Halaman jurnal: http://jurnal.yudharta.ac.id/v2/index.php/EXPLORE-IT/

Explore it!
JURNAL KEILMUAN DAN APLIKASI TEKNIK INFORMATIKA

Akreditasi Sinta 5 (S5) – SK No. 36/E/KPT/2019
p-ISSN 2086-3489 (Print)- e-ISSN 2549-354X (Online)
DOI https://doi.org/10.35891/explorit

Perbandingan TF-IDF dengan *Count Vectorization* Dalam *Content-Based Filtering* Rekomendasi Mobil Listrik

Muhammad Zaynurroyhan[1], Asriyanik[2], Agung Pambudi[3]

[1,2,3] Teknik Informatika, Universitas Muhammadiyah Sukabumi, Kota Sukabumi, Indonesia

email: [1] royhan114@ummi.ac.id, [2] asriyanik263@ummi.ac.id, [3] agungpambd@ummi.ac.id

INFO ARTIKEL

Sejarah Artikel:
Diterima 15 Mei 2023
Direvisi -
Disetujui 17 Juni 2023
Dipublikasi 18 Juni 2023

Katakunci:
Content-Based Filtering
K-Nearest Neighbor
Mobil Listrik

ABSTRAK

Mobil listrik mulai menjadi pilihan beberapa tahun terakhir ini, karakternya yang lebih ramah lingkungan serta biaya pemeliharaan yang lebih rendah daripada mobil konvensional menjadi alasan utama konsumen lebih memilihnya. Seiring meningkatnya minat konsumen, perusahaan besar banyak yang mulai memproduksi mobil listrik dengan berbagai spesifikasi seperti kapasitas baterainya juga jarak tempuhnya. Hal tersebut membuat konsumen diberikan banyak pilihan dalam memilih mobil listrik yang sesuai preferensinya. Penelitian ini ditujukan untuk mempermudah konsumen dalam memilih mobil listrik yang sesuai dengan preferensinya. Metode yang digunakan adalah metode Content-Based Filtering dari sistem rekomendasi yang berfokus memberikan rekomendasi berdasarkan deskripsi barang serta hal yang disukai konsumen, dari sisi pembentukan modelnya, untuk melihat metode pemodelan yang menghasilkan akurasi lebih baik, peneliti membandingkan metode TF-IDF dengan Count Vectorization. Dimanfaatkan algoritma K-Nearest Neighbor dalam menguji akurasi dari model sistem rekomendasi mobil listrik yang terbentuk. Hasil dari penelitian ini menunjukkan bahwa sistem rekomendasi dapat digunakan untuk merekomendasikan mobil listrik terhadap konsumen. Dari sisi akurasi, model Content-Based Filtering yang dibentuk menggunakan TF-IDF menunjukkan akurasi yang lebih kecil yaitu sebesar 64% dibanding model yang memanfaatkan Count Vectorization yaitu sebesar 75%..

Keyword:
Content-Based Filtering
K-Nearest Neighbor
Electric Cars

DOI Artikel:

ABSTRACT

Electric cars have become an option in recent years, with their more environmentally friendly character and lower maintenance costs than conventional cars being the main reasons consumers prefer them. Along with increasing consumer interest, many large companies have started producing electric cars with various specifications such as battery capacity and mileage. This makes consumers given many choices in choosing an electric car that suits their preferences. This research is intended to facilitate consumers in choosing an electric car that suits their preferences. The method used is the Content-Based Filtering method of the recommendation system which focuses on providing recommendations based on the description of goods and things that consumers prefer, in terms of model formation, to see which modeling method produces better accuracy, researchers compare the TF-IDF method with Count Vectorization. The K-Nearest Neighbor algorithm is used to test the accuracy of the electric car recommendation system model formed. The results of this study show that the recommendation system can be used to recommend electric cars to consumers. In terms of accuracy, the Content-Based Filtering model formed using TF-IDF shows a smaller accuracy of 64% than the model that utilizes Count Vectorization which is 75%.

1. Pendahuluan

Mobil listrik mulai menapaki kenaikan kepopulerannya beberapa tahun terakhir. Ada sekitar 10 juta mobil listrik di dunia saat ini, dimana 50% persen pasarnya terdapat di China. Hal ini masih bisa bertambah dikarenakan penjualannya meningkat setiap tahunnya [1]. Di Indonesia sendiri, pada survei yang dilakukan pada tahun 2022 menyebutkan bahwa sebanyak 31% dari 1002 responden berencana membeli mobil listrik, angka tersebut akan meningkat setiap tahunnya. Dalam survei tersebut responden memberikan pula alasan utama mengapa akan memilih mobil listrik di masa mendatang, yaitu lebih ramah lingkungan serta biaya pemeliharaannya yang lebih rendah [2].

