

Komparasi Naïve Bayes dan Support Vector Machine dalam Klasifikasi Jenis Citrus

Comparison of Naïve Bayes and Support Vector Machine in Citrus Species Classification

I Wayan Rangga Pinastawa¹, Nurul Afifah Arifuddin²

^{1,2} Informatika, Universitas Pembangunan Nasional Veteran Jakarta

E-mail: ¹rangga@upnvj.ac.id, ² nurulafifaharifuddin@upnvj.ac.id

Abstrak

Citrus merupakan pohon berbunga dan tergolong dalam kelompok Rutaceae. Pohon Citrus menghasilkan buah jeruk dengan berbagai jenis buah-buahan. Karena kesamaan spesies sehingga antar jenisnya memiliki kemiripan satu sama lain, dan tidak semua dapat melakukan identifikasi secara jelas setiap jenis buahnya. Cara yang dapat dilakukan untuk melakukan identifikasi dan pengelompokan adalah mengelompokkan data sesuai dengan kelas label aslinya menggunakan bantuan data mining. Pendekatan data mining yang dapat diterapkan salah satunya dengan teknik klasifikasi, dengan melakukan pengelompokan berdasarkan kriteria atau kategori tertentu. Pada hal ini, klasifikasi didasarkan pada diameter, dan citra warna Red, Green, Blue atau RGB untuk mendapatkan pengelompokan sesuai dengan kelasnya. Algoritma yang digunakan ada 2 yakni, Support Vector Machine (SVM) dan Naïve Bayes, keduanya akan dilakukan perbandingan dalam melakukan klasifikasi pengelompokan jenis buah citrus. Teknik komparasi dilakukan dengan mengamati hasil akurasi dari setiap algoritma klasifikasi, penelitian ini menyimpulkan bahwa akurasi algoritma Support Vector Machine (SVM) mencapai prosentase sebesar 96,36 % dan algoritma Naïve Bayes memiliki akurasi sebesar 92 %. Algoritma paling optimal dalam penelitian klasifikasi citrus ini adalah algoritma Support Vector Machine (SVM).

Kata kunci: Citrus, Klasifikasi, Support Vector Machine, Naïve Bayes

Abstract

Citrus is a flowering tree and belongs to the Rutaceae group. Citrus trees produce citrus fruits with various types of fruits. Due to the similarity of species, the types of fruit are similar to each other, and not all of them can identify each type of fruit. The way that can be done to carry out identification and grouping is to group data according to the original label class using the help of data mining. One of the data mining approaches that can be applied is a classification technique, by grouping based on certain criteria or categories. In this case, the classification is based on diameter, and red, green, blue or RGB ,color images to get grouping grouped to class. There are 2 algorithms used, namely, Support Vector Machine (SVM) and Naïve Bayes, both of which will be compared in classifying citrus fruit types. The comparison technique is carried out by observing the results of the accuracy of each classification algorithm. This study concludes that the accuracy of the Support Vector Machine (SVM) algorithm reaches a percentage of 96.36% and the Naïve Bayes algorithm has an accuracy of 92%. The most optimal algorithm in this citrus classification research is the Support Vector Machine (SVM) algorithm.

Keywords: Citrus, Classification, Support Vector Machine, Naïve Bayes

1. PENDAHULUAN

Citrus merupakan pohon berbunga dan tergolong dalam kelompok Rutaceae. Pohon Citrus menghasilkan buah jeruk dengan berbagai jenis seperti jeruk, lemon, jeruk bali, pomelo dan limau. Dikarenakan masih dalam satu keluarga atau spesies yang sama, maka jenis jenis

dari pohon citrus ini memiliki kedekatan dan kemiripan, meskipun ada hal yang bisa menjadi penciri dan pembeda di setiap jenisnya. Bagi sebagian orang, mengidentifikasi antar jenis dari citrus ini memerlukan waktu dan tidak semua orang mengetahui jenis dari buah pohon citrus. Salah satu cara identifikasi untuk menggolongkan jenis dari pohon citrus adalah dari citra warnanya. Citra warna yang digunakan adalah Red, Green, Blue untuk mendefinisikan dan mengidentifikasi klasifikasi jenis citrus.

