CLASSIFICATION APPROACH TO PREDICT CUSTOMER DECISION BETWEEN PRODUCT BRANDS BASED ON CUSTOMER PROFILE AND TRANSACTION

LAURA LAHINDAH[1], IVAN DIRYANA SUDIRMAN [2]

[1]Program Studi Manajemen, Sekolah Tinggi Ilmu Ekonomi Harapan Bangsa, Bandung, Indonesia
[2]Entrepreneurship Department, BINUS Business School Undergraduate Program,
Bina Nusantara University, Bandung Campus, Bandung, Indonesia, 40181
E-mail: [1]laura@ithb.ac.id, [2]ivan.diryana@binus.ac.id,

ABSTRACT

Businesses need to be able to anticipate what products their customers will buy so that they can better respond to changing market demands and consumer tastes. The purpose of this study is to employ several machine learning models that can reliably estimate the customer's likelihood of purchasing the product given a customer's profile, transaction date, and other transaction information. This was achieved by training and evaluating different machine learning techniques, such as naive bayes, linear models, deep learning, and decision trees, on a dataset consisting of actual transaction data from three months of sales at a medium-scale grocery store in Bandung. Results indicated that naive bayes performed best as a prediction algorithm, this study shows that data mining can be used to predict grocery store datasets. This research provides insights into how machine learning can be used to improve businesses' ability to anticipate consumer behavior and respond to changing market demands. We also found that demographic factors like age and location, as well as contextual factors like time of week, significantly influenced customers' propensity to buy.

Keywords: *Data Mining*, *Machine Learning, Classification, Naïve Bayes, Customer Product Decision*

1. INTRODUCTION

The ability to anticipate and meet the needs and preferences of customers depends on businesses being able to guess which product a customer will choose to buy. Having more reliable insight into the future can help organizations make better decisions about product offerings and marketing strategies, which in turn can boost customer satisfaction and retention. It is crucial in retail to analyze and anticipate customer behavior because it can significantly affect a company's ability to turn a profit and bring in money.

Successful retail operations rely heavily on anticipating customers' needs and preferences. Businesses can better match consumer demand if they know in advance which things will be popular and so sell well. This can have far-reaching effects on a business' bottom line by increasing client happiness and loyalty.

Keeping up with ever-shifting retail market trends and customer preferences necessitates data-driven decision making. Businesses can't make educated decisions regarding product offers and marketing tactics without first understanding and accurately forecasting customer behavior. Missed chances, dissatisfied customers, and poorer revenues could result from failing to do so.

This study's utilize three months' of real transactions data from a medium-sized Bandung grocery store from August 2022 to October 2022. Then train a machine learning model that can accurately predict the customer decision given the customer's profile, the date of the transaction, and other information regarding the customer's previous transactions. We will focus on classification algorithms, a subset of machine learning algorithms, because of their ability to predict a categorical result, such as which product a buyer would buy based on a set of options.

This study collects a dataset consisting of customer profiles, transaction dates, and information about customers' transactions for a significant number of consumers in order to accomplish this objective. Following that, we will make use of this dataset to train and test several machine learning models. These models will include a variety of classification methods. In this study, the performance of these models is tested using a variety of metrics, such as accuracy, precision, and recall, and then compare the findings to decide which strategy is the most effective when it comes to forecasting the product that a customer would choose.

The results of this study may be of use to businesses and researchers in understanding and forecasting customer behavior and preferences, as well as in developing more effective product and marketing strategies. This study has the potential to inform business decisions and enhance the shopping experience for customers at a medium-sized grocery store by providing data on the types of products purchased.

To achieve our objective, we will address the following research questions: Which classification algorithms are most effective for predicting customer choice of product based on customer profile, and transaction date? How does the model's performance? What are the most important features in the data for predicting customer choice of product?