Dengan kecenderungan pengguna untuk memilih mobil listrik meningkat, banyak produsen mobil berbahan bakar konvensional mulai memproduksi mobil listrik, dengan spesifikasi yang berbagai macam, seperti jarak tempuhnya, kapasitas baterai, juga daya isi ulangnya. Semakin banyaknya pilihan inilah, membuat pengguna memiliki banyak perbandingan dalam menentukan pilihan mobil listrik yang sesuai dengan kebutuhannya.

Oleh karena itu, agar mempermudah pengguna dalam melakukan pemilihan mobil listrik agar sesuai dengan kebutuhannya dapat dimanfaatkan sistem rekomendasi. Sistem rekomendasi bekerja dengan cara menghimpun informasi minat pengguna, informasi tersebut akan digunakan acuan oleh sistem yang terbentuk untuk memberikan prediksi barang atau produk yang mungkin diminati pengguna [3]. Salah satu metode dalam pemanfaatan sistem rekomendasi yaitu *Content-Based Filtering*, metode ini memberikan rekomendasi berdasarkan suatu barang yang disukai pengguna, berbeda dengan *Collaborative Filtering* yang memberikan rekomendasi atas penilaian yang diberikan pengguna lain [4].

Terdapat beberapa penelitian terdahulu yang telah melakukan pembahasan mengenai sistem rekomendasi, seperti yang dilakukan oleh Wijaya dan Alfian yang memanfaatkan sistem rekomendasi pada rekomendasi laptop, kesimpulan dari penelitian ini menunjukkan bahwa sistem rekomendasi hibrid

menunjukkan sistem rekomendasi yang lebih baik, namun dari sisi pemrosesan lebih baik metode *Content-Based Filtering* [5]. Adapun penelitian lain yang dilakukan Fahmi menunjukkan pemanfaatan *Content-Based Filtering* dengan TF-IDF (*Term Frequency-Inverse Document Frequency*) serta *Cosine Similarity* dalam merekomendasikan *event online* yang akurasinya diuji menggunakan algoritma SVM (*Support Vector Machine*), adapun hasil uji akurasinya menunjukkan angka sebesar 86% [6]. Selain itu, dalam penelitian yang dilakukan oleh Muliawan untuk memberikan rekomendasi hotel dengan data sebanyak 50 data, yang akurasinya diuji menggunakan algoritma KNN (K-*Nearest Neighbor*) menunjukkan akurasi sebesar 84.50%.

Dilihat dari tinjauan jurnal-jurnal di atas, dalam menjawab permasalahan yang dipaparkan yaitu sistem rekomendasi mobil listrik, dapat digunakan *Content-Based Filtering* serta metode K-*Nearest Neighbor* untuk menguji akurasi dari model yang terbentuk.

2. Kajian Teori

2.1 Content-Based Filtering

Metode ini merupakan salah satu jenis sistem rekomendasi dimana yang merekomendasikan barang/hal dari fitur, atribut, atau karakteristiknya. Cara kerjanya, ketika suatu atribut disukai atau berkorelasi dengan pengguna, maka atribut tersebut digunakan untuk menjadi acuan mencari hal yang serupa yang kemudian dijadikan rekomendasi terhadap pengguna bersangkutan. Dalam perhitungan kemiripan, metode ini dapat menggunakan *Cosine Similarity* [7]. *Cosine Similarity* akan menghasilkan angka yang menunjukkan kemiripan antara 2 dokumen ketika semakin mirip akan mendekati nilai 1 apabila tidak mirip mendekati nilai 0 [8]. Proses perhitungan dengan *Cosine Similarity* dilakukan setelah kata-kata dalam dokumen dihitung menggunakan TF-IDF atau *Count Vectorization* [6,9]. Berikut persamaannya.

$$Cos\ a\ = \frac{Q\ .\ D}{|Q|\ |D|} = \frac{\sum_{i=1}^{n} wqi\ \times\ wdij}{\sqrt{\sum_{i=1}^{n}(wqi)^2}\ \times\ \sqrt{\sum_{i=1}^{n}(wdij)^2}} \quad (1)$$

Keterangan:
Q = Vektor Q, dibandingkan kemiripannya
D = Vektor D, dibandingkan kemiripannya.
Q • D = dot product antara vektor Q dan D
|Q| = panjang vektor Q
|D| = panjang vektor D
|Q||D| = cross product antara |Q| dan |D|

2.2 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF ialah metode yang digunakan untuk memperoleh tingkat kepentingan suatu kata dalam dokumen [10]. Adapun cara kerjanya dimulai dengan menghitung jumlah kata yang muncul pada dokumen. Kemudian dihitung tingkat kepentingan dari kata-kata tersebut, apabila kata bersangkutan sering muncul, maka kata tersebut tingkat kepentingannya tinggi [7]. Berikut rumus perhitungannya.