Cara yang dapat dilakukan untuk melakukan identifikasi dan pengelompokan adalah dengan menggunakan pendekatan data mining [1]. Ada banyak jenis pendekatan yang dapat dilakukan dengan data mining, salah satunya adalah dengan teknik klasifikasi [2]. Pada penelitian ini akan dilakukan teknik klasifikasi untuk mengelompokkan data berdasarkan jenis atau kelasnya, adapun data yang dikelompokkan adalah data pohon citrus. Pohon citrus diidentifikasi berdasarkan data citra RGB dan diameternya untuk melakukan pengelompokan sesuai dengan jenisnya. Dalam klasifikasi ada banyak jenis algoritma dengan berbagai kegunaan seperti untuk melakukan tugas pengelompokan data, adapun algoritma terpilih untuk penelitian ini adalah Naïve Bayes Classifier dan Support Vector Machine (SVM). Algoritma klasifikasi terpilih secara keseluruhan memiliki hasil dan presisi dengan kategori baik dalam melakukan klasifikasi.

Dalam rangka menguatkan pemilihan algoritma dan tema penelitian dilakukan rujukan referensi terdahulu dari beberapa rujukan literatur referensi sejenis. Rujukan penelitian pertama adalah komparasi antara algoritma SVM dan Naïve Bayes pada kasus klasifikasi peluang penyakit jantung dengan hasil akhir penelitian berupa akurasi yang dihasilkan metode Naïve Bayes sebesar 84% dan metode Support Vector Machine sebesar 87% [3]. Hal senada juga terjadi pada penelitian Hennie Tuhuteru dengan membandingkan kedua algoritma ini dalam melakukan klasifikasi terhadap sentiment yang terjadi di social media twitter, hasilnya SVM memiliki nilai akurasi sebesar 76,42% dan Naïve Bayes Classifier akurasi terbaik pada K-Fold 2 sebesar 67,20% [4]. Referensi selanjutnya masih menunjukkan hal yang serupa, dengan keunggulan pada metode SVM dibandingkan dengan Naïve Bayes dengan hasil akurasi sebesar 80% untuk deteksi Intrusion Detection System (IDS) [5]. Terkait deteksi dengan Intrusion Detection System (IDS) ada penelitian lain yang menunjukkan bahwa kombinasi kedua algoritma dapat menghasilkan akurasi yang baik pada ketiga dataset yang digunakan dengan besaran paling tinggi pada Dataset NSL-KDD dengan hasil sebesar 99,35% [6].

Pada penelitian yang dilakukan pada tahun 2018 untuk melakukan prediksi keberhasilan imunoterapi pada penyakit kutil menghasilkan kesimpulan yang berbeda dari referensi sebelumnya, hasil akhir penelitian ini menghasilkan akurasi yang lebih baik oleh algoritma Naïve Bayes dibandingkan dengan algoritma Support Vector Machine sebesar 80% [7]. Pada penelitian yang lain untuk klasifikasi jalur minat di SMA menghasilkan akurasi terbaik pada algoritma Support Vector Machine. Adapun detail penggunaan algoritma pada penelitian ini berhasil melakukan klasifikasi lebih baik dibandingkan dengan Naïve Bayes [8]. Selain dilakukan komparasi kedua metode ini juga dapat dikombinasikan dengan contoh penelitian Hajah T Sueno dengan hasil tulisan bahwa Naïve Bayes mampu memberikan dampak pada data tekstual dan menjadikan proses transformasi ke SVM menjadi lebih efektif [9]. Komparasi juga memiliki hasil yang menunjukkan SVM memiliki tingkat akurasi yang tinggi dari penulis Yogendra Narayan dengan hasil akurasi sebesar 98,58% [10].

