Seminar Nasional Pendidikan, Sains dan Teknologi **ISBN : 978-602-61599-6-0**
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Muhammadiyah Semarang

KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA (IPM) DENGAN PENDEKATAN K-NEARSET NEIGHBOR (K-NN)

[1]Moh. Yamin Darsyah
[1]Prodi Statistika FMIPA, Universitas Muhammadiyah Semarang, Semarang
Email: mydarsyah@unimus.ac.id

Abstract

Human development index (HDI) is one of measuring instrument of achieving quality of life of one region even country. There are three basic components of the Human Development Index compilers: health dimension, knowledge dimension, and decent living dimension. To measure the health dimension, we use life expectancy at birth, knowledge dimension is used combination of indicator of old school expectation and mean of school length, and life dimension suitable for use indicator ability of people purchasing power to some basic requirement seen from mean of expense per customized capita. Data mining works to gather information from a large amount of data. Jobs that are closely related to data mining are prediction models, group analysis, association analysis, and anomaly detection. One of the classification methods contained in data mining and is often used and produces a fairly good accuracy is the K-Nearset Neighbor (k-NN) method. The absence of research on the classification or grouping of Human Development Index with K-Nearset Neighbor (k-NN) method will be done by using k-NN method with k value of 1, 5, and 10. With the ultimate goal of comparing the accuracy of kaslifikasi between value k on the k-NN method. The result of classification of IPM by using k-NN method with k value of 5 and 10 obtained classification accuracy of 91.43% which is the best classification accuracy, with sensitivity of 100% and 83.33%.

Key words: *HDI , Classification, K-Nearset Neighbor, Accuracy*

1. PENDAHULUAN

Indeks pembangunan manusia adalah salah satu alat ukur pencapaian kualitas hidup satu wilayah bahkan negara. Terdapat 3 komponen dasar peyusun Indeks Pembangunan Manusia(IPM) yaitu dimensi kesehatan, dimensi pengetahuan, dan dimensi hidup layak (BPS, 2014). Untuk mengukur dimensi kesehatan, digunakan angka harapan hidup waktu lahir. Selanjutnya untuk mengukur dimensi pengetahuan digunakan gabungan indikator harapan lama sekolah dan rata-rata lama sekolah. Dimensi hidup layak digunakan indikator kemampuan daya beli masyarakat terhadap sejumlah kebutuhan pokok yang dilihat dari rata-rata besarnya pengeluaran per kapita disesuaikan.

Indeks Pembangunan Manusia (IPM) menurut Badan Pusat Statistik (BPS) dibagi menjadi 4 kategori atau golongan yaitu Indeks Pembangunan Manusia (IPM) rendah jika <60, sedang 60≤IPM<70, tinggi 70≤IPM<80, dan ≥80 sangat tinggi(BPS:2014). Karena pembangunan di Indonesia tidak merata maka Indeks Pembangunan Manusia(IPM) di wilayah-wilayah terutama kabupaten/kota sangatlah beragam.

Data mining adalah serangkaian proses untuk menggali suatu nilau tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data yang sangat besar (Prambudiono, 2007). Data mining bekerja mengumpulkan informasi dari sejumlah data yang besar. Pekerjaan yang berkaitan erat dengan data mining adalah model prediksi (*prediction modelling*), analisis kelompok (*cluster analysis*), analisis asosiasi (*association analysis*), dan deteksi anomaly (Prasetyo, 2012). Metode yang terdapat dalam data mining dantaranya *Artifical Neural Network*(ANN), *Support Vector Machine* (SVM), *Support Vector Regression* (SVR), dan *k Nearset Neighbor* (k-NN).

k-nearset neighbor (k-NN) merupakan salah satu metode klasifikasi yang terdapat dalam data mining dan termasuk dalam kelompok *instace-beased learning.*kNN dilakukan dengan mencari dengan mencari k objek dalam data training yang paling dekat(mirip)

dengan objek pada data testing (Wu, 2009). Data training diproyeksikan kedalam ruang yang berdimensi banyak, dimana masing-masing dimensi mempresentasikan fitur dari data. Dekat atau jauhnya tetangga/data bisa dihitung berdsarkan jarak Euclidean.