$$W_{dt}\ =\ tf_{dt} \times idf_t \quad (2)$$

Keterangan:
Wdt = Bobot dokumen ke-d terhadap kata ke-t
tfdt = Banyaknya kata yang dicari pada sebuah dokumen
idft = *Inversed Document Frequency*
N = Total dokumen
df = Banyak dokumen yang mengandung kata yang dicari

2.3 *Count Vectorization*

Count Vectorization merupakan metode yang hanya menghitung kemunculan kata dalam dokumen, hal ini berbeda dengan TF-IDF yang juga melakukan pembobotan kata selain menghitung jumlah katanya. Adapun kekurangan dari metode ini ketika kata terkait jarang muncul maka bobot nya rendah, padahal dalam beberapa kasus kata yang jarang muncul pun bisa mempunyai kepentingan yang tinggi [11].

2.4 KNN (*K-Nearest Neighbor)*

KNN merupakan suatu metode klasifikasi. Metode klasifikasi merupakan metode yang mengelompokkan data berdasarkan label yang telah ditentukan sebelumnya [12]. KNN sendiri merupakan metode yang bekerja dengan mencari tetangga paling dekat dari data yang dievaluasi terhadap data latih [13]. Jarak data didefinisikan dengan angka 1 jika sangat mirip, apabila nilainya mendekati 1 artinya kemiripannya rendah [14]. Adapun persamaan untuk menghitung jarak antar datanya sebagai berikut.

$$\mathrm{D(x,y)}\ =\ \sqrt{\sum_{\mathrm{k=1}}^{\mathrm{n}}(V_{\mathrm{training}}\ -\ W_{\mathrm{testing}})^2}. \quad (3)$$

Keterangan:
$V_{training}$: Data latih.
$W_{testing}$ = Data uji.

2.5 *Confusion Matrix*

Metode ini merupakan metode yang digunakan untuk evaluasi seberapa baik model klasifikasi yang terbentuk dari algoritma yang digunakan. Dalam metode memuat beberapa hal seperti TP (*True Postive*) berarti nilai sebenarnya serta prediksinya benar, TN (*True Negative*) berarti nilai yang negatif diprediksi secara negatif, FP (*False Positive*) berarti nilai aktualnya melupakan negatif tetapi diprediksi secara positif, dan FN (*False Negative*) data aktualnya ialah positif tetapi diprediksi negatif [15]. Adapun ketika evaluasi dilakukan, *confusion matrix* menghasilkan beberapa hal, seperti *Accuracy*, *precision*, *recall*, dan f1-*score.* Adapun persamaan perhitungannya sebagai berikut.

$$Accuracy\ = \frac{\sum TP}{\sum Support} \quad (4)$$

$$Precision\ = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score = 2\frac{(Precision \times Recall)}{(Precision + Recall)} \quad (7)$$

3. Metodologi Penelitian

Metodologi penelitian yang dimanfaatkan yaitu menggunakan CRISP-DM (*Cross Industry Standard Process for Data Mining*). Metode memiliki lingkup cakupan yang luas, adapun metode ini digunakan dalam pengembangan model data mining [16], ada 6 tahapan dalam metode ini yaitu, *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, serta *deployment* [17]. Berikut penjelasan dari tahapan-tahapan penelitian tersebut.

3.1 *Business Understanding*

Dalam Tahapan ini peneliti melakukan analisis mengenai keadaan bisnis yang dihadapi juga memuat proses identifikasi hal yang dibutuhkan serta yang tersedia. Seperti metode data mining yang cocok juga menentukan metode untuk menentukan kriteria keberhasilan data mining.

3.2 *Data Understanding*

Proses pengumpulan data serta memuat proses untuk memahami data yang terhimpun.

3.3 *Data Preparation*

Proses pembersihan data, seperti membersihkan data dari nilai null, menghapus karakter huruf yang tidak diperlukan, dan menggabungkan seluruh data menjadi satu kolom.

3.4 *Modelling*

Dalam tahapan ini memuat pemodelan TF-IDF serta *Count Vectorization*. Selain kedua metode tersebut, digunakan *Cosine Similarity* untuk menghitung kemiripan dari dokumen yang dimodelkan dengan TF-IDF ataupun *Count Vectorization*.

3.5 *Evaluation*

Proses identifikasi kesesuaian antara model yang terbentuk dengan permasalahan yang dihadapi. Pada tahapan ini dilakukan perhitungan akurasi, presisi, *recall* serta f1-*score* dengan memanfaatkan *confusion matrix* untuk meninjau hasil klasifikasi yang terbentuk dari model TF-IDF juga *Count Vectorization* untuk melihat model mana yang menghasilkan akurasi yang lebih baik. Disini pula dimanfaatkan algoritma KNN untuk melakukan klasifikasi dari model TF-IDF serta *Count Vectorization*.

