Collabits Journal
Vol 1 No. 1 January 2024 : 21-26
E-ISSN : 3046-6709
https://publikasi.mercubuana.ac.id/index.php/collabits

Prediksi Merek Pasta Gigi Berdasarkan Analisa Kondisi Gigi dan Preferensi Harga Menggunakan Algoritma Random Forest

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

[1,3] Fakultas Ilmu Komputer, Universitas Mercu Buana, Jakarta, Indonesia
[2] Fakultas Ilmu Komputer, Institut Teknologi PLN, Jakarta, Indonesia

*Coressponden Author: afiyati.amiludin@mercubuana.ac.id

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;

Article History:
Received: 21-11-2023
Revised: 13-12-2023
Accepted: 23-01-2024

Article DOI : 10.22441/collabits.v1i1.25560

1. INTRODUCTION

Pasta gigi adalah produk semi padat yang terdiri dari campuran bahan penggosok, bahan pembersih, dan bahan tambahan yang digunakan untuk membantu membersihkan gigi tanpa merusak gigi maupun membran mukosa mulut [1]. Fungsi utama dari pasta gigi adalah menghilangkan pengotor dari permukaan gigi dengan efek buruk yang kecil terhadap gigi. Mitsui dalam Fatkan (2017) mengatakan bahwa fungsi lain dari pasta gigi adalah untuk mencegah kerusakan gigi dan mengurangi bau mulut [2].

Mendukung fungsi sikat gigi tersebut, produsen pasta gigi berupaya untuk meningkatkan kualitas dengan melakukan formulasi pasta gigi sesuai dengan preferensi konsumen. Hal ini selaras dengan kemunculan sejumlah merek pasta gigi hasil survei Katadata Insight Center (KIC) pada tahun 2023. Survei menunjukkan bahwa Pepsodent merupakan merek pasta gigi yang paling sering digunakan dan juga popular di kalangan Masyarakat. Posisinya disusul oleh Ciptadent, CloseUp, dan Sensodyne [3].

Pada penelitian ini, prediksi merek Pasta Gigi berdasarkan analisa kondisi gigi dan preferensi harga dilakukan untuk mendapatkan faktor-faktor yang berkaitan erat dengan pemilihan merek Pasta gigi. Kemudian, terhadap faktor-faktor tersebut dilakukan perancangan analisa sistem prediksi merek pasta gigi untuk mendapatkan hasil akurasi yang komprehensif.

2. RELATED WORKS

Pratama, dkk [4] dalam penelitiannya yang berjudul Penerapan *Data Mining* untuk Prediksi Merek Pakaian yang Paling Diminati dengan Metode K-Nearest Neighbor (Studi Kasus: PT. Matahari Departement Store Binjai) berhasil melakukan prediksi pada sistem yang dibangun. Sistem dianalisa menggunakan menggunakan metode klasifikasi dengan algoritma *K-Nearest Neighbor* menggunakan *Software Visual Basic*. Hasil uji coba sistem bertujuan untuk mengetahui prediksi penjualan pada masa mendatang berdasarkan variabel tahun, bulan, produk dan jumlah penjualan yang paling dekat serta mengetahui merek pakaian mana saja yang paling banyak diminati.

Sedangkan Rismala, dkk [5] dalam penelitiannya yang berjudul Penerapan Metode *K-Nearest Neighbor* Untuk Prediksi Penjualan Sepeda Motor Terlaris, menghasilkan akurasi terhadap data penjualan sepeda motor dari bulan Januari sampai dengan Desember tahun 2022 di PT. Sumber Rejeki Jabar sebesar 96,15. Selain itu, terdapat penelitian Muhammad, dkk [6] berjudul Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM yang memanfaatkan proses data mining dengan framework CRISP-DM.

