Published online on the journal website: https://publikasi.mercubuana.ac.id/index.php/fasilkom/

Collabits Journal

Vol 1,No.1 (2024) ISSN :

Toothpaste Brand Prediction Based on Analysis of Teeth Condition and Price Preferences Using the Random Forest Algorithm

Afiyati Amiludin[1], Rahma Farah Ningrum[2], Faaza Naima[3]

[1,2]Department of Computer Science; Mercubuana University; Jakarta
[3]Department of Computer Science; PLN Institute of Technology; Jakarta
e-mail:*[1]afiyati.amiludin@mercubuana.ac.id [2] rahmafarah@itpln.ac.id,[3]ifafaaza@gmail.com

Abstract

This study aimed to predict toothpaste brands based on an analysis of dental conditions and price preferences using the Random Forest algorithm and the CRISP-DM approach. The research results indicated that the variables of tooth color range and frequency of toothache had the highest influence, suggesting that consumers were more likely to choose a brand based on tooth color and sensitivity. Evaluation using the Confusion Matrix and Classification Report models demonstrated good performance with an accuracy of 91.3%. Based on the result, the model could serve as a robust foundation for developing a GUI-based Toothpaste Brand Prediction Application using the **tkinter** library, assisting users in making more informed decisions.

Keywords : *Prediction, CRISP-DM, Random Forest, Tkinter*

1. Introduction

Toothpaste is a semi-solid product consisting of a mixture of scrubbing agents, cleaning agents and additional ingredients used to help clean teeth without damaging the teeth or oral mucous membranes (1). The main function of toothpaste is to remove impurities from the surface of the teeth with minimal adverse effects on the teeth. Mitsui in Fatkan (2017) said that another function of toothpaste is to prevent tooth decay and reduce bad breath (2). Supports the function of the toothbrush, Toothpaste manufacturers strive to improve quality by formulating toothpaste according to consumer preferences. This is in line with the emergence of a number of toothpaste brands as a result of the Katadata Insight Center (KIC) survey in 2023. The survey shows that Pepsodent is the toothpaste brand that is most frequently used and is also popular among the public. Its position is followed by Ciptadent, CloseUp, and Sensodyne (3). In this research, Toothpaste brand predictions based on analysis of tooth condition and price preferences were carried out to obtain factors that are closely related to selection.Toothpaste brand. Then, based on these factors, a toothpaste brand prediction system analysis was designed to obtain comprehensive accuracy results.

2. Library Study

Pratama, et al (4) in their research entitled Application*Data Mining* for Predicting the Most Popular Clothing Brands using the K-Nearest Neighbor Method (Case Study: PT. Matahari Department Store Binjai) succeeded in making predictions on the system that was built. The system is analyzed using a classification method with an algorithm*K-Nearest Neighbor* use*Software Visual Basic*. The results of the system trials aim to determine future sales predictions based on the closest variables of year, month, product and number of sales as well as finding out which clothing brands are most in demand.

Meanwhile, Rismala, et al (5) in their research entitled Application of Methods*K-Nearest Neighbor*

For the Best Selling Motorcycle Sales Prediction, produce accurate motorbike sales data from January to December 2022 at PT. West Java's Source of Fortune is 96.15. Apart from that, there is research by Muhammad, et al (6) entitled Predicting the Sales Performance of Musical Works Using the CRISP-DM Framework which utilizes the data mining process with the CRISP-DM framework.

From previous research, brand predictions on goods based on certain criteria can be done by implementing data mining. However, in this study, Toothpaste Brand Prediction Based on Analysis of Teeth Condition and Price Preferences was carried out using the Random Forest Algorithm to overcome overfitting and datasets are available that have many features.

3. Method

The flow of this research is adjusted to the methodology*Cross-Industry Standard Process for Data Mining* or CRISP-DM which is the standard form*data mining* compiled by Daimler Chrysler (Daimler-Benz), SPSS (ISL), and NCR which were developed in various workshops in 1997-1999 (7). The stages in CRISP-DM include*Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* and*Deployment.*

This research uses an algorithm supervised *learning* in the form of a Random Forest method which is tasked with classifying data into certain categories or classes. The category in question is the toothpaste brand based on the available dataset.

