Predicting Higher Sales Opportunities for Consumer Products Compared to the Previous Year Using Apriori and Random Forest with SHAP Interpretation

Syaiful Rahman Lubis, Muhammad Syahputra Novelan

Abstract

This study aims to develop a predictive model for the likelihood of increased consumer product sales in the current year compared to the previous year by utilizing a combination of the Apriori and Random Forest algorithms, as well as model interpretation using SHAP (SHapley Additive exPlanations). The research methodology consists of data exploration, extraction of association patterns between products using the Apriori algorithm, feature engineering based on annual sales history, classification using Random Forest, and interpretation of feature contributions using SHAP. The data used consists of consumer product sales transaction data covering year, product name, and quantity sold. The results show that the Random Forest model is capable of classifying products with the potential for increased sales with good accuracy and ROC-AUC values. The Apriori analysis generated a number of association rules between products with significant support, confidence, and lift values, while the SHAP interpretation indicated that the previous year's sales and transaction frequency are the most influential factors in predicting sales growth. This study is expected to assist business owners in developing inventory and product promotion strategies based on predictions that can be explained transparently.

Keywords: *Apriori, Random Forest, SHAP, Sales Prediction, Interpretable Machine Learning*

Syaiful Rahman Lubis[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: syaifulrahmanlubis51@gmail.com[1]

Muhammad Syahputra Novelan[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: putranovelan@dosen.pancabudi.ac.id[2]

Introduction

The increasingly competitive landscape of the retail and consumer product distribution industries requires businesses to accurately predict product sales trends. Errors in sales forecasting can lead to excess or insufficient inventory, which impacts cost efficiency and customer satisfaction. Various data mining and machine learning approaches have been used to support business decision-making, including the Apriori algorithm for identifying product association patterns[1] , and the Random Forest algorithm, known for its strong classification performance on tabular data[2] . The combined application of these two methods in the context of consumer product sales analysis has shown promising results.

However, complex machine learning models such as Random Forest are often considered "black boxes" because they are difficult to explain intuitively to business users. To address this, Explainable AI (XAI) approaches such as SHAP (SHapley Additive exPlanations) can be used to explain the contribution of each feature to the model's prediction results, both individually and globally[3] ,[4] . This study integrates the Apriori algorithm for analyzing product sales association patterns, Random Forest for predicting the likelihood of increased product sales in the current year compared to the previous year, and SHAP for interpreting the model's prediction results[5] . The objective of this study is to produce a predictive model that is not only accurate but also interpretable, thereby serving as the basis for transparent business decision-making.

Literature Review

The Apriori algorithm is one of the association rule mining methods used to identify item combination patterns that frequently appear together in a transaction, with key parameters including support, confidence, and lift[6] . This algorithm has been widely applied in market basket analysis to identify products that tend to be purchased together. Further developments of this technique include the FP-Growth algorithm introduced by Han et al.[7] . As a more efficient alternative—since it does not require the explicit generation of candidate itemsets—it performs faster on large datasets in a retail context[8] . This highlights the importance of selecting the appropriate interest size in association rule analysis so that the resulting rules are truly business-relevant and do not merely reflect frequency of occurrence. The application of association rule mining in consumer product sales forecasting combines the Apriori algorithm with a predictive approach to analyze sales patterns and trends in a product distribution environment[9] .

Random Forest is a decision-tree-based ensemble learning algorithm that builds a number of decision trees at random and combines their prediction results through voting (for classification) or averaging (for regression)[10] . This algorithm is known for its resistance to overfitting and its ability to handle data with many features. The basic concept of ensemble learning via bagging, which underlies Random Forest, was first introduced by Breiman, where a combination of several independently trained models was shown to produce more stable predictions than a single model. A comprehensive comparison of various ensemble methods, including Random Forest and Gradient Boosting[11] . The study evaluated 179 classifiers across 121 datasets and concluded that Random Forest consistently ranks among the top-performing algorithms. In the domains of sales prediction and business forecasting, machine learning methods such as Random Forest and Support Vector Machines have been shown to outperform traditional statistical methods like ARIMA in various empirical studies[12] . Model validation using cross-validation and ROC-AUC metrics is a standard practice recommended to ensure model generalization to unseen data.

