

Klasifikasi Indeks Pembangunan Manusia dengan Metode *K-Nearest Neighbor* dan *Support Vector Machine* di Pulau Jawa

Ida Ayu Ade Sita Pratiwi^{a1}, Arie Wahyu Wijayanto^{a2}

^aPoliteknik Statistika STIS
Jl. Otto Iskandardinata No.64C Jakarta 13330
¹211709745@stis.ac.id
²ariewahyu@stis.ac.id

Abstrak

Pembangunan merupakan suatu upaya yang dilakukan oleh pemerintah untuk mewujudkan kesejahteraan dan kemakmuran masyarakat. Salah satu indikator yang dapat digunakan untuk mengukur hasil pembangunan adalah Indeks Pembangunan Manusia (IPM). Komponen-komponen IPM terdiri dari tiga bidang, yaitu bidang pendidikan, kependudukan, dan kesehatan. Pulau Jawa merupakan salah satu wilayah yang berada di Kawasan Barat Indonesia yang didominasi adanya pembangunan infrastruktur. Dengan adanya pembangunan tersebut diharapkan taraf hidup masyarakat menjadi lebih sejahtera sehingga kualitas sumber daya manusia menjadi lebih baik. Akan tetapi, antarwilayah di Pulau Jawa memiliki kualitas sumber daya manusia yang bervariasi. Hal tersebut dikarenakan terdapat tidak meratanya pembangunan. Pada penelitian ini metode yang digunakan adalah *K-Nearest Neighbor* dan *Support Vector Machine* yang bertujuan untuk membandingkan akurasi klasifikasi Indeks Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa tahun 2019. Hasil penelitian menunjukkan metode yang lebih baik adalah *Support Vector Machine* dengan parameter terbaik yang digunakan yaitu *kernel-linear*, *gamma* sebesar 1, dan *cost* sebesar 5. Akurasi yang dihasilkan adalah sebesar 88,89 persen dengan nilai AUC sebesar 0,870.

Kata kunci: Indeks Pembangunan Manusia, Klasifikasi, *K-Nearest Neighbor*, *Support Vector Machine*

Abstract

Development is an effort made by the government to realize the welfare and prosperity of the community. One of the indicators that can be used to measure development outcomes is the Human Development Index (HDI). The HDI components consist of three fields are education, population, and health. Java Island is one of the areas located in the Western Region of Indonesia which is dominated by infrastructure development. With this development, it is hoped that the people's standard of living will be more prosperous so that the quality of human resources will be better. However, between regions in Java Island, the quality of human resources varies. This is because there is an uneven development. In this study the method used is the K-Nearest Neighbor and Support Vector Machine which aims to compare the accuracy of the classification of the Human Development Index (HDI) of regencies/municipalities in Java 2019. The results show that the better method is the Support Vector Machine with the best parameters used is kernel-linear, gamma of 1, and cost of 5. The resulting accuracy is 88,89 percent with an AUC value of 0,870.

Keywords: Human Development Index, Classification, *k-Nearest Neighbor*, *Support Vector Machine*

1. Pendahuluan

Pembangunan merupakan suatu upaya yang dilakukan oleh pemerintah untuk mewujudkan kesejahteraan dan kemakmuran masyarakat. Salah satu indikator yang dapat digunakan untuk mengukur hasil pembangunan adalah Indeks Pembangunan Manusia (IPM). IPM menjelaskan bagaimana penduduk dapat mengakses hasil pembangunan dalam memperoleh pendapatan, kesehatan, pendidikan, dan sebagainya. IPM diperkenalkan oleh *United Nations Development Programme* (UNDP) pada tahun 1990 dan dipublikasikan secara berkala dalam laporan tahunan Human Development Report (HDR) (Badan Pusat Statistik). IPM dapat menentukan peringkat atau level pembangunan suatu wilayah atau negara. Bagi Indonesia, IPM merupakan data strategis karena selain sebagai ukuran kinerja Pemerintah, IPM juga digunakan sebagai salah satu alokator penentuan Dana Alokasi Umum (DAU). IPM merupakan indeks komposit yang dihitung dari indeks harapan hidup, indeks pendidikan, dan indeks standar hidup layak. Karena dalam perhitungan indeks harapan hidup, indeks pendidikan, dan indeks standar hidup layak, melibatkan komponen ekonomi maupun nonekonomi, seperti kualitas pendidikan, kesehatan, dan kependudukan. Dimana dalam bidang pendidikan menggunakan komponen angka harapan lama sekolah dan rata-rata lama sekolah. Pada bidang kesehatan menggunakan komponen capaian umur panjang dan sehat. Sedangkan pada bidang kependudukan menggunakan komponen kemampuan daya beli masyarakat terhadap sejumlah kebutuhan pokok yang dilihat dari rata-rata besarnya pengeluaran perkapita sebagai pendekatan pendapatan. Dengan demikian IPM dianggap telah relevan untuk dijadikan tolak ukur dalam menentukan keberhasilan pembangunan. Sejauh mana variabel ekonomi maupun nonekonomi tersebut dapat menunjang IPM (Meliana dan Zain, 2013). Ketiga hal tersebut saling mempengaruhi satu sama lain, selain itu IPM dapat dipengaruhi oleh faktor-faktor lain, seperti ketersediaan kesempatan kerja yang ditentukan oleh pertumbuhan pembangunan ekonomi, infrastruktur, dan kebijakan pemerintah sehingga IPM akan meningkat. Dengan nilai indeks IPM yang tinggi menandakan keberhasilan pembangunan ekonomi suatu wilayah atau negara.

Menurut Badan Pusat Statistik (BPS), Indeks Pembangunan Manusia dibagi menjadi empat kategori yaitu : 1) <60 (rendah); 2) $60 \leq \text{IPM} < 70$ (sedang); 3) $70 \leq \text{IPM} < 80$ (tinggi); 4) >80 (sangat tinggi). Tinggi rendahnya nilai IPM suatu daerah dipengaruhi oleh keadaan atau situasi pada daerah tersebut. Serta pembangunan di Indonesia masih belum merata sehingga IPM di wilayah- wilayah terutama kabupaten/kota sangatlah beragam. Pada umumnya, wilayah yang memiliki IPM tinggi hanya terdapat di kabupaten/kota besar di Indonesia, karena kabupaten/kota besar di Indonesia memiliki fasilitas kesehatan, pendidikan, dan kebutuhan yang mencukupi (Fauzi, 2017). Pulau Jawa merupakan salah satu wilayah yang berada di Kawasan Barat Indonesia yang didominasi adanya pembangunan infrastruktur. Dengan adanya pembangunan tersebut diharapkan taraf hidup masyarakat menjadi lebih sejahtera sehingga kualitas sumber daya manusia menjadi lebih baik. Akan tetapi, antarwilayah di Pulau Jawa memiliki kualitas sumber daya manusia yang bervariasi. Hal tersebut dikarenakan terdapat tidak meratanya pembangunan.