Selain dari sisi akurasi, dari sisi waktu juga dapat dilakukan komparasi seperti penelitian yang melakukan perbandingan klasifikasi opini di social media twitter terkait topik vaksin, hasilnya Naïve Bayes memiliki waktu yang lebih cepat dibandingkan dengan SVM [11]. Selain kecepatan, Naïve Bayes juga beberapa kali memiliki tingkat akurasi yang lebih baik pada beberapa kasus seperti, pada penelitian sentiment analisis dari kandidat Gubernur Jawa Barat periode 2018-2023 dengan hasil akurasi tertinggi Naïve Bayes Classifier sebesar 98% [12]. Namun naïve bayes juga tidak selalu menghasilkan akurasi terbaik dan waktu eksekusi lebih singkat seperti pada komparasinya dengan KNN dan C.45 bahwa KNN memiliki akurasi yang lebih baik [13] begitu juga C.45 memiliki hasil yang lebih superior [14]. Pada kasus di Amazon group sentiment analisis yang dilakukan kedua algoritma menghasilkan kesimpulan akhir

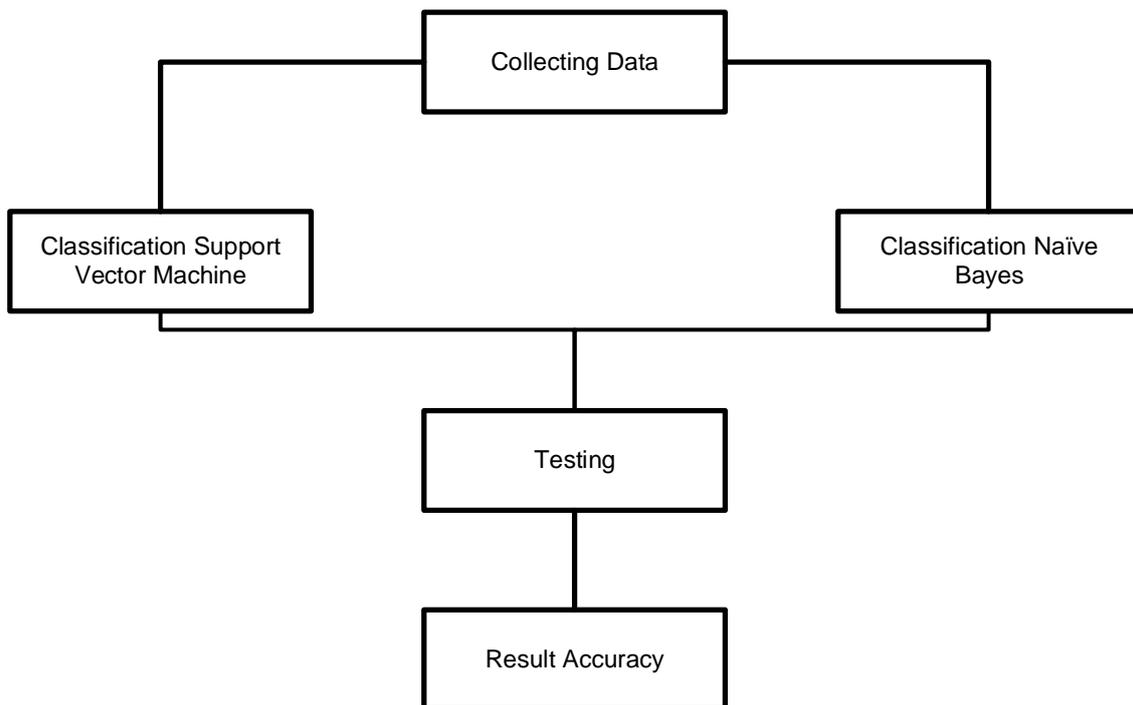
berupa SVM memiliki hasil polarisasi feedback produk amazon dengan hasil lebih tinggi [15]. Pada penelitian yang akan dilakukan, proses identifikasi berdasarkan citra warna dapat dilakukan dengan menghitung nilai kedekatan dari masing-masing komponen warna dan dilakukan komparasi hasil performansi dari kedua algoritma. Hal ini menjadi kontribusi penelitian ini, dengan raatan rujukan klasifikasi dalam konteks teks, klasifikasi dalam hal ini dilakukan pada data yang dapat dilakukan pada data yang mengarah ke klasifikasi data citra.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Langkah dalam penelitian dari awal sampai akhir ditunjukkan pada

Gambar 1.