2. LITERATURE REVIEW

Data mining is an interdisciplinary field of computer science and statistics that seeks to uncover patterns within a data repository. Data mining's primary goal is to extract valuable insights from massive data sets and present them in an understandable format for future application. [1]. There are at least six distinct types of activities that can be done during the data mining process. Anomaly detection, association rule, cluster analysis, classification, regression, and prediction are just some of the areas explored. The process of classification is central to data mining and has numerous practical applications. Machine learning technique known as classification is used in data mining to make inferences about the group membership of data instances.[2]. An important goal of any classification algorithm is to maximize the model's predicted accuracy. In this sense, the classification task can be seen as an illustration of a supervised method, in which each instance is guaranteed to fall into one of several predetermined classes [3]. There is a wide variety of machine learning methods that may be utilized for categorization problems, some of which include the following[4]–[7]:

1. Logistic regression is a linear model used to predict a binary outcome, such as whether or not a consumer will buy a product. The practice of using logistic regression has been around for quite some time. It has many applications and is used extensively in fields as diverse as medicine, advertising, and finance.
2. Decision trees are one type of tree-based model that can be put to use in classifying tasks. In order to make predictions, they segment the feature space into regions and then use a tree-like structure to take into account the feature values as inputs.
3. SVMs, or support vector machines, are a type of linear model that can be used for classification. Machines that use support vectors are also known by this name. These tools find the hyperplane in the feature space that most effectively divides the various classes.
4. Neural networks are a specific category of model that takes their cues from the physical make-up and operational logic of the human brain. They may be put to work for a broad variety of classification tasks, but shine when applied to issues with intricate, non-linear interactions.
5. Naive Bayes. The Naive Bayes model is a probabilistic framework predicated on the assumption that no two features are interdependent. Used frequently for classification tasks, especially text categorization problems.
6. K-nearest Neighbors (KNN): K-nearest neighbors is a simple model that can be used for classification tasks without the need for any parameters. To make a prediction, it finds the K neighbors in the feature space that are most similar to the new example and then extrapolates to the class to which the majority of those neighbors belong.

Finding patterns within large databases is the goal of data mining, a field at the intersection of computer science and statistics. Data mining's goal is to glean useful insights from data and present them in an easily consumable format. Data mining uses a wide variety of methods, such as pattern recognition, cluster analysis, classification, regression, summarization, and anomaly detection. Classification is a common data mining technique that uses machine learning methods to make predictions about the membership of groups to which data instances belong. Many different types of machine learning techniques, such as logistic regression, decision trees, support vector machines, neural networks, Naive Bayes, and K-nearest neighbors, can be applied to classification problems. Medical care, advertising, and finance are just some of the many industries that make use of these algorithms.

Customer choice modeling is the process of attempting to construct a model that can accurately forecast which product (or brand or service) a person will buy based on a number of characteristics of both the customer and the product, as well as the circumstance in which the purchase will take place[8]. Feldman, Zhang, Liu and Zhang [9] examine two methods for choosing the best products to show shoppers on Alibaba's two online marketplaces, Tmall and Taobao. They randomly assigned 10,421,649 client visits over a week to one of the two ways and measured revenue per customer visit. When both approaches were given access to the identical collection of the 25 most relevant attributes, the MNL-based strategy yielded 5.17 renminbi (RMB) each client visit, while the machine-learning approach generated 4.04 RMB.

Lhéritier, Bocamazo, Delahaye, Agost [10] discusses the problem of predicting customer choice of flight itineraries in the travel industry. In the past, the problem has been addressed using Multinomial Logit (MNL) models. The authors propose an alternative modeling approach based on non-parametric machine learning (ML). They test the models on a dataset of flight searches and bookings in Europe and find that the ML approach outperforms the MNL models in terms of accuracy and computation time, with less modeling effort required.

From the point of view of Machine Learning, a panel data set containing customer preferences could be analyzed using the mentioned techniques from matrix completion. Similar consumer purchase patterns within a product and between products would be used by the model to draw conclusions about individuals and their preferences. These models are useful for comparing one thing to another, but they are not always suited to examining counterfactuals or the degree to which two things are equivalent.

In the case of online shopping with a wide selection of items, for instance, Jacobs et al.[11] suggest utilizing a similar latent factorization technique to flexibly describe customer heterogeneity. Instead of modeling customer reactions to price changes or substituting between comparable items, Jacobs et al measure performance by predicting which new things a customer would purchase.