Data mining adalah serangkaian proses untuk mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Menurut Turban dalam Gunadi (2016), data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar. Pekerjaan yang berkaitan erat dengan data mining adalah model prediksi (*prediction modelling*), analisis kelompok (*cluster analysis*), analisis asosiasi (*association analysis*), klasifikasi (*classification*), dan deteksi anomaly (Prasetyo dalam Darsyah, 2017). Data mining terdiri dari *unsupervised learning* dan *supervised learning*. Menurut Ridwan, dkk (2013), *unsupervised learning* adalah metode yang penerapannya tanpa adanya latihan (training) dan tanpa ada guru (*teacher*) yang disebut sebagai label dari suatu data. Sedangkan *supervised learning* adalah metode yang penerapannya dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi digunakan untuk beberapa contoh data yang mempunyai output atau label selama proses training. Metode yang termasuk ke dalam *supervised learning* seperti *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM).

Menurut Wu dalam Darsyah (2017), *K-Nearest Neighbor* (k-NN) adalah salah satu metode klasifikasi yang terdapat dalam data mining dan termasuk ke dalam kelompok instance-based

learning. K-NN dilakukan dengan mencari k-objek dalam data training yang paling dekat (mirip) dengan objek pada data testing. Data training diproyeksikan ke dalam ruang yang berdimensi banyak, dimana masing-masing dimensi mempresentasikan fitur dari data. Dekat atau jauhnya tertangga/data dapat dihitung dengan menggunakan perhitungan jarak Euclidean. *Support Vector Machine* (SVM) juga merupakan suatu metode klasifikasi yang dikenal oleh Vapnik pada tahun 1995 (Prangga, 2017). SVM termasuk dalam kelas *Artificial Neural Network* (ANN). Dalam melakukan klasifikasi, SVM perlu adanya tahapan training dan tahapan testing. Tujuan dari SVM adalah menemukan fungsi pemisah (*klasifier*) yang optimal yang bisa memisahkan dua set data yang berbeda. Fungsi pemisah terbaik adalah yang terletak di tengah-tengah dua objek dari kedua kelas. Kedua metode tersebut merupakan metode pengklasifikasian untuk jumlah data yang besar.

Oleh karena itu, peneliti ingin meneliti tentang pengklasifikasian Indeks Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa tahun 2019 dengan menggunakan dua metode, yaitu metode *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM). Penelitian ini juga bertujuan untuk membandingkan tingkat akurasi dari *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM).

2. Metode Penelitian

Pada penelitian ini langkah-langkah yang dilakukan meliputi pengumpulan data, analisis data, dan kesimpulan. Data yang digunakan pada penelitian adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS). Dimana data yang digunakan, yaitu Indeks Pembangunan Manusia (IPM), Rata-rata Lama Sekolah (RLS), Produk Domestik Regional Bruto per Kapita, tingkat pengangguran, dan Umur Harapan Hidup (UHH) sejak lahir pada tahun 2019 di seluruh kabupaten/kota yang ada di Pulau Jawa. Dengan struktur data terdapat pada Tabel 1 di bawah ini.

Tabel 1. Struktur Data

Bidang (1)	Variabel (2)	Notasi (3)	Keterangan (4)
Indeks Pembangunan Manusia	Indeks Pembangunan Manusia (IPM)	Y	Kategorik : R = rendah (<70) T = tinggi (≥ 70)
Pendidikan	Rata-rata Lama Sekolah (RLS)	X1	Numerik
Kependudukan	PDRB per Kapita	X2	Numerik
	Pengangguran	X3	Numerik
Kesehatan	Umur Harapan Hidup (UHH)	X4	Numerik

Pada penelitian ini menggunakan dua teknik klasifikasi data mining, yaitu teknik *K-Nearest Neighbor* (k-NN) dan metode *Support Vector Machine* (SVM).

1. *K-Nearest Neighbor* (k-NN)

Algoritma *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada kecocokan bobot dari sejumlah fitur yang ada. *K-Nearest Neighbor* (k-NN) adalah suatu metode yang menggunakan algoritma supervised learning. Perbedaan antara supervised learning dengan unsupervised learning adalah pada supervised learning bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada unsupervised learning bertujuan untuk menemukan pola dalam sebuah data. *K-Nearest Neighbor* sering digunakan adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Data pembelajaran atau training diproyeksikan ke dalam ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Tujuan dari algoritma k-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Dimana

hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada k-NN. Pada proses pengklasifikasian, algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Algoritma ini menggunakan klasifikasi ketetapan sebagai nilai prediksi dari sampel uji yang baru (Krisandi, 2013). Jarak yang digunakan adalah Euclidean Distance. Adapun rumus dari Euclidean Distance, yaitu:

$$D(a, b) = \sqrt{\sum_{k=1}^d (a_k - b_k)^2} \dots\dots\dots (1)$$

Dimana matriks D(a,b) adalah jarak scalar dari kedua vector a dan b dari matriks dengan ukuran d dimensi (Rohman, 2019). Pada fase training, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data training samples. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk testing data (klasifikasinya belum diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor training sample dihitung, dan sejumlah k buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksikan termasuk pada klasifikasi terbanyak dari titik-titik tersebut. Nilai k yang terbaik untuk algoritma ini tergantung pada data. Pada umumnya nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai k yang bagus dapat dipilih dengan optimasi parameter. Ketepatan algoritma K-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberikan bobot terhadap fitur, sehingga performa klasifikasi menjadi lebih baik.

K-Nearest Neighbor memiliki keunggulan dan kelemahan. Dimana keunggulan dari K-NN, yaitu teknik klasifikasi yang sangat sederhana sehingga mudah diimplementasikan, kuat dalam hal pencarian (misalnya, kelas tidak harus linear dipisahkan), efektif untuk menghitung data dalam skala kecil, dan memiliki beberapa parameter untuk acuan (jarak metric dan k). Sedangkan kelemahan dari K-NN adalah perlu untuk menentukan nilai k yang optimal sehingga untuk menyatakan jumlah tetangga terdekatnya lebih mudah serta biaya komputasi yang cukup tinggi karena perhitungan jarak harus dilakukan pada setiap query instance.