4. Experiment Preparation

4.1 Dataset

Data collection on toothpaste brand usage based on analysis of tooth condition and price preferences was obtained from a survey via Google Form with the following list of questions.

No.	Question	Variable
1	What type of toothpaste model do you use?	type
2	Which brand of toothpaste do you use? (if using)	brand
3	Length of Use of Toothpaste Brand (according to number 2)	old
4	Current condition of your teeth?	condition
5	Do you have a history of illness? tooth?	toothache
6	If the answer to number 5 is yes, how much Do you often experience toothache?	sick_frequency
7	Have you ever had other dental problems? (e.g., bleeding gums, tooth loss, etc.)	dental_problems
8	Do you have teeth? hollow?	cavities
9	Do you have regular dental checkups at least 6 months once?	routine_check
10	How often do you brush your teeth in a day?	tooth_brushing_intensity
11	Do you use that technique right when brushing your teeth?	tooth_brushing_technique
12	Frequency in changing pasta brands the gear you use	frequency_replace
13	Do you smoke?	smoking
14	Does price influence the choice Are you buying toothpaste?	price
15	Have you ever experienced a negative reaction after using a certain toothpaste? (e.g. gum irritation, burning sensation, etc.)	negative_reaction
16	Your current tooth color range Example answers:1	tooth_color_range

4.2 Proposed Classification Model

a. Business Understanding

Andriawan in Ilham (2023) said that at this stage it is carried out by interpreting the objectives and requirements of the business side, which are then translated into knowledge to define the main problems that can be overcome through data *mining.* (8)

a) Determine business objectives, namely predicting toothpaste brands based on analysis of tooth condition and price preferences.
b) Assessing the situation, several available variables have different levels of influence in determining the toothpaste brand, so it is necessary to know the factors that most influence consumer preferences in choosing a toothpaste brand.
c) Determine the purpose of data mining, namely to increase knowledge about the characteristics of toothpaste brands based on analysis of tooth condition and consumer price preferences with predetermined variables.

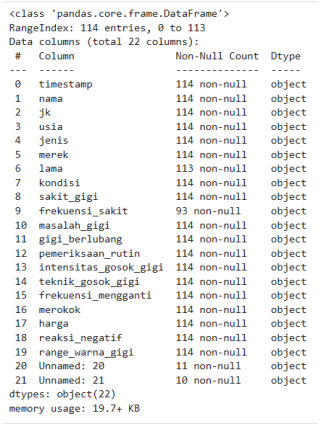
b. Data Understanding

Pinto in Ilham (2023) explains that at this stage the data identification process is carried out, understanding the quality of the data, getting initial insights from the data, and obtaining several hypotheses to find hidden information in the data. At this stage, data visualization is also carried out to understand and clean the data and handle problematic features to get a better and more general machine learning model. (8)

The dataset has input and output that determines the brand of toothpaste in the form of a certain value, namely 'Ciptadent': 1, 'CloseUp': 2, 'Colgate': 3, 'Pepsodent': 4, 'Sensodyne': 5, and 'Other' : 6. There are 114 observation sample data with 22 attributes, namely timestamp, name, jk, age, type, brand, duration, condition, tooth_pain, frequency_of_pain, tooth_problem, tooth_cavity, routine checkup, tooth_brushing_intensity,

tooth_brushing_technique, frequency_of_replacement, smoking, price, negative_reaction , tooth_color_range, Unnamed:20, Unnamed:21. In implementing this research using python.