Gambar 1. Alur Penelitian

Dalam detail alur penelitian adalah sebagai berikut :

Proses pengelompokan data sebagai dataset yang akan dilakukan untuk pengujian kedua algoritma, kemudian dilanjutkan dengan proses klasifikasi pada kedua algoritma, kedua algoritma akan melakukan klasifikasi berdasarkan kelas yang sudah tersedia dengan dataset yang dimiliki. Dalam melakukan klasifikasi kedua algoritma akan menghasilkan nilai akurasi dengan seberapa besar algoritma berhasil dalam menentukan kelas label sesuai dengan kelas aslinya.

2.2 Proses Klasifikasi Naïve Bayes

Proses yang dilakukan dalam melakukan klasifikasi Naïve Bayes dengan menggunakan rumus berikut ini :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Dari perumusan diatas perhitungan dilakukan dengan mengidentifikasi kelas yang belum diketahui dalam variable X, dan hipotesis atau dugaan sementara pada H dan probabilitas dari Hipotesis berdasar kondisi variable X. Langkah yang dilakukan pada dataset adalah dari gambaran dataset semula dalam bentuk asli seperti ditunjukkan pada Gambar 2.

	diameter	weight	red	green	blue
0	2.96	86.76	172	85	2
1	3.91	88.05	166	78	3
2	4.42	95.17	156	81	2
3	4.47	95.60	163	81	4
4	4.48	95.76	161	72	9

Gambar 2. Dataset Awal

Proses selanjutnya adalah dengan melakukan proses penentuan variable dependent dan independent. Proses dilakukan dengan drop variable independent dimana field name sebagai field dari label kelas, sehingga didapatkan hasil pada Gambar 3.

```
0    orange
1    orange
2    orange
3    orange
4    orange
Name: name, dtype: object
```

Gambar 3. Drop Variable

Setelah beberapa tahapan berikutnya berhasil dilakukan, maka berikut ini adalah langkah yang dilakukan dengan merubah ke dalam array pada Gambar 4.

```
Out[7]: array([3994., 4006.] )
```

Gambar 4. Array Visualisasi

Proses berikutnya tinggal dilakukan klasifikasi dan pengelompokan dari label kelas oleh algoritma Naïve Bayes.

2.3 Proses Klasifikasi Support Vector Machine

Tahapan klasifikasi dalam Support Vector Machine juga menggunakan dataset yang sama sebagai sumber data. Adapun persamaan yang digunakan dalam SVM adalah sebagai berikut :

$$f : w \cdot x + b = 0 \quad (2)$$

Dimana parameter dari hyperlane yang dicari (garis tegak lurus antara hyperlance dan titik support vector akan dikalikan dengan data input SVM dan dijumlahkan dari parameter hyperlane yang dicari atau nilai biasanya.

Kemudian, proses yang dilalui sedikit berbeda dengan menggunakan Naïve Bayes. Pertama, akan diidentifikasi mengenai tipe dari dataset tersebut seperti pada Gambar 5.

```

RangeIndex: 10000 entries, 0 to 9999
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   name        10000 non-null  object
 1   diameter    10000 non-null  float64
 2   weight      10000 non-null  float64
 3   red         10000 non-null  int64
 4   green       10000 non-null  int64
 5   blue        10000 non-null  int64
dtypes: float64(2), int64(3), object(1)
memory usage: 468.9+ KB

name        0
diameter    0
weight      0
red         0
green       0
blue        0
dtype: int64
    
```

Gambar 5. Tipe Data

Setelah proses tersebut dapat dilakukan visualisasi dan proses klasifikasi menggunakan Support Vector Machine.