2. SVM

Support Vector Machine (SVM) merupakan metode supervised learning yang dapat digunakan untuk prediksi, klasifikasi, ataupun regresi. *Support Vector Machine* (SVM) termasuk dalam kelas *Artificial Neural Network* (ANN). Dalam melakukan klasifikasi *Support Vector Machine* (SVM) perlu adanya tahapan training dan tahapan testing. Tujuan dari metode *Support Vector Machine* (SVM) menemukan fungsi pemisah (klasifier) yang optimal yang bisa memisahkan dua set data yang berbeda (Vapnik dalam Fauzi, 2017). Pada saat ini SVM telah diaplikasikan dan memiliki performansi yang baik pada beberapa bidang. Konsep dasar dari SVM adalah bagaimana menemukan atau mencari fungsi pemisah (hyperplane) terbaik yang memisahkan dua kelas pada input space atau ruang input. Fungsi pemisah (hyperplane) terbaik antara dua kelas dapat ditemukan dengan mencari margin hyperplane terbesar. Tujuan dari usaha ini adalah untuk meningkatkan probabilitas pengelompokan secara benar pada data testing. Margin pada konsep ini adalah jarak antara titik terdekat dari masing-masing kelas terhadap hyperplane. Titik terdekat dari masing-masing kelas ini disebut support vector. Dengan kata lain, tugas utama dari SVM adalah menemukan hyperplane yang terletak tepat di tengah-tengah antara dua support vector dari kelompok kelas yang berbeda, dan besar atau jarak margin terhadap masing-masing support vector adalah sama. Secara matematika, kasus klasifikasi secara linear dengan SVM dapat dituliskan sebagai berikut.

$$g(x) := \text{sgn}(f(x)) \dots\dots\dots (2)$$

Dengan,

$$f(x) = wx + b \dots\dots\dots (3)$$

Atau,

$$g(x) = \begin{cases} +1, & \text{jika } (wx + b) \geq +1 \\ -1, & \text{jika } (wx + b) \leq -1 \end{cases} \dots\dots\dots (4)$$

Dimana $x, w \in R^n$ dan $b \in R$.

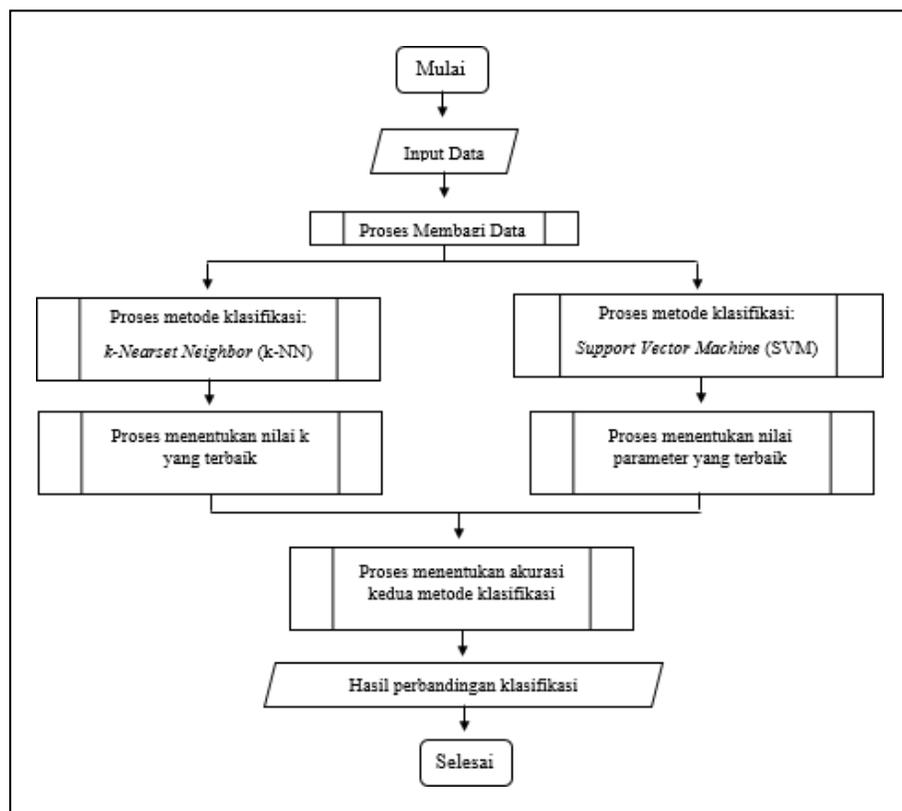
Secara umum, kasus-kasus di dunia nyata adalah kasus yang tidak linear. Data ini sulit dipisahkan secara linear. Metode kernel adalah salah satu cara untuk mengatasi hal tersebut. Fungsi kernel yang biasanya dipakai dalam literature SVM adalah:

- a. Linear : $x^T x$
- b. Polynomial : $(x^T x_i + 1)^p$
- c. Radial Basis Function (RBF) : $\exp(-\frac{1}{2\sigma^2} \|x - x_i\|^2)$

d. Tangent Hyperbolic (sigmoid) : $\tanh(\beta x^T x_i + \beta_1)$, dimana $\beta, \beta_1 \in R$.

Metode *Support Vector Machine* memiliki beberapa keunggulan, yaitu generalisasi didefinisikan sebagai kemampuan suatu metode untuk mengklasifikasi suatu pattern atau pola, yang tidak termasuk data yang digunakan dalam fase pembelajaran metode tersebut. Selain itu, SVM dapat diimplementasikan relative lebih mudah, karena proses penentuan support vector dapat dirumuskan dalam *Quadratic Programming* (QP) problem. Sedangkan kelemahan dari SVM adalah sulit dipakai pada kasus yang berskala besar. Dalam hal ini dimaksudkan dengan jumlah sampel yang diolah. Serta SVM secara teoritik dikembangkan untuk problem klasifikasi dengan dua kelas.

Berikut ini alur proses dalam penelitian ini, dapat dilihat pada Gambar 1 di bawah ini.



Gambar 1. Alur Proses Penelitian

Berdasarkan Gambar 1 di atas, adapun langkah-langkah analisis data penelitian ini, adalah sebagai berikut.

- 1) Melakukan pengumpulan data yang akan digunakan. Pada penelitian ini, data yang digunakan adalah Indeks Pembangunan Manusia (IPM), Rata-rata Lama Sekolah (RLS), Produk Domestik Regional Bruto (PDRB) per Kapita, pengangguran, dan Umur Harapan Hidup (UHH) menurut kabupaten/kota di Pulau Jawa Tahun 2019.
- 2) Data diinput dan diolah dengan menggunakan *software* seperti *Software R-studio*.
- 3) Melakukan pengkodean terhadap variabel prediktor.
- 4) Membagi data menjadi dua bagian yaitu data training dan data testing dengan menggunakan *split validation*.
- 5) Melakukan pengklasifikasian *K-Nearest Neighbor* (k-NN) dengan algoritma sebagai berikut:

- a. Menentukan jumlah nilai k untuk *K-Nearest Neighbor* (k-NN), yaitu 1,5, dan 10.
- b. Menghitung *Confusion matrix* untuk melihat akurasi klasifikasi.
- 6) Melakukan pengklasifikasian *Support Vector Machine* (SVM) dengan algoritma sebagai berikut:
 - a. Menentukan fungsi kernel untuk pengklasifikasian.
 - b. Menentukan parameter kernel dan parameter cost untuk optimasi.
 - c. Memilih nilai parameter terbaik untuk optimasi data training dan data testing.
 - d. Menghitung *Confusion matrix* untuk melihat akurasi klasifikasi.
- 7) Membandingkan ketepatan klasifikasi yang diperoleh dari metode klasifikasi *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM).
- 8) Membuat kesimpulan dari hasil yang diperoleh.

Perbandingan antara dua teknik klasifikasi data mining dapat dilakukan dengan mengevaluasi dan memvalidasi, yaitu *Confusion matrix* dan kurva ROC/nilai AUC.