Dari penelitian terdahulu, prediksi merek pada barang

berdasarkan kriteria tertentu dapat dilakukan dengan mengimplementasikan suatu data mining. Namun demikian, pada penelitian ini, Prediksi Merek Pasta Gigi Berdasarkan Analisa Kondisi Gigi dan Preferensi Harga dilakukan menggunakan Algoritma Random Forest untuk mengatasi *overfitting* dan dataset tersedia yang memiliki banyak fitur.

3. PROPOSED METHOD

Alur pada penelitian ini disesuaikan dengan metodologi *Cross-Industry Standard Process for Data Mining* atau CRISP-DM yang merupakan bentuk standar *data mining* yang disusun oleh Daimler Chrysler (Daimler-Benz), SPSS (ISL), dan NCR yang dikembangkan di berbagai workshops pada tahun 1997-1999 [7]. Tahapan dalam CRISP-DM meliputi *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* dan *Deployment.*

Penelitian ini menggunakan algoritma *supervised learning* berupa metode Random Forest yang bertugas melakukan klasifikasi data ke dalam kategori atau kelas tertentu. Kategori dimaksud yaitu merek pasta gigi berdasarkan dataset tersedia.

4. EXPERIMENTAL SETUP

Experimental

4.1 Dataset

Pengumpulan data penggunaan merek pasta gigi berdasarkan analisa kondisi gigi dan preferensi harga diperoleh dari survei melalui Google Form dengan daftar pertanyaan sebagai berikut.

No.	Pertanyaan	Variabel
1	Jenis Model Pasta Gigi yang digunakan?	jenis
2	Merk Pasta Gigi yang Anda gunakan? (jika menggunakan)	merek
3	Lama Penggunaan Merk Pasta gigi tersebut (sesuai nomor 2)	lama
4	Kondisi gigi anda saat ini?	kondisi
5	Apakah anda mempunyai riwayat sakit gigi?	sakit_gigi
6	Jika jawaban no 5 adalah iya, seberapa sering Anda mengalami sakit gigi?	frekuensi_sakit
7	Pernah mengalami masalah gigi lainnya? (misalnya, gusi berdarah, gigi tanggal, dll.)	masalah_gigi
8	Apakah anda memiliki gigi yang berlubang?	gigi_berlubang
9	Apakah Anda melakukan pemeriksaan gigi rutin setidaknya 6 bulan sekali?	pemeriksaan_rutin
10	Intensitas gosok gigi dalam sehari?	intensitas_gosok_gigi
11	Apakah Anda menggunakan teknik yang benar saat menyikat gigi?	teknik_gosok_gigi
12	Frekuensi dalam mengganti merk pasta gigi yang anda gunakan	frekuensi_mengganti
13	Apakah anda merokok?	merokok
14	Apakah harga mempengaruhi pilihan Anda dalam membeli pasta gigi?	harga
15	Apakah Anda pernah mengalami reaksi negatif setelah menggunakan pasta gigi tertentu? (misalnya, iritasi gusi, rasa panas, dll.)	reaksi_negatif
16	Range warna gigi anda saat ini Contoh jawaban:1	range_warna_gigi

4.2 Proposed Classification Model

a. Business Understanding

Andriawan dalam Ilham (2023) mengatakan bahwa pada tahap ini dilakukan dengan penafsiran tujuan dan persyaratan sisi bisnis, yang kemudian diterjemahkan ke dalam pengetahuan untuk mendefinisikan masalah utama yang dapat diatasi melalui *data mining.* [8]

a) Menentukan tujuan bisnis, yaitu memprediksi merek pasta gigi berdasarkan analisa kondisi gigi dan preferensi harga.
b) Menilai situasi, beberapa variabel tersedia memiliki tingkat pengaruh yang berbeda dalam penentuan merek pasta gigi, sehingga perlu mengetahui faktor yang paling berpengaruh terhadap preferensi konsumen dalam memilih merek pasta gigi.
c) Menentukan tujuan data mining, yaitu untuk meningkatkan pengetahuan tentang karakteristik merek pasta gigi berdasarkan analisa kondisi gigi dan preferensi harga konsumen dengan variabel yang telah ditentukan.