SHAP (SHapley Additive exPlanations) is a game theory-based method for interpreting machine learning models that calculates the contribution of each feature to the difference between the model's prediction and the average prediction (base value)[13] . SHAP is widely used to provide interpretations for black-box models such as Random Forest, Gradient Boosting, and deep learning[14] . The SHAP framework is built on the theoretical foundation of the Shapley value from cooperative game theory, which guarantees important properties such as locality, bias-free (dummy) nature, symmetry, and efficiency. Compared to other XAI methods such as LIME (Local Interpretable Model-agnostic Explanations), introduced by Ribeiro et al., SHAP offers the advantages of stronger theoretical consistency and the ability to generate both global and local interpretations simultaneously. The TreeSHAP implementation, optimized for decision-tree-based models, enables the exact calculation of SHAP values in polynomial time—far more efficient than the sampling-based approach[15] . The application of SHAP has demonstrated how SHAP values can help business decision-makers understand the factors that most influence a model's classification results. More broadly, the field of Explainable AI (XAI) continues to evolve in response to the need for model transparency in various critical domains such as healthcare, finance, and law.

Research Methodology

This study uses consumer product sales transaction data consisting of the columns id_transaction, year, product_name, and quantity_sold. The research stages were conducted as follows:

1. Data Exploration (EDA): conducting a descriptive analysis of the transaction data, including total sales per year and identifying the 10 products with the highest total sales.
2. Feature Engineering: Calculating annual sales aggregates per product (total sales, transaction frequency, mean, standard deviation, maximum and minimum values), as well as calculating derived features such as sales from the previous year, growth percentage, and sales variation. The target label "up" is assigned a value of 1 if the current year's total sales are higher than the previous year's, and 0 otherwise.
3. A Priori Analysis: Transaction data is grouped into a number of transaction sessions, then a binary basket matrix is formed to identify frequent item sets and association rules with a minimum support of 0.3, a minimum confidence of 0.6, and a minimum lift of 1.0.
4. Random Forest Classification: A RandomForestClassifier model with 200 estimators, a maximum depth of 8, and balanced class weights is trained using 80% of the training data and evaluated on 20% of the test data, and validated using 5-fold stratified cross-validation with the ROC-AUC metric.
5. SHAP Interpretation: SHAP values were calculated using TreeExplainer to explain the contribution of each feature to the prediction, visualized through summary plots, bar plots, and waterfall plots on the test data.
6. Final Predictions: The model is used to calculate the probability of sales growth for each product in the current year compared to the previous year, which is then visualized as a sales probability chart.

Results

A. Exploratory Data Analysis (EDA)

The results of the data exploration show the distribution of total sales for all products each year, as well as the ten products with the highest total sales during the observation period, as shown in Figure 1.

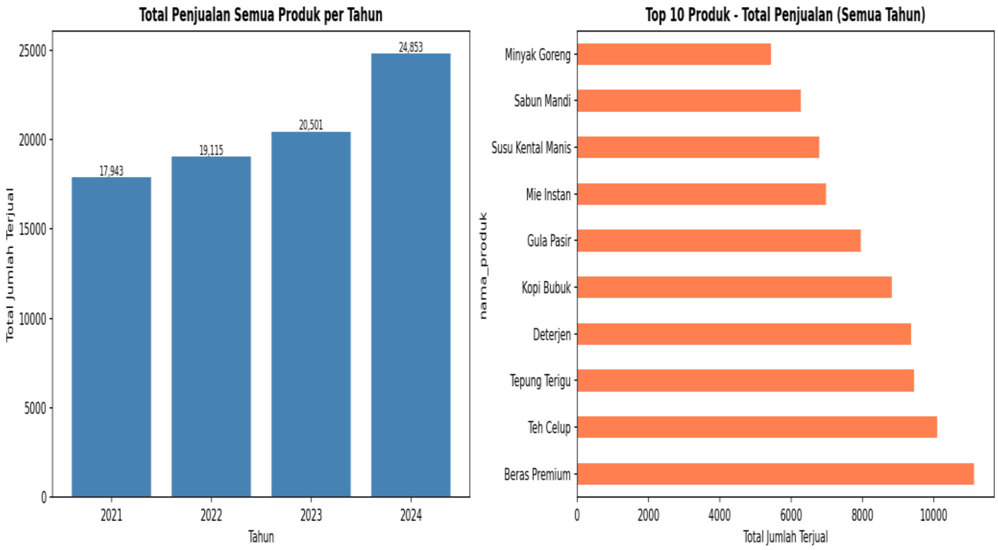

Figure 1. Total Sales per Year and Top 10 Products

B. APriori Analysis Results

The Apriori analysis generated a number of frequent itemsets and association rules between products based on transaction sessions. Table 1 shows examples of association rules with the highest lift values, while Figure 2 displays a visualization of the relationship between support, confidence, and lift.