1. *Confusion matrix*

Confusion matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi pada bidang data mining. *Confusion matrix* memberikan keputusan yang diperoleh dalam *training* dan *testing* yang mana memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah (Mayadewi dan Rosely, 2015). Berikut ini merupakan bentuk perhitungan dari *Confusion matrix*.

Tabel 2. Bentuk *Confusion Matrix*

	Actual	
Predict	True	False
(1)	(2)	(3)
True	True Positif	False Negative
False	False Positif	True Negative

Keterangan:

- *True positif* (TP) : nilai prediksi positif dan aktual positif.
- *False positif* (FP) : nilai prediksi positif dan aktual negatif.
- *False negative* (FN) : nilai prediksi negatif dan aktual positif.
- *True negative* (TN) : nilai prediksi negatif dan aktual negatif.

Tabel *Confusion matrix* digunakan untuk mempermudah menghitung nilai *accuracy*, *precision*, *recall*, dan *F-measure*.

- *Accuracy* : perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \dots \dots \dots (5)$$

- *Precision* : proporsi kasus dengan hasil positif yang diidentifikasi benar.

$$Precision = \frac{TP}{TP+FP} \times 100\% \dots \dots \dots (6)$$

- *Recall* : proporsi kasus positif yang diidentifikasi dengan benar.

$$Recall = \frac{TP}{TP+FN} \times 100\% \dots \dots \dots (7)$$

- *F-measure* : kombinasi pengukuran dari *precision* dan *recall* menjadi suatu metrik.

$$Fmeasure = \frac{2 \times precision \times recall}{precision+recall} \dots \dots \dots (8)$$

2. Kurva ROC (nilai AUC)

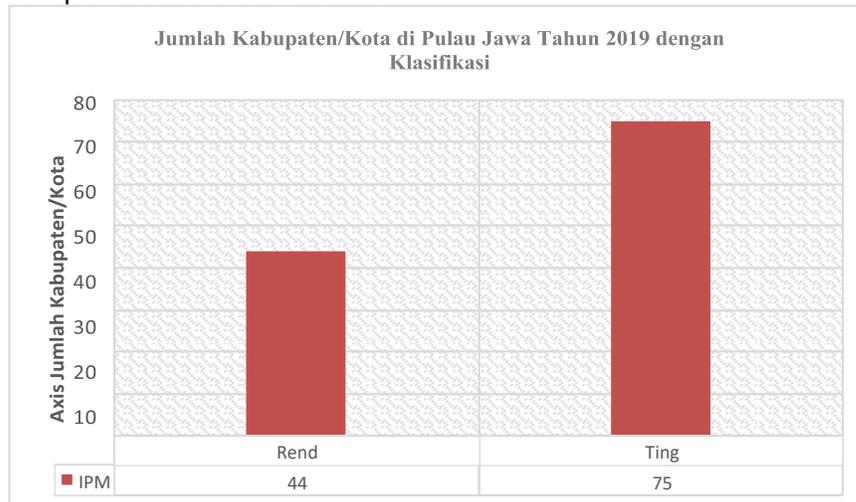
Kurva ROC (*Receiver Operating Characteristic*) adalah salah satu teknik yang dapat memvisualisasikan, mengorganisasi, dan memilih *classifier* berdasarkan performanya. ROC merupakan hasil dari pengukuran klasifikasi dalam bentuk dua dimensi, dimana garis horizontal menggambarkan nilai *false positif* dan garis vertikal menggambarkan nilai *true positif* (Umaidah dan Purwanto, 2019). Untuk menentukan klasifikasi mana yang lebih baik digunakan metode yang menghitung luas daerah di bawah kurva ROC yang disebut AUC (*Area Under the ROC Curve*) yang diartikan sebagai probabilitas. Nilai AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari sampel yang dipilih secara acak dari populasi positif atau negatif. Semakin besar nilai AUC maka semakin kuat klasifikasi

yang digunakan, karena nilai AUC adalah bagian dari daerah unit persegi nilainya akan selalau berada direntang nilai 0 dan 1.

3. Hasil dan Pembahasan

a. Analisis Deskriptif

Rata-rata Indeks Pembangunan Manusia (IPM) di kabupaten/kota yang ada di Pulau Jawa tahun 2019 adalah sebesar 72,61 atau tergolong tinggi dengan nilai Indeks Pembangunan Manusia (IPM) tertinggi adalah sebesar 86,65 dan yang terendah sebesar 61,94. Jumlah kabupaten/kota di Pulau Jawa dengan klasifikasi menurut Badan Pusat Statistik (BPS) dapat dilihat pada Gambar 2 di bawah ini.



Sumber : Badan Pusat Statistik, data diolah

Gambar 2. Jumlah Kabupaten/Kota di Pulau Jawa Tahun 2019 dengan Klasifikasi Indeks Pembangunan Manusia (IPM) menurut Badan Pusat Statistik (BPS)

Berdasarkan Gambar 2 di atas memperlihatkan jumlah kabupaten/kota yang ada di Pulau Jawa dengan klasifikasi Indeks Pembangunan Manusia (IPM) menurut Badan Pusat Statistik (BPS) pada tahun 2019 bahwa terdapat 44 kabupaten/kota yang termasuk ke dalam kelompok Indeks Pembangunan Manusia (IPM) terendah dan 75 kabupaten/kota termasuk ke dalam kelompok tertinggi dengan total keseluruhan daerah di Pulau Jawa adalah sebanyak 119 kabupaten/kota. Hal ini menunjukkan bahwa di Pulau Jawa sebagian besar wilayahnya memiliki IPM yang bagus.

Indeks Pembangunan Manusia (IPM) adalah suatu indeks yang didasarkan pada tiga bidang, yaitu bidang pendidikan, bidang kependudukan, dan bidang kesehatan. Pada bidang pendidikan menggunakan rata-rata lama sekolah, pada bidang kependudukan menggunakan PDRB per kapita, dan pada bidang kesehatan menggunakan umur harapan hidup yang dicapai. Berikut ini hasil statistic deskriptif pada ketiga bidang tersebut, dapat dilihat pada Tabel 3 di bawah ini.