b. Data Understanding

Pinto dalam Ilham (2023) menjelaskan bahwa pada tahap ini dilakukan proses identifikasi data, memahami kualitas data, mendapatkan wawasan awal dari data, dan memperoleh beberapa hipotesis untuk menemukan informasi tersembunyi dalam data. Pada tahap ini juga dilakukan visualisasi data untuk memahami dan membersihkan data serta menangani fitur bermasalah untuk mendapatkan model pembelajaran mesin yang lebih baik dan lebih umum. [8]

Pada dataset memilki input dan output yang menetapkan merek pasta gigi berupa nilai tertentu, yaitu ‘Ciptadent': 1, 'CloseUp': 2, 'Colgate': 3, 'Pepsodent': 4,

'Sensodyne': 5, dan 'Lainnya': 6. Ada 114 data sampel observasi dengan 22 atribut yaitu timestamp, nama, jk, usia, jenis, merek, lama, kondisi, sakit_gigi, frekuensi_sakit, masalah_gigi, gigi_berlubang, pemeriksaan rutin, intensitas_gosok_gigi, teknik_gosok_gigi, frekuensi_mengganti, merokok, harga, reaksi_negatif, range_warna_gigi, Unnamed:20, Unnamed:21. Dalam penerapan penelitian ini menggunakan python.

Berikut rincian dari dataset Pasta Gigi :

```
<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
```

Sumber: Hasil Penelitian (2023)

c. Data Preparation

Golvances dalam Ilham (2023), pada tahap ini merupakan kegiatan yang melibatkan semua aktivitas membangun Kumpulan data untuk digunakan dalam sebuah model. [8]

Pada tahap ini data dipersiapkan untuk dilakukan proses pelatihan dengan beberapa pengolahan data. Tahapan pengolahan data mencakup pemilihan variabel, proses pembersihan dan regulasi data, dan mengatur posisi variabel dalam tabel.

Pada proses ini juga menghilangkan beberapa missing value dan menangani fitur bermasalah. Pada tahap ini pula, dilakukan penghapusan jenis pasta gigi berupa ‘siwak’. Dari 22 variabel yang terdapat di dalam dataset, variabel yang digunakan hanya 12 varianel yang disebut ‘selected_columns’ dan variabel yang dijadikan target atau ‘target_variable’ yaitu ‘merek’. Kemudian, data entri yang digunakan dalam penelitian yaitu sebanyak 113 data.