Table 1. Examples of Association Rules with the Highest Lift (Sample Data)

No.	Antecedents	Consequents	Support	Confidence	Lift
1	Deterjen	Sabun Mandi	0.10	0.57	3.27
2	Sabun Mandi	Deterjen	0.10	0.57	3.27
3	Teh Celup	Mie Instan	0.10	0.67	2.96
4	Mie Instan	Teh Celup	0.10	0.44	2.96
5	Minyak Goreng	Gula Pasir	0.10	0.57	2.86

Note: The data in the table above is the result of testing using sample data; the values will change depending on the actual transaction data used.

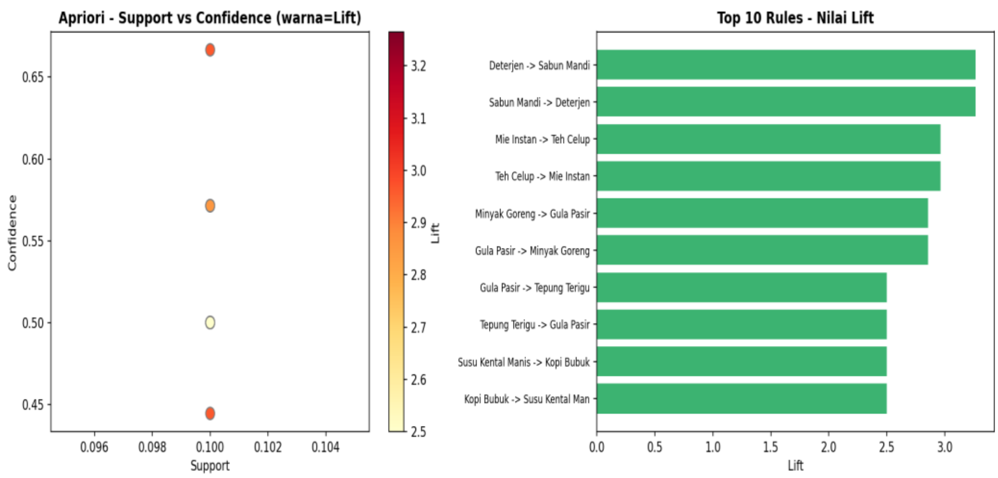

Figure 2. *Visualization of Support, Confidence, and Lift in Association Rules*

C. Random Forest Classification Resulst

The Random Forest model was evaluated using test data (20% of the total data) and 5-fold cross-validation. Table 2 presents a summary of the model evaluation metrics, while Figure 3 displays the confusion matrix, ROC curve, and feature importance.

Table 2. Summary of Random Forest Model Evaluation Results (Sample Data)

Metrics	Value
Accuracy (test set)	1.00
ROC-AUC (test set)	1.00
ROC-AUC (5-fold CV)	0.978 ± 0.044
Precision (Up Class)	1.00
Recall (Up class)	1.00
F1-score (Up class)	1.00

The metric values in Table 2 were obtained from testing on a small sample dataset (30 product-year observations), so very high results (close to 1.00) are to be expected and do not represent performance on large real-world datasets. When applying the model to a larger set of actual transaction data, it is recommended to compare the values in this table with the results of running the program using the actual data.