Tabel 3. Hasil Ringkasan Bidang Pendidikan, Kependudukan, dan Kesehatan di Pulau Jawa Tahun 2019

Komponen	N	Minimum	Maksimum	Mean
(1)	(2)	(3)	(4)	(5)
RLS	119	4,550	11,800	8,205
PDRB per Kapita	119	19367	760271	63778
Pengangguran	119	0,950	10,650	4,480
UHH	119	64,47	77,55	72,56

Sumber : Badan Pusat Statistik, data diolah

Berdasarkan Tabel 3 di atas menunjukkan hasil statistik deskriptif pada bidang pendidikan, kependudukan, dan kesehatan yang meliputi, Rata-rata Lama Sekolah (RLS), Produk Domestik Regional Bruto per Kapita, tingkat pengangguran, dan Umur Harapan Hidup (UHH) sejak lahir. Dimana jumlah individu masing-masing komponen sebanyak 119 kabupaten/kota yang ada di Pulau Jawa pada tahun 2019. Berikut ini penjelasan lebih rinci mengenai hasil pada Tabel 3 untuk setiap komponen IPM.

a. Rata-Rata Lama Sekolah (RLS)

Rata-rata lama sekolah menggambarkan jumlah yang digunakan oleh masyarakat dalam menjalani pendidikan formal. Berdasarkan pada Tabel 3 di atas menunjukkan nilai rata-rata untuk rata-rata lama sekolah kabupaten/kota di Pulau Jawa pada tahun 2019 adalah sebesar 8,205. Artinya bahwa jumlah yang digunakan masyarakat dalam menempuh pendidikan formal adalah 8 tahun. Wilayah yang memiliki rata-rata lama sekolah tertinggi adalah Kota Tangerang Selatan sebesar 11,800 yang artinya jumlah yang digunakan masyarakat dalam menempuh pendidikan formal di wilayah tersebut adalah 11 tahun. Sedangkan wilayah yang memiliki rata-rata lama sekolah terendah adalah Kabupaten Sampang dengan sebesar 4,550 yang artinya jumlah yang digunakan masyarakat dalam menempuh pendidikan formal di wilayah tersebut adalah 4 tahun.

b. Produk Domestik Regional Bruto

Produk Domestik Regional Bruto (PDRB) per kapita merupakan pembagian dari nilai Produk Domestik Regional Bruto dengan jumlah penduduk dalam suatu wilayah per periode tertentu. Produk Domestik Regional Bruto (PDRB) per kapita digunakan untuk mengetahui pertumbuhan nyata ekonomi per kapita penduduk di suatu wilayah. Berdasarkan hasil statistik deskriptif pada Tabel 3 di atas, dapat dilihat bahwa rata-rata PDRB per kapita kabupaten/kota di Pulau Jawa pada tahun 2019 adalah sebesar 63778. Artinya bahwa pendapatan setiap penduduk di Pulau Jawa tahun 2019 secara rata-rata sebesar Rp 63.778. Kabupaten/kota yang memiliki PDRB per kapita tertinggi adalah Kota Jakarta Pusat dengan nilainya sebesar 760271 yang artinya pendapatan tiap penduduk di wilayah tersebut secara rata-rata sebesar Rp 760.271. Sedangkan kabupaten/kota yang memiliki PDRB per kapita terendah berada di Kabupaten Pamekasan dengan nilainya sebesar 19367 yang artinya bahwa pendapatan tiap penduduk di wilayah tersebut secara rata-rata sebesar Rp 19.367.

c. Pengangguran

Pengangguran adalah suatu keadaan dimana seseorang yang tergolong dalam angkatan kerja ingin mendapatkan pekerjaan tetapi belum mendapatkannya (Sukirno dalam Franita, 2016). Pengangguran adalah keadaan dimana orang ingin bekerja namun tidak mendapatkan pekerjaan. Pada penelitian ini menggunakan tingkat pengangguran yang merupakan jumlah pengangguran terhadap jumlah angkatan kerja. Berdasarkan hasil statistik deskriptif pada Tabel 3 di atas, dapat dilihat bahwa tingkat pengangguran di Pulau Jawa tahun 2019 secara rata-rata sebesar 5,273. Artinya persentase jumlah pengangguran terhadap jumlah angkatan kerja adalah sebesar 5,27 persen. Kabupaten/kota yang memiliki tingkat pengangguran tertinggi adalah Kabupaten Serang dengan nilainya sebesar 10,650 yang artinya persentase jumlah pengangguran terhadap jumlah angkatan kerja di wilayah tersebut sebesar 10,65 persen. Sedangkan kabupaten/kota yang memiliki tingkat pengangguran terendah berada di Kabupaten Pacitan dengan nilainya sebesar 0,950 yang artinya persentase jumlah pengangguran terhadap jumlah angkatan kerja di wilayah tersebut adalah sebesar 0,95 persen.

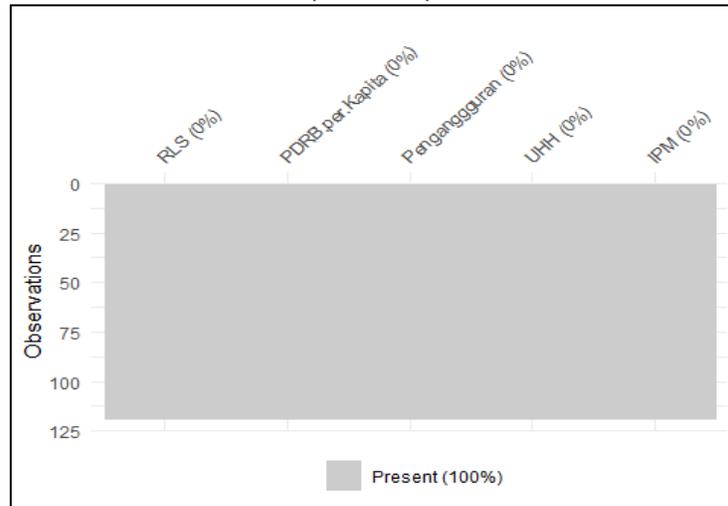
d. Umur Harapan Hidup (UHH)

Umur harapan hidup sejak lahir menggambarkan kualitas kesehatan masyarakat di suatu wilayah. Umur harapan hidup merupakan rata-rata lama umur atau masa hidup di suatu wilayah. Berdasarkan hasil statistik deskriptif pada Tabel 3 di atas, dapat dilihat bahwa rata-rata umur harapan hidup kabupaten/kota di Pulau Jawa pada tahun 2019 adalah sebesar 72,56. Artinya bahwa rata-rata lama umur atau masa hidup kabupaten/kota di Pulau Jawa adalah berkisar 72 tahun. Kabupaten/kota yang memiliki umur harapan hidup tertinggi adalah Kabupaten Sukoharjo dengan nilainya

77,55 yang artinya lama umur atau masa hidup di wilayah tersebut berkisar 77 tahun. Sedangkan wilayah yang memiliki umur harapan hidup yang terendah berada di Kabupaten Serang dengan nilainya 64,47 yang artinya bahwa lama umur atau masa hidup di wilayah tersebut berkisar 64 tahun.

b. Preprocessing dan Proses Training

Sebelum melakukan ke tahap pengujian baik metode *K-Nearest Neighbor* (k-NN) maupun *Support Vector Machine* (SVM), dilakukan data preprocessing. Dimana data dibersihkan dan ditransformasikan ke dalam bentuk kategori untuk data yang memiliki nilai kategori. Pada dataset yang meliputi Indeks Pembangunan Manusia (IPM), Rata-rata Lama Sekolah (RLS), Produk Domestik Regional Bruto per Kapita, tingkat pengangguran, dan Umur Harapan Hidup (UHH) sejak lahir tidak terdapat missing value atau data yang hilang di dalam suatu dataset. Hal tersebut dapat dilihat pada Gambar 3 di bawah ini.