3.6 *Deployment*

Luaran yang berupa aplikasi berbasis web yang memuat model sistem rekomendasi yang telah terbentuk.

4. Hasil Uji Coba Dan Pembahasan

4.1 *Business Understanding*

Pada langkah ini, peneliti mencoba memahami masalah yang teridentifikasi pada penelitian, yaitu semakin meningkatnya minat pengguna untuk menggunakan mobil listrik, maka banyak produsen mulai memproduksi mobil listrik dengan berbagai klasifikasi, sehingga membuat konsumen mempunyai banyak pilihan. Dapat dimanfaatkan *Content-Based Filtering* pada mobil listrik untuk membantu konsumen dalam memilih mobil yang sesuai dengan preferensinya. KNN digunakan untuk melakukan evaluasi terhadap model yang terbentuk, untuk melihat akurasi klasifikasinya dimanfaatkan *Confusion Matrix.*

4.2 *Data Understanding*

Pada tahapan ini peneliti melakukan *web scraping* dari situs motorwatt.com didapatkan data mobil listrik sebanyak 319 data. Sampelnya dapat dilihat pada gambar 1. terdapat beberapa kolom data dimulai dari *Manufacturer, Model, Price, Manufactured_in, Range, Max_speed, release_year, 0 to 100, Horsepower, Battery_capacity, Car_type, dan Drive_type.*

	A	B	C	D	E	F	G	H	I	J	K	L
1	Manufacturer	Model	Price	Manufactured_in	Range	Max_speed	Release_year	0 to 100	Horsepower	Battery_capacity	Car_type	Drive_type
2	Wuling	Wuling Air EV RWD 26.7 kWh	9700,00 $	China	300	100	2022	unknown	68	26.7	hatchback / 3 doors	RWD
3	Vauxhall	Vauxhall Mokka FWD 50 kWh	39700,00 $	United Kingdom	336	150	2021	9	134	50	SUV / 5 doors	FWD
4	Vauxhall	Vauxhall Corsa Electric FWD 50 kWh	36000,00 $	United Kingdom	357	150	2020	7.6	136	50	hatchback / 3 doors	FWD
5	WELTMEISTER	Weltmeister W6 FWD 69 kWh	30100,00 $	China	620	160	2021	8.8	218	69	SUV / 5 doors	FWD
6	WELTMEISTER	Weltmeister EX6 FWD 69 kWh	37000,00 $	China	505	160	2019	9.5	214	69	SUV / 5 doors	FWD
7	WELTMEISTER	Weltmeister E5 FWD 58.6 kWh	21000,00 $	China	505	170	2021	8.9	215	58.6	sedan	FWD
8	Estrema	Estrema Fulminea AWD 100 kWh	2300000,00 $	Italy	520	415	2023	1.8	2040	100	coupe	AWD
9	XEV	XEV YOYO FWD 10.3 kWh	13500,00 $	China	150	80	2021	unknown	20	10.3	roadster	FWD
10	SMART	SMART #1 RWD 64 kWh	42700,00 $	China	440	180	2022	6.7	286	64	SUV / 5 doors	RWD

Gambar 1. Sampel Dataset Mobil Listrik

Setelah data terhimpun, untuk lebih memahami data, dilakukan analisis atribut data.

Tabel 1: Analisis Atribut Data

No	Atribut	Keterangan
1	*Manufacturer*	Nama produsen mobil
2	*Model*	Nama model mobil
3	*Price*	Harga mobil
4	*Manufactured_in*	Negara asal mobil dibuat
5	*Range*	Jarak tempuh maksimal mobil
6	*Max_speed*	Kecepatan maksimal mobil
7	*Release_year*	Tahun rilis mobil
8	*0 to 100*	Waktu tempuh mobil dari 0 ke 100
9	*Horsepower*	Kecepatan kuda mobil
10	*Battery_capacity*	Kapasitas baterai mobil
11	*Car_type*	Jenis mobil
12	*Drive_type*	Jenis penggerak roda mobil

4.3 *Data Preparation*

Tahapan ini memuat proses pembersihan data, kemudian dilanjutkan dengan menggabungkan seluruh atribut mulai dari atribut Manufacturer hingga Car_type. Di bawah ini ditunjukkan hasil penggabungan seluruh atribut data yang telah dibersihkan.