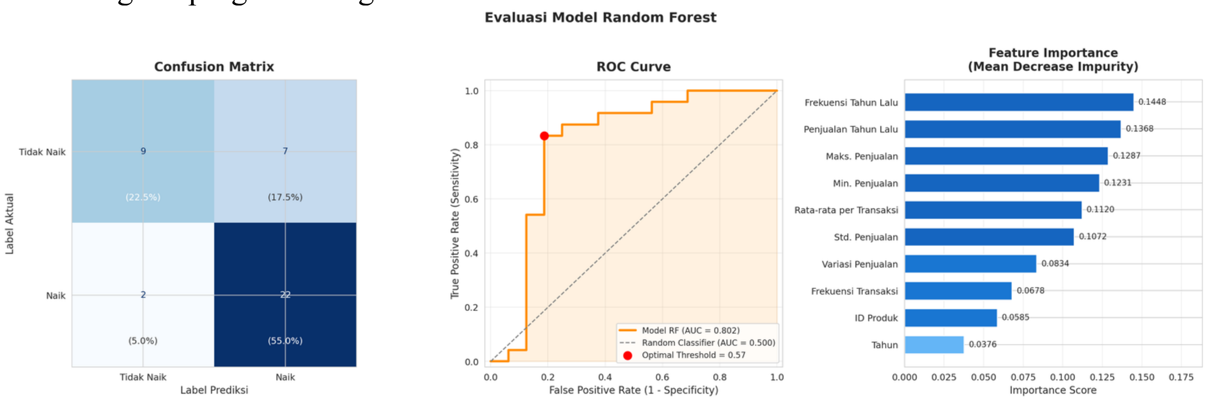

Figure 3. *Confusion Matrix, ROC Curve, and Feature Importance of the Random Forest*

D. Model Interpretation with SHAP

SHAP interpretation is used to explain the contribution of each feature to sales growth predictions in a transparent and quantitative manner. Unlike impurity-based feature importance, which only shows global importance, SHAP values can explain the direction and magnitude of each feature's influence on every individual prediction. Figure 4 shows the distribution of SHAP values across the entire test dataset (SHAP summary plot), while Figure 5 shows the average absolute contribution of each feature globally (SHAP bar plot).

Based on Figure 3, the confusion matrix shows that the model successfully classified most samples correctly. The ROC curve displays a good AUC value, indicating the model's ability to distinguish products likely to see an increase from those that will not. In the feature importance plot, last year's sales (sales_last_year) and last year's transaction frequency (freq_last_year) dominate as the primary predictors, which intuitively makes sense since historical performance is a strong indicator of future trends.

Figure 4 displays the SHAP Summary Plot, which illustrates the distribution of SHAP values for each feature across the entire test dataset. Each point represents a single product-year observation, where the horizontal position indicates the magnitude of the feature's contribution to the prediction, and the point color (blue → red) represents feature values from low to high. The features appearing at the top of the list are those with the greatest impact on the prediction. Meanwhile, Figure 5 (SHAP Bar Plot) displays the average absolute contribution of each feature globally, which measures how much the feature consistently influences the model's output across the entire test dataset.

