Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 788-796
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Perbandingan *Naïve Bayes Classifier* Dan *Support Vector Machine* Dalam Mengklasifikasikan Tingkat Pengangguran Terbuka Di Indonesia

Dhita Diana Dewi[1], Ivana Lucia Kharisma[2], Nida Aulia Salsa Bila[3]
[1,2,3]Fakultas Teknik Komputer dan Desain, Program Studi Teknik Informatika, Universitas Nusa Putra, Sukabumi, Indonesia
Email: Dhita.dianadewi@nusaputra.ac.id[1], ivana.lucia@nusaputra.ac.id[2], nida.aulia_ti20@nusaputra.ac.id[3]

Abstract

Unemployment is one of the factors of problems in the economic field, this will have an impact on the balance of the economy. A person can be said to be unemployed if the person does not meet the requirements as a workforce. Open unemployment is a workforce that does not actually have a job. Therefore, this study will classify the Open Unemployment Rate (TPT) in Indonesia in the 2020-2023 period. This study will use the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) algorithms. In the SVM algorithm method, for the negative class consists of a precision value of 62%, a recall of 80%, an F1 Score of 70%. While for the positive class consists of a precision value of 87%, a recall of 72%, an F1 Score of 79%. In the NBC algorithm method, for the negative class consists of a precision value of 71%, a recall of 50%, an F1 Score of 59%. While for the positive class consists of a precision value of 76%, a recall of 89%, an F1 Score of 82%. Based on these calculations, the accuracy value of each algorithm has the same accuracy value, which is 75%.

Keywords*: Unemployment, Open Unemployment, Naïve Bayes Classifier and Support Vector Machine*

Abstrak

Pengangguran merupakan salah satu faktor permasalahan dalam bidang ekonomi, ini akan berdampak pada keseimbangan perekonomian. Seseorang bisa dikatakan sebagai pengangguran apabila seseorang tersebut tidak memenuhi syarat sebagai angkatan kerja. Pengangguran terbuka adalah angkatan kerja yang sebenarnya tidak mempunyai pekerjaan. Maka dari itu, pada penelitian ini akan dilakukan pengklasifikasian Tingkat Pengangguran Terbuka (TPT) di Indonesia pada rentang tahun 2020-2023. Penelitian ini akan menggunakan algoritma Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). Pada metode algoritma SVM, untuk kelas negative terdiri dari nilai precision yaitu sebesar 62%, recall sebesar 80%, F1 Score sebesar 70%. Sedangkan untuk kelas positive terdiri dari nilai precision yaitu sebesar 87%, recall sebesar 72%, F1 Score sebesar 79%. Pada metode algoritma NBC, untuk kelas negative terdiri dari nilai precision yaitu sebesar 71%, recall sebesar 50%, F1 Score sebesar 59%. Sedangkan untuk kelas positive terdiri dari nilai precision yaitu sebesar 76%, recall sebesar 89%, F1 Score sebesar 82%. Berdasarkan perhitungan tersebut, nilai akurasi dari masing-masing algoritma mempunyai nilai akurasi yang sama yaitu 75%.

Kata kunci*: Pengangguran, Pengangguran Terbuka, Naïve Bayes Classifier dan Support Vector Machine*

1. PENDAHULUAN

Pengangguran merupakan salah satu faktor permasalahan dalam bidang ekonomi, ini akan berdampak pada keseimbangan perekonomian. Seseorang bisa dikatakan sebagai pengangguran apabila seseorang tersebut tidak memenuhi syarat sebagai angkatan kerja. Di negara berkembang seperti negara Indonesia ini biasanya mendapati masalah yang berhubungan dengan pembangunan nasional