3. HASIL DAN PEMBAHASAN

3.1 Datasets

Penghimpunan data atau data sumber pada penelitian ini adalah citrus yang memiliki kelas orange dan grapefruit. Dengan field pendukung seperti diameter, berat dan juga citra warna Red, Green, Blue. Adapun detail data ditunjukkan pada memiliki field pada Tabel 1.

Tabel 1. Dataset Penelitian

No	diameter	weight	red	green	blue	class
1	2.96	86.76	172	85	2	Orange

2	3.91	88.05	166	78	3	Orange
3	4.42	95.17	156	81	2	Orange
4	10.81	186.88	145	76	27	Grapefruit
5	10.92	188.85	156	64	13	Grapefruit
.....						
10.000	16.45	261.51	152	74	2	Grapefruit

Dari dataset yang dimiliki, dilakukan proses klasifikasi menggunakan masing-masing algoritma. Klasifikasi dilakukan dengan memprediksi kelas label sesuai dengan kelas label asli. Masing-masing algoritma akan memiliki tingkat presisi dan kebenaran masing-masing untuk setiap prediksi kelas yang dilakukan.

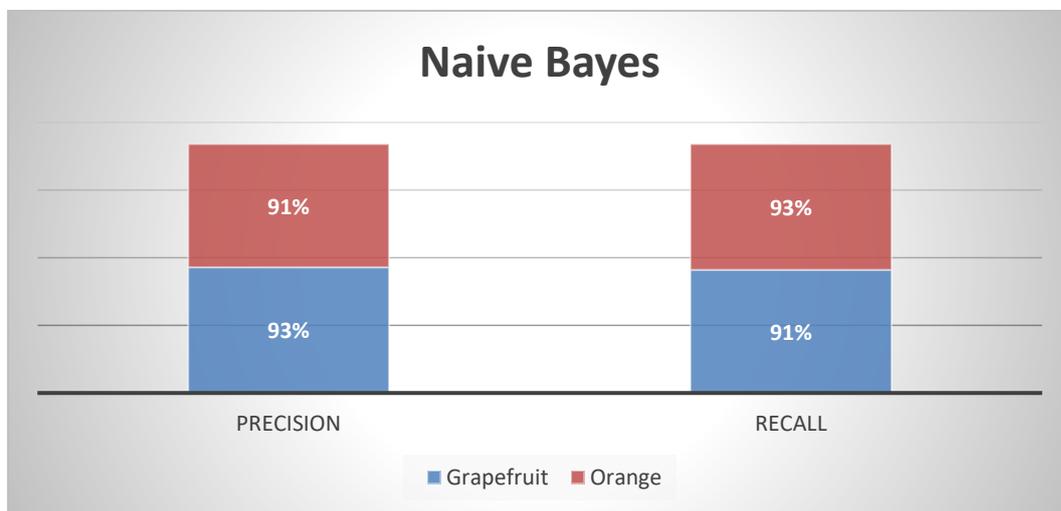
3.2 Naïve Bayes

Algoritma atau metode Naïve Bayes dalam melakukan klasifikasi melakukan proses penghitungan nilai probabilitas dalam bentuk array berikut ini :

```
array([[2.91609738e-03, 9.97083903e-01],
       [3.73046732e-07, 9.99999627e-01],
       [9.99997219e-01, 2.78116538e-06],
       ...,
       [5.50293494e-05, 9.99944971e-01],
       [9.99456946e-01, 5.43054300e-04],
       [1.58064899e-06, 9.99998419e-01]])
```