Tabel 2: Hasil Penggabungan Atribut Data

No	Atribut
1	vauxhall vauxhall mokka 39700 united kingdom 336.0 150.0 2021.0 9.0 134.0 50.0 suv 5 doors fwd
2	vauxhall vauxhall corsa electric 36000 united kingdom 357.0 150.0 2020.0 7.6 136.0 50.0 hatchback 3 doors fwd
3	weltmeister weltmeister w6 30100 china 620.0 160.0 2021.0 8.8 218.0 69.0 suv 5 doors fwd
4	weltmeister weltmeister ex6 37000 china 505.0 160.0 2019.0 9.5 214.0 69.0 suv 5 doors fwd
5	weltmeister weltmeister e5 21000 china 505.0 170.0 2021.0 8.9 215.0 58.6 sedan fwd

4.4 *Modelling*

Pada tahapan ini dilakukan pemodelan dengan menggunakan TF-IDF serta *Count Vectorization*, dimana tingkat kesamaannya dihitung dengan menggunakan *Cosine Similarity*. Berikut perhitungannya.

1. Perhitungan Menggunakan TF-IDF

   Perhitungan ini dimulai dengan menghitung kata yang muncul pada setiap dokumen, disini digunakan dokumen 1, 2 serta 3 dari tabel 2 berikut hasilnya. Dapat dilihat bahwa peneliti menghitung tingkat kesamaan dokumen 2 dan 3 terhadap dokumen 1 dimana hasilnya menunjukkan bahwa dari kedua dokumen tersebut dokumen 2 lebih mirip dengan dokumen 1 dengan tingkat kemiripan 0.516.

Tabel 3: Perhitungan TF-IDF

No	Kata	Dokumen		DF	IDF	TF.IDF		Wqi*Wdij
		1	2			Dokumen		
						1	2	
1	134	1	0	1	1.693147	1.693147	0	0
2	136	0	1	1	1.693147	0	1.693147	0
3	150	1	1	2	1	1	1	1
4	2020	0	1	1	1.693147	0	1.693147	0
5	2021	1	0	1	1.693147	1.693147	0	0
6	336	1	0	1	1.693147	1.693147	0	0
7	357	0	1	1	1.693147	0	1.693147	0
8	36000	0	1	1	1.693147	0	1.693147	0
9	39700	1	0	1	1.693147	1.693147	0	0
10	50	1	1	2	1	1	1	1
11	corsa	0	1	1	1.693147	0	1.693147	0
12	doors	1	1	2	1	1	1	1
13	electric	0	1	1	1.693147	0	1.693147	0
14	fwd	1	1	2	1	1	1	1
15	hatchback	0	1	1	1.693147	0	1.693147	0
16	kingdom	1	1	2	1	1	1	1
17	mokka	1	0	1	1.693147	1.693147	0	0
18	suv	1	0	1	1.693147	1.693147	0	0
19	united	1	1	2	1	1	1	1
20	vauxhall	2	2	2	1	2	2	4
						27.20048	30.06723	10
						5.215408	5.48336	28.59796
						Cosine Similarity =		0.349675

2. Perhitungan Menggunakan *Count Vectorization*

   Proses perhitungan dengan menggunakan *Count Vectorization* dapat dilakukan dengan cara seperti TF-IDF, dimana perbedaannya hanya menjumlahkan kata yang muncul pada dokumen lalu menjadikannya vektor. Maka dari itu bisa dimanfaatkan perhitungan sebelumnya yang ditunjukkan tabel 3. Pada perhitungan ini hanya digunakan dokumen 1 dan dokumen 2. Dari tabel di atas dapat dikatakan bahwa vektor yang terbentuk ialah. D1 = [1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 2] dan D2 = [0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 2]. Selanjutnya dihitung dengan *Cosine Similarity*, prosesnya dapat dilihat di bawah ini.

$$D1 \times D2 = (1\times0)+(0\times1)+(1\times1)+(0\times1)+(1\times0)+(1\times0)+(0\times1)+(0\times1)+(1\times0)+(1\times1)+(0\times1)+(1\times1)+(0\times1)+(1\times1)+(0\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times1)+(2\times2)$$

$$= 10$$

$$\sqrt{D1^2} = \sqrt{1^2+0^2+1^2+0^2+1^2+1^2+0^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2+0^2+1^2+1^2+1^2+1^2+2^2}$$

$$= 4$$

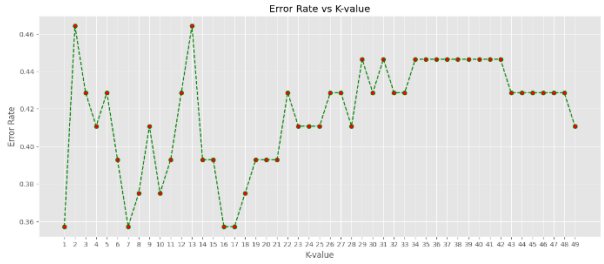
$$\sqrt{D2^2} = \sqrt{0^2+1^2+1^2+1^2+0^2+0^2+1^2+1^2+0^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0^2+0^2+1^2+2^2}$$

$$= 4.123$$

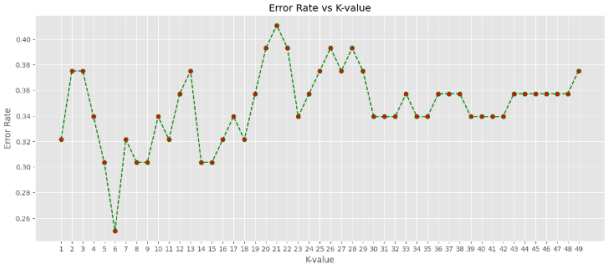
$$Similarity\ (A,B) = \frac{10}{4 \times 4.123} = 0.606$$