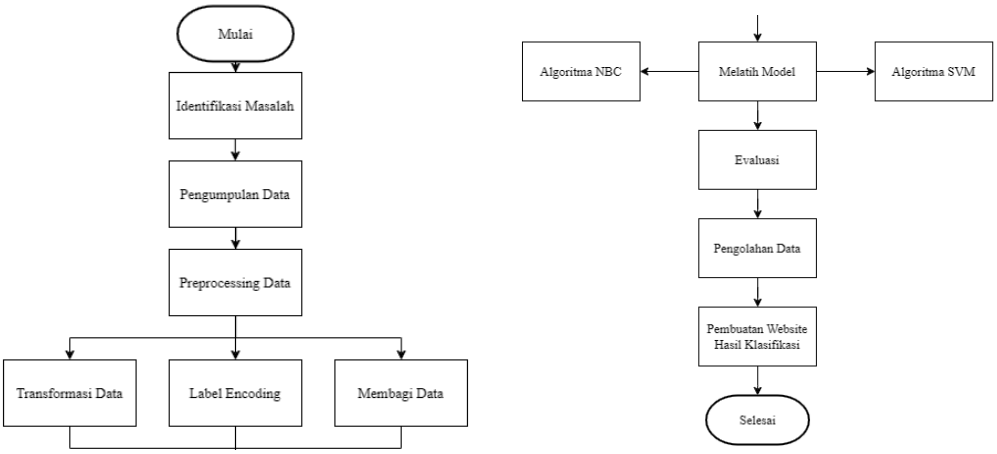
yaitu pengangguran. Jumlah angkatan kerja Sakernas pada Agustus 2023 yaitu sebanyak 147,71 juta orang dan dapat mempengaruhi pada Tingkat Partisipasi Angkatan Kerja (TPAK) yang naik sebesar 0,85% daripada Agustus 2022. Komposisi angkatan kerja pada Agustus 2023 berjumlah 139,85 juta orang bekerja dan 7,86 juta orang menganggur. Dibandingkan Agustus 2022, angkatan kerja bertambah 3,99 juta orang, penduduk bekerja bertambah 4,55 juta orang, dan pengangguran berkurang 0,56 juta orang[1].

Penelitian yang membahas tentang pengangguran sudah pernah dilakukan pada penelitian sebelumnya, salah satunya penelitian yang dilakukan oleh Inggit Fatika DKK yang berjudul "Klasifikasi Tingkat Pengangguran Terbuka Di Indonesia Dengan Algoritma *Classification And Regression Tree* (Cart) Dan C4.5". Berdasarkan hal tersebut, pada penelitian ini akan dilakukan pengklasifikasian pengangguran terbuka dengan menerapkan algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) serta perbandingan antara dua metode tersebut. NBC merupakan algoritma *machine learning* untuk pengklasifikasian, ini yang paling mudah dan cepat, yang cocok untuk sejumlah besar data. NBC juga dikenal sebagai pengklasifikasi probabilitas karena didasarkan pada *Teorema Bayes*. Algoritma SVM menawarkan beberapa keuntungan, seperti menggunakan fungsi kernel untuk menerapkan *hyperlayer* ke data masukan berdimensi tinggi yang kompleks, sehingga meningkatkan kinerjanya[2]. Penelitian ini akan dilakukan dengan menggunakan data pada rentang waktu 2020-2023.

Dalam melakukan klasifikasi pengangguran terbuka ini terdapat berbagai macam alat bantu yang digunakan, salah satunya yaitu menggunakan bahasa pemograman *python*. *Python* merupakan bahasa pemograman yang sering digunakan dalam Data *Science* dan *Machine Learning*. Dengan dilakukannya penelitian ini, diharapkan dapat bermanfaat bagi pembaca dan menghasilkan klasifikasi tingkat pengangguran terbuka serta perbandingan antara algoritma NBC dan SVM.

2. METODOLOGI PENELITIAN

Berikut merupakan tahapan alur dari metode penelitian:



Gambar 1. Alur Penelitian

2.1. Identifikasi Masalah

Pada penelitian ini, penulis akan mengidentifikasi masalah yang terjadi pada faktor tingkat pengangguran terbuka di Indonesia. Masalah yang ada tersebut akan dijadikan dasar dalam membuat solusi nantinya.