Gambar 6. Proses Array pada Naïve Bayes

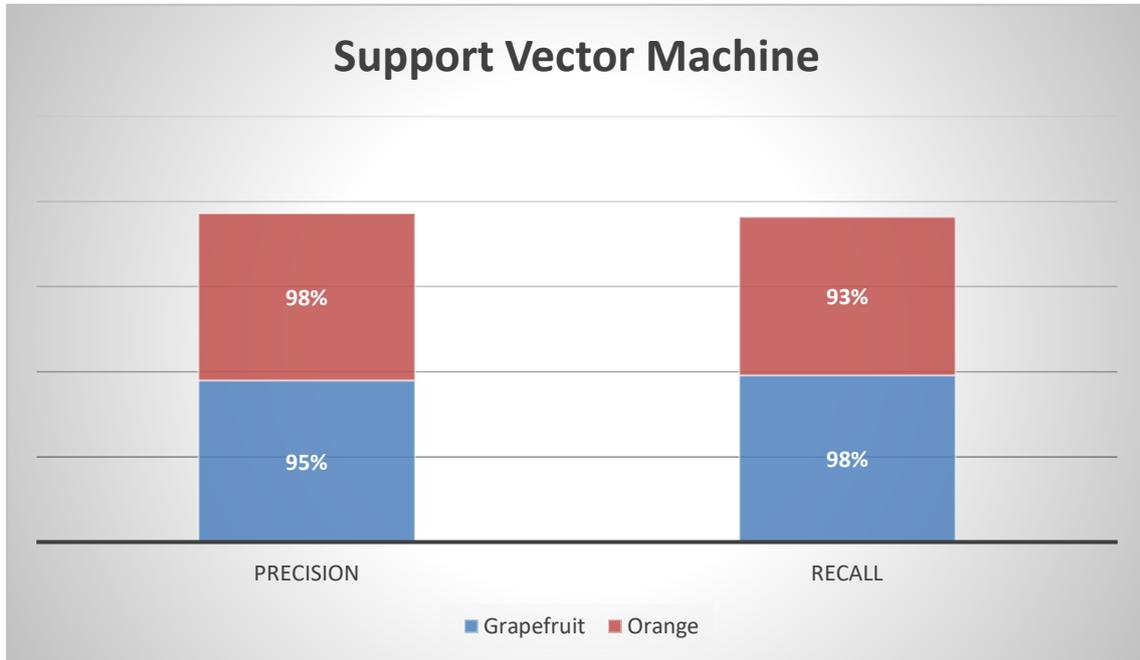
Setelah tahapan pengolahan data di algoritma naïve bayes maka akan menghasilkan akurasi dengan detail sebagai berikut.



Gambar 7. Nilai Precision Recall Naïve Bayes

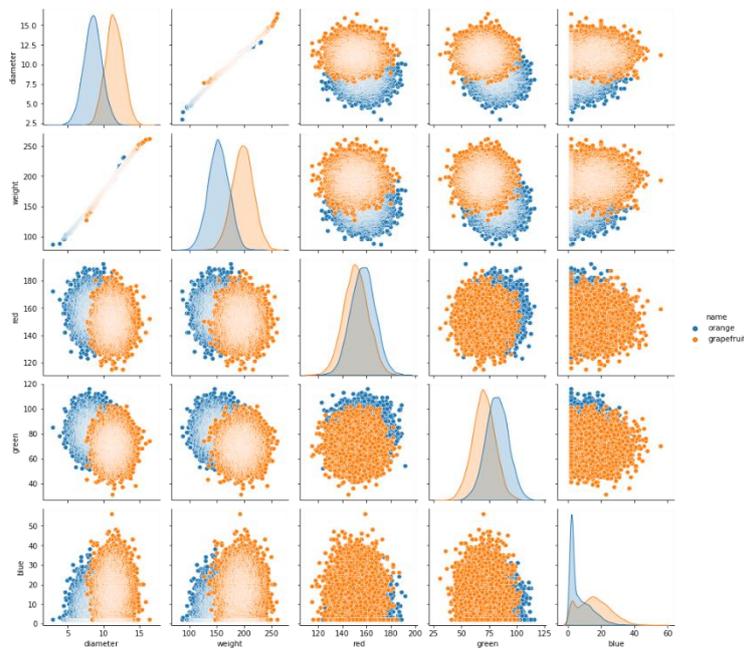
3.2 Support Vector Machine

Hasil penelitian ini dengan penerapan menggunakan algoritma Support Vector Machine menghasilkan akurasi dengan detail sebagai berikut.