Penelitian terdahulu tentang metode *k Nearset Neighbor* (k-NN) *dilakukan* oleh Leidiyana (2013) yang berjudul "Algoritma K-Nearset Neighebor untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor" memberikan akurasi klasifikasi sebesar 81.46%. Penelitian SVM oleh Darsyah dan Darmawati (2017) memberikan akurasi 98% pada pasien TB Paru. Penelitian mengenai IPM pernah dilakukan oleh Darsyah dan Wasono (2013) dengan judul "Pendugaan IPM Pada Area Kecil Di Kota Semarang dengan Pendeketakan Nonparametrik". Fauzi, Darsyah, Utami (2017) " Smooth Support Vector Machine Untuk Pengklasifikasian IPM Se Indonesia. Belum adanya penelitian tentang kasus klasifikasi Indeks Pembangunan Manusia(IPM) dengan metode *k Nearset Neighbor* (k-NN). Penelitian ini bertujuan mengetahui akurasi klasifikasi IPM dengan metode *k*-NN dengan nilai *k* sebesar 1, 5, dan 10, kemudian mencari nilai *k* manakah yang menghasilkan akurasi klasfikasi IPM paling akurat di Jawa Tengah.

2. TINJAUAN PUSTAKA

Data Mining

Data mining adalah suatu teknolgi terbaru dimana menggabungkan metode analisis tradisional dengan algoritma yang mampu memroses data dengan jumlah besar. Data mining adalah suatu istiliah yang digunakan untuk mencari informasi yang tersembunyi didalam data yang besar. Data mining merupakan metode yang yang mengabungkan tida ilmu yaitu statistik, metematika, kecerdasan buatan, dan machine learning. Beberapa definsi data mining antara lain data mining disebut juga serangkaian proses untuk menggali suatu nilau tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data yang sangat besar (Prambudiono, 2007). Data mining atau *Machine Learning* adalah strategi semi empiris yang menggunakan data tentang sifat/*properties* dan deskripor (Lu*et al*., 2017).

k Nearset Neighbor (kNN)

k-nearset neighbor (kNN) adalah suatu metode klasifikasi yang terdapat dalam data mining selain *Support Vector Machine* (SVM).kNN dilakukan dengan mencari dengan mencari k objek dalam data *training* yang paling dekat(mirip) dengan objek pada data testing (Wu, 2009).

Adapun cara untuk mengukur *kedekatan* antara data baru dengan data yang lama (data trining), diantaranya Euclidean distance dan mahattan distance. Yang paling seing digunakan untuk mengukur kedekatan antar data adalah euclidean distance(Bramer, 2007) yaitu:

*Dimana*mewakili n nilai atribut dari dua *record*.

Untuk *megukur* jarak dari atribut yang mempunyai nilai besar, maka dilakukan normalisasi. Normalisasi bisa menggunakan *min-max normalization* atau *Z-score standardization* (Larose,2005). Untuk menghitung kemiripan kasus digunakan rumus:

keterangan:
p = kasus baru
q = kasus yang ada dalam penyimpanan
n = jumlah atribut dalam tiap khasus
i = atirbut individu antara 1 sampai dengan n
f = fungsi similarity atribut i antara kasus p dan q
w = bobot yang diberikan pada atribut ke-i

K-*fold cross validation* (KCV)

K-fold cross validation merupaka teknik untuk membagi dokumen menjadi k bagian. Pembagian tersebut merupakan pembagian sebagi data *training* dan *test set*. Seluruh data secara acak dibagi menjadi K buah subset dengan ukuran yang sama dimana himpunan bagian dari {1, …, n}, sedemikian sehingga dan . Kemudian dilakukan tahap iterasi sebanyak k kali, kemudian pada iterasi ke k subset menjadi *test test* sedangkan sisanya menjadi *training set*. Kelebihan dari metode k-*fold cros validation* adalah tidak adanya permasalahan dalam pembagian data.