2.2. Pengumpulan Data

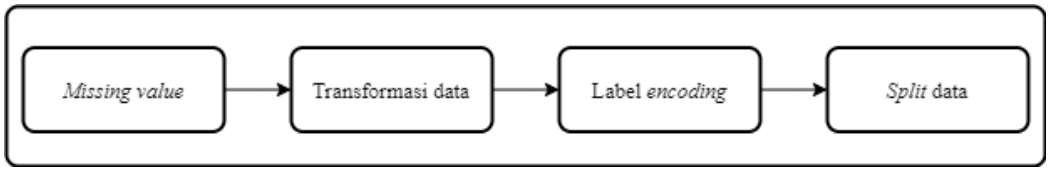
Pengumpulan data untuk penelitian ini yaitu menggunakan data sekunder yang didapatkan dari situs *website* Badan Pusat Statistik (https://www.bps.go.id/id). Data yang dikumpulkan yaitu meliputi 8 (delapan) variabel faktor penyebab pengangguran, dan 1 (satu) variabel terikat yang terdiri dari 34 provinsi dari rentang tahun 2020-2023. Variabel-variabel yang digunakan dalam penelitian ini yaitu sebagai berikut:

Tabel 1. Variabel yang digunakan dalam penelitian

No	Variabel	Kode
1	Angka Partisipasi Sekolah (APS)	X1
2	Persentase Penduduk Miskin (PPM)	X2
3	Rata - Rata Lama Sekolah (RLM)	X3
4	Indeks Pembangunan Manusia (IPM)	X4
5	Tingkat Partisipasi Angkatan Kerja (TPAK)	X5
6	Tenaga Kerja Formal (TKF)	X6
7	Proporsi Lapangan Kerja Informal (PLKI)	X7
8	Upah Minimum Provinsi (UMP)	X8
9	Tingkat Pengangguran Terbuka (TPT)	Y

2.3. *Preprocessing* Data

Dalam proses ini dilakukan beberapa tahapan, yaitu pengumpulan data, pembagian data, dan memberikan label pada data. *Preprocessing* data adalah langkah yang sangat penting yang melibatkan persiapan data sebelum melatih model. Tujuan pengolahan data adalah untuk menata dan menyiapkan data agar sesuai untuk digunakan dalam proses pelatihan dan evaluasi model. Berikut gambar tahapan pada *preprocessing* data:



Gambar 2. Tahap *Preprocessing*

Data pada tahap *preprocessing* seperti Gambar 2, dilakukan tahapan lagi yaitu sebagai imputasi data, transformasi data, label *encoding*, dan *split* data.

2.4. Klasifikasi Data dengan Algoritma *Naïve Bayes Classifier* (NBC)

Dalam tahapan ini dilakukan klasifikasi data dengan menggunakan algoritma NBC untuk mendapatkan hasil berupa kategori rendah atau tinggi. Pada tahap ini

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 788-796
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

akan diketahui berapa besar nilai akurasi yang didapatkan terhadap pengklasifikasian data.

Dalam konteks penelitian klasifikasi, metode yang sering digunakan yaitu metode *naïve bayes*[3]. Berikut rumus *Naïve Bayes Classifier*:

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

Keterangan:

X : Data dengan *class* yang belum diketahui
H : Hipotesis data merupakan suatu *class* spesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (Posterori Probabilitas)
P(H) : Probabilitas hipotesis H (Prior Probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi hipotesis H
P(X) : Probabilitas X

2.5. Klasifikasi Data dengan Algoritma *Support Vector Machine* (SVM)

Dalam tahapan ini akan dilakukan pembagian data menggunakan algoritma SVM dengan kernel linear untuk membuat dan melatih model menggunakan data latih. Pada tahap ini hasil akhir dari klasifikasi yaitu berupa kategori rendah atau tinggi. Dengan dilakukan tahapan-tahapan yang dilakukan akan menghasilkan nilai akurasi dalam pengklasifikasian menggunakan algoritma svm terhadap data yang ada. Saat mencari *hyperplane* dengan SVM, dapat menggunakan persamaan berikut[4]:

$$(w.x_i) + b = 0 \quad (2)$$
$$(w.x_i + b) \leq 1, y_i = -1 \quad (3)$$
$$(w.x_i + b) \geq 1, y_i = 1 \quad (4)$$