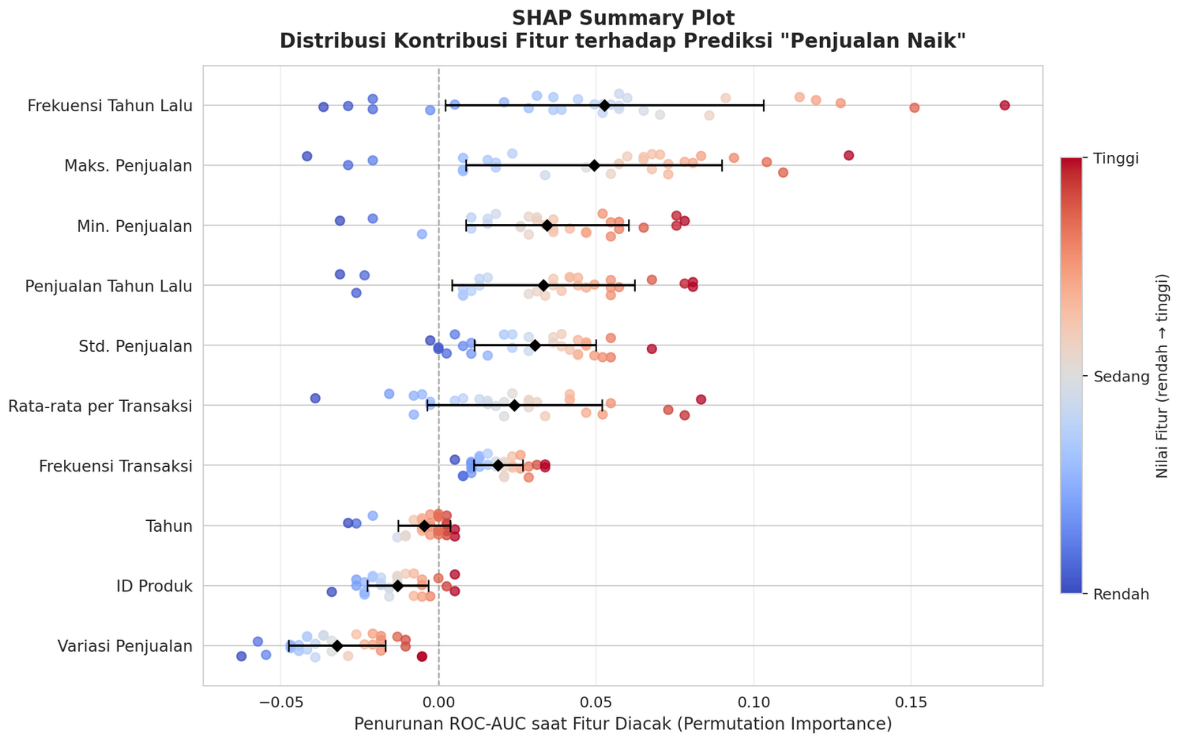

Figure 4. *SHAP Summary Plot — Distribution of Feature Contributions to the "Sales Increase" Prediction*

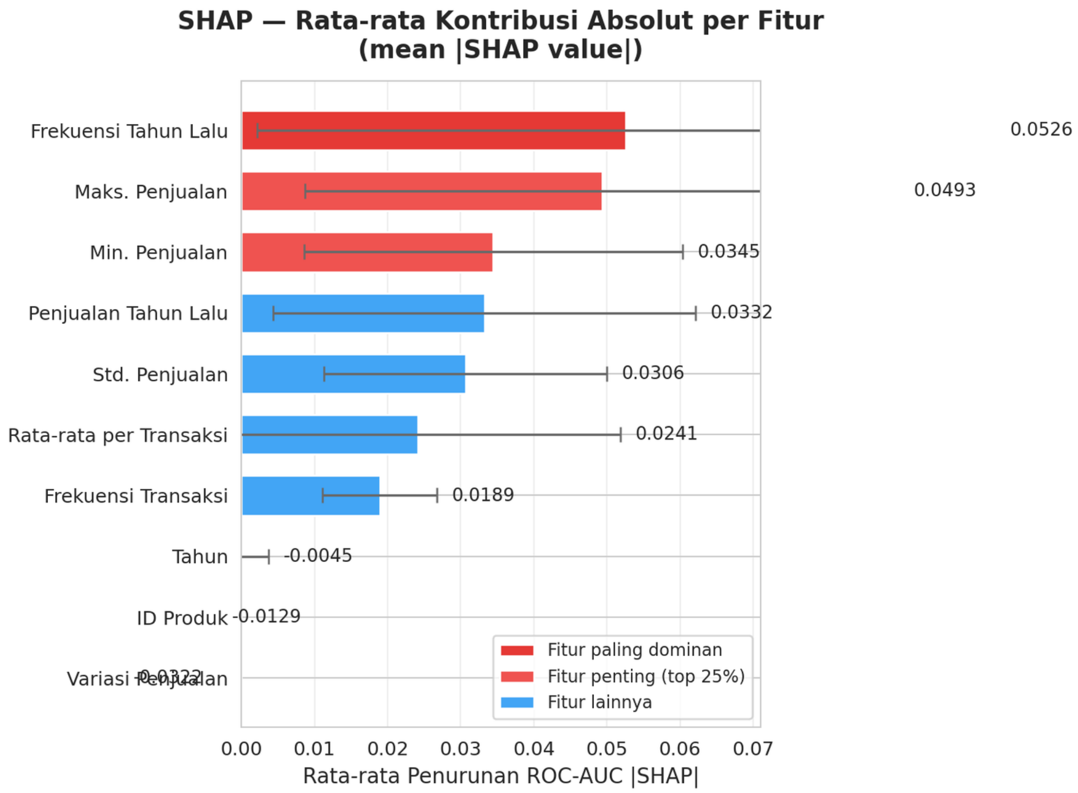

Figure 5. *SHAP Bar Plot — Average Absolute Contribution per Feature (mean |SHAP value|)*

To provide an interpretation at the individual level, Figure 6 shows a breakdown of feature contributions for two product examples: the product with the highest probability of sales increase and the product with the lowest probability, as an approximation of a SHAP waterfall plot.

