oversample the minority class, ensuring that the model does not favor the majority class.

2.4 Support Vector Machine

The SVM algorithm is employed to classify the chocolate consumption patterns based on the features present in the dataset. SVM is chosen due to its effectiveness in high-dimensional spaces and its ability to handle both linear and non-linear classification problems.

The SVM model is trained using the training set and then evaluated using the test set. The classification process involves:

- Training: The SVM algorithm is trained on the labeled dataset to learn the decision boundaries between different consumption categories.
- Testing: The model is tested on unseen data to evaluate its ability to correctly classify the chocolate consumption patterns.

2.5 Evaluation Metrics

The performance of the SVM algorithm is evaluated using several metrics:

- Accuracy: The proportion of correctly classified instances among all instances.
- Precision: The proportion of true positive instances among all instances predicted as positive for each class.
- Recall: The proportion of true positive instances among all actual positive instances.
- F1-Score: The harmonic mean of precision and recall, providing a balanced measure of classification performance.

These metrics are computed for each of the seven classes to assess the model's overall performance and identify any potential areas for improvement.

2.6 Experimental Setup

The dataset is split into nine partitions, with each partition serving as the testing set once, while the remaining partitions are used for training. This technique, known as cross-validation, helps to ensure that the model's performance is generalizable and not dependent on a particular data split. Each experiment involves running the SVM algorithm with different configurations, adjusting hyperparameters such as the kernel type and regularization parameter to optimize performance.

2.7 Tools and Software

The research utilizes the Python programming language for data preprocessing, model development, and evaluation. Key libraries include:

- Scikit-learn: For implementing the SVM algorithm and other machine learning tools.
- Pandas: For data manipulation and preprocessing.
- Imbalanced-learn: For applying the SMOTE technique to balance the dataset.

3. Results And Discussion

3.1 Performance Metrics

The performance of the Support Vector Machine (SVM) classification model was evaluated using several performance metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's effectiveness in classifying chocolate consumption patterns.

The results for each partition of the training and testing data are summarized in Table 1. The table shows the accuracy, precision, recall, and F1-score values for various training-to-testing splits.

Table 1. Results of Accuracy, Precision, Recall, and F1-Score for Each Data Partition

Data Train	Data Test	Accuracy	Precision	Recall	F1-Score
90%	10%	0.67	0.65	0.67	0.64
80%	20%	0.65	0.62	0.65	0.62
70%	30%	0.64	0.62	0.64	0.61
60%	40%	0.62	0.60	0.62	0.59
50%	50%	0.62	0.61	0.62	0.60
40%	60%	0.61	0.57	0.61	0.58
30%	70%	0.58	0.56	0.58	0.56
20%	80%	0.55	0.53	0.55	0.52
10%	90%	0.49	0.45	0.49	0.45

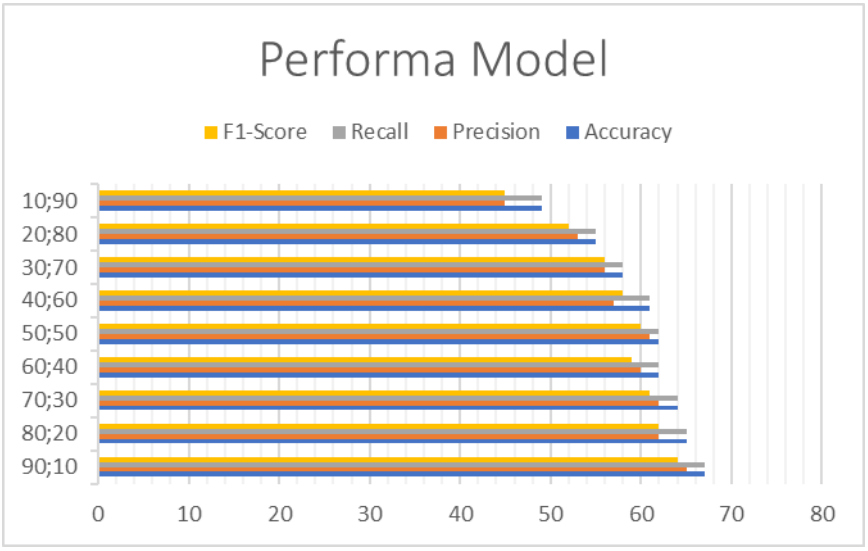

Figure 1. Comparison of Accuracy, Precision, Recall, and F1-Score of Each Classification Method in Each Data Partition

From the table, it can be observed that the model's performance improves as the percentage of training data increases. The highest performance is achieved with 90% training data and 10% testing data, where the accuracy is 67%, precision is 65%, recall is 67%, and F1-score is 64%.