Berikut rincian dari dataset Merek Pasta Gigi setelah dilakukan *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
```

Sumber: Hasil Penelitian (2023)

d. Modelling

Ilham (2023) mengatakan bahwa pada tahap ini merupakan pemilihan calon model yang melibatkan peninjauan literatur di masa lampau dan mengidentifikasi model prediksi yang umum digunakan sebelumnya. [8]

Peneliti memutuskan untuk menggunakan algoritma Random Forest dalam memecahkan masalah prediksi merek pasta gigi. Adapun karakteristik data yang dapat ditangani dengan algoritma Random Forest yaitu dataset dengan banyak fitur, ketidakseimbangan kelas, campuran tipe variabel, dan adanya outliers. Random Forest cenderung menghasilkan akurasi tinggi karena membangun beberapa pohon keputusan dan menggabungkan hasil dengan mengatasi overfitting, mampu menangani data tidak seimbang, tipe data campuran, fitur yang banyak tanpa perlu pemrosesan, dan cenderung memiliki stabilitas model yang cukup meskipun terdapat perubahan pada dataset.

Sebelum menuju tahap pembuatan model klasifikasi *random forest*, dataset akan dibagi terlebih dahulu. Pembagian data mengacu berdasarkan prinsip *pareto* dengan menerapkan 80/20, yang berarti sebesar 80% dari data untuk data training (data latih) dan 20% dari data untuk data testing (data uji). Pada dasarnya, teori ini menetapkan bahwa 80% dari output atau hasil berasal dari 20% efek atau aliansi yang tidak proporsional antara input dengan output. [9]

e. Evaluation

Van der Voort dalam Ilham (2023), pada tahap ini dilakukan analisa untuk memastikan bahwa model yang dipilih memiliki kualitas terbaik yang dapat mencapai tujuan dari masalah bisnis yang ada. [8]

Pada tahap evaluasi dilakukan pengumpulan informasi yang berkaitan dengan kinerja atau performa yang dihasilkan yang dapat digunakan untuk menentukan alternatif terbaik dalam membuat keputusan. [9] Evaluasi pada penelitian ini menggunakan metode *Confusion Matrix* yang menunjukkan kinerja algoritma klasifikasi suatu model. Hasil evaluasi *Confusion Matrix* menampilkan beberapa informasi seperti *True Negative, True Positive, False Negative dan False Positive* yang mana dari informasi tersebut dapat diketahui nilai akurasi, *recall*, presisi, dan *f-measure*.

Nilai akurasi dapat dihitung melalui persamaan sebagai berikut.

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

Nilai precision dapat dihitung melalui persamaan sebagai berikut.

$$\boldsymbol{Precision} = \boldsymbol{TN}/(\boldsymbol{TP}+\boldsymbol{FP}) \quad [9]$$

Nilai recall dapat dihitung melalui persamaan sebagai berikut.

