Optimal model choice using AIC Method and Naive Bayes Classification

Dam Sao Mai¹, Phan Hong Hai², Bui Huy Khoi³*

Industrial University of Ho Chi Minh City, 700000, Vietnam

*buihuykhoi@iuh.edu.vn

Abstract. The paper purposes to evaluate the optimal model choice of trademark, nutrition, price, packaging design, product flavor, utility, convenience, promotion, and safety of the purchasing decision in Vietnam. The method was implemented through quantitative research. Data was collected Vietnam consumers who are living in Ho Chi Minh City, usually eat products from the soursop. We used R 4.0.2 for analyzing statistics. The research model was suggested from the studies of the product purchasing decisions. We used R 4.0.2 for analyzing statistics. The results of this study indicate that the product flavor, safety, nutrition, and convenience had an important and positive effect on purchasing decisions the products from the soursop is 37.45%. Trademark, price, packaging design, utility, and promotion are not related to the purchasing decision. Implications were giving the necessity to develop the soursop trees to boost the economy in Soc Trang, Vietnam. Previous studies revealed that using linear regression. This study uses the optimal choice by the AIC method.

1. Introduction
Vietnam is recognized as one of Asia's major agricultural countries, resulting in a strong fruit industry in the region. Moreover, as their revenue has risen dramatically, the fresh fruit industry is one of the first targets in consumer demand [1]. Every customer in the world, in the age of globalization, they have an opportunity and a chance to enjoy domestic goods. When choosing the necessary items to meet their needs, the majority of customers would like and aspire, to give the first choice over domestic goods [2]. An indispensable theme in all countries is economic integration. Nevertheless, there are, in addition to the benefits that global integration brings. There are also drawbacks, such as rivalry between firms.

Vietnam is a developing country with many benefits [3], especially in the fruit-producing sector. Soursop has the scientific name Annona muricata; they have a fast growth rate. They adapt to the tropical climate, are less resistant to drought and cold than soursop. This plant prefers heat so the tree commonly grown in the South is often used as a smoothed and effective treatment. Agricultural products include, in particular, vegetables, berries, plants, essence, milk, eggs, sea products, and other crucial production products. Agrarian products have a direct effect on our social health, so we should pay more attention to the agricultural products industry. Choice in Purchasing Decision of Products of agrarian products may boost the earnings of manufacture and eating of farming products [4].

Worldwide, Soursop consists of about 130 genera and 2300 species [5]. They are native to the tropics of South and North America. Currently, custard-apple has been widely distributed in tropical
and subtropical regions worldwide, such as India, Malaysia, and Nigeria [6]. In Vietnam, Soursop is
grown mainly in the Southern and Central regions. Soursop trees live in areas with high humidity,
about 1800 mm rainfall, and are less resistant to cold. A suitable pH range is 5.0 - 6.5 [7]. In salty or
salty soils and under the influence of tides like in some areas of our country, people plant Soursop
grafted on the base of the bowl.

The Soursop tree is a fruit tree with a height of 5-8 m, thick, dark leaves, shiny, and green all year
round. Soursop fruit is green, heart-shaped, average 15-20cm in diameter, weighing up to 4kg, the
pulp consists of the white part surrounding the long core [8]. Soursop seeds are black, slippery,
depending on the growing area that the seeds can range from 5 to 200 seeds/fruit [7]. Because the
thorns custard-apple is difficult to preserve for long, it is usually harvested when it is ripe enough, the
fruit is gradually turning from green to yellow, sweet, slightly sour, characteristic aroma.

Soc Trang is an agricultural province in Vietnam, the economic structure and the production
structure in each industry, each region has moved slowly, the production scale is small, scattered, and
the efficiency of production and business has improved. still low, production costs are still high. The
international and regional economic integration of enterprises in the province is still quite far from the
requirements, in the condition of increasing competitive pressure in the world and the region.
The big challenge for Soc Trang in the country's general integration process is the issue of resources;
first of all, human resources for the implementation of industrialization and modernization; the gap in
development levels among regions, especially in remote areas, is populated with Khmer ethnic
minorities.