Following are the details of the Toothpaste dataset:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114 entries, 0 to 113
Data columns (total 22 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   timestamp              114 non-null    object
 1   nama                   114 non-null    object
 2   jk                     114 non-null    object
 3   usia                   114 non-null    object
 4   jenis                  114 non-null    object
 5   merek                  114 non-null    object
 6   lama                   113 non-null    object
 7   kondisi                114 non-null    object
 8   sakit_gigi             114 non-null    object
 9   frekuensi_sakit        93 non-null     object
 10  masalah_gigi           114 non-null    object
 11  gigi_berlubang         114 non-null    object
 12  pemeriksaan_rutin      114 non-null    object
 13  intensitas_gosok_gigi  114 non-null    object
 14  teknik_gosok_gigi      114 non-null    object
 15  frekuensi_mengganti    114 non-null    object
 16  merokok                114 non-null    object
 17  harga                  114 non-null    object
 18  reaksi_negatif         114 non-null    object
 19  range_warna_gigi       114 non-null    object
 20  Unnamed: 20            11 non-null     object
 21  Unnamed: 21            10 non-null     object
dtypes: object(22)
memory usage: 19.7+ KB
```

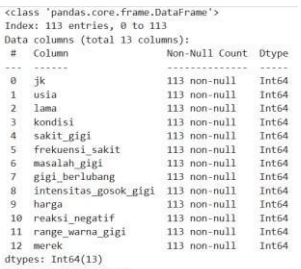
Source: Research Results (2023)

c. Data Preparation

Golvances in Ilham (2023), at this stage is an activity that involves all activities to build a collection of data to be used in a model. (8)

At this stage the data is prepared for the training process with several data processing. Data processing stages include variable selection, data cleaning and regulation processes, and arranging variable positions in the table.

This process also removes some missing values and addressing problematic features. At this stage, the type of toothpaste in the form of 'siwak' is removed. Of the 22 variables contained in the dataset, the variables used are only 12 variants called 'selected_columns' and the variable used as the target or 'target_variable' is 'brand'. Then, the entry data used in the research was 113 data. Following are the details of the Toothpaste Brand dataset after completion data *preparation*:



```
<class 'pandas.core.frame.DataFrame'>
Index: 113 entries, 0 to 113
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   jk                     113 non-null    Int64
 1   usia                   113 non-null    Int64
 2   lama                   113 non-null    Int64
 3   kondisi                113 non-null    Int64
 4   sakit_gigi             113 non-null    Int64
 5   frekuensi_sakit        113 non-null    Int64
 6   masalah_gigi           113 non-null    Int64
 7   gigi_berlubang         113 non-null    Int64
 8   intensitas_gosok_gigi  113 non-null    Int64
 9   harga                  113 non-null    Int64
 10  reaksi_negatif         113 non-null    Int64
 11  range_warna_gigi       113 non-null    Int64
 12  merek                  113 non-null    Int64
dtypes: Int64(13)
memory usage: 13.8 KB
```

Source: Research Results (2023)

d. Modelling

Ilham (2023) stated that this stage involves the selection of candidate models, which includes reviewing past literature and identifying commonly used predictive models. (8) The researcher decided to use the Random Forest algorithm to address the toothpaste brand prediction problem. The data characteristics suitable for handling with the Random Forest algorithm include datasets with many features, class imbalances, a mix of variable types, and the presence of outliers. Random Forest tends to achieve high accuracy by building multiple decision trees and combining results to overcome overfitting. It can handle imbalanced data, mixed data types, numerous features without the need for preprocessing, and generally exhibits model stability even with changes in the dataset.

Before proceeding to the stage of creating the random forest classification model, the dataset will be divided first. The data division follows the pareto principle by applying 80/20, meaning 80% of the data for training (training data) and 20% of the data for testing (test data). In essence, this theory

e. Evaluations

Van der Voort in Ilham (2023), at this stage an analysis is carried out to ensure that the selected model has the best quality that can achieve the objectives of the existing business problem. (8)

At the evaluation stage, information is collected relating to the resulting performance or performance which can be used to determine the best alternative in making decisions.

(9) Evaluation in this research uses methods*Confusion Matrix* which shows the performance of a model's classification algorithm. Evaluation result*Confusion Matrix* displays some information such as*True Negative, True Positive, False Negative dan False Positive* from which information can be known the value of accuracy,*recall,* precision, and *f-measure*.

Mark accuracy can be calculated through the following equation.

$$\boldsymbol{Akurasi} = \boldsymbol{TP}+\boldsymbol{TN}/(\boldsymbol{TP}+\boldsymbol{FN}+\boldsymbol{FP}+\boldsymbol{TN}) \quad (9)$$

Mark precision can be calculated through the following equation.

$$\boldsymbol{Precision} = \boldsymbol{TN}/(\boldsymbol{TP}+\boldsymbol{FP})$$

(9)

The recall value can be calculated using the following equation.

$$Recall = TP/(TP+FN)$$

(9)

The specificity value can be calculated using the following equation.

$$\text{F1-Score} = 2\times((\text{Precision}+\text{Recall})/(\text{Precision}\times\text{R ecall}))$$ (9)

f. Deployment

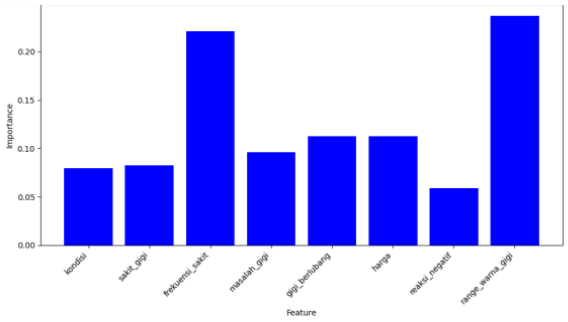
In Ilham (2023), this stage is the result of testing and implementation in real applications using real data in the field. (8)

This research uses literature **tkinter** (ttk) to create a user interface (GUI) implementing a more modern thematic widget in displaying Toothpaste Brand prediction

5. Results and Analysis

5.1 Research result

The research was carried out by testing the model that had been created using training data and test data. Before testing, the factors that most influence the performance of the model are tested. Testing the most influential features using the scikit-learn library and importing classes**RandomForestClassifier** from the module ensemble with a data generation and model fitting approach.



Feature Importance Table Based on the Feature Importance Table above,

It can be seen that the variables tooth_color_range and pain_frequency have the highest values, namely 0.24 and 0.21 respectively. Then followed by the cavity_tooth variable and price with the same value, namely 0.11.

Analysis Feature *importance* shows that in the prediction of Toothpaste brands, variables `tooth_colour_range' is the most influential factor and shows significant implications in brand selection by consumers. Furthermore, 'sick_frequency' and 'price' are also important variables which show that economic aspects, especially the product price range, have a quite significant impact on brand preferences by consumers.

Based on this analysis, the variables selected as 'selected_column' are as follows, namely 'condition', 'tooth_pain', 'sick_frequency', 'tooth_problem', 'cavities', 'price', 'negative_reaction', and 'tooth_color_range'. The justification for research using 'selected_colomn' is that when all columns are used, the accuracy results only show 43% and the most influential feature is the 'old' variable, which should not be a cause, but an impact.

5.2 Evaluation

Methods Random *forest* used is set with the parameter random_state=42. The following is an evaluation matrix for the results of model training using the random forest algorithm.

```
Confusion Matrix:
[[ 1  0  0  0  0  0]
 [ 0  2  0  1  0  0]
 [ 0  0  1  0  0  0]
 [ 0  0  0 12  0  0]
 [ 0  0  0  0  4  0]
 [ 0  1  0  0  0  1]]
```

Picture*Confusion Matrix*

Source: Processed by the Author

Based on the picture confusion *matrix* above, it is known that the 1st row 1st column (1) shows the number of samples that should go into the first class (*true class*) and correctly predicted by the model as first class (*predicted class*). Then, in the 2nd row the 2nd column (2) shows the number of samples that should fall into the second class and are correctly predicted by the model as the second class. The 2nd row of the 4th column (1) shows the number of samples that should fall into the second class, but were incorrectly predicted as the fourth class. The 2nd row of the 4th column (12) shows the number of samples that should fall into the fourth class and are correctly predicted by the model as the fourth class. The 6th row of the 2nd column (1) shows the number of samples that should belong to the sixth class, but were incorrectly predicted as the second class.