Pengukuran Kinierja Klasifikasi

Diharapkan dalam suatu kaslifikasi semua data dapat diklasifikasi dengan benar, tetapi terkadang tidak bisadiga klasifikasi yang didapat tidak mencapai 100% benear sehingga sebuah sistem kalsifikasi juga harus diukur kinerjanya. Dalam pengukuran kenerja klasifikasi umunya dilakukan dengan matriks konfusi (*confusion matrix*).

Pada tabel 2.3 Merupakan contoh dari tabel konfungsi untuk mengukur hasil kerja klasifikasi dengan masalah 2 kelas (*biner*). Hanya ada dua kelas yaitu kelas 0 dan kelas 1. Setiap dalam matriks menyatakan jumlah *record*/ data dari kelas yang prediksinya masuk kedalam kelas. Sedangkan rumus untuk menghitung Sensitivitas dan sepesifitas berdasarkan tabel 1 sebagai berikut:

Tabel 1. Matriks Konfusi untuk Dua Kelas(Eko Prasetyo, 2012)

		Kelas hasil prediksi(	
		Kelas = 1	Kelas = 0
Kelas asli(	Kelas = 1		
	Kelas = 0		

Untuk mengitung hasil akurasi menggunakan formula (Prasetyo, 2012)

3. METODE

Secara umum tahapan penelitian ini yaitu pengumpulan data, analisis data, dan kesimpulan. Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Jawa Tengah tentang Indeks Pembangunan Manusia, Angka Harapan Hidup,Rata-rata Lama Sekolah, Harapan Lama Sekolah, dan pengeluaran perkapita yang disesuaikan tahun 2016 dengan jumlah 35 data. Dengan struktur data terdapat pada tabel 2 berikut:

Tabel 2. Struktur Data

Variabel	Definisi	Kategori	Sekala
	Indeks Pembangunan Manusia (IPM)	1 = rendah (IPM rendah, IPM sedang) 2 = tinggi (IPM tinggi, IPM sangat tinggi)	Nominal
	Angka Harapan Hidup (AHH)	-	Rasio
	Rata-rata Lama Sekolah (RLS)	-	Rasio
	Harapan Lama Sekolah(HLS)	-	Rasio
	Pengeluaran Perkapita yang Disesuaikan(PPD)	-	Rasio

ISBN : 978-602-61599-6-0

Langkah-langkah analisis data dalam penelitian ini adalah sebagi berikut

1. Melakukan pengumpulan data sekunder, yaitu Indeks Pembangunan Manusia(IPM), Angka Harapan Hidup(AHH), Rata-rata Lama Sekolah(RLS), Harapan Lama sekolah(HLS), dan Pengeluaran Perkapita yang Disesuaikan(PPD) provinsi Jawa Tengah tahun 2016.
2. Melakukan pengkodingan terhadap variabel prediktor.
3. Membagi data menjadi data training dan data testing dengan menggunakan *stratified* 10 *fold-cross validation*.
4. Melakukan pengklasifikasian *k nearset neighbor* (kNN) dengan algoritma sebagai berikut:
    a. Menentukan jumlah nilai k untuk *k nearset neighbor* yaitu 1, 5, dan 10.
    b. Menghitung akurasi klasifikasi.

4. HASIL DAN PEMBAHASAN

Deskriptif Statistik

Rata-rata Indeks Pembangunan Manusia(IPM) di Provinsi Jawa Tengah Tahun 2016 adalah sebesar 70.61, IPM tertinggi yaitu 81.19, dan terendah 63.98. Untuk jumlah penduduk dengan kriteria kurang baik, dan baik dapat dilihat dalam pada gambar 1:

Gambar 1. Deskriptif IPM

Berdarkan gambar 1 Indeks Pembangunan Manusia dengan kriteria kurang baik terdapat 17 kabupaten/kota sedangkan kriteria baik terdapat 18 kabupaten/kota.