3.2 Effect of Training Data Size

The analysis of the results clearly shows that as the size of the training data increases, the performance of the model improves. For instance, when the model is trained with 90% of the

data, it achieves the highest accuracy and other metrics. This suggests that larger training datasets allow the model to learn more effectively and generalize better to unseen data.

On the other hand, when the proportion of training data is reduced (e.g., 50% training and 50% testing), the model's performance drops. For example, with 50% training data, the accuracy drops to 62%. This indicates that the model struggles to generalize effectively when a smaller portion of the data is used for training.

Additionally, with only 10% of the data used for training, the performance significantly declines, with accuracy dropping to 49%, and both precision and recall also showing similar reductions. This emphasizes the importance of having sufficient training data for effective model training and classification.

3.3 Impact of Oversampling with SMOTE

To address the issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset. SMOTE generates synthetic samples for the minority classes, which helps to ensure that the classifier is not biased toward the majority class. After applying SMOTE, the data was re-split into training and testing sets as shown in Table 2. The increased balance in the dataset allowed the model to better handle the minority classes, resulting in improved overall performance.

Table 2. Data Split After SMOTE Application

Data Train	Data Test
5084	565
4519	1130
3954	1695
3389	2260
2824	2825
2259	3390
1694	3955
1129	4520
564	5085

The application of SMOTE led to more balanced class distributions and helped the model achieve better results across all performance metrics.

3.4 Kernel Comparison and Parameter Tuning

The choice of kernel in the SVM model is crucial for achieving optimal performance. The model was evaluated with three different kernel types: Radial Basis Function (RBF), Polynomial, and Linear. Figure 2 shows that the RBF kernel performed the best in terms of accuracy, precision, recall, and F1-score. This result aligns with previous research that suggests the RBF kernel is often the most effective for non-linear classification problems, such as the classification of chocolate consumption patterns in this study. The RBF kernel was particularly effective in capturing complex, non-linear relationships in the data, leading to better overall performance compared to the Polynomial and Linear kernels.

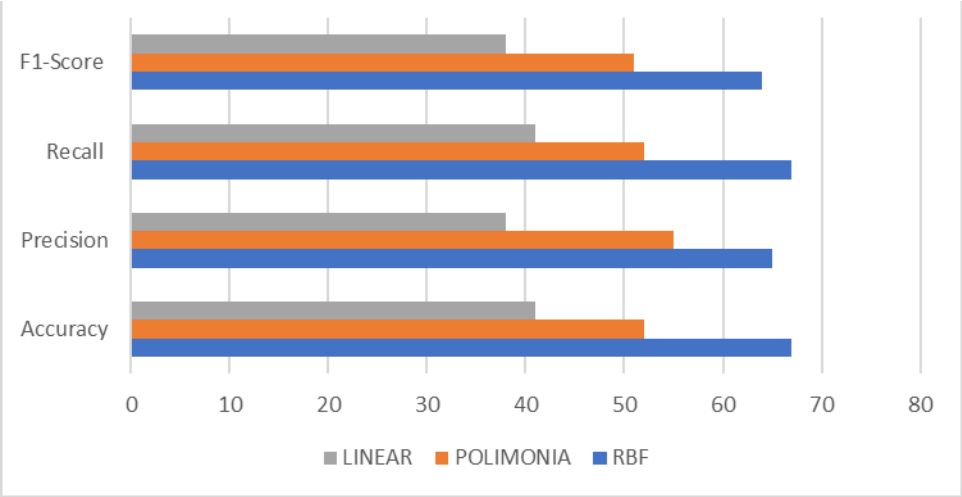

Figure 2. Comparison of RBF, Polymonia and Linear Kernels.

3.5 Interpretation and Future work

In summary, the Support Vector Machine (SVM) model, combined with SMOTE for addressing class imbalance, proved to be effective in classifying chocolate consumption patterns. The model's performance was best when 90% of the data was used for training, achieving an accuracy of 67%, along with good precision (65%) and recall (67%).