Customer decision modeling is an approach to understanding and predicting which products or services a customer will choose to purchase based on a variety of factors, such as the individual's characteristics, the characteristics of the products or services themselves, and the circumstances surrounding the purchase. Several methods, including multinomial logit (MNL) models and machine learning strategies, can be used to attempt to predict what a client will want. The MNL modeling strategy is based on statistical analysis, and it requires the formulation of hypotheses about the relationships between the various factors that influence consumers' choices. Machine learning-based methods, on the other hand, involve training a model with a large dataset of consumer decisions and giving the model permission to learn the relationships between the factors that influence those decisions.

Machine learning strategies have been demonstrated to outperform MNL models in terms of accuracy and efficiency when it comes to predicting customers' preferences. Machine learning-based methods, for instance, have been shown to outperform MNL models in predicting travelers' preferences. Other methods, such as latent factorization, necessitate making predictions about the new products that consumers would buy and characterizing the diversity of the client base. These methods function best when a wide variety of products are available for purchase from a single online store.

3. METHOD

Research design will involve a quantitative approach utilizing machine learning techniques. In this researech CRISP-DM process is used to achieve the objective. CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely-used framework for data mining and data analysis projects. It consists of six steps:

1. Business understanding: In this step, the goals and objectives of the data mining project are defined, and the scope of the project is determined. This includes identifying the business problem that the project aims to solve, and the data that will be used to solve it.
2. Data understanding: In this step, the data is explored and analyzed to understand its characteristics and quality. This includes gathering and describing the data, checking for missing or invalid values, and identifying any potential problems or issues with the data.
3. Data preparation: In this step, the data is cleaned and transformed into a format that is suitable for analysis. This includes selecting the relevant

data, dealing with missing or invalid values, and formatting the data in a way that is suitable for the analysis methods to be used.

4. Modeling: In this step, statistical or machine learning models are created and evaluated to identify the most appropriate model for the data. This includes selecting the appropriate modeling techniques, building and testing the models, and comparing the performance of different models.
5. Evaluation: In this step, the results of the modeling process are evaluated to determine whether the project has achieved its goals and objectives. This includes comparing the performance of the model to other models and to the business objectives, and identifying any areas for improvement.
6. Deployment: In this step, the final model is deployed and put into use. This includes deciding how the model will be used, how the results will be communicated, and how the model will be maintained over time.

CRISP-DM provides a structured approach to data mining and data analysis projects, and helps to ensure that the project is focused on achieving its business objectives.

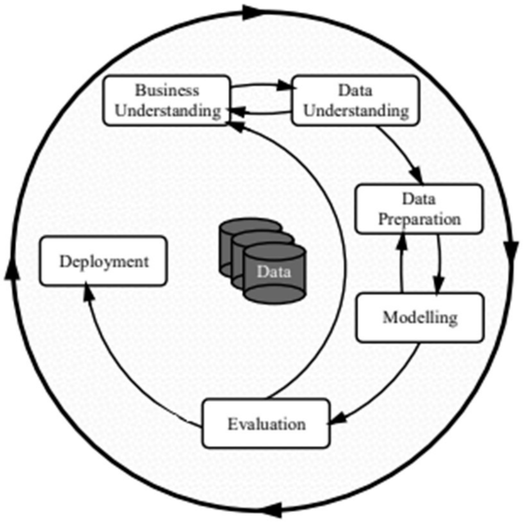

Figure 1: CRISP-DM Process Model for Data Mining

By using the CRISP DM method, in this study the process is as follows:

3.1 Business Understanding

The data for this study came from a medium-sized grocery store in Bandung Regency. There are numerous boarding students near the grocery shop because the area is home to several well-known state institutions in Indonesia. As a result, the people surrounding it is highly diversified, hailing from many parts of Indonesia. The store is aimed towards intermediate to lower income people who are well suited to the neighborhood.

As a result, supermarket sales transactions are relatively high. The Grocery also has a loyalty program, and members have the option to receive various rewards. Every time a member-turned-consumer makes a purchase, his member number is recorded. The consumer profile that was previously captured while registering as a member is then referenced by this member number.

However, because this study involves predicting consumer preferences, another business that will be explored is the product offered by this grocery. In this study, two goods in the same category with many transactions from different brands were chosen. The two goods investigated were instant noodles, specifically Indomie instant noodles with the Mie Goreng variant and Sedaap Mie instant noodles with the Mie Goreng variant.