2.6. Evaluasi

Tahapan evaluasi dilakukan untuk mengevaluasi kinerja model yang telah dilatih. Pada tahap evaluasi ini dilakukan menggunakan *confusion matrix* yang digunakan sebagai metrik evaluasi untuk mengukur seberapa baik model melakukan prediksi, model yang digunakan adalah algoritma NBC dan SVM. *confusion matrix* akan menghitung *nilai precision, recall, accuracy, F1-Score*[5]. Berikut adalah rumus perhitungan *confusion matrix*:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$
$$Recall = \frac{TP}{TP+FN} \quad (6)$$
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Keterangan:

TP : *True Positive*
TN : *True Negative*
FP : *False Positive*
FN : *False Negative*

2.7. Pembuatan *Website*

Pada tahap ini akan dibuat *website* untuk menampilkan hasil dari pengklasifikasian data menggunakan algoritma NBC dan SVM. Secara garis besar alur dari sistem *website* yang nantinya akan menjadi tempat untuk menampilkan hasil dari model latih algoritma NBC dan SVM dalam pengklasifikasian TPT di Indonesia berdasarkan provinsi. Di dalamnya terdapat menu-menu serta kebutuhan lainnya seperti melihat hasil klasifikasi TPT berdasarkan provinsi algoritma NBC dan SVM.

3. HASIL DAN PEMBAHASAN

3.1. Identifikasi Masalah

Pada tahap ini, sebelum dilakukannya penelitian maka akan dilakukan identifikasi masalah yang terjadi. Pada penelitian ini ditemukan beberapa masalah dalam tingginya TPT di berbagai provinsi di Indonesia yang dapat mempengaruhi kesejahteraan sosial di Indonesia. Berdasarkan masalah yang ada, maka akan dilakukan penelitian ini yaitu pengklasifikasian TPT dan mencari faktor pengangguran yang mempengaruhi terhadap TPT di Indonesia.

3.2. Pengumpulan Data

Pada penelitian ini data yang digunakan yaitu faktor pengangguran terbuka di Indonesia berdasarkan provinsi dari rentang tahun 2020-2023. Setelah dilakukannya pengumpulan data, kurang lebih data yang didapatkan yaitu 136 data.

3.3. *Preprocessing* Data

Pada tahap ini sebelum data latih, akan dilakukan *preprocessing* data agar data lebih sesuai dengan model yang akan dilakukan. Di dalam *preprocessing* data terdapat 4 tahapan lagi yaitu imputasi data, transformasi data, label *encoding*, dan *split* data. Berikut tahapannya:

a) Imputasi Data
   Pada tahap ini akan dilakukan imputasi data yaitu untuk mengisi data yang hilang dari data yang telah dikumpulkan mulai dari tahun 2020-2023. *Missing value* pada data yang telah dikumpulkan, terdapat sebanyak 34 data atau pada kolom TKF yang kosong. Setelah dilakukan imputasi data menggunakan metode median, data yang kosong sudah terisi yaitu pada kolom TKF pada tahun 2023.
b) Transformasi Data
   Pada tahap ini akan dilakukan pengubahan data TPT yang awalnya berbentuk persentase menjadi kategori rendah atau tinggi. Nilai tersebut akan diubah dengan melakukan perhitungan median dari seluruh data atau kolom TPT dan hasil tersebut akan menjadi landasan nilai dalam kategori TPT. Setelah dilakukan median maka akan menjadi kategori rendah atau tinggi seperti yang tertera pada kolom TPT *Actual*. Hasil perhitungan median tersebut yaitu 4,96%, maka jika nilai lebih

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 788-796
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

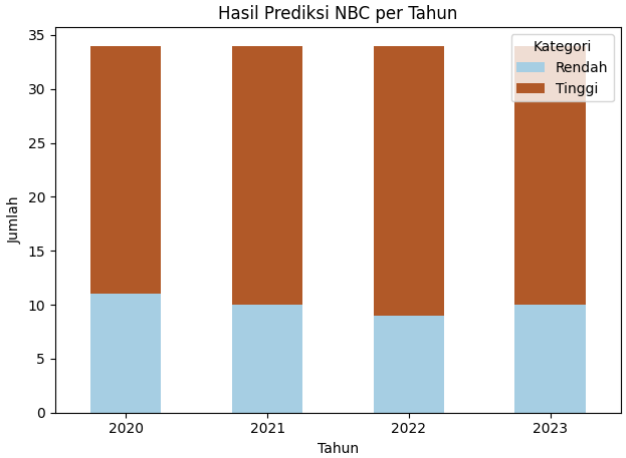
besar dari 4,96% maka akan dikategorikan tinggi. Sedangkan jika kurang dari 4,96% maka akan dikategorikan rendah.