Along with the difficulties of an agricultural province, with a low starting point, the infrastructure,
although initially improved, has not yet met the requirements of socio-economic development, and has
to regularly deal with the problems, benefits of weather and natural disasters. The rate of poor
households is still high. From the high economic efficiency of the soursop tree on drought and salty
land, many farmers in Soc Trang, Vietnam have developed the garden area of soursop.
In this research, we have designed an empirical study in the context of purchasing decision the
products from the soursop in Vietnam to examine its factors to develop farming the soursop in Soc
Trang Province, Vietnam.

Naive Bayes (NB) is a Bayes-oriented learning method that is very useful for learning involving
high-dimensional data[1]–[3], such as text classification[1], [4]–[6], Searching[7]–[10] and web
mining[11]. In general, the Bayesian classification method has a conditional dependency between
random variables. This problem is often called the independent assumption of attribute which assumes
all independent attributes so that the effect is time consuming because it examines the relationship
between all random variables in which this task is a combinatorial optimization task[1]. Alternatively,
the NB relaxes the restriction of the dependency structure between attributes by simply assuming that
attributes are independent by class labeling. Consequently, examining relationships between attributes
is no longer necessary and derivation of NB methods can be linearly measured by training data.

Many data sets are used to solve the problem of NB attribute independence attributes, different
methods and frameworks[12]–[14], so a through overview of the assumptions of independent attribute
solutions is necessary. This literature review aims to identify and analyze the research trends, data sets,
methods and frameworks used in the study of attribute independent assumptions on NB between 2010
and 2017.

2. Literature Review
Prediction of behavior based on the behavior of decisions and using the expectations. One of the main
concerns is outcomes to optimize one's gains [9]. Purchaser food selection is the outcome of the
quality outlooks before and quality experience after the buying [10]. A significant number of studies
have shown that when purchasing, buyers frequently choose a product with a trademark (X1) that
assurances a convincing quality. They also show that the nutrition aspect (X2) is an important factor
affecting the procurement of food [11].
The first, the division into dimensions of search, experience, and credence is the utility (X6) of quality dimensions for food. Before the actual purchase of the product, search attributes, such as color, price (X3), and size, are measured and experience points are those that customers interpret after the buying and use of the product [10]. Second, packaging design is worth the measure in which it helps to improve the long-term loyalty of the customer associated with the trademark owner of the company-producer and to build a positive image [11].

Third, product flavor (X5) is the sensory impression perceived by taste and smell buds and is a leading factor in the consumer's buying decision and product consistency determination. [12]. Fourth, convenience (X7) is sometimes mentioned as a motivating element, and the choice of goods has also been established. White and Manning [13] have taken convenience into account in the relation of time, space, and effort. Several separate sub-categories for price and goods are also listed. Convenience-related problems were most commonly mentioned as related to the decision to buy.

Fifth, with Promotion (X8), buyers are supposed to want to try these goods and inspire current customers to buy products more frequently so that a company's products can be repurchased and sales volume will increase. Promotional campaigns can encourage customers to be interested in purchasing goods and to be able to have a stable response for buyers, dramatize product deals, and raise sales in the short term. If the promotion is performed continuously and improved, the promotion is supposed to influence buying decisions; it will encourage the purchase of goods and increase the number of products purchased by customers. Companies may attract new customers through discounts, inspire customers to purchase more and influence their clients to try new products [14].

Lastly, food safety (X9) is a big concern for customers, the food industry, and the government. Their view of food safety is, in part, a matter of trust in the food sequence, considering that consumers are not readily able to recognize food safety threats themselves. The goal of the study is to focus on livestock farmers and to investigate the causal link between the factors that determine consumer trust, food safety, and, in turn, their likelihood of purchase [15].

Over time, some authors in Vietnam showed that the product (X5), price (X3), place, and promotion (X8) - have demonstrated their superior impact in speeding up sales [1], growing business, and competing well in the marketplace [16]. This demonstrates that companies are increasingly willing to invest in marketing initiatives [17]. The positive effect of the product (X5), price (X3), place, and promotion (X8) on the decision to buy from consumers are, therefore, evident [1; 18].

In this paper, we have designed an empirical study in the context of purchasing decision the products from the soursop in Vietnam to examine factors as below function:

\[ Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 + \beta_9X_9 \]

Code: Trademark (X1), Nutrition (X2), Price (X3), Packaging design (X4), Product flavor (X5), Utility (X6), Convenience (X7), Promotion (X8), Safety (X9), Purchasing decision (Y).