4.5 *Evaluation*

Pada langkah ini peneliti membagi data menjadi data *train* sebesar 80% juga data uji sebanyak 20% dengan kolom data target adalah Car_type. Selain itu, evaluasi yang memanfaatkan KNN ini dilakukan terhadap model yang terbentuk dengan TF-IDF serta *Count Vectorization*, akan tetapi sebelum menghitung nilai evaluasinya, peneliti meninjau pemilihan nilai tetangga atau K paling kecil nilai errornya.

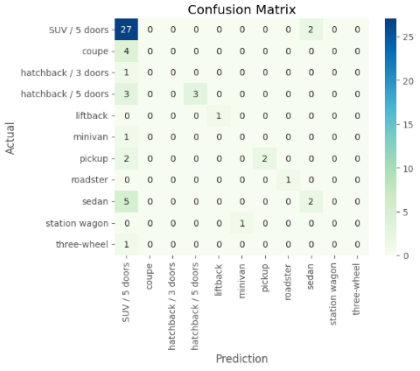


Gambar 2. Peninjauan Nilai K TF-IDF



Gambar 3. Peninjauan Nilai K *Count Vectorization*

Dapat dilihat gambar 2 menunjukkan bahwa nilai k paling kecil errornya ditunjukkan oleh k=7, sedangkan gambar 3 menunjukkan nilai k pada model *Count Vectorizaiton* paling kecil errornya ditunjukkan oleh k=6. Jika nilai k sudah ditentukan maka evaluasi hasil klasifikasinya dapat dilaksanakan. Berikut hasilnya ditunjukkan dalam bentuk *confusion matrix*.



Gambar 4. Hasil Klasifikasi TF-IDF

Dari *confusion matrix* pada gambar 4 di atas, dapat dilakukan perhitungan *accuracy*, *precision*, *recall*, juga *f1-score.* berikut tahapan perhitungannya.

$$Accuracy = \frac{27+0+0+3+1+0+2+1+2+0+0}{56}$$

$$= 0.64$$

$$Precision\ label\ pickup = \frac{2}{2+0+0+0+0+0+0+0+0+0+0}$$

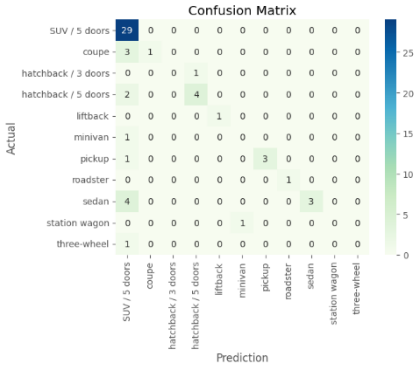
$$= \frac{2}{2} = 1$$

$$Recall\ label\ pickup = \frac{2}{2+2} = 0.5$$

$$F1 - Score\ label\ pickup = 2\frac{(1 \times 0.5)}{(1+0.5)}$$

$$= 2 \times 0.33 = 0.66$$

Di bawah ini pada gambar 5 merupakan hasil evaluasi klasifikasi pada model yang terbentuk menggunakan *Count Vectorization*.



Gambar 5. Hasil Klasifikasi *Count Vectorization*

Dari *confusion matrix* pada gambar 5 di atas, dapat dilakukan perhitungan *accuracy*, *precision*, *recall*, juga *f1-score.* berikut tahapan perhitungannya.

$$Accuracy = \frac{29+1+0+4+1+0+3+1+3+0+0}{56}$$

$$= 0.75$$

$$Precision\ label\ coupe = \frac{1}{1+0+0+0+0+0+0+0+0+0+0}$$

$$= \frac{1}{1} = 1$$

$$Recall\ label\ coupe = \frac{1}{1+3} = 0.25$$

$$F1 - Score\ label\ coupe = 2\frac{(1 \times 0.25)}{(1+0.25)}$$

$$= 2 \times 0.2 = 0.4$$

Hasil keseluruhan dari perhitungan perhitungan sebelumnya ditunjukkan pada gambar 6 untuk model TF-IDF serta gambar 7 untuk model *Count Vectorization*. Dapat dilihat pada gambar-gambar tersebut, akurasi yang ditunjukkan berbeda dimana model yang menggunakan TF-IDF akurasinya adalah 0.64, sedangkan yang menggunakan *Count Vectoriuzation* 0.75.