Indomie products are well known in Indonesia and have been there for a long time, but Sedaap noodle products are not well known because they have only been on the market for a short period.

According to GoodStats [12] from the Indonesia Top Brand Index 2022, the first instant noodle brand is Indomie with 72.9%. Mie Sedaap the second for instant noodle with 15.5%. As a result, Mie Sedaap will be the focus of this classification. Because it will be more intriguing to discover what aspects will assist this brand in competing with more dominating brands.

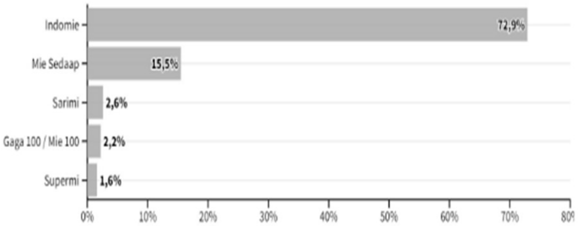

Figure 2: Instant Noodle Brand Favorite

3.2 Data understanding

The data obtained initially consisted of two tables. The first table was a transaction table with properties such as transaction date, business day, store, product name, buy amount, net sell, and member number. The second table contains attributes in the form of consumer data for participants of the loyalty program. This table includes the attributes age, gender, date of birth, province, city, and district. When viewed from the city, consumer data is highly diverse because many members come from outside Bandung. To get a better undestanding of the data, both of the tables were merge together. Member number attribute act as the primary key.

3.3 Data Preparation

The following step is to prepare the data by cleaning it. First, purchasing information for Indomie Goreng and Sedaap Mie Goreng is chosen. The Province, City, and District data then have 0 values. The 0 value in the column is deemed a missing value because it has no meaning. Missing value records are then removed. The age attribute information shows that the minimum value is 0 and the highest value is 2020, which is not possible. This attribute is then cleaned up so that it only shows ages ranging from 18 to 75 years old. After this process, the dataset consists of 2,836 examples with 0 special attributes and 7 regular attributes. Here are the statistics of the dataset after cleaning.

Table 1: Dataset Statistic

Name	Missing	Least	Most
Brand	0	Sedaap	Indomie
Date	0	10/09/22	02/10/22
Gender	0	Male	Female
Age	0	18	75
Province	0	Temanggung	Jawa Barat
City	0	Temanggung	Sumedang
District	0	Weru	Jatinagor

As seen in the statistical table of the dataset used, there is no missing value. Most of the location data are from the province of West Java, the city of Sumedang, and the district of Jati Nanor. This is reasonable because the store is in that area. However, it is also seen that there are consumers from outside the area, although the number is not many.

The number of samples for SedaapMie was only 237, compare to Indomie which has 2,000 samples. This can cause an imbalance classification. Thus in order balance between those classes, we chose to undersampling the majority class, which is Indomie. As a result, the total dataset examples used in this research is 474.

3.4 Modeling

In this research, the models of Naive Bayes, Generalized Linear Models, Deep Learning, and Decision Trees will be studied and contrasted. To determine which model is the most effective at making predictions. For this process we use RapidMiner.

RapidMiner is a suite of software that can be used to get the data ready for analysis, train the machine learning models, and then deploy the models. Designed for easily use by the users, it eliminates the need for coding in order to build, test, and deploy data-driven models. Data pipelines and machine learning models can be built with a visual interface in RapidMiner. It includes a wide variety of algorithms for tasks like classification, regression, clustering, and anomaly detection. Included are techniques for cleaning and transforming data, creating and evaluating features, visualizing results, and assessing the quality of models. RapidMiner is utilized by the financial, medical, advertising, and retail sectors.

3.5 Evaluation and Deployment

We will examine various findings, including accuracy as well as classification error. This section will be discussed further in the results section. After selecting a model based on the various parameters that have been discussed, the prediction results on the selected machine learning will be explored. Futher more the factors that most influence predictions will be studied further for business purposes.

4. RESULT AND DISCUSSION

Classification error measures how well a model can predict the classes in a given dataset. It is calculated by splitting the number of wrong predictions in half. When assessing the quality of a classification model, the classification error is a common metric to employ.