Y: the importance levels of the characteristics that lead to the decision to purchase the soursop product (color, flavor, smell, design, and pylorus).

3. Research Approach Method

In this research, questionnaires were used to collect data for the quantitative analysis of the trip. The respondents’ survey responses were the primary method for gathering information. The research was conducted in the year 2020. Participants replied to questions from the survey, which was the primary data collection method. Questions about the status of the determinants that influenced the buying decision were included in the survey. Respondents were selected using conventional methods, and a total of 201 consumers who are living in Ho Chi Minh City, Vietnam. Our study population consisted of 70 (34.8%) men and 131 (65.2%) women. Their ages, income, and jobs are shown in Table 1 and Table 2.

### Table 1. Age Groups and Income
There are seven age groups in our research. The group from 18 to 25 account for the highest proportion. This is the main object of paper. They are studying at the university. The income is divided into ten groups. The dependent income mill VND group accounts for the highest rate.

| Age Groups | Number | Percentage (%) | Income          | Number | Percentage (%) |
|------------|--------|----------------|-----------------|--------|----------------|
| <18        | 3      | 1.5            | Dependent income| 53     | 26.4           |
| >65        | 2      | 1.0            | < 5 mill VND    | 20     | 10.0           |
| 18-25      | 89     | 44.3           | 5-10 mill VND   | 51     | 25.4           |
| 26-35      | 66     | 32.8           | 10-15 mill VND  | 44     | 21.9           |
| 36-45      | 37     | 18.4           | > 15 mill VND   | 33     | 16.4           |
| 46-55      | 3      | 1.5            |                 |        |                |
| 56-65      | 1      | 0.5            |                 |        |                |
| Total      | 201    | 100.0          | Total           | 201    | 100.0          |

Our study is composed of nine Jobs. The Officer accounts for the highest proportion. A Likert-scale survey was used to evaluate the determinants of interest, with estimations ranging from (1) “No care to product” to (4) “Strongly care to product”.

Blinding

All study personnel and participants were blinded for the duration of the study. No one from the outside world had any contact with the study participants.

4. Result

4.1. Akaike Information Criterion Selection

AIC (Akaike’s Information Criteria) was utilized to choose the best model by R software. AIC has been used in the theoretical context for model selection. And when multicollinearity occurs, the AIC approach can handle multiple independent variables. AIC may be applied from a set of one or more independent variables as a regression model, estimating one or more dependent variables. An important and useful measurement for deciding a simple and complete model is the Akaike Information Criterion (AIC). The model will stop when the minimum AIC value in table 3 [19].

| Model                                                                 | AIC   |
|-----------------------------------------------------------------------|-------|
| $Y \sim X_1 \ + \ X_2 \ + \ X_3 \ + \ X_4 \ + \ X_5 \ + \ X_6 \ + \ X_7 \ + \ X_8 \ + \ X_9$ | -464.04 |
| $Y \sim X_1 \ + \ X_2 \ + \ X_3 \ + \ X_5 \ + \ X_6 \ + \ X_7 \ + \ X_8 \ + \ X_9$ | -465.99 |
| $Y \sim X_1 \ + \ X_2 \ + \ X_3 \ + \ X_5 \ + \ X_6 \ + \ X_7 \ + \ X_8 \ + \ X_9$ | -467.86 |
In the above results, R reports show every step of searching the optimal model. The first step is to start with all 09 independent variables with AIC = -464.04. The second step consists of 08 variables with AIC = -465.99. The sixth step is to find a model, R stops with a model of 04 independent variables (X\(_2\), X\(_5\), X\(_7\), X\(_9\)) in table 3 with AIC = -470.19.

| Table 4. Results of Regression Analysis |
|----------------------------------------|
| Y | Beta | SE  | t     | P-value | Decision |
|---|------|-----|-------|---------|----------|
| Cons | 1.11480 | 0.17848 | 6.246 | 0.000000 | Supported |
| X\(_2\) | 0.11602 | 0.03789 | 3.062 | 0.002506 | Supported |
| X\(_5\) | 0.18743 | 0.04401 | 4.259 | 0.000003 | Supported |
| X\(_7\) | 0.10627 | 0.03607 | 2.946 | 0.003605 | Supported |
| X\(_9\) | 0.17974 | 0.04759 | 3.777 | 0.000211 | Supported |

All variables have p-value> 0.05 [20], so independent variables are correlated with independent variables in table 4.