Classification Report TF-IDF Matrix :

	precision	recall	f1-score	support
SUV / 5 doors	0.61	0.93	0.74	29
coupe	0.00	0.00	0.00	4
hatchback / 3 doors	0.00	0.00	0.00	1
hatchback / 5 doors	1.00	0.50	0.67	6
liftback	1.00	1.00	1.00	1
minivan	0.00	0.00	0.00	1
pickup	1.00	0.50	0.67	4
roadster	1.00	1.00	1.00	1
sedan	0.50	0.29	0.36	7
station wagon	0.00	0.00	0.00	1
three-wheel	0.00	0.00	0.00	1
accuracy			0.64	56
macro avg	0.46	0.38	0.40	56
weighted avg	0.59	0.64	0.58	56

Gambar 6. Keseluruhan Evaluasi TF-IDF

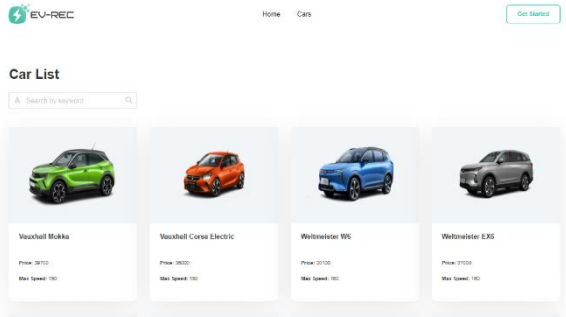
Classification Report :

	precision	recall	f1-score	support
SUV / 5 doors	0.71	1.00	0.83	29
coupe	1.00	0.25	0.40	4
hatchback / 3 doors	0.00	0.00	0.00	1
hatchback / 5 doors	0.80	0.67	0.73	6
liftback	1.00	1.00	1.00	1
minivan	0.00	0.00	0.00	1
pickup	1.00	0.75	0.86	4
roadster	1.00	1.00	1.00	1
sedan	1.00	0.43	0.60	7
station wagon	0.00	0.00	0.00	1
three-wheel	0.00	0.00	0.00	1
accuracy			0.75	56
macro avg	0.59	0.46	0.49	56
weighted avg	0.76	0.75	0.71	56

Gambar 7. Keseluruhan Evaluasi *Count Vectorization*

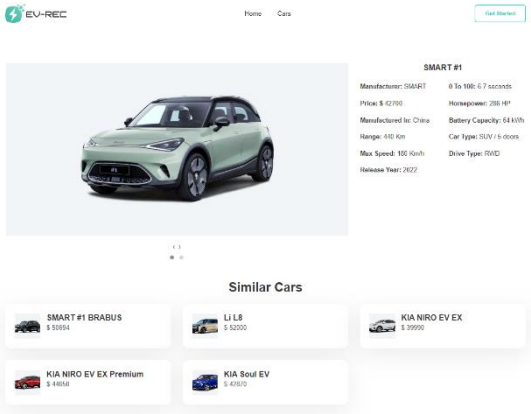
4.6 *Deployment*

Pada tahapan ini peneliti membuat suatu aplikasi berbasis *website* sebagai implementasi model yang terbentuk. Aplikasi ini dapat memberikan rekomendasi terhadap pengguna mengenai mobil listrik yang sesuai dengan preferensinya. Pada gambar 8 ditunjukkan halaman yang memuat seluruh mobil yang ada pada dataset.



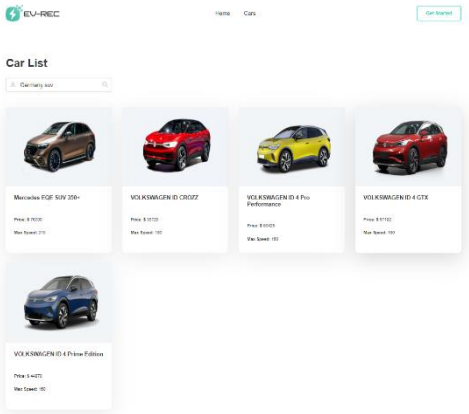
Gambar 8. Halaman Mobil Listrik

Sistem akan memberikan rekomendasi ketika pengguna memilih satu dari sekian banyak mobil listrik yang dimuat, ketika salah satu mobil listrik dipilih, aplikasi akan mengarahkan ke halaman yang memuat detail mobil listrik yang dipilih serta pada bagian bawahnya ditunjukkan beberapa mobil listrik yang memiliki tingkat kemiripan serupa dengan mobil listrik yang dipilih seperti yang ditunjukkan gambar 9.