Gambar 8. Nilai Precision Recall Support Vector Machine

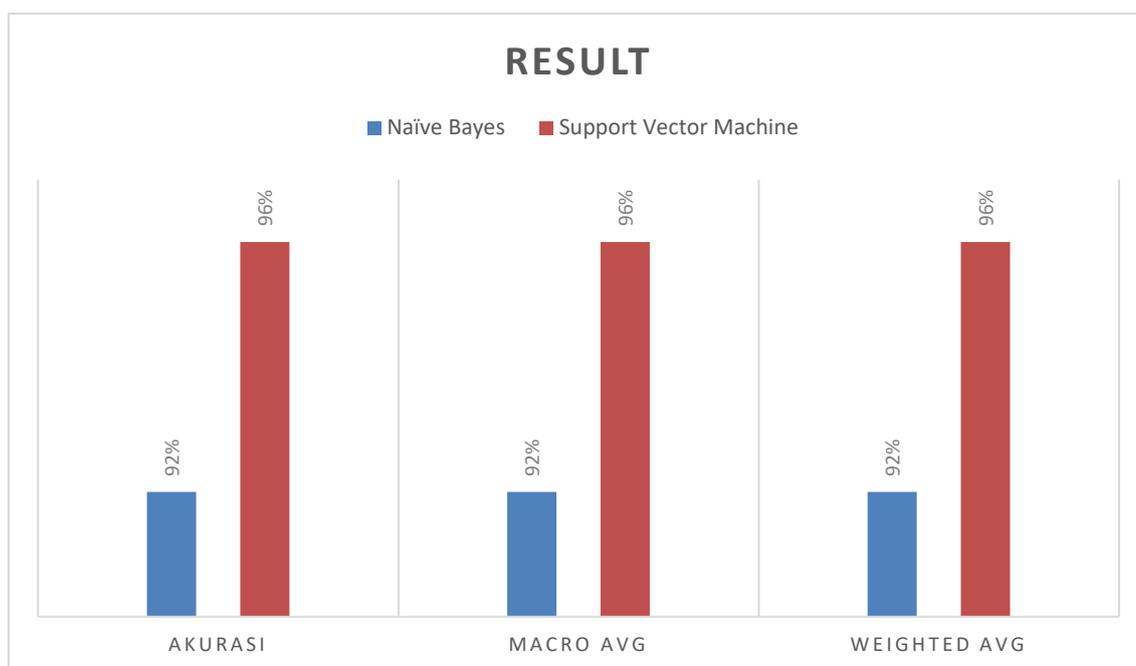
Data dapat pula divisualisasikan tiap field variable dalam scatter plot seperti pada gambar berikut ini.



Gambar 9. Scatter Plot Variable

3.3 Komparasi Hasil

Berdasarkan skenario uji yang sudah dilakukan maka hasil dari perbandingan kedua algoritma. Proses perbandingan dilakukan dengan mencari nilai akurasi dari masing-masing algoritma. Masing-masing algoritma akan menghasilkan nilai akurasi, precision dan recall. Hal ini dilakukan perbandingan satu sama lain yang di visualisasikan dalam bentuk grafik pada hasil berikut ini. Gambar 10.



Gambar 10. Komparasi Hasil Algoritma

Dari perbandingan yang dilakukan dan sudah divisualisasikan, maka algoritma Support Vector Machine memiliki hasil yang lebih baik dari algoritma Naïve Bayes dari sisi akurasi.