*K Nearset Neighbor*dengan K=1

Hasil klasifikasi IPM metode k-NN dengan nilai k=1 dirangkum dalam *confusion matrix* pada tabel 3 sebagai berikut:

Tabel 3. *Confusion matrix* k=1

		Kelas hasil prediksi(		Sensitivitas	Spesivitas	Akurasi
		Kelas = kurang baik	Kelas = baik			
Kelas asli(	Kelas = kurang baik	12	3	70,69%	83,33%	77,14%
	Kelas = baik	5	15			

Berdsarkan tabel 3 kelompok IPM kurang baik yang diprediksi secara tepat berjumlah 12 kabupaten/kota dan 3 kabupaten/kota diperdiksi tidak tepat. Sedangkan untuk kelas IPM baik 15 kabupaten/kota diprediksi secara secara tepat, dan 5 kabupaten/kota diprediksi tidak tepat. Didapat akurasi sebesar 77,14% untuk k-NNdengan nilai k=1. Sedangkan untuk sensitivitas dan spesivitas didapat sebesar 70,69% dan 83,33%, artinya 77,14% hasil prediksi kelompok IPM kurang baik dengan menggunakan k-NN k=1 sesuai dengan kelompok awal atau sebelum diprediksi, sedangkan untuk sepesitivitassebesar 83,33% artinya 83,33% hasil prediksi IPM baik menggunakan metode k-NN k=1 sesuai dengan kelompok awal sebelum diprediksi.

*K Nearset Neighbor*dengan K=5

Hasil klasifikasi IPM metode k-NN dengan nilai k=5 dirangkum dalam *confusion matrix* pada tabel 3 sebagai berikut:

Tabel 4. *Confusion matrix* k=5

		Kelas hasil prediksi		Sensitivitas	Spesivitas	Akurasi
		Kelas = kurang baik	Kelas = baik			
Kelas asli	Kelas = kurang baik	17	3	100%	83,33%	91,43%
	Kelas = baik	0	15			

Berdasarkan tabel 4 kelompok IPM kurang baik yang diprediksi secara tepat berjumlah 17 kabupaten/kota dan 3 kabupaten/kota diperdiksi tidak tepat. Sedangkan untuk kelas IPM baik 15 kabupaten/kota diprediksi secara secara tepat. Sensitivitas dan spesivitas didapat sebesar 100% dan 83,33%, artinya 100% hasil prediksi kelompok IPM kurang baik dengan menggunakan k-NN k=5 sesuai dengan kelompok awal, sedangkan untuk sepesitivitassebesar 83,33% artinya 83,33% hasil prediksi IPM baik menggunakan metode k-NN k=5 sesuai dengan kelompok awal sebelum diprediksi. Akurasi klasifikasi didapat sebesar 91.43%.

*K Nearset Neighbor*dengan K=10

Hasil klasifikasi IPM metode k-NN dengan nilai k=10 dirangkum dalam *confusion matrix* pada tabel 3 sebagai berikut:

Tabel 5. *Confusion matrix* k=10

		Kelas hasil prediksi(		Sensitivitas	Spesivitas	Akurasi
		Kelas = kurang baik	Kelas = baik			
Kelas asli(	Kelas = kurang baik	17	3	100%	83,33%	91,43%
	Kelas = baik	0	15			

Berdasarkan tabel 5 kelompok IPM kurang baik yang diprediksi secara tepat berjumlah 17 kabupaten/kota dan 3 kabupaten/kota diperdiksi tidak tepat. Sedangkan untuk kelas IPM baik 15 kabupaten/kota diprediksi secara secara tepat. Akurasi yang didapat untuk k-NN k=10 sebesar 91.43%.Sensitivitas dan spesivitas didapat sebesar 100% dan 83,33%, artinya 100% hasil prediksi kelompok IPM kurang baik dengan menggunakan k-NN k=10 sesuai dengan kelompok awal, sedangkan untuk sepesitivitassebesar 83,33% artinya 83,33% hasil prediksi IPM baik menggunakan metode k-NN k=10 sesuai dengan kelompok awal sebelum diprediksi.