Gambar 9. Halaman Detail

Selain rekomendasi bisa didapat dengan cara memilih salah satu mobil listrik, aplikasi juga menyediakan halaman pencarian untuk mencari mobil listrik yang sesuai dengan kata kunci yang dimasukkan oleh pengguna, pada gambar di 10 dapat dilihat bahwa peneliti memasukkan kata Germany suv.



Gambar 10. Rekomendasi KataKunci

5. Kesimpulan

Berdasarkan hasil dan pembahasan yang dipaparkan di atas, sistem rekomendasi menggunakan metode *Content-Based Filtering* bisa digunakan untuk merekomendasikan mobil listrik sesuai dengan preferensi konsumen. Akan tetapi dari sisi performa, yang memanfaatkan *Count Vectorizaiton* lebih baik dari sisi akurasi yang dihitung menggunakan KNN serta *Confusion Matrix* dimana menunjukkan akurasi sebesar 75% dibanding dengan yang menggunakan TF-IDF yang akurasinya hanya 64%.

6. Daftar Pustaka

[1] Dik A, Omer S, Boukhanouf R. Electric Vehicles: V2G for Rapid, Safe, and Green EV Penetration. Energies 2022;15. https://doi.org/10.3390/en15030803.

[2] Team E. Indonesians Consider Buying Electric Vehicles in 5 Years: Survey. D-Insights 2022. https://dinsights.katadata.co.id/read/2022/03/02/indonesians-consider-buying-electric-vehicles-in-5-years-survey?__cf_chl_tk=sszJMTYnO8X6AzQUptnOEU9W1fo52ayIXDO_UMEPF1Q-1674389343-0-gaNycGzNCP0 (accessed September 4, 2023).

[3] Anggela SH, Santoso LW, Andjarwirawan J. Sistem Rekomendasi Pembelian Laptop dengan K-Nearest Neighbor (KNN). J Infra 2022;10:254–60.

[4] Mondi RH, Wijayanto A, Winarno. Recommendation System With Content-Based Filtering Method for Culinary Tourism in Mangan Application. ITSMART J Ilm Teknol Dan Inf 2019;8:65–72.

[5] Wijaya AE, Alfian D. Sistem Rekomendasi Laptop Menggunakan Collaborative Filtering dan Content-Based Filtering. J Comput Bisnis 2018;12:11–27.

[6] Nurfalah F, Asriyanik A, Pambudi A. Sistem Rekomendasi Event Online Menggunakan Metode Content Based Filtering. J Ilm Elektron DAN Komput 2022;15:271–9.

[7] Pramesti DAPD, Santiyasa IW. Penerapan Metode Content-Based Filtering dalam Sistem Rekomendasi Video Game. JNATIA 2022;1:229–34.

[8] Suryani L, Edy K. Pengembangan Aplikasi “Lost & Found” Berbasis Android Dengan Menggunakan Metode Term Frequency – Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity. Electro Luceat 2020;6:190–204. https://doi.org/10.32531/jelekn.v6i2.232.

[9] Pradana DS, Prajoko P, Hartawan GP. Perbandingan Algoritma Content-Based Filtering dan Collaborative Filtering dalam Rekomendasi Kegiatan Ekstrakurikuler Siswa. Progresif J Ilm Komput 2022;18:151. https://doi.org/10.35889/progresif.v18i2.854.

[10] Sjarif NNA, Azmi NFM, Chuprat S, Sarkan HM, Yahya Y, Sam SM. SMS Spam Sessage Detection Using Term Fequency-Inverse Document Frequency and Random Forest Algorithm. Procedia Comput Sci 2019;161:509–15. https://doi.org/10.1016/j.procs.2019.11.150.

[11] Wendland A, Zenere M, Niemann J. Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique. Springer 2021;28:289–300.

[12] Putry NM, Sari BN. Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus. EVOLUSI J Sains Dan Manaj 2022;10. https://doi.org/10.31294/evolusi.v10i1.12514.

[13] Nikmatun IA, Waspada I. Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. J SIMETRIS 2019;10:421–32.

[14] Hidayatullah T, Wibisono S. Pembobotan Atribut Dengan Pairwise Comparison Pada Case Based Reasoning Deteksi Dini Penyakit Gigi Menggunakan KNN. Explore IT 2022;5:17–23.

[15] Amaliah S, Nusrang M, Aswi A. Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. VARIANSI J Stat Its Appl Teach Res 2022;4:121–7. https://doi.org/10.35580/variansiunm31.

[16] Martinez-Plumed F, Contreras-Ochando L, Ferri C, Hernandez-Orallo J, Kull M, Lachiche N, et al. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans Knowl Data Eng 2021;33:3048–61. https://doi.org/10.1109/TKDE.2019.2962680.

[17] Schröer C, Kruse F, Gómez JM. A systematic literature review on applying CRISP-DM process model. Procedia Comput Sci 2021;181:526–34. https://doi.org/10.1016/j.procs.2021.01.199.