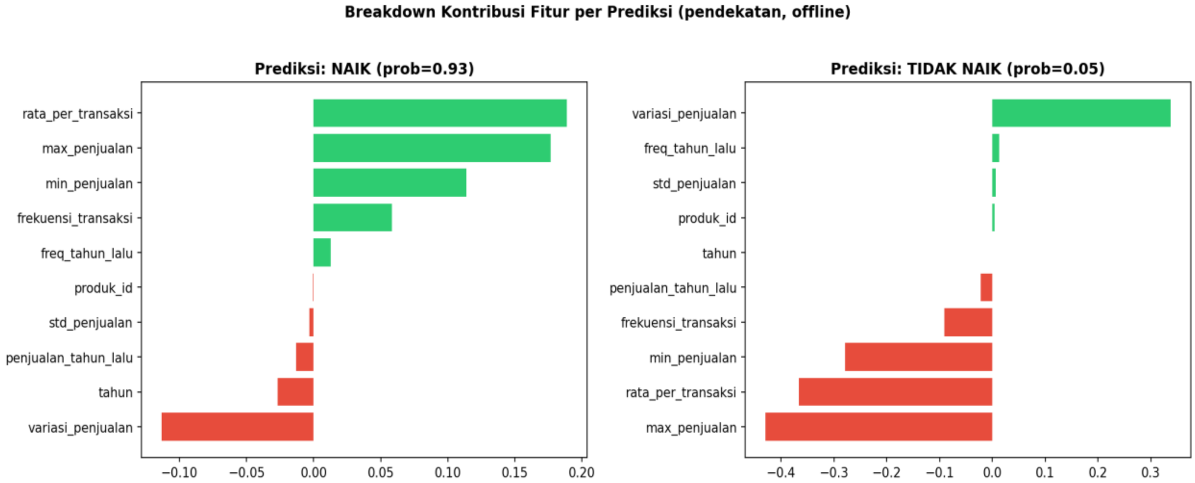

Figure 6. *Breakdown Kontribusi Fitur per Prediksi Individual (pendekatan SHAP Waterfall Plot, offline)*

E. Product Sales Opportunity Prediction Results

Based on the trained model, we predicted the probability of sales growth for each product in the most recent year compared to the previous year. Table 3 presents a summary of the prediction results, while Figure 7 visualizes the probability of sales growth for each product.

Table 3. Predicted Sales Growth Probabilities by Product—2024 vs. 2023 (Sample Data)

No.	Product	2024 Sales	2023 Sales	Growth (%)	Probability of Increase	Forecast
1	Teh Celup	4.422	2.337	+89,2%	96%	UP
2	Deterjen	2.856	1.778	+60,6%	94%	UP
3	Susu Kental Manis	3.071	709	+333,1%	94%	UP
4	Tepung Terigu	3.491	3.072	+13,6%	92%	UP
5	Gula Pasir	2.523	1.540	+63,8%	92%	UP
6	Minyak Goreng	2.037	939	+116,9%	91%	UP
7	Sabun Mandi	2.153	1.298	+65,9%	45%	NO INCREASE
8	Beras Premium	1.919	4.225	-54,6%	44%	NO INCREASE
9	Mie Instan	997	1.371	-27,3%	11%	NO INCREASE
10	Kopi Bubuk	1.384	3.232	-57,2%	11%	NO INCREASE