c) Label *Encoding*
Pada tahap ini akan diubah nilai pada kolom 'Kategori' yang berisi data berupa kategori 'rendah' atau 'tinggi' yang merupakan hasil dari transformasi data. Mengubah label kategori 'rendah' atau 'tinggi' ini menggunakan *LabelEncoder*. Setelah dilakukan label *encoding* menggunakan *LabelEncoder* maka akan menghasilkan nilai 1 atau 0 seperti yang tertera pada kolom. Nilai 1 untuk kategori tinggi, sedangkan 0 untuk kategori rendah.

d) *Split* Data
Pada tahap ini, data akan dibagi menjadi 2 subset yaitu data latih (*train*) dan data uji (*test*). Data dibagi menjadi data training sebanyak 108 dan data testing sebanyak 28 atau perbandingannya yaitu 80% : 20%.

3.4. Melatih Model Algoritma *Naïve Bayes Classifier* (NBC)

Setelah melalui tahap *preprocessing*, pada tahap ini yaitu melatih model algoritma NBC terhadap data untuk pengklasifikasian TPT di Indonesia berdasarkan provinsi. Setelah dilakukan *preprocessing* data, maka data sudah terbagi menjadi dua bagian yaitu data *training* 80% dan *testing* 20% untuk melakukan pengujian model algoritma NBC. Berikut merupakan grafik hasil klasifikasi:



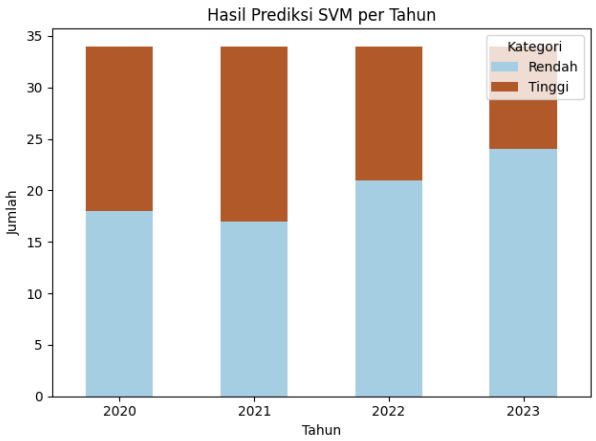
Gambar 3. Grafik Hasil NBC

Berdasarkan grafik hasil klasifikasi TPT menggunakan algoritma NBC terlihat bahwa kategori TPT tinggi lebih banyak daripada kategori rendah. Dari tiap tahunnya, kategori tinggi menjadi yang lebih unggul.

3.5. Melatih Model Algoritma *Support Vector Machine* (SVM)

Pada tahap ini akan dilakukan pengklasifikasian menggunakan algoritma SVM. Setelah dilakukan *preprocessing* data agar lebih sesuai dengan model algoritma. Berikut merupakan grafik hasil klasifikasi:

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 788-796
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

