PERBANDINGAN AKURASI ANTARA METODE *K-NEAREST NEIGHBOR* (KNN) DAN *ARTIFICIAL NEURAL NETWORK* (ANN) UNTUK KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA KABUPATEN/KOTA DI PULAU JAWA

Garwita Widyadhana Putri[1)], Mochamad Husni[2)] , Rahayu Widayanti[3)]
Sistem Informasi, STMIK PPKIA Pradnya Paramita Malang
witawidyaa12@gmail.com [1)], husni@stimata.ac.id [2)], rahayu@stimata.ac.id [3)]

Abstract

In 2021, 56.1% of Indonesia's population was on the Java Island, so the government needs to do mapping for policy making in various fields. Classifying the Human Development Index (HDI) can be done to assist the government in measuring the results of human resource development. The purpose of this research is to compare the accuracy results of two methods, namely K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) to classify the HDI of districts/cities in Java Island. The results showed that the application of KNN and ANN methods on the same data resulted in different accuracy values. In the KNN method, using 80%-20% training and testing data, the K=7 value shows the highest accuracy rate [95.83%]. While the ANN method with the split 70%-30%, resulting in the highest accuracy value [94.44%]. By calculation, KNN method produces a higher accuracy value. However, the evaluation results using Fold Cross Validation show that the best KNN model is at K=3, with a mean score 84.85%. While in the model of the highest accuracy value of the ANN method, there is overfitting. Based on this comparison, it can be concluded that the highest accuracy value of both KNN and ANN methods both have weaknesses.

Keywords: *Akurasi*, *K-Nearest Neighbor (KNN)*, *Artificial Neural Network (ANN)*

1. PENDAHULUAN

Saat ini Indonesia telah memasuki fase bonus demografi, dimana sebagian besar penduduk Indonesia merupakan penduduk usia kerja (15-64 tahun) [1]. Terjadinya masa bonus demografi ini, membuat pemerintah harus meningkatkan pemerataan pembangunan pada seluruh wilayah Indonesia untuk menyesuaikan dengan pertumbuhan penduduknya. Pemerataan pembangunan merupakan sebuah upaya dari pemerintah untuk menyejahterakan masyarakat dan mengatasi ketimpangan [2]. Salah satu pemerataan pembangunan yang dilakukan oleh pemerintah adalah pada pembangunan manusia.

Pembangunan manusia telah dilakukan di wilayah-wilayah Indonesia, salah satunya di Pulau Jawa. Pada tahun 2021, sekitar 56,01% dari jumlah penduduk Indonesia berdomisili di Pulau Jawa [3]. Peningkatan jumlah penduduk mengharuskan pemerintah lebih memperhatikan pembangunan di Pulau Jawa. Oleh karena itu, diperlukan penyusunan Indeks Pembangunan Manusia (IPM) untuk mengukur hasil dari proses peningkatkan kualitas hidup penduduk di kabupaten/kota di Pulau Jawa. Selain itu, IPM dapat digunakan oleh pemerintah untuk menentukan kebijakan pada daerah-daerah di Pulau Jawa.

IPM sendiri dibentuk dari 3 (tiga) dimensi dasar, yaitu umur panjang dan hidup sehat, standar hidup layak, dan pengetahuan [4]. Berdasarkan 3 (tiga) dimensi dasar tersebut, diperoleh 4 (empat) indikator yang menjadi perhitungan IPM, yaitu indikator Umur Harapan Hidup Saat Lahir yang merujuk pada dimensi umur panjang dan hidup sehat, Rata-Rata Lama Sekolah dan Harapan Lama Sekolah yang merujuk pada dimensi pengetahuan, dan Pengeluaran Per Kapita Disesuaikan yang merujuk pada dimensi standar hidup layak. BPS juga menjelaskan bahwa indikator-indikator tersebut kemudian diolah sesuai dengan metode yang digunakan hingga mendapatkan hasil yang selanjutnya dikelompokkan menjadi 4 (empat) kategori berdasarkan UNDP, yaitu Indeks Pembangunan Manusia (IPM) rendah untuk hasil (< 60), sedang ($60 \leq IPM < 70$), tinggi ($70 \leq IPM < 80$), dan sangat tinggi (≥ 80) [5]. Dalam perhitungan IPM, terdapat beberapa metode klasifikasi

yang telah digunakan oleh peneliti sebelumnya , salah satunya pada penelitian yang menyimpulkan bahwa metode *Artificial Neural Network* (ANN) lebih unggul dengan nilai akurasi 97,4% sedangkan metode *Support Vector Machine* (SVM) menghasilkan nilai akurasi 53,25% [6]. Selain itu, terdapat penelitian lain yang menggunakan metode *K-Nearest Neighbor* (KNN) dan SVM untuk mengklasifikasi IPM di Jawa Tengah [7]. Dalam penelitiannya, menunjukkan hasil bahwa metode SVM lebih unggul dibanding metode KNN dengan nilai akurasi SVM 91,64%.