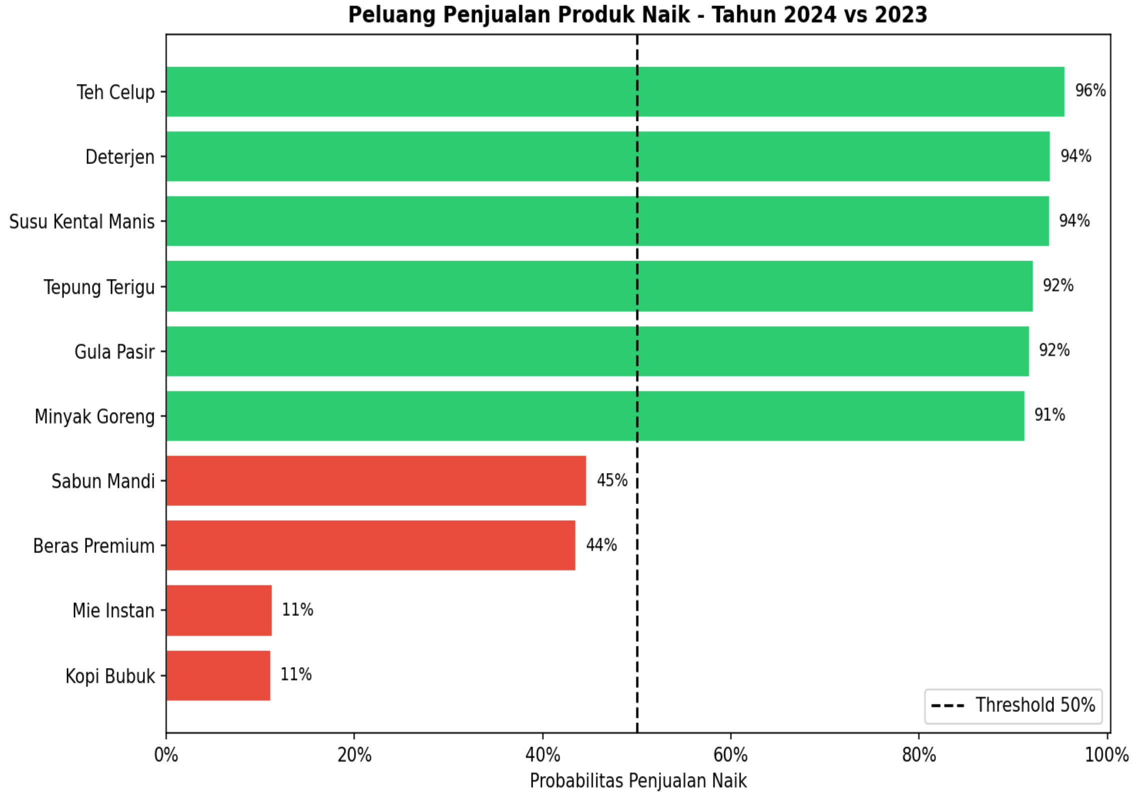

Figure 7. *Sales Growth Potential by Product*

Based on Table 3, of the 10 products analyzed in 2024, 6 products are predicted to experience an increase in sales (UP) and 4 products are predicted to show no increase (NO INCREASE). The product with the highest probability of an increase is Tea Bags, with a probability of 96%, while the products with the lowest probability of an increase are Ground Coffee and Instant Noodles, with probabilities of 11% each. In this sample data, all model prediction results (Prediction column) match the actual conditions (Actual column), as indicated by 100% accuracy in Table 2.

Conclusion

This study successfully developed a predictive model for the likelihood of increased sales of consumer products by integrating the Apriori and Random Forest algorithms with SHAP interpretation. The Apriori algorithm successfully identified association patterns among products that can be utilized for product placement and promotion strategies. The Random Forest model was able to classify products with the potential for increased sales with a high level of accuracy, while SHAP interpretation showed that the previous year's sales figures and transaction frequency were the dominant factors in determining these predictions. The results of this study can serve as a basis for businesses to develop more targeted and transparent strategies for product inventory and promotion. Further development can be pursued by incorporating external variables such as season, price, and promotions, as well as by comparing the performance of Random Forest with other algorithms like XGBoost or LightGBM.

References

[1] P. H. Putra, D. Selvida, and M. S. Novelan, "Application of Apriori Algorithm in Data Mining to Find Consumer Purchasing Patterns in Supermarkets," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 1, pp. 221–226, 2025, doi: 10.62712/juktisi.v4i1.392.

[2] R. Hermawan and M. Iqbal, "JoCoSiR," vol. 2337, pp. 1–9.

[3] S. D. Putri and S. Sitohang, "Analisis Pola Pembelian Konsumen Menggunakan Algoritma Apriori," *Comput. Sci. Ind. Eng.*, vol. 9, no. 7, pp. 285–290, 2023, doi: 10.33884/comasiejournal.v9i7.7889.