4.2. Variance Inflation Factor

| Table 5. Variance Inflation Factor (VIF) |
|----------------------------------------|
| Indicator | X\(_2\) | X\(_5\) | X\(_7\) | X\(_9\) |
| VIF | 1.274312 | 1.286762 | 1.078414 | 1.440017 |

According to table 5, VIF for the independent variables is smaller than 10 [21], so there is no collinearity between the independent variables.

4.3. Heteroscedasticity

| Table 6. Heteroscedasticity |
|----------------------------|
| Source | chi2 | df | p     |
| White  | 12.623 | 4 | 0.01327 |

White Test shows that P-value = 0.01327 and lower than 0.05 [22], so it can be concluded that there is no heteroscedasticity in table 6.

4.4. Autocorrelation

| Table 7. Autocorrelation Test |
|----------------------------|
| Durbin-Watson | test for autocorrelation |
| 2.0156 | p-value = 0.5363 |

H\(_0\): no serial correlation

Durbin-Watson Test shows that there is no autocorrelation from the model in table 7 [23].

4.5. Model Test
Table 8. Model Test

| Indicator             | Value   |
|-----------------------|---------|
| Multiple R-squared    | 0.387   |
| Adjusted R-squared    | 0.3745  |
| F-statistic           | 30.94   |
| p-value               | 0.0000  |

According to the results from table 6, the product flavor (X5), Safety (X9), Nutrition (X2), and Convenience (X7) are related to the purchasing decision (Y) of the products from the soursop is 37.45% in table 8. The above analysis shows the regression equation below is statistically significant [24].

\[ Y = 1.11480 + 0.11602X_2 + 0.18743X_5 + 0.10627X_7 + 0.17974X_9 \]

5. Discussion

X2: Positive sign (+): positive relationship: When the nutrition (X2) increases by 1 point, the purchasing decision (Y) of the products from the soursop will increase by 0.11602 points.
X5: Positive sign (+): positive relationship: When the product flavor (X5) increases by 1 point, the purchasing decision (Y) of the products from the soursop will increase by 0.18743 point.
X7: Positive sign (+): positive relationship: When the convenience (X7) increases by 1 point, the purchasing decision (Y) of the products from the soursop will increase by 0.10627 points.
X9: Positive sign (+): positive relationship: When the safety (X9) increases by 1 point, the purchasing decision (Y) of the products from the soursop will increase by 0.17974 points.

6. Conclusions

6.1. Conclusion

The goal of this article is to concentrate on two key issues: the theoretical structure of quantitative models and the results of the application for the purchasing decision (Y) to create the consumption market for many products from the soursop. Through the tests, it can be affirmed: The factors influencing purchasing decision (Y) in the order of importance are the product flavor (X5), the safety (X9), the nutrition (X2), and the product flavor (X5) will increase purchasing decision (Y).  

6.2. Implications

Regarding the purpose of using products from the soursop, more than 50% of participants said it was good for health, nearly 20% said it was because of preference. Trademark (X1), Price (X3), Packaging design (X4), Utility (X6), and Promotion (X8) are not the main reasons to lead the purchasing decision (Y) of the products from the soursop. Therefore, it can be seen that the product flavor (β = 0.18743) is an important factor in choosing the products from the soursop of this group of consumers. In addition, safety (β = 0.17974) is the second influential factor in the survey respondents' decisions to buy products. Nutrition (β = 0.11602) is the third important factor. Convenience is the final important factor (β = 0.10627). These four factors play an important role for consumers when choosing to use products from the soursop, but not equal to the five rejected factors mentioned above. This result shows that the need to pay attention to factors of Product flavor (X5) and Safety (X9) when developing products from soursop.

Besides, ensuring Nutrition (X2) and Convenience (X7) will help ensure the product's success in the market. Summarily, the Soc Trang province needs to develop the soursop trees to boost the economy. Soursop is a large, thorny tropical fruit. It has a sweet taste, so it is often used to make ice
creams, candies, and many foods and drinks. Besides its role as a food, the soursop contains many valuable ingredients that have therapeutic effects.