Penentuan Model *K Nearset Neighbor* terbaik

Berikut perbandingan akurasi klasifikasi Indeks Pembangunan Manusia (IPM) dengan metode *kNearset Neighbor*(k-NN) dengan nilai k=1, 5, dan 10. Tabel 6. ketepatan prediksi berdsarkan nilai k pada *kNearset Neighbor*(k-NN)

Tabel 6. Model k-NN terbaik

nilai K	Sensitivitas	Spesivitas	Akurasi
k=1	70,69%	83,33%	77.14%
k=5	100%	83,33%	**91.43%**
k=10	100%	83,33%	**91.43%**

Berdasasrkan tabel 6.bahwa akurasi hasil prediksi *kNearset Neighbor*(k-NN) paling tinggi adalah dengan nilai k=5, 10, keakuratan mencapai 91.64% dengan sensitivitas dan spesivitas . Untuk klasifikasi data mining, nilai akurasi dapat dibagi menjadi beberapa kelompok(Gorunescu, 2011).

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

Berdsarkan teori di atas metode yaitu *kNearset Neighbor*(k-NN) dengan =k=1 tergolong klasifikasi cukup baik sedangkan k=5 dan 10 tergolong sangat baik.

5. KESIMPULAN

Berdasarkan pembahasan di atas dapat disimpulkan bahwa rata-rata Indeks Pembangunan Manusia(IPM) Provinsi Jawa Tengah tahun 2016 adalah70.61, dengan IPM tertinggi di Kota Semarang sebesar 81.19, dan IPM terendah di Kabupaten Brebes sebesar 63.98. Hasil akurasi tertinggi/terbaik klasifikasi IPM dengan metode *kNearset Neighbor*(k-NN) dengan nilai k=5 dan 10 dengan tingkat akurasi mencapai 91.43%, dengan sensitivitas 100% dan spesivitas 83,33%.

6. REFERENSI

BPS.2014. *Indeks Pembangunan Manusia Metode Baru*.Jakarta:BPS.

-----. 2017. *Jawa Tengah dalam Angka 2017*. Semarang:BPS Jawa Tengah.

Bramer, M. 2007. *Principles of Data Mining*. London:Springer.

Darsyah, M.Y. dan Wasono, R. 2013. *Pendugaan IPM Pada Area Kecil Di Kota Semarang dengan Pendeketakan Nonparametrik. Prosiding 10th Seminar Nasional Staitstika 2013*. Semarang: Universitas Diponegoro.

Darsyah, M.Y. dan Darmawati, S. 2017. *Support Vector Machine For Classification Of Pulmonary Tuberculosis In Semarang* . Journal Of Advanced Science Letters. American Scientific Publishers. USA

Fauzi, F., Darsyah, M.Y., Utami, T.W. 2017. *Smooth Support Vector Machine Untuk Pengklasifikasian Indek Pembangunan Manusia (IPM) Se Indonesia. Jurnal Statistika*. Semarang: UNIMUS

ISBN : 978-602-61599-6-0

Gorunescu, Florin. 2011. *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg:Springer.

Larose, D.T. 2005. *Discovering Knowladge in Data*. New Jersey:John Willey & Sons, Inc.

Leidiyana, H. 2013. *Penerapan Algoritma K-Nearset Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor*. *Jurnal Penelitian Ilmu Komputer*, 1(1), 65-76.

Lu, W., R Xiao, J. Yang, H. Li, dan W. Zhang. 2017. *Data mining-aided materals discovery and optimization*. *J Materomics*.1-11.

Pramudiono, I. 2007. *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data (Online)*. (http://www.ilmukomputer.org/wp-content/uploads/2006/08/iko-datamining.zip, diakses tanggal 16 April 2017).

Prasetyo, E.2012. *Data Mining Konserp dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Andi.

Wu, Xindong & Kumar, V.2009. *The Top Ten Algorithms in Data Mining*. (V. Wu, Xindong & Kumar, Ed.). USA: Taylor & Francis Group.