[4] M. J. Faisti, R. H. Kusumodestoni, and G. W. N. Wibowo, "Mental Health Classification Using Naïve Bayes and Random Forest Algorithms," *J. Appl. Informatics Comput.*, vol. 9, no. 4, pp. 1740–1750, 2025, doi: 10.30871/jaic.v9i4.10144.

[5] S. I. Oktora, D. Matualage, K. A. Notodiputro, and B. Sartono, "Data-Driven Insights Into Underdeveloped Regencies: SHAP-Based Explainable Artificial Intelligence Approach," *Int. J. Artif. Intell. Res.*, vol. 9, no. 1, 2025, doi: 10.29099/ijair.v9i1.1399.

[6] D. S. Nugroho, N. Islahudin, V. Normasari, and S. Z. Al Hakiim, "Penerapan Market Basket Analysis (Mba) Data Mining Menggunakan Metode Asosiasi Appriori Dan Fp-Growth Pada Wan Caffeine Addict Yogyakarta," *JISI J. Integr. Sist. Ind.*, vol. 11, no. 1, pp. 121–134, 2024, doi: 10.24853/jisi.11.1.121-134.

[7] R. Wilda, D. Saripurna, and O. K. Sulaiman, "Implementasi Algoritma Frequent Pattern Growth (FP-Growth) untuk Pola Penjualalan Tiket Travel pada PT Taxi Kita Bersama," *Hello World J. Ilmu Komput.*, vol. 3, no. 3, pp. 138–145, 2025, doi: 10.56211/helloworld.v3i3.588.

[8] D. Alfitra, M. Afdal, M. Fronita, and E. Saputra, "Analisa Keranjang Belanja untuk Menentukan Tata Letak Barang Menggunakan Algoritma FP-Growth," *J. Sist. Inf.*, vol. 13, no. 4, pp. 1651–1661, 2024, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[9] Eko Budianto and Muhammad Iqbal, "Model Predictive Analysis of Performance in Training and Course Institutions Using Naive Bayes and K-Means Clustering," *J. Comput. Sci. Res.*, vol. 3, no. 1, pp. 10–16, 2025, doi: 10.65126/jocosir.v3i1.68.

[10] N. D. Ariyanta, A. N. Handayani, J. T. Ardiansah, and K. Arai, "Ensemble learning approaches for predicting heart failure outcomes: A comparative analysis of feedforward neural networks, random forest, and XGBoost," *Appl. Eng. Technol.*, vol. 3, no. 3, pp. 173–184, 2024, doi: 10.31763/aet.v3i3.1750.

[11] A. Ernawati, Z. Sitorus, M. Iqbal, and D. Nasution, "Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan K-Nearest Neighbors," *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 23–35, 2025, doi: 10.47065/bit.v5i2.1783.

[12] M. S. Novelan, S. Efendi, P. Sihombing, and H. Mawengkang, "Vehicle Routing Problem Optimization With Machine Learning in Imbalanced Classification Vehicle Route Data," *Eastern-European J. Enterp. Technol.*, vol. 5, no. 3(125), pp. 49–56, 2023, doi: 10.15587/1729-4061.2023.288280.

[13] J. Anggara, F. R. Ramadhan, A. F. Syofian, A. A. Panjaitan, and R. F. Fithra, "Pengembangan Sistem Prediksi Harga dan Rekomendasi Mobil Bekas Berbasis Machine Learning," *J. Technol. Informatics*, vol. 7, no. 1, pp. 11–23, 2025, doi: 10.37802/joti.v7i1.987.

[14] S. Darma, A. Jihad, A. Fayed, S. M. P. Pardede, and M. H. Aqsha, "Predictive Analysis of Flood Risk Factors Based on a Machine Learning Approach : Comparative Study of SVM and XGBoost Algorithms," vol. 3, no. 1, pp. 24–33, 2026.

[15] N. Alfarizi, P. Lydia, M. S. Novelan, A. Putra, and S. Sinurat, "Comparative Machine Learning Analysis for Sentiment Classification of Sumatra Disaster 2025," vol. 3, no. 1, pp. 68–76, 2026.