$$\boldsymbol{Recall} = \boldsymbol{TP}/(\boldsymbol{TP}+\boldsymbol{FN}) \quad [9]$$

Nilai specificity dapat dihitung melalui persamaan sebagai berikut.

$$\boldsymbol{F1\text{-}Score} = 2\times((\boldsymbol{Precision}+\boldsymbol{Recall})/(\boldsymbol{Precision}\times\boldsymbol{Recall})) \quad [9]$$

f. Deployment

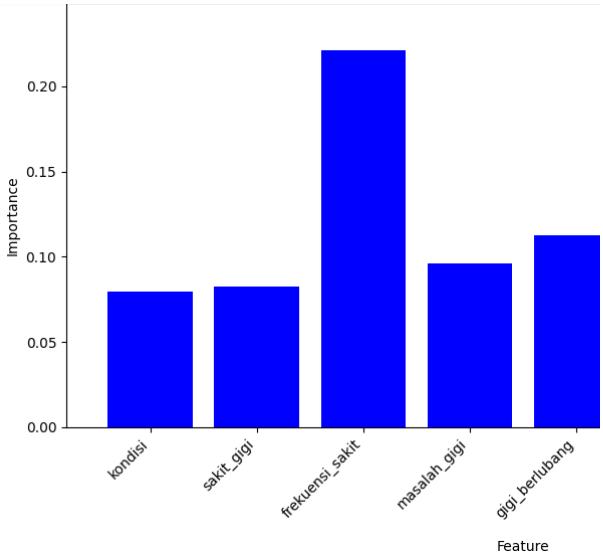
Dalam Ilham (2023), tahap ini merupakan hasil pengujian dan implementasi dalam aplikasi nyata menggunakan data nyata di lapangan. [8]

Penelitian ini menggunakan pustaka **tkinter** (ttk) untuk membuat antarmuka pengguna (GUI) mengimplementasikan widget tematik yang lebih modern dalam menampilkan prediksi Merek Pasta Gigi.

5. RESULT AND ANALYSIS

Hasil Penelitian

Penelitian dilakukan dengan pengujian pada model yang telah dibuat dengan data latih dan data uji. Sebelum dilakukan pengujian, dilakukan pengujian faktor-faktor yang paling berpengaruh terhadap kinerja model tersebut. Pengujian terhadap fitur paling berpengaruh menggunakan pustaka scikit-learn dan mengimport kelas **RandomForestClassifier** dari modul **ensemble** dengan pendekatan data generation dan model fitting.



Tabel Feature Importance

Berdasarkan Tabel Feature Importance di atas, dapat diketahui bahwa variabel rang_warna_gigi dan frekuensi_sakit memiliki nilai tertinggi yaitu masing-masing 0.24 dan 0.21. Kemudian diikuti dengan variabel gigi_berlubang dan harga dengan nilai yang sama yaitu 0.11.

Analisa *feature importance* menunjukkan bahwa dalam prediksi merek Pasta Gigi, variabel `range_warna_gigi' menjadi faktor yang paling berpengaruh dan menunjukkan implikasi signifikan dalam pemilihan merek oleh konsumen. Selanjutnya, 'frekuensi_sakit' dan 'harga' menjadi variabel penting juga yang menunjukkan bahwa aspek ekonomi, terutama kisaran harga produk, memiliki dampak yang cukup signifikan terhadap preferensi merek oleh konsumen.

Berdasarkan analisa ini pula, variabel terpilih sebagai ‘selected_column’ adalah sebagai berikut, yaitu 'kondisi', 'sakit_gigi', 'frekuensi_sakit', 'masalah_gigi', 'gigi_berlubang', 'harga', 'reaksi_negatif', dan 'range_warna_gigi'. Justifikasi bahwa penelitian menggunakan ‘selected_colomn’ adalah ketika semua kolom digunakan, maka hasil akurasi hanya menunjukkan 43% dan fitur paling berpengaruh adalah variabel ‘lama’ yang mana hal tersebut seharusnya bukan merupakan sebab, tetapi sebuah dampak.

5.1 Evaluasi

Metode *random forest* yang digunakan diatur dengan parameter random_state=42. Berikut adalah matrik evaluasi hasil pelatihan model menggunakan algoritma random forest.

Gambar 1. *Confusion Matrix*