Key findings include:

- Larger training data results in better model performance.
- SMOTE was essential in balancing class distributions and improving model predictions.
- The RBF kernel was the most suitable for this classification problem.

Future work could focus on further optimizations, such as hyperparameter tuning using grid search, and potentially incorporating additional features to improve model performance. Additionally, exploring different machine learning algorithms may provide further insights into the most effective approaches for classifying chocolate consumption behaviors.

4. Conclusions

This study aimed to classify chocolate consumption patterns using the Support Vector Machine (SVM) algorithm with benchmark data from the UCI Repository. The results show that the SVM algorithm proved to be effective in classifying chocolate consumption, with the model achieving an accuracy of up to 67%. The precision, recall, and F1-score were also relatively good when 90% of the data was used for training. The study also highlighted the significant impact of training data size on model performance, where increasing the percentage of training data improved the accuracy and other evaluation metrics. Furthermore, the application of the Synthetic Minority Oversampling Technique (SMOTE) played a crucial role in addressing class imbalance, ensuring that the model could classify minority classes more accurately. In terms of kernel selection, the Radial Basis Function (RBF) kernel performed the best, effectively handling non-linear relationships in the data. This study recommends further research to explore more parameter optimization, alternative machine learning techniques, and additional features to enhance model performance. Overall, the SVM algorithm, combined with SMOTE for class imbalance handling, was shown to be a promising approach in classifying chocolate consumption patterns with satisfactory results.

References

Cova, I., Leta, V., Mariani, C., Pantoni, L., & Pomati, S. (2019). Exploring cocoa properties: is

theobromine a cognitive modulator? *Psychopharmacology*, *236*(2), 561–572. https://doi.org/10.1007/S00213-019-5172-0

Dala-Paula, B., Deus, V., Tavano, O., Chemistry, M. G.-F., & 2021, undefined. (n.d.). In vitro bioaccessibility of amino acids and bioactive amines in 70% cocoa dark chocolate: What you eat and what you get. *Elsevier*. Retrieved January 10, 2024, from https://www.sciencedirect.com/science/article/pii/S0308814620322597

Fauzi, F. A., Islami, S., Pembangunan, E., & Tidar, U. (2022). ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI VOLUME EKSPOR KAKAO INDONESIA KE AMERIKA SERIKAT: ANALYSIS OF FACTORS AFFECTING THE. *Ojs.Untika.Ac.IdFA Fauzi, FS IslamiJurnal Ilmiah Mahasiswa Fakultas Pertanian, 2022•ojs.Untika.Ac.Id*, *2*(2), 2775–3646. https://doi.org/10.52045/jimfp.v2i2.348

Febriansyah, E., Nuha, K., & Kamal, S. (2021). PENGARUH COKELAT HITAM TERHADAP INTENSITAS NYERI DISMENORE PRIMER PADA MAHASISWI AKADEMI KEBIDANAN SALEHA BANDA ACEH. *Sel Jurnal Penelitian Kesehatan*, *8*(2), 96–106. https://doi.org/10.22435/SEL.V8I2.5108

Fusar-Poli, L., Gabbiadini, A., Ciancio, A., Vozza, L., Signorelli, M. S., & Aguglia, E. (2022). The effect of cocoa-rich products on depression, anxiety, and mood: A systematic review and meta-analysis. *Critical Reviews in Food Science and Nutrition*, *62*(28), 7905–7916. https://doi.org/10.1080/10408398.2021.1920570

Gopalakrishnan, B., SKA, A., … R. S.-A. of the, & 2021, undefined. (2021). Antioxidant Activity in Toffees and Selected Medicinal Plants. *Annalsofrscb.RoB Gopalakrishnan, AKS SKA, R Selvam, G LakshmananAnnals of the Romanian Society for Cell Biology, 2021•annalsofrscb.Ro*, *25*(2), 1294–1300. http://annalsofrscb.ro/index.php/journal/article/view/1080