Sumber : Badan Pusat Statistik, data diolah

Gambar 3. Pengecekan *Missing Value* Pada Setiap Data

Berdasarkan Gambar 3 di atas, menunjukkan bahwa untuk setiap data, yaitu Indeks Pembangunan Manusia (IPM), Rata-rata Lama Sekolah (RLS), Produk Domestik Regional Bruto per Kapita, tingkat pengangguran, dan Umur Harapan Hidup (UHH) sejak lahir tidak terdapat missing value. Hal ini dapat dilihat dari besarnya persentase yang diperoleh, dimana sebesar 100 persen.

Setelah data dilakukan preprocessing, kemudian data dibagi menjadi 2 bagian, yaitu data training dan data testing. Pada proses training dilakukan dengan menggunakan split validation. Menurut Untari (2014), split validation adalah teknik validasi yang membagi data menjadi dua bagian secara acak, yang mana sebagian data sebagai data training dan sebagian lainnya sebagai data testing. Dengan menggunakan split validation akan dilakukan percobaan training yang berdasarkan split ratio yang telah ditentukan sebelumnya. Kemudian sisa dari split ratio data training akan dianggap sebagai data testing. Data training adalah data yang akan digunakan dalam pembelajaran sedangkan data testing adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data pengujian kebenaran atau keakurasian hasil pembelajaran.

Pada penelitian ini, split ratio yang digunakan adalah 7:3, yang artinya 70 persen dari dataset digunakan sebagai data training dan sisanya yaitu 30 persen dari dataset digunakan sebagai data testing.

c. Klasifikasi dengan Metode *K-Nearest Neighbor*(k-NN)

Pada penelitian, untuk mengklasifikasi Indeks Pembangunan Manusia (IPM) termasuk tergolong rendah maupun tinggi digunakan metode *K-Nearest Neighbor* (k-NN). *K-Nearest Neighbor* (k-NN) adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama. Pada metode ini, menggunakan nilai k, yaitu 1, 5, dan 10. Tabel 4 di bawah ini menunjukkan hasil beberapa percobaan pada metode *K-Nearest Neighbor* (k-NN).

Tabel 4. Hasil Akurasi Berdasarkan Nilai k Pada Metode *K-Nearest Neighbor* (k-NN)

Nilai k (1)	Sensitivity (2)	Specificity (3)	Akurasi (4)
k = 1	0,7143	0,8182	77,78%
k = 5	0,6429	0,9545	83,88%
k = 10	0,7857	0,9091	86,11%

Sumber : Badan Pusat Statistik, data diolah

Berdasarkan Tabel 4 di atas yang memperlihatkan hasil dari beberapa percobaan nilai k pada metode *K-Nearest Neighbor* (k-NN) menunjukkan bahwa, nilai akurasi yang paling tinggi adalah dengan nilai k=10 yang keakuratannya mencapai 86,11 persen. Nilai sensitivity pada nilai k= 10 adalah sebesar 0,7857. Artinya, sebesar 78,57 persen hasil prediksi kelompok Indeks Pembangunan Manusia (IPM) yang tergolong rendah dengan menggunakan *K-Nearest Neighbor* (k-NN) sesuai dengan kelompok awal atau sebelum diprediksi. Sedangkan nilai specificity sebesar 0,9091. Artinya, sebesar 90,91 persen hasil prediksi Indeks Pembangunan Manusia (IPM) yang tergolong tinggi dengan menggunakan *K-Nearest Neighbor* (k-NN) sesuai dengan kelompok awal sebelum diprediksi.

Berikut ini hasil dari *Confusion matrix* dengan menggunakan nilai k=10, dapat dilihat pada tabel 5 di bawah ini.

Tabel 5. Hasil *Confusion Matrix*

Prediction (1)	Reference	
	Rendah (2)	Tinggi (3)
Rendah	11	2
Tinggi	3	20

Sumber : Badan Pusat Statistik, data diolah

Confusion matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matriks seperti pada Tabel 5 di atas yang menunjukkan hasil *Confusion matrix* pada nilai k=10 dengan menggunakan metode *K-Nearest Neighbor* (k-NN). Berdasarkan Tabel 4, memperlihatkan bahwa terdapat 11 data benar diprediksi dan sebanyak 2 data salah yang di prediksi pada Indeks Pembangunan Manusia (IPM) yang tergolong rendah. Dan terdapat 20 data benar yang diprediksi dan sebanyak 3 data salah yang di prediksi pada Indeks Pembangunan Manusia (IPM) yang tergolong tinggi. Dengan menggunakan *Confusion matrix*, diperoleh nilai presisi, recall, dan F-1 score. Nilai presisi yang diperoleh adalah sebesar 0,7857 yang menunjukkan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Sedangkan nilai recall adalah sebesar 0,8461 dengan nilai F-1 score sebesar 0,8148.

d. Klasifikasi dengan Metode *Support Machine Learning* (SVM)

Penelitian ini mengklasifikasi Indeks Pembangunan Manusia kabupaten/kota di Pulau Jawa pada tahun 2019 juga menggunakan metode *Support Vector Machine* (SVM). Konsep dasar dari SVM adalah bagaimana menemukan atau mencari fungsi pemisah (*hyperplane*) terbaik yang memisahkan dua kelas pada *input space* atau ruang input. Untuk menentukan *hyperplane* yang terbaik pada penelitian ini, dapat menggunakan *function tune*, dapat dilihat pada Gambar 4 di bawah ini.

```
library(e1071)
model<-data$IPM~HLS+Pengeluaran.per.Kapita.Disesuaikan+RLS+UHH

hyper<- tune(svm, train.x = training_set[,1:4], train.y = training_set[,5], type = "C",
            ranges = list(gamma = 1:10, cost = c(0.1,5,10,15,25,50,100), kernel = c
            ("linear", "radial")),
            tunecontrol = tune.control(sampling = "fix"),
            validation.x = test_set[,1:4],
            validation.y = test_set[,5])

hyper
```

Sumber : Badan Pusat Statistik, data diolah

Gambar 4. Syntax Function Tune

Berdasarkan hasil sintaks pada Gambar 4 di atas, menghasilkan bahwa hyperplane yang terbaik adalah kernel linear dengan nilai parameter yang terbaik yaitu gamma sebesar 1 dan cost sebesar 5. Setelah mendapatkan hyperplane yang terbaik, maka selanjutnya membuat model dan melakukan klasifikasi serta menentukan nilai akurasi. Berikut ini Tabel 6 yang menunjukkan parameter yang digunakan serta hasil akurasi, sensitivity, dan specificity.