```
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]]
```

Sumber: Diolah Penulis

Berdasarkan gambar *confusion matrix* di atas, diketahui bahwa baris ke-1 kolom ke-1 (1) menunjukkan jumlah sampel yang seharusnya masuk ke kelas pertama (*true class*) dan secara benar diprediksi oleh model sebagai kelas pertama (*predicted class*). Kemudian, pada baris ke-2 kolom ke-2 (2) menunjukkan jumlah sampel yang seharusnya masuk ke kelas kedua dan secara benar diprediksi oleh model sebagai kelas kedua. Baris ke-2 kolom ke-4 (1) menunjukkan jumlah sampel yang seharusnya masuk ke kelas kedua, tetapi salah diprediksi sebagai kelas keempat. Baris ke-2 kolom ke-4 (12) menunjukkan jumlah sampel yang seharusnya masuk ke kelas keempat dan secara benar diprediksi oleh model sebagai kelas keempat. Baris ke-6 kolom ke-2 (1) menunjukkan jumlah sampel yang seharusnya masuk ke kelas keenam, tetapi salah diprediksi sebagai kelas kedua.

Gambar 2. *Classification Report*

```
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
```

Sumber: Diolah Penulis

Berdasarkan *classification report* di atas, penelitian menghasilkan akurasi model sebesar 91.3% yang berarti bahwa sebagian besar prediksi benar.

Precision yang baik (93.2%) menunjukkan bahwa sebagian besar dari yang diprediksi sebagai positif memang positif. Pada kelas 1.0, kelas 3.0, kelas 5.0, dan kelas 6.0 menunjukkan presisi 1.00 yang berarti bahwa semua prediksi *instance* masing-masing kelas tersebut benar. Sedangkan kelas 2.0 menghasilkan presisi sebesar

0.67 yang menunjukkan bahwa 92% dari prediksi instance kelas 2.0 adalah benar.

Recall yang baik (86.1%) menunjukkan bahwa sebagian besar dari yang sebenarnya positif berhasil diprediksi. Pada kelas 1.0, kelas 3.0, kelas 5.0, dan kelas 6.0 menunjukkan *recall* sebesar 1.00 yang berarti model berhasil mengidentifikasi semua *instance* yang sebenarnya masuk ke dalam kelas tersebut. Namun demikian, masih terdapat *instance* yang tidak berhasil diidentifikasi model yaitu sebesar 43% pada kelas 2.0 dan 50% pada kelas 6.0.

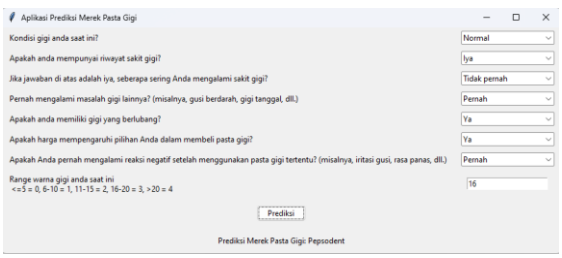
F1-Score yang baik (88.2%) menunjukkan keseimbangan antara *precision* dan *recall.* Pada kelas 1.0, kelas 3.0, dan kelas 5.0 menunjukkan *f1-score* masing-masing sebesar 1.00 yang berarti bahwa model telah berhasil menimimalkan kedua jenis kesalahan yaitu *false positives* dan *false negatives* dengan baik. Adapun *f1-score* pada kelas 4.0 juga memiliki keseimbangan yang baik antara *precision* dan *recall* yaitu 0.96. Namun demikian, masih terdapat *instance* yang tidak berhasil diidentifikasi model yaitu sebesar 33% pada kelas 2.0 dan kelas 6.0.

Pada gambar juga ditunjukkan hasil rata-rata m acroyang menghitung rata-rata kinerja dari semua kelas tanpa mempertimbangkan ketidakseimbangan kelas. Hasilnya adalah presisi rata-rata Macro sebesar 0.93, recall sebesar 0.86, dan F1-Score sebesar 0.88. Sedangkan rata-rata tertimbang menghasilkan presisi rata-rata tertimbang sebesar 0.92, recall sebesar 0.91, dan F1-Score sebesar 0.91. Rata-rata tertimbang mempertimbangkan kontribusi setiap kelas berdasarkan proporsinya dalam dataset. Dalam penelitian ini, kelas dalam dataset menghadapi ketidakseimbangan kelas, sehingga rata-rata tertimbang memberikan gambaran yang lebih akurat tentang kinerja model.

Selain metrik evaluasi di atas, terdapat nilai support yang menunjukkan bobot kontribusi terhadap rata-rata tertimbang. Support yang lebih tinggi memiliki kontribusi lebih besar dibandingkan dengan kelas yang memiliki support rendah. Hal tersebut dimiliki kelas 4.0 yang memiliki instance sebanyak 12 poin yang dikategorikan sebagai *ground truth*.

5.2 Perancangan Aplikasi