Jain, A., Mishra, A., Shakkarpude, J., Ijcs, P. L.-, & 2019, undefined. (2019). Beta endorphins: the natural opioids. *Researchgate.NetA Jain, A Mishra, J Shakkarpude, P LakhaniIjcs, 2019•researchgate.Net*. https://www.researchgate.net/profile/Anand-Jain-10/publication/343850641_Beta_endorphins_The_natural_opioids/links/5f44adb2299bf13404f0d30e/Beta-endorphins-The-natural-opioids.pdf

Jurnal Kesehatan Masyarakat, P., Ikawati, N., & Studi IV kebidanan Fakultas Keperawatan dan Kebidanan Universitas Megarezky Makassar, P. D. (n.d.). PENGARUH PEMBERIAN COKELAT HITAM TERHADAP PENURUNAN INTENSITAS DISMENORRHEA PRIMER PADA REMAJA PUTRI DI SMA NEGERI 3. *Journal.Universitaspahlawan.Ac.IdN Ikawati, S SyamsuryanitaPREPOTIF: JURNAL KESEHATAN MASYARAKAT, 2022•journal.Universitaspahlawan.Ac.Id*. Retrieved January 10, 2024, from http://journal.universitaspahlawan.ac.id/index.php/prepotif/article/view/5178

Kanova, M., & Kohout, P. (2021). Tryptophan: A unique role in the critically ill. *International Journal of Molecular Sciences*, *22*(21). https://doi.org/10.3390/IJMS222111714

Kesehatan, R. A.-J. A., & 2020, undefined. (n.d.). Pengaruh Konsumsi Coklat Hitam (Theobroma cacao) Terhadap Dismenore Pada Mahasiswa Kebidanan. *Download.Garuda.Kemdikbud.Go.IdRF AdriJurnal Amanah Kesehatan, 2020•download.Garuda.Kemdikbud.Go.Id*. Retrieved January 10, 2024, from http://download.garuda.kemdikbud.go.id/article.php?article=2730178&val=24846&title=Pengaruh Konsumsi Coklat Hitam Theobroma cacao Terhadap Dismenore Pada Mahasiswa Kebidanan

Kusmawanto, M. (2022). *Pembibitan pada Budidaya Tanaman Kakao (Theobroma cacao L.) Bulk di PTPN XII Kebun Kendeng Lembu Glenmore–Banyuwangi*. https://sipora.polije.ac.id/15162/

Mannem, M., Mehta, T. R., Murala, S., & Bollu, P. C. (2022). Endorphins. *Neurochemistry in Clinical Practice*, 239–245. https://doi.org/10.1007/978-3-031-07897-2_12

Murphy, K. (2022). *Probabilistic machine learning: an introduction*.

https://books.google.com/books?hl=id&lr=&id=OyYuEAAAQBAJ&oi=fnd&pg=PR27&dq=K.+Murphy,+Probabilistic+machine+learning:+an+introduction.+2022.&ots=A7kLWCCx_8&sig=0jwi6W5BSdaRIs9RJk75l16PP5k

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. https://books.google.com/books?hl=id&lr=&id=sKXIDwAAQBAJ&oi=fnd&pg=PP1&dq=S.+Raschka+and+V.+Mirjalili,+Python+machine+learning:+Machine+learning+and+deep+learning+with+Python,+scikit-learn,+and+TensorFlow+2.+2019.&ots=VaCluOWIFs&sig=v2DF-RuQzMij2B9L0-7hGhgYVV0

Ross, M. (2021). *Kratom is Medicine: Natural Relief for Anxiety, Pain, Fatigue, and More*. https://books.google.com/books?hl=id&lr=&id=RAwhEAAAQBAJ&oi=fnd&pg=PA1&dq=M.+Ross,+Kratom+is+Medicine:+Natural+Relief+for+Anxiety,+Pain,+Fatigue,+and+More.+2021.&ots=WoFRB2zF_i&sig=m5cqiCzpFQtYRloEdS9nzp4QFJU

Swidan, S., & Bennett, M. (2020). *Advanced Therapeutics in Pain Medicine*. https://books.google.com/books?hl=id&lr=&id=wxEIEAAAQBAJ&oi=fnd&pg=PT6&dq=S.+Swidan+and+M.+Bennett,+Advanced+Therapeutics+in+Pain+Medicine.+2020.&ots=B1Hj9CsD8S&sig=3plkrBywW65lyXDWOIVx6VoyoZQ