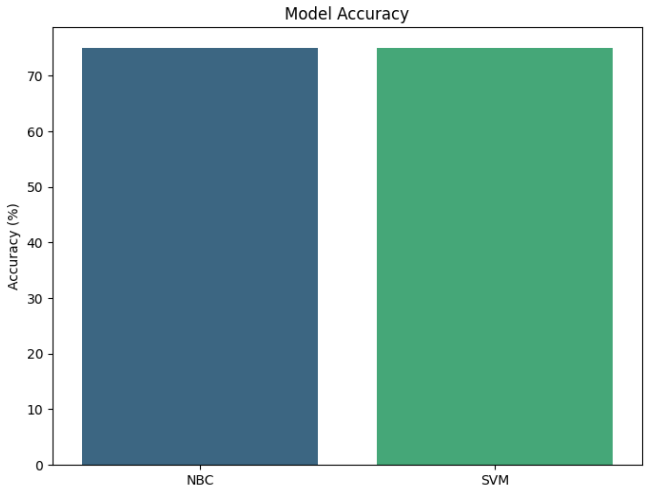


Gambar 4. Grafik Hasil NBC

Berdasarkan grafik hasil TPT menggunakan algoritma SVM terlihat bahwa kategori TPT rendah yang paling banyak, seperti pada tahun 2020, 2022, dan 2023 sedangkan pada tahun 2021 TPT kategori rendah dan tinggi sama banyak.

3.6. Perbandingan Nilai Akurasi

Berdasarkan perbandingan nilai akurasi menggunakan metode *confusion matrix* maka menghasilkan persentase dari keakuratan metode NBC dan SVM seperti berikut:



Gambar 5. Grafik Perbandingan Algoritma

Berdasarkan gambar bahwa tingkat akurasi algoritma NBC dan SVM yaitu sama-sama sebesar 75%. Berdasarkan hasil tersebut bahwa pada penelitian ini, algoritma SVM dengan algoritma NBC pada penelitian ini nilai akurasinya sama, pada penelitian ini digunakan evaluasi dengan menggunakan metode *confusion matrix*. Berikut merupakan tabel hasil dari klasifikasi TPT algoritma NBC dan SVM:

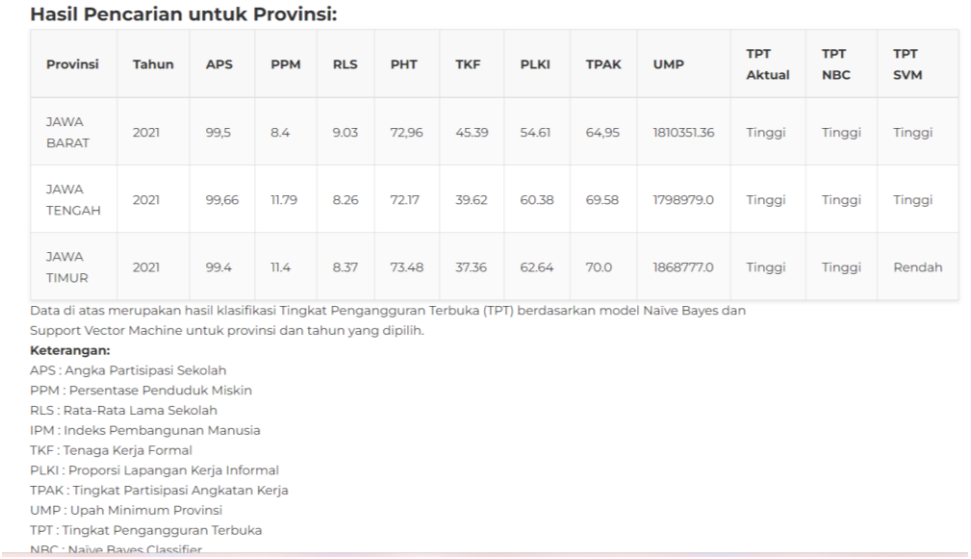
Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 788-796
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Tabel 2. Hasil Klasifikasi Algoritma NBC dan SVM

No.	Provinsi	Tahun	TPT	TPT *Actual*	TPT NBC	TPT SVM
1	ACEH	2020	6,59	Tinggi	Tinggi	Tinggi
2	SUMATERA UTARA	2020	6,91	Tinggi	Tinggi	Rendah
3	SUMATERA BARAT	2020	6,88	Tinggi	Tinggi	Rendah
4	RIAU	2020	6,32	Tinggi	Tinggi	Tinggi
5	JAMBI	2020	5,13	Tinggi	Tinggi	Rendah
...	...	...	...	...	...	...
132	SULAWESI BARAT	2023	2,27	Rendah	Rendah	Rendah
133	MALUKU	2023	6,31	Tinggi	Tinggi	Tinggi
134	MALUKU UTARA	2023	4,31	Rendah	Tinggi	Rendah
135	PAPUA BARAT	2023	5,38	Tinggi	Rendah	Tinggi
136	PAPUA	2023	2,67	Rendah	Rendah	Rendah