Table 2: Classification Error Result

Model	Classification Error	Standard Deviation
Naïve Bayes	38.3%	±3.8%
Linear Model	39.7%	±3.5%
Deep Learning	41.9%	±8.0%
Decision Tree	50.0%	±1.9%

When applied to a particular dataset, a classification error of 38% for Naive Bayes implies that the algorithm is generating inaccurate predictions for 38% of the cases contained within the dataset. This may be viewed as good prediction in this context. In general, a classification error that is smaller implies a model that is more accurate, while a classification error that is larger indicates a model that is less accurate.

The degree to which a model correctly predicts the labels assigned to each class in a given dataset is referred to as its accuracy. Accuracy in classification models can be calculated by dividing the proportion of correct predictions by the total number of predictions.

Table 3: Accuracy Result

Model	Accuracy	Standard Deviation
Naïve Bayes	61.7%	±3.8%
Linear Model	60.3%	±3.5%
Deep Learning	58.1%	±8.0%
Decision Tree	50.0.%	±1.9%

As can be seen in the table comparing the various models' levels of accuracy, Naive Bayes has the highest level of accuracy. Naive Bayes accuracy of 61.7% indicates that this model has the best accuracy for this dataset.

True-positive-prediction rate (TPR) and false-positive-prediction rate (FPR) are two measures of prediction accuracy. Area under the curve is calculated by plotting the true and false positive rates (TPR and FPR) against the cutoffs used for classification (AUC). AUC is calculated as the area under the TPR-FPR curve. AUC can range from 0 to 1, where 0.5 indicates an untrained or random model and 1 indicates a perfect model. Given its insensitivity to changes in the threshold used for classification, AUC is widely used to evaluate classification models.

Table 4: AUC Result

Model	AUC	Standard Deviation
Naïve Bayes	0.66	±0.04
Linear Model	0.672	±0.037
Deep Learning	0.678	±0.087
Decision Tree	0.567	±0.033

As seen in the table of AUC results, Deep Learning is the model with the highest AUC, which is 0.678.

A classification model's precision is the percentage of correct predictions relative to the total number of correct predictions it makes. If the model's precision is high, it is less likely to produce false positive results, while if it is low, it is more likely to do so.

Table 5: Precision Result

Model	Precision	Standard Deviation
Naïve Bayes	60.3%	±4.0%
Linear Model	60.2%	±4.4%
Deep Learning	58.9%	±7.8%
Decision Tree	50.0%	±1.9%

For precision, Naive Bayes has the highest value of 60.3%. In most cases suggests a more accurate model since it is making fewer predictions that turn out to be incorrectly positive.

Recall is a measure used to assess a classification model's efficacy; it is the proportion of correct positive predictions to the total number of positive examples in the training data.

Table 6: Recall Result

Model	Recall	Standard Deviation
Naïve Bayes	65.5%	±5.4%
Linear Model	60.1%	±5.7%
Deep Learning	59.6%	±10.3%
Decision Tree	100%	±0.0%

Decision Tree has the highest Recall, it is an indication that it can accurately anticipate a higher proportion of positive events.

Specificity is a measure of a classification model's accuracy, specifically the ratio of true negative predictions to the total number of negative cases in the dataset. One possible counter-example to "sensitivity" is "specificity.".

Table 7: Specificity Result

Model	Specificity	Standard Deviation
Naïve Bayes	57.8%	±7.4%
Linear Model	60.2%	±4.6%
Deep Learning	56.8%	±11.6%
Decision Tree	0.0	±0.0%

Linear Model has the highest Specificity at 60.2%. Higher specificity suggests the model is generating fewer incorrect positive predictions, whereas lower specificity indicates a lesser proportion of negative events are accurately predicted.

A mapping of the findings is created depending on the model employed and the results of the measurements obtained in order to gain an overall view of the results of the comparison of the models analyzed.

Table 8: The Resul Mapping

	NB	LM	DL	DT
CE	*			
Acc	*			
AUC			*	
Pre	*			
Rec				*
Spec		*		

Explanation for the table 7 is as follows. NB is for Naïve Bayes, LM is for Linear Model, DL is for Deep Learning, and DT is for Decision Tree. CE is for Classification error, Acc is for Accuracy, AUC is Area Under Curve, Pre is for Precision, Rec is for Recall, and Spec is for Specificity.