Berdasarkan hasil dari penelitian sebelumnya, maka penelitian ini akan menggunakan dua metode yang sebelumnya telah dipakai pada dua penelitian berbeda yaitu metode *K-Nearest Neighbor* (KNN) dan *Artificial Neural Network* (ANN) untuk mengklasifikasikan IPM kabupaten/kota di Pulau Jawa tahun 2021. Tujuan dari penelitian ini adalah membandingkan hasil akurasi model KNN dan (ANN).

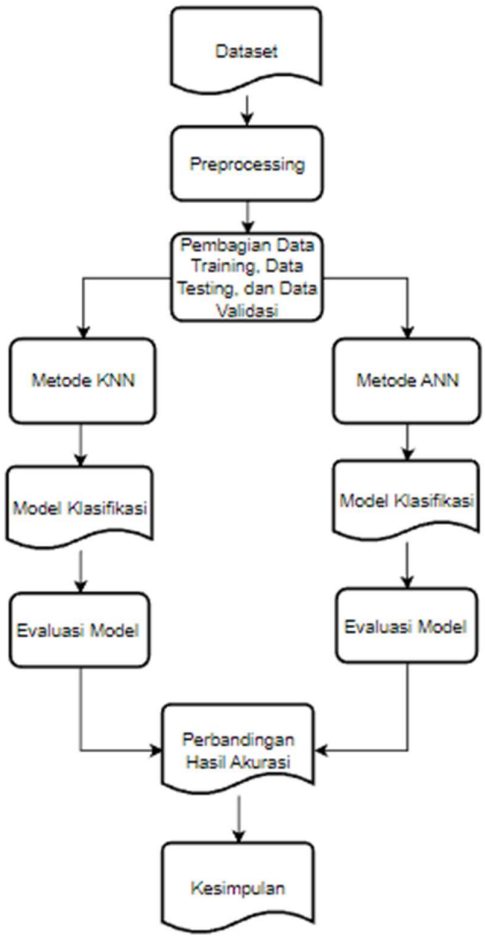
2. KAJIAN LITERATUR

Data Mining merupakan cabang ilmiah baru yang digunakan untuk mempermudah pengambil keputusan dalam menganalisis dan mengekstrasi data [8]. Sedangkan menurut pendapat lain, data mining adalah kegiatan penggalian atau penambangan informasi dari data yang besar/banyak [9].

K-Nearest Neighbor (KNN) merupakan metode pengklasifikasian objek berdasarkan contoh yang paling mirip dengan karakteristik objek tersebut [10]. Selain itu, KNN merupakan sebuah kasus dimana *learning* algoritmanya berdasarkan jarak atau kesamaan fungsi untuk berbagai pengamatan seperti fungsi *Euclidean Distance* [11].

Aritificial Neural Network (ANN) adalah sistem untuk pemrosesan data yang mengambil inspirasi dari konfigurasi otak manusia yang terbuat dari blok-blok bangunan berbasis neuron yang diakui sebagai komponen dari hubungan yang saling memperkuat untuk memecahkan masalah saat ini. [12].

3. METODE PENELITIAN



Gambar 1. *Framework* Penelitian

A. DATASET

Dataset yang digunakan dalam penelitian ini adalah data yang tersedia secara online di website Badan Pusat Statistik (BPS). Data yang diambil berupa 4 indikator perhitungan IPM, yaitu Umur Harapan Hidup Saat Lahir (UHH), Rata-Rata Lama Sekolah (RLS), Harapan Lama Sekolah (HLS), Pengeluaran yang Disesuaikan, dan Keterangan Label Status IPM kabupaten/kota di Pulau Jawa tahun 2021. Dataset yang digunakan sebagai dataset berjumlah 119 kabupaten/kota di Pulau Jawa.