```
Classification Report:
              precision    recall  f1-score   support

         1.0       1.00      1.00      1.00         1
         2.0       0.67      0.67      0.67         3
         3.0       1.00      1.00      1.00         1
         4.0       0.92      1.00      0.96        12
         5.0       1.00      1.00      1.00         4
         6.0       1.00      0.50      0.67         2

    accuracy                           0.91        23
   macro avg       0.93      0.86      0.88        23
weighted avg       0.92      0.91      0.91        23
```

Picture*Classification Report*
Source: Processed by the Author

Based on classification *report* above, the research produced a model accuracy of 91.3%, which means that most of the predictions were correct.

Good precision (93.2%) indicates that the majority of those predicted as positive are indeed positive. In class 1.0, class 3.0, class 5.0, and class 6.0 show a precision of 1.00 which means that all predictions instance each of these classes is correct. Meanwhile class 2.0 produces a precision of 0.67 which shows that 92% of class 2.0 instance predictions are correct. *Recall* The good score (86.1%) shows that most of the actual positives were successfully predicted. In class 1.0, class 3.0, class 5.0, and class 6.0 shows recall of 1.00, which means the model successfully identified all*instance* who actually go into the classroom the said. However, there are still instance those who failed to identify the model were 43% in class 2.0 and 50% in class 6.0.

F1-Score Which Good (88.2%) shows a balance between precision and recall. In class 1.0, class 3.0, and class 5.0 shows*f1-score* each of 1.00 which means that the model has succeeded in minimizing both types of errors, namely false *positives* and false *negatives* well. As for*f1-score* in class 4.0 also has a good balance between precision and recall namely 0.96. However, it still exists instance those who failed to identify the model were 33% in class 2.0 and class 6.0. The figure also shows the average m acro results which calculate the average performance of all classes without considering class imbalance. The result is an average Macro precision of 0.93, recall of 0.86, and F1-Score of 0.88. Meanwhile, the weighted average produces a weighted average precision of 0.92, recall of 0.91, and F1-Score of 0.91. Weighted average takes into consideration contribution each classes based on their proportion in the dataset. In this research, the classes in the dataset face imbalance class, so that The weighted average provides a more accurate picture about model performance.

Apart from the evaluation metrics above, there is a support value which shows the weight of the contribution to the weighted average. Higher support has a greater contribution compared to classes that have low support. This is owned by class 4.0 which has instances of 12 points which are categorized as ground *truth*.

5.3 Application Planning

Applications are built using Libraries Tkinter in Python as a user interface to display questions like a questionnaire. Interface also available button predictions as widget which will call the function*predict_brand* and display the results by calling the function*result_label* which is dependent on function predict *brand*.

As*feature_importance* that the most influential factors in brand prediction research are based on analysis of tooth condition and price preferences.

when input is done*tooth_color_range*which tends to be high, then the toothpaste brand that emerged was "Pepsodent". Then, there is a tendency that when individuals have*tooth_condition*sensitive without any*history_of_tooth_ache, tooth_cavity*, or*negative_reaction*, the brand “Sensodyne” will appear as a prediction of the Toothpaste brand that the individual will choose in*tooth_color_range*up to equal to 20. Meanwhile for*tooth_condition*sensitive with*history_of_tooth_pain*less significant regardless of price preferences, at level*tooth_color_range*less than equal to 5, then the prediction for the toothpaste brand is "Close Up".

Applying the fuzzy wuzzy technique with the DFS (Depth-First Search) Algorithm to File Search on an FTP Server yields numerous significant findings:

1. **Search Efficiency:** When looking for files on the FTP server, the DFS algorithm is employed. This method traverses the full directory structure to provide thorough searches, but it may be less effective for particularly large directory structures.