According to the mapping results in table 7, Naive Bayes excels in three measurements: Classification Error, Accuracy, and Precision. All three of these metrics are frequently used to compare machine learning models.

To evaluate the efficacy of a classification scheme, researchers create a table called a confusion matrix. It is used to summarize the model's predictions for a given dataset in terms of the proportion of correct, incorrect, or null predictions.

Table 9: Confusion Matrix

	True Indomie Goreng	True Sedaap Mie Grg	Class precision
Pred Indomie Goreng	40	23	63.49%
Pred Sedaap Mie Grg	29	44	60.27%
Class Recall	57.97%	65.67%	

This classification study focuses on the Sedaap Mie Goreng. From the table above, the class recall for Sedaa Mie is 65.67% and class precision is 60.27%.

One common application of Naive Bayes algorithms, a subset of machine learning techniques, is classification. It is based on the idea of using Bayes' theorem to estimate the likelihood that an instance belongs to a certain class, with the estimate being derived from the probabilities of the characteristics of the instance.

Weights in a machine learning model are a shorthand way of referring to the relative significance of each feature or attribute in generating accurate predictions. The weights used in this analysis were learned as part of the training process.

Table 10: Naïve Bayes Weights Result

Attribute	Weight
City	0.145
Date:day_of_week = 6	0.095
District	0.093
Date:day_of_week = 3	0.069
Date:month_of_quarter = 2	0.059
Age	0.048
Date:month_of_quarter = 1	0.042
Date:day_of_week = 7	0.038
Gender	0.037
Date:month_of_quarter = 3	0.034
Date:day_of_week = 2	0.025
Date:day_of_week = 5	0.024
Date:day_of_week = 4	0.024
days_diff(Date, Today)	0.021
Date:day_of_week = 1	0.015
Date:day_of_month	0.013

According to the weighted findings in table 8, the ten most important factors are city, day 6, district, day 3, month quarter 2, age, quarter 1, day 7, and gender. These attributes may be the most relevant or influential in the model's predictions, and they may have a substantial influence on the model's accuracy.

5. CONCLUSSION

In this study, we use machine learning classification algorithms to the problem of predicting what products shoppers will select from a mid-sized grocery store. Similar methods in other fields have also been investigated in the existing literature. Example: Feldman et al. [9] investigated two approaches for product recommendations on

Alibaba's marketplaces and found that the machine-learning approach resulted in lower revenue per customer visit than the multinomial logit-based strategy. However, it should be noted that the study was conducted on online marketplaces and may not be immediately applicable to your grocery store scenario.

However, Lhéritier et al.[10] took a non-parametric machine learning method to the same issue of forecasting customers' flight itinerary preferences in the tourism sector. Machine learning was found to be more accurate and take less time to calculate than conventional multinomial logit models in their research. This result bolsters the application of machine learning algorithms to forecasting consumer preferences.

One of the main goals of this research was to develop a machine learning model that could reliably classify data and make predictions. To achieve this goal, first a dataset is collected consisting of features and labels, which will be used to trained and tested on several machine learning algorithms. Algorithms like naive bayes, linear models, deep learning, and decision trees were among those used. The effectiveness of these algorithms was evaluated and compared in this study using a wide range of metrics, including classification error, accuracy, precision, recall, AUC, and specificity.

The results of this study indicated that the naive bayes algorithm was the most effective prediction approach, ranking first in terms of both accuracy and precision. Evidence presented here suggests the naive bayes approach is a promising strategy for this classification problem, and it may be a feasible option for future applications.

The naïve bayes approach's effectiveness in this investigation could be attributable to a number of factors. It's possible that this is because the data was especially well-suited for this method. Naive bayes is well-known for its efficacy on datasets with few features and a moderate number of examples. Both of these were true of this data collection. Because naive bayes is a fast and easy-to-implement algorithm that can be run with a minimum of data and computational resources, this may be a contributing factor.

Overall, the results of our research suggest that naive bayes is a good method for predicting a category outcome based on a set of data, and they emphasize the potential of this algorithm for a wide variety of classification tasks.