Tabel 6. Parameter dan Hasil Akurasi Pada Metode *Support Vector Machine* (SVM)

Parameters		Sensitivity	Specificity	Akurasi
(1)	(2)	(3)	(4)	(5)
SVM-Type	C-classification	0,7857	0,9545	88,89%
SVM-Kernel	Linear			
Cost	5			

Sumber : Badan Pusat Statistik, data diolah

Berdasarkan Tabel 6 di atas menunjukkan bahwa jenis yang digunakan pada metode *Support Vector Machine* (SVM) adalah klasifikasi dengan fungsi pemisah (hyperplane) yang digunakan adalah kernel linear serta cost sebesar 5. Dengan fungsi pemisah (hyperplane) tersebut, diperoleh hasil nilai akurasi sebesar 88,89 persen. Hal tersebut menunjukkan model ini dapat dikatakan sangat bagus karena memiliki nilai akurasi yang besar. Nilai sensitivity yang diperoleh adalah sebesar 0,7857. Artinya, sebesar 78,57 persen hasil prediksi kelompok Indeks Pembangunan Manusia (IPM) yang tergolong rendah dengan menggunakan *Support Vector Machine* (SVM) sesuai dengan kelompok awal atau sebelum diprediksi. Sedangkan nilai specificity sebesar 0,9545. Artinya, sebesar 95,45 persen hasil prediksi Indeks Pembangunan Manusia (IPM) yang tergolong tinggi dengan menggunakan *Support Vector Machine* (SVM) sesuai dengan kelompok awal sebelum diprediksi.

Berikut ini hasil *Confusion matrix* pada fungsi pemisah kernel linear dengan cost sebesar 5, dapat dilihat pada tabel 7 di bawah ini.

Tabel 7. Hasil *Confusion Matrix* Pada Metode *Support Vector Machine* (SVM)

Prediction	Reference	
	Rendah	Tinggi
(1)	(2)	(3)
Rendah	11	1
Tinggi	3	21

Sumber : Badan Pusat Statistik, data diolah

Confusion matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matriks seperti pada Tabel 7 di atas yang menunjukkan hasil *Confusion matrix* dengan menggunakan metode *Support Vector Machine* (SVM). Berdasarkan Tabel 6, memperlihatkan bahwa terdapat 11 data benar diprediksi dan sebanyak 1 data salah yang di prediksi pada Indeks Pembangunan Manusia (IPM) yang tergolong rendah. Dan terdapat 21 data benar yang diprediksi dan sebanyak 3 data salah yang di prediksi pada Indeks Pembangunan Manusia (IPM) yang tergolong tinggi. Dengan menggunakan *Confusion matrix*, diperoleh nilai presisi, recall, dan F-1 score. Nilai presisi yang diperoleh adalah sebesar 0,7857 yang menunjukkan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Sedangkan nilai recall adalah sebesar 0,9167 dengan nilai F-1 score sebesar 0,8462.

e. Perbandingan Metode *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM)

Nilai *k* pada metode *K-Nearest Neighbor* (k-NN) yang digunakan dan nilai parameter pada metode *Support Vector Machine* (SVM) yang digunakan dijadikan input untuk proses klasifikasi sehingga menghasilkan tingkat akurasi. Berikut ini, diperoleh perbandingan tingkat akurasi pada metode *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM).

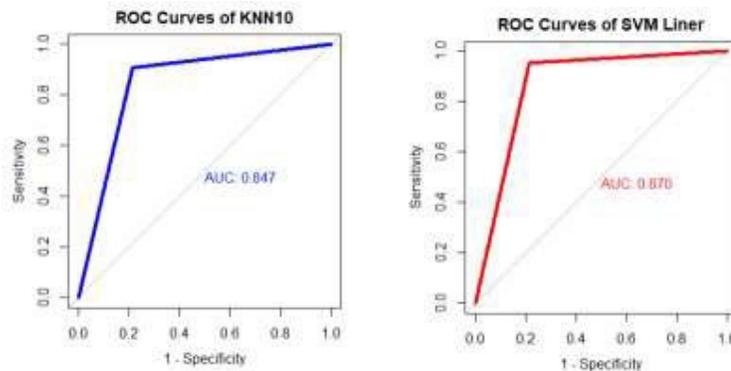
Tabel 8. Perbandingan Hasil Akurasi 2 Metode

Metode	Akurasi
(1)	(2)
<i>K-Nearest Neighbor</i> (k-NN)	86,11%
<i>Support Vector Machine</i> (SVM)	88,89%

Sumber : Badan Pusat Statistik, data diolah

Pada Tabel 8 di atas, memperlihatkan bahwa pengklasifikasian dataset Indeks Pembangunan Manusia (IPM) dengan menggunakan metode *K-Nearest Neighbor* (k-NN) diperoleh hasil akurasinya sebesar 86,11 persen dengan nilai *k* yang dipilih adalah 10. Sedangkan pengklasifikasian dengan menggunakan metode *Support Vector Machine* (SVM) diperoleh hasil akurasi sebesar 88,89 persen dengan fungsi pemisah (hyperplane) yang digunakan adalah kernel-linear serta cost sebesar 5 dan gamma sebesar 1. Perbandingan dua metode tersebut dengan hasil akurasi menunjukkan bahwa metode yang terbaik untuk pengklasifikasian Indeks Pembangunan Manusia (IPM) adalah metode *Support Vector Machine* (SVM).

Selain menggunakan hasil akurasi, kurva ROC juga dapat menentukan metode mana yang terbaik. Berikut ini hasil kurva ROC baik pada metode *K-Nearest Neighbor* (k-NN) maupun *Support Vector Machine* (SVM), dapat dilihat pada Gambar 5 di bawah ini.



Sumber : Badan Pusat Statistik, data diolah

Gambar 5. Kurva ROC Pada Metode *K-Nearest Neighbor* (k-NN) maupun *Support Vector Machine* (SVM)

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC merupakan grafik dua dimensi dengan false positives sebagai garis horizontal dan true positive sebagai garis vertikal. Sedangkan nilai AUC (Area Under Curve) adalah area di bawah kurva ROC, suatu kurva yang menggambarkan probabilitas dengan variabel sensitivitas dan kekhususan (specificity). AUC adalah salah satu jenis statistik akurasi untuk model probabilitas dalam penilaian/analisis (Nefeslioglu drr dalam Sadisun dkk, 2018). Berdasarkan Gambar 5 di atas menunjukkan bahwa hasil Kurva ROC dan nilai AUC dari masing-masing metode *K-Nearest Neighbor* (k-NN) maupun *Support Vector Machine* (SVM). Pada metode *K-Nearest Neighbor* (k-NN) dengan nilai $k=10$ diperoleh nilai AUC sebesar 0,847. Sedangkan pada metode *Support Vector Machine* (SVM) diperoleh nilai AUC sebesar 0,870.

Untuk klasifikasi data mining, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu dalam Leidiana, 2013).