3.7. *Website*

Pada tahap ini akan dilakukan *deployment website* agar hasil perhitungan klasifikasi dari algoritma NBC dan SVM dapat dilihat umum. *Website* yang dibuat terdiri dari menu *Home*, *Naïve Bayes Classifier*, dan *Support Vector Machine*. Berikut merupakan hasil pencarian untu klasifikasi TPT di Indonesia berdasarkan provinsi:



Hasil Pencarian untuk Provinsi:

Provinsi	Tahun	APS	PPM	RLS	PHT	TKF	PLKI	TPAK	UMP	TPT Aktual	TPT NBC	TPT SVM
JAWA BARAT	2021	99,5	8.4	9.03	72,96	45.39	54.61	64,95	1810351.36	Tinggi	Tinggi	Tinggi
JAWA TENGAH	2021	99,66	11.79	8.26	72.17	39.62	60.38	69.58	1798979.0	Tinggi	Tinggi	Tinggi
JAWA TIMUR	2021	99.4	11.4	8.37	73.48	37.36	62.64	70.0	1868777.0	Tinggi	Tinggi	Rendah

Data di atas merupakan hasil klasifikasi Tingkat Pengangguran Terbuka (TPT) berdasarkan model Naive Bayes dan Support Vector Machine untuk provinsi dan tahun yang dipilih.

Keterangan:

APS : Angka Partisipasi Sekolah
PPM : Persentase Penduduk Miskin
RLS : Rata-Rata Lama Sekolah
IPM : Indeks Pembangunan Manusia
TKF : Tenaga Kerja Formal
PLKI : Proporsi Lapangan Kerja Informal
TPAK : Tingkat Partisipasi Angkatan Kerja
UMP : Upah Minimum Provinsi
TPT : Tingkat Pengangguran Terbuka
NBC : Naive Bayes Classifier

Gambar 6. Tampilan Hasil Pencarian

4. SIMPULAN

Berdasarkan pengujian data yang sudah dilakukan menggunakan algoritma NBC dan SVM, maka diperoleh kesimpulan bahwa dalam mengklasifikasikan TPT di Indonesia pada rentang tahun 2020-2023 tersebut mendapatkan hasil keakuratan yang berbeda di setiap variabel kelasnya. Seperti pada metode algoritma SVM, untuk kelas *negative* terdiri dari nilai *precision* yaitu sebesar 62%, *recall* sebesar 80%, F1 *Score* sebesar 70%. Sedangkan untuk kelas *positive* terdiri dari nilai *precision* yaitu sebesar 87%, *recall* sebesar 72%, F1 *Score* sebesar 79%. Pada metode algoritma NBC, untuk kelas *negative* terdiri dari nilai *precision* yaitu

sebesar 71%, *recall* sebesar 50%, F1 *Score* sebesar 59%. Sedangkan untuk kelas *positive* terdiri dari nilai *precision* yaitu sebesar 76%, *recall* sebesar 89%, F1 *Score* sebesar 82%. Berdasarkan perhitungan tersebut, nilai akurasi dari masing-masing algoritma mempunyai nilai akurasi yang sama yaitu 75%.

DAFTAR PUSTAKA

[1] Badan Pusat Statistik, "Berita Resmi Statistik: Keadaan Ketenagakerjaan Indonesia Agustus 2023," *Badan Pus. Stat.*, vol. 11, no. 84, pp. 1–28, 2023.

[2] H. Andreansyah, "Klasifikasi Sentimen Positif dan Negatif pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine (SVM)," vol. 9, pp. 329–336, 2024.

[3] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[4] A. Sentimen, P. Maskapai, H. C. Husada, and A. S. Paramita, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM) Sentiment Analysis of Airline on Twitter Platform Using Support Vector Machine (SVM) Algorithm," vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.

[5] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal Ris. Komputer)*, vol. 8, no. 6, p. 219, 2021, doi: 10.30865/jurikom.v8i6.3655.