This research found that, of the various prediction algorithms tested, the naive bayes approach gave the most consistent outcomes. At its highest, its accuracy was up to 61.7%, significantly higher than that of any of the competing algorithms. It's worth noting that the naive bayes model performed quite well in terms of precision, scoring 60.3%. This indicates that the model made few mistakes in its positive predictions. Since the dataset includes information about customer behavior, which is notoriously tricky to predict. As an addition, the risk from miss prediction is low. So, it seems reasonable to conclude that the results produced by the measuring model are satisfactory.

The study also found that buyers were more likely to purchase Indomie Goreng if they were older, while buyers were more likely to purchase Sedaap Mie Goreng if they were younger. This could be due to the fact that Sedaap Mie is still a new brand while Indomie has been available for quite some time. This data also demonstrates the viability of targeting youth audiences with marketing efforts.

This research also found that the customer's home city played a significant role in determining which products they ultimately purchased. This may be better revealed by additional research with larger datasets. The study also found that the day of the week had a significant impact on consumers' buying habits. The likelihood that a young person will purchase Mie Sedaap during the weekday is higher, for example. Potential applications include using the data to plan sales and marketing initiatives and forecast sales.

Further investigation into the causes of these tendencies will help businesses better understand their target markets and refine their advertising strategies. Finally, the paper notes that under sampling was utilized to achieve statistical parity, thus the sample size is smaller than usual. Further research using larger datasets may yield more convincing results and increase the generalizability of the findings.

The limitation of this study is that the data used is small because it uses undersampling when making the sample balanced. More data will give better results

REFERENCES

[1] S. Agarwal, “Data Mining: Data Mining Concepts and Techniques,” in *2013 International Conference on Machine Intelligence and Research Advancement*, Dec. 2013, pp. 203–207. doi: 10.1109/ICMIRA.2013.45.

[2] G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Jul. 2013, pp. 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[3] N. Putta, “A STUDY OF SOME DATA MINING CLASSIFICATION TECHNIQUES,” 2018. Accessed: Dec. 25, 2022. [Online]. Available: https://www.semanticscholar.org/paper/A-STUDY-OF-SOME-DATA-MINING-CLASSIFICATION-Putta/484bae65c48ec28af00776d23e62ab6d2e10ccb1

[4] T. N. Phyu, “Survey of Classification Techniques in Data Mining,” *Hong Kong*, 2009.

[5] S. Gupta, D. Kumar, and A. Sharma, “Performance Analysis Of Various Data Mining Classification Techniques On Healthcare Data,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, Aug. 2011, doi: 10.5121/ijcsit.2011.3413.

[6] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic Kidney Disease analysis using data mining classification techniques,” in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 300–305. doi: 10.1109/CONFLUENCE.2016.7508132.

[7] V. Rajeswari and Dr. P. K. A. Arunesh, “Analysing Soil Data using Data Mining Classification Techniques,” *Indian J. Sci. Technol.*, vol. 9, May 2016, doi: 10.17485/ijst/2016/v9i19/93873.

[8] M. van Wezel and R. Potharst, “Improved customer choice predictions using ensemble methods,” *Eur. J. Oper. Res.*, vol. 181, no. 1, pp. 436–452, Aug. 2007, doi: 10.1016/j.ejor.2006.05.029.

[9] J. Feldman, D. J. Zhang, X. Liu, and N. Zhang, “Customer Choice Models vs. Machine Learning: Finding Optimal Product Displays on Alibaba,” *Oper. Res.*, vol. 70, no. 1, pp. 309–328, Jan. 2022, doi: 10.1287/opre.2021.2158.

[10] A. Lhéritier, M. Bocamazo, T. Delahaye, and R. Acuna-Agost, “Airline itinerary choice modeling using machine learning,” *J. Choice Model.*, vol. 31, pp. 198–209, Jun. 2019, doi: 10.1016/j.jocm.2018.02.002.

[11] B. Jacobs, B. Donkers, and D. Fok, “Model-based Purchase Predictions for Large Assortments.” Rochester, NY, Feb. 18, 2016. doi: 10.2139/ssrn.2443455.

[12] GoodStats, “5 Merek Mi Instan Pilihan Masyarakat Indonesia 2022,” *GoodStats*. https://goodstats.id/article/5-merek-mi-instan-pilihan-masyarakat-indonesia-2022-HeS3T (accessed Dec. 25, 2022).