References

[1]. TRAN, T. A., PHAM, N. T., Van PHAM, K., & NGUYEN, L. C. T. (2020). The Roles of Health Consciousness and Service Quality toward Customer Purchase Decision. *The Journal of Asian Finance, Economics and Business (JAFEB)*, 7(8), 345-351.

[2]. POTLURI, R. M., & JOHNSON, S. (2020). An Exploratory Research on Country-of-Origin and Its Impact on the UAE Consumers Buying Decisions. *The Journal of Asian Finance, Economics, and Business*, 7(9), 455-466.

[3]. NGUYEN, M. T., & KHOA, B. T. (2020). Improving the Competitiveness of Exporting Enterprises: A Case of Kien Giang Province in Vietnam. *The Journal of Asian Finance, Economics, and Business*, 7(6), 495-508.

[4]. Shen, H. (2019). An original design of fresh agricultural products e-commerce platform based on Mobile Internet. Paper presented at the IOP Conference Series: Materials Science and Engineering.

[5]. Mishra, S., Ahmad, S., Kumar, N., & Sharma, B. (2013). Annona muricata (the cancer killer): a review. *Glob J Pharma Res*, 2(1), 1613-1618.

[6]. Adewole, S. O., & Caxton-Martins, E. A. (2006). Morphological changes and hypoglycemic effects of Annona muricata linn. ( annonaceae) leaf aqueous extract on pancreatic β-cells of streptozotocin-treated diabetic rats. *African Journal of Biomedical Research*, 9(3).

[7]. Morton, J. F. (1987). *Fruits of warm climates*: JF Morton.

[8]. da Silva, R. R. (2009). Enhanced extraction yields and mobile phase separations by solvent mixtures for the analysis of metabolites in Annona muricata L. leaves. *analysis*, 9, 10.

[9]. Vij, S., Mukhopadhyay, D., & Agrawal, A. (2019). Automated Negotiation in E Commerce: Protocol Relevance and Improvement Techniques. *CMC-COMPUTERS MATERIALS & CONTINUA*, 61(3), 1009-1024.

[10]. Mondelaers, K., Verbeke, W., & Huynlenbroeck, G. V. (2009). Importance of health and environment as quality traits in the buying decision of organic products. *British Food Journal*, 111(10), 1120-1139.

[11]. Ljumović, I., & Lečovski-Milojkić, I. (2018). Organic product labelling: consumer attitudes and impact on purchasing decision (pp. 26-39): Institute of Economic Sciences, Serbia.

[12]. Khan, M. I., Jo, C., & Tariq, M. R. (2015). Meat flavor precursors and factors influencing flavor precursors—A systematic review. *Meat Science*, 110, 278-284.

[13]. White, G. K., & Manning, B. J. (2001). Convenience, price, product: motivators for online specialty food consumers. *Journal of Food Products Marketing*, 7(1-2), 53-65.

[14]. Limpo, L., Rahim, A., & Hamzah, H. (2018). Effect of product quality, price, and promotion to purchase decision. *International Journal on Advanced Science, Education, and Religion*, 1(1), 9-22.

[15]. Yee, W. M., Yeung, R. M., & Morris, J. (2005). Food safety: building consumer trust in livestock farmers for potential purchase behaviour. *British Food Journal*.

[16]. Constantinides, E. (2002). The 4S web-marketing mix model. *Electronic commerce research and applications*, 1(1), 57-76.

[17]. Svendsen, M. F., Haugland, S. A., Grønhaug, K., & Hammervoll, T. (2011). Marketing strategy and customer involvement in product development. *European Journal of Marketing*.

[18]. Srinivasan, S., Rutz, O. J., & Pauwels, K. (2016). Paths to and off purchase: quantifying the impact of traditional marketing and online consumer activity. *Journal of the academy of marketing science*, 44(4), 440-453.

[19]. Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261-304.
[20]. Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of econometrics*: John Wiley & Sons.

[21]. Miles, J. (2014). Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online*.

[22]. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.

[23]. Durbin, J., & Watson, G. S. (1971). Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1), 1-19.

[24]. Greene, W. H. (2003). *Econometric analysis*: Pearson Education India.