2. **Matching Accuracy:** More adaptable matching is possible with the fuzzy wuzzy approach. This enables the adjustment of typos, small variations in file names, or spelling variances in the context of file searches, allowing relevant results to be found despite variations in writing or spelling..
3. **Accuracy of Results:** By taking into account the similarity between the searched string and the file name on the server, the DFS and fuzzywuzzy combo improves the accuracy of search results. Even if there are small variations in the file name or mistakes in the search query, this aids in locating the needed file.
4. **Performance Limitations:** Using the fuzzy wuzzy approach might raise the computing strain on the server, especially when searching on huge datasets, even though it helps boost search accuracy. Improper management of this could have an impact on server performance.
5. **Adaptability:** In an FTP Server environment, changes or variations in file names are frequently encountered. This approach enables good adaption to these situations. It is a more dynamic

approach because of its capacity to handle mistakes or variances in file naming.

In conclusion, using the fuzzy wuzzy approach in conjunction with the DFS algorithm on the FTP server helps improve file discovery; nevertheless, system performance must be balanced with search accuracy.

6. Conclusion

The conclusion of this research is that the model succeeded in showing good accuracy, namely 91.3%. This research focuses on the weighted average results, namely precision of 0.92,*recall* of 0.91, nd*f1-score* of 0.91. This shows that balance precision and recall also shows good results so that the model can be used to design Toothpaste Brand Prediction Applications with **tkinter.** The results of this research provide a strong foundation for applying this model in real-world situations and help users make more informed decisions.

However, further research needs to be done regarding the accuracy of models that accommodate dataset imbalances by utilizing**StratifiedKFold**divide the dataset into multiple folds. Apart from that, it is necessary to evaluate each class that still exists instance classified under-sampling namely 2 instance in grade 6 which is still not.

Preferences

[1] Widodo, Rahmah R., Rachmadi P. (2014). "Comparison of the Effectiveness of Herbal Toothpaste and Non Herbal Toothpaste in Reducing Plaque Index in Angsau Elementary School Students 4 Pelaihari." Lambung Mangkurat University. Journal of Dentino Dentistry, Vol. 2.

[2] Sofyan, Van Fatkhan. (2017). "Use of Na - CMC (Gelling Agent) in Toothpaste Preparation of Miswak Wood Extract (Salvadora persica) and Red Betel Leaf Extract (Piper crocatum)." Muhammadiyah University Purwokerto.

[3] Annur, Cindy Grass. (2023). "Consumer Products." Media Network Catadata. [Online] Katadata. March 24, 2023. [Cited: 12 01, 2023.] https://databoks.katadata.co.id/datapublish/2023/03/24/pepsodent-brand-pasta-gigi-yang-most-frequently-used- Indonesian consumers.

[4] Andrean Pratama, Budi Serasi Ginting, Nurhayati. (2021). "Application of Data Mining to Predict the Most Popular Clothing Brands Using the K-Nearest Neighbor Method (Case Study: PT. Matahari Department Store Binjai)." Panca Budi, Vol. 14, pp. 54-64. ISSN.

[5] Rismala, Irfan Ali, Ade Rizki Rinaldi. (2023). "Application of the K-Nearest Neighbor Method for Best Selling Motorcycle Sales Predictions." Malang Institute of Technology, Vol. 7, pp. 585-590. ISSN.

[6] Muhammad Zain Imtiyaz, Muhammad Nasrun S.Si, M.T., Umar Ali Ahmad S.T, M.T. (2015). "Analysis and Implementation of the CRISP-DM Framework to Understand the Behavior of Customer Transaction Data." Telkom University, Vol. 2. ISSN: 2355-9365.

[7] Purwarianti, A. A. Prajitno, A. (2011). "Prediction of Sales Performance of Musical Works Using the CRISP-DM Framework (Case Study: X Music Indonesia)." Bandung Institute of Technology.

[8] Ilham Kurniawan, Duwi Cahya Puri Buani, Abdussomad, Widya Apriliah, Rizal Amegia Saputra. (2023). "Implementation of the Random Forest Algorithm to Determine Raskin Assistance Recipients." Journal of Information Technology and Computer Science, Vol. 10. ISSN.

[9] Luthfiyah Amatullah, Yuni Widiastiwi, Nurul Chamidah. (2022). "Application of Random Forest Classification to Data on Autism Spectrum Disorders (ASD) in Children Using Principal Component Analysis Feature Selection." National Seminar for Computer Science Students and Their Applications (SENAMIKA)