B. PREPROCESSING

Preprocessing merupakan proses mengolah dataset murni menjadi bentuk informasi yang lebih mudah dipahami [13]. *Preprocessing* pada penelitian ini akan melalui 3 tahap, yaitu tahap *cleaning* data, transformasi data, dan normalisasi data dengan menggunakan *Jupyter Notebook*.

C. PEMBAGIAN DATA

Data hasil *preprocessing* akan dibagi menjadi 2 (dua) yaitu data *training* dan data *testing*. Penelitian ini akan melakukan 3 (tiga) pembagian yang berbeda pada pembagian data *training* dan *testing* untuk mengetahui pembagian data yang memiliki hasil akurasi paling optimal di antara kedua metode tersebut. Pembagian pertama adalah 60% data *training* dan 40% data *testing*. Pembagian kedua, yaitu 70% data *training* dan 30% data *testing*. Pembagian terakhir, yaitu 80% data *training* dan 20% data *testing*. Setelah dilakukan pembagian data tersebut, kemudian dilakukan pembagian data validasi sebesar 10% yang diambil dari data *training*.

D. METODE *K-NEAREST NEIGHBOR*

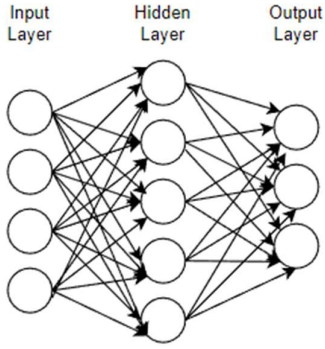
Nilai *K* yang akan dihitung ada 3 (tiga) angka ganjil, yaitu 3, 5, dan 7. Hal ini karena nilai *K* yang ganjil diketahui digunakan untuk menghindari munculnya kesamaan jarak saat proses *K-Nearest Neighbor* (KNN) dijalankan [14].

Selain itu, pada metode KNN, teknik validasi data yang digunakan dalam penelitian ini menggunakan metode *K-Fold Cross-Validation*. *K-Fold Cross-Validation* dilakukan untuk mencegah *overfitting* atau *underfitting* model, sehingga memberikan hasil yang lebih akurat dalam mengukur perfoma model. Adapun nilai *k* yang pada penelitian ini adalah *5-fold cross-validation* untuk menghilangkan bias pada data [15]. Data akan dijalankan sebanyak 5 (kali) kali dengan subset data akan bergantian menjadi data training dan data validasi pada setiap iterasi.

Tabel 1. *K-Fold Cross Validation*

K-Fold	Cross Validation				
1	***Validation***	Training	Training	Training	Training
2	Training	***Validation***	Training	Training	Training
3	Training	Training	***Validation***	Training	Training
4	Training	Training	Training	***Validation***	Training
5	Training	Training	Training	Training	***Validation***

E. METODE *ARTIFICIAL NEURAL NETWORK*



Gambar 2. *Multi Layer Percepteron*

Terdapat 3 (tiga) kelas dasar arsitektur *Artificial Neural Network* (ANN), yaitu *Single Layer Feedforward Network*, *Multi Layer Feedforward Network*, dan *Recurrent Network* [16]. Pada penelitian ini digunakan metode ANN dengan *Multi Layer Feedforward Network*. Jika pada *Single Layer Feedforward Network* memiliki 2 layer yaitu *input layer* dan *output layer*, maka pada *Multi Layer Feedforward Network* terdapat 1 layer tambahan yang disebut *hidden layer*. Pada arsitektur *multi layer*, jumlah hidden layer yang digunakan bisa lebih dari satu, tergantung kasus atau masalah yang ingin dipecahkan.

F. EVALUASI

Teknik evaluasi model yang digunakan adalah *Confusion Matrix*. *Confusion Matrix* digunakan untuk membandingkan hasil klasifikasi dari model yang telah dibuat dengan hasil klasifikasi yang sebenarnya [17].

Nilai akurasi berupa perbandingan antara data yang berhasil terklasifikasi benar dengan total data yang diprediksi. Akurasi sendiri digunakan sebagai ukuran seberapa sering algoritma klasifikasi membuat prediksi yang benar [18]. Adapun rumus perhitungan nilai akurasi yang digunakan adalah sebagai berikut.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\% \quad (1)$$

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

Tabel 2. *Confusion Matrix*