Aplikasi dibangun menggunakan Pustaka **tkinter** pada Python sebagai antarmuka pernggguna untuk menampilkan pertanyaan-pertanyaan sebagaimana kuisioner. Pada *interface* juga tersedia *button* prediksi sebagai *widget* yang akan memanggil fungsi *predict_brand* dan menampilkan hasil dengan memanggil fungsi *result_label* yang bersifat dependent terhadap fungsi *predict brand*.



Sebagaimana *feature_importance* bahwa faktor paling berpengaruh dalam penelitian prediksi merek berdasarkan analisa kondisi gigi dan preferensi harga, apabila dilakukan input *range_warna_gigi* yang cenderung tinggi, maka merek pasta gigi yang muncul adalah "Pepsodent". Kemudian, terdapat kecenderungan bahwa ketika individu memiliki *kondisi_gigi* sensitif tanpa ada *riwayat_sakit_gigi*, *gigi_berlubang*, ataupun *reaksi_negatif*, merek "Sensodyne" akan muncul sebagai prediksi merek Pasta Gigi yang akan dipilih individu dalam *range_warna_gigi* hingga sama dengan 20. Sedangkan untuk *kondisi_gigi* sensitive dengan *riwayat_sakit_gigi* kurang signifikan tanpa memperhatikan preferensi harga, pada tingkat *range_warna_gigi* kurang dari sama dengan 5, maka prediksi merek pasta gigi yaitu "Close Up".

6. CONCLUSION

Kesimpulan pada penelitian ini yaitu bahwa model berhasil menunjukkan akurasi yang baik yaitu sebesar 91.3%. Penelitian ini menitikberatkan pada hasil rata-rata tertimbang yaitu *precision* sebesar 0.92, *recall* sebesar 0.91, dan *f1-score* sebesar 0.91. Hal ini menunjukkan bahwa keseimbangan *precision* dan *recall* juga menunjukkan hasil yang baik sehingga model dapat digunakan perancangan Aplikasi Prediksi Merek Pasta Gigi dengan **tkinter.** Hasil penelitian ini memberikan landasan yang kuat untuk menerapkan model ini dalam situasi dunia nyata dan membantu pengguna membuat keputusan yang lebih tepat.

Namun demikian, perlu dilakukan penelitian lebih lanjut yaitu terkait akurasi model yang mengakomodasi ketidakseimbangan dataset dengan memanfaatkan **StratifiedKFold yang** membagi dataset menjadi beberapa lipatan. Selain itu, perlu melakukan evaluasi pada setiap kelas yang mana masih terdapat *instance* yang tergolong *under-sampling* yaitu sebanyak 2 *instance* pada kelas 6 yang masih belum ditangani dengan baik.

REFERENCES

[1] R. R. R. P. Widodo, "Perbandingan Efektivitas Pasta Gigi Herbal Dengan Pasta Gigi Non Herbal Terhadap Penurunan Indeks Plak Pada Siswa Sdn Angsau 4 Pelaihari," Vols. Dentino Jurnal Kedokteran Gigi, 2, no. 2, 2014.

[2] V. F. Sofyan, "PENGGUNAAN Na - CMC (GELLING AGENT) DALAM SEDIAAN PASTA GIGI EKSTRAK KAYU SIWAK (Salvadora persica) DAN EKSTRAK DAUN SIRIH MERAH (Piper crocatum)," Purwokerto, 2017.

[3] C. M. Annur, "Produk Konsumen," Katadata Media Network, 24 Maret 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/03/24/pepsodent-merek-pasta-gigi-yang-paling-sering-digunakan-konsumen-indonesia. [Accessed 01 12 2023].

[4] B. S. G. N. Andrean Pratama, "PENERAPAN DATA MINING UNTUK PREDIKSI MEREK PAKAIAN YANG PALING DIMINATI DENGAN METODE K-NEAREST NEIGHBOR (STUDI KASUS : PT. MATAHARI DEPARTEMENT STORE BINJAI)," *Jurnal Panca Budi,* vol. 14, no. 2, pp. 54-64, 2021.

[5] I. A. A. R. R. Rismala, "PENERAPAN METODE K-NEAREST NEIGHBOR UNTUK PREDIKSI PENJUALAN SEPEDA MOTOR TERLARIS," Vols. 7, pp. 585-590, no. 1, 2023.

[6] M. N. S. M. U. A. A. S. M. Muhammad Zain Imtiyaz, "ANALISIS DAN IMPLEMENTASI FRAMEWORK CRISP-DM UNTUK MENGETAHUI PERILAKU DATA TRANSAKSI PELANGGAN," vol. 2, no. 1, 2015.

[7] A. A. P. d. A. Purwarianti, ""Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM (Studi Kasus: X Music Indonesia)," Vols. Jurnal Institut Teknologi Bandung bidang Teknik Elektro dan Informatika,, 2011.

[8] D. C. P. B. A. W. A. R. A. S. Ilham Kurniawan, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," vol. 10, no. 2, 2023.

[9] Y. W. N. C. Luthfiyah Amatullah, "Penerapan Klasifikasi Random Forest Terhadap Data Gangguan Spektrum Autisme (ASD) Pada Anak – Anak Menggunakan Seleksi Fitur Principal Component Analysis," 2022.