0,90-1,00	= klasifikasi sangat baik
0,80-0,90	= klasifikasi baik
0,70-0,80	= klasifikasi cukup
0,60-0,70	= klasifikasi buruk
0,50-0,60	= klasifikasi salah

Berdasarkan kelompok klasifikasi di atas, untuk metode *K-Nearest Neighbor* (k-NN) dengan nilai $k=10$ termasuk ke dalam kelompok klasifikasi baik dan untuk metode *Support Vector Machine* (SVM) termasuk ke dalam kelompok klasifikasi baik. Sehingga dengan melihat hasil akurasi dan kurva ROC serta nilai AUC, diperoleh bahwa metode klasifikasi Indeks Pembangunan Manusia (IPM) yang terbaik pada penelitian ini adalah metode *Support Vector Machine* (SVM).

4. Kesimpulan dan Saran

Berdasarkan hasil dan pembahasan di atas, dapat disimpulkan bahwa klasifikasi Indeks Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa pada tahun 2019 dengan menggunakan metode *K-Nearest Neighbor* (k-NN) diperoleh nilai k yang memiliki akurasi tertinggi adalah $k=10$ dengan nilai akurasinya sebesar 86,11 persen. Sedangkan klasifikasi Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa pada tahun 2019 dengan menggunakan metode *Support Vector Machine* (SVM) diperoleh parameter yang terbaik adalah kernel linear dengan nilai gamma sebesar 1 dan nilai cost sebesar 10 dengan nilai akurasi yang diperoleh sebesar 88,89 persen. Selain melihat nilai akurasi, perbandingan antara kedua metode dapat dilihat dari kurva ROC dan nilai AUC. Dimana pada metode *K-Nearest Neighbor* (k-NN) diperoleh nilai AUC sebesar 0,847 dengan klasifikasi baik sedangkan pada metode *Support Vector Machine* (SVM) sebesar 0,870 dengan klasifikasi sangat baik. Sehingga metode yang lebih baik untuk mengklasifikasi Indeks Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa pada tahun 2019 adalah dengan menggunakan metode *Support Vector Machine* (SVM). Adapun manfaat dari penelitian ini bagi pemerintah adalah dapat mengetahui apakah suatu wilayah memiliki Indeks Pembangunan Manusia yang rendah atau tinggi, sehingga pemerintah dapat mengambil kebijakan atau keputusan khususnya bagi wilayah yang memiliki Indeks Pembangunan Manusia yang rendah. Selain itu, manfaat bagi penelitian lainnya dapat dijadikan sebagai referensi khususnya dalam kasus mengklasifikasi Indeks Pembangunan Manusia.

Berdasarkan hasil penelitian ini disarankan untuk penelitian selanjutnya perlu dilakukan studi lebih lanjut dalam kasus klasifikasi Indeks Pembangunan Manusia, seperti dengan menggunakan metode klasifikasi yang berbeda serta dilakukan perbandingan dengan beberapa algoritma lainnya sehingga diperoleh metode klasifikasi dengan tingkat akurasi yang sangat baik. Selain itu, dapat melibatkan variabel lainnya, seperti angka kemiskinan, kepadatan penduduk, pengeluaran per kapita, dan lain sebagainya.

Daftar Pustaka

- [1] Badan Pusat Statistik (BPS). (2020). Produk Domestik Regional Bruto Kabupaten/Kota di Indonesia 2015 – 2019. BPS, Indonesia.
- [2] Badan Pusat Statistik (BPS). (2020). Indeks Pembangunan Manusia menurut Kabupaten/Kota [Metode Bary], 2010-2019. BPS, Indonesia.
- [3] Darsyah, M. Y, "Kasifikasi Indeks Pembangunan Manusia (IPM) dengan Pendekatan *K-Nearest Neighbor*(K-NN)" *In Prosiding Seminar Nasional & Internasional*, 2017.
- [4] Fauzi, F, "*K-Nearest Neighbor*(K-NN) dan *Support Vector Machine*(SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah" *Jurnal Mipa*, 40(2), 118-124, 2017.
- [5] Franita, R. (2016). Analisa pengangguran di Indonesia. *Jurnal Ilmu Pengetahuan Sosial*, 1, 88-93.
- [6] Gunadi, G., & Sensuse, D. I, Penerapan Metode Data Mining Market Basket Analysis terhadap Data Penjualan Produk Buku dengan Menggunakan Algoritma Apriori dan Frequent Pattern Growth (fp-growth): Studi Kasus percetakan pt. Gramedia, *Telematika MKOM*, 2016, 4(1), 118-132.
- [7] Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Semantik*, 2(1).
- [8] Khomarudin, A. N, "Teknik data mining: Algoritma k-meansclustering", *Ilmu Komputer*, 2016.
- [9] Krisandi, N., & Helmi, B. P, "Algoritma *K-Nearest Neighbor* dalam Klasifikasi Data Hasil Produksi Kelapa Sawit pada PT. MinamasKecamatan Parindu", *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, 2(1), 33-38, 2013.
- [10] Leidiana, H. (2013). Penerapan algoritma *K-Nearest Neighbor* untuk penentuan resiko kredit kepemilikan kendaraan bermotor. *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, 1(1), 65-76.
- [11] Mayadewi, P., & Rosely, E. (2015). Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining. *SESINDO 2015*, 2015.
- [12] Melliana, A., & Zain, I, "Analisis Statistika Faktor yang Mempengaruhi Indeks Pembangunan Manusia di Kabupaten/Kota Provinsi Jawa Timur dengan Menggunakan Regresi Panel", *Jurnal Sains dan Seni ITS*, 2(2), D237-D242, 2013.
- [13] Prangga, S. (2017). Optimasi Parameter pada *Support Vector Machine* menggunakan Pendekatan Metode Taguchi untuk Data High-Dimensional (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).
- [14] Pratowo, N. I, "Analisis Faktor-faktor yang Berpengaruh terhadap Indeks Pembangunan Manusia", *Jurnal Studi Ekonomi Indonesia*, 1(1), 15-31, 2012.
- [15] Ridwan, M., Suyono, H., & Sarosa, M, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier", *Jurnal EECCIS*, 7(1), 59-64, 2013.
- [16] Purba, M. A. (2019). Implementasi Algoritma Nearest Neighbor dalam Mengukur Tingkat Kepuasan Masyarakat pada Pelayanan Samsat Medan Selatan. *Pelita Informatika: Informasi dan Informatika*, 8(2), 551-556.
- [17] Rohman, A., & Rochcham, M, "Komparasi Metode Klasifikasi Data Mining untuk Prediksi Kelulusan Mahasiswa", *Neo Teknika*, 5(1), 2019.
- [18] Sadisun, I. A., Kartiko, R. D., & Arifianti, Y, "Metode Kombinasi Weight of Evidence (WoE) dan Logistic Regression (LR) untuk Pemetaan Kerentanan Gerakan Tanah di Takengon", *Jurnal Lingkungan dan Bencana Geologi*, 9(2), 77-86, 2018.
- [19] Untari, D, "Data Mining untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4. 5", *Universitas Dian Nuswantoro*, 2014.
- [20] Umaidah, Y., & Purwanto, P. (2019). Penerapan algoritma *K-Nearest Neighbor* (K-Nn) dPencarian Optimal untuk Prediksi Prestasi Siswa. *Journal of Information System, Informatics and Computing*, 3(2), 1-8.