4. KESIMPULAN DAN SARAN

Penelitian yang sudah dilakukan menghasilkan kesimpulan kedua algoritma memiliki hasil yang cenderung baik dalam melakukan klasifikasi karena memiliki akurasi diatas 90%. Adapun detail akurasi pada Naïve Bayes Classifier sebesar 92% dan pada Support Vector Machine sebesar 96%. Adapun saran dalam penelitian ini, dapat ditambahkan variasi skenario uji lain untuk memperkaya hasil dan dapat ditindaklanjuti sebagai penelitian lanjutan.

DAFTAR PUSTAKA

- [1] V. Kale, *Enterprise Performance Intelligence and Decision Patterns*. CRC Press, 2018.
- [2] P. H. S. Musthofa Galih Pradana, Azriel Christian Nurcahyo, "PENGARUH SENTIMEN DI SOSIAL MEDIA DENGAN HARGA SAHAM PERUSAHAAN," *J. Ilm. Educ.*, vol. 6, no. 2, 2020.
- [3] P. H. S. Musthofa Galih Pradana, Dhina Puspasari Wijaya, "KOMPARASI METODE SUPPORTVECTORMACHINE DANNAÏVE BAYES DALAMKLASIFIKASI PELUANG PENYAKIT SERANGAN JANTUNG," *Indones. J. Bus. Intell.*, vol. 5, no. 2, pp. 87–91, 2022.
- [4] H. Tuhuteru and A. Iriani, "Analisis Sentimen Perusahaan Listrik Negara Cabang

- Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier,” *J. Inform. J. Pengemb. IT*, vol. 3, no. 3, pp. 394–401, 2018, doi: 10.30591/jpit.v3i3.977.
- [5] I. Riadi, R. Umar, and F. D. Aini, “Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (Svm),” *Ilk. J. Ilm.*, vol. 11, no. 1, pp. 17–24, 2019, doi: 10.33096/ilkom.v11i1.361.17-24.
- [6] J. Gu and S. Lu, “An effective intrusion detection approach using SVM with naïve Bayes feature embedding,” *Comput. Secure.*, vol. 103, p. 102158, 2021, doi: 10.1016/j.cose.2020.102158.
- [7] A. Supriyatna and W. P. Mustika, “Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 2, no. 2, p. 152, 2018, doi: 10.30645/j-sakti.v2i2.78.
- [8] O. Arifin and T. B. Sasongko, “Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier,” in *Seminar Nasional Teknologi Informasi dan Multimedia 2018*, 2018, vol. 6, no. 1, pp. 67–72.
- [9] H. T. Sueno, “Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3937–3944, 2020, doi: 10.30534/ijatcse/2020/216932020.
- [10] Y. Narayan, “Comparative analysis of SVM and Naive Bayes classifier for the SEMG signal classification,” *Mater. Today Proc.*, vol. 37, no. Part 2, pp. 3241–3245, 2020, doi: 10.1016/j.matpr.2020.09.093.
- [11] F. Fitriana, E. Utami, and H. Al Fatta, “Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes,” *J. Komtika (Komputasi dan Inform.*, vol. 5, no. 1, pp. 19–25, 2021, doi: 10.31603/komtika.v5i1.5185.
- [12] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, “Comparison of SVM Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter,” *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citism 2018, pp. 1–6, 2019, doi: 10.1109/CITSM.2018.8674352.
- [13] K. Kaharudin, Musthofa Galih Pradana, “PREDIKSI CUSTOMER CHURN PERUSAHAAN TELEKOMUNIKASI MENGGUNAKAN NAÏVE BAYES DAN K-NEAREST NEIGHBOR,” *Inf. Interaktif*, vol. 4, no. 3, 2019.
- [14] M. G. Pradana and P. H. Saputro, “Komparasi Metode Naïve Bayes Dan C4.5 Dalam Klasifikasi Loyalitas Pelanggan Terhadap Layanan Perusahaan,” *Indones. J. Bus. Intell.*, vol. 3, no. 1, p. 20, 2020, doi: 10.21927/ijubi.v3i1.1205.
- [15] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, “A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews,” *2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020*, pp. 217–220, 2020, doi: 10.1109/IC3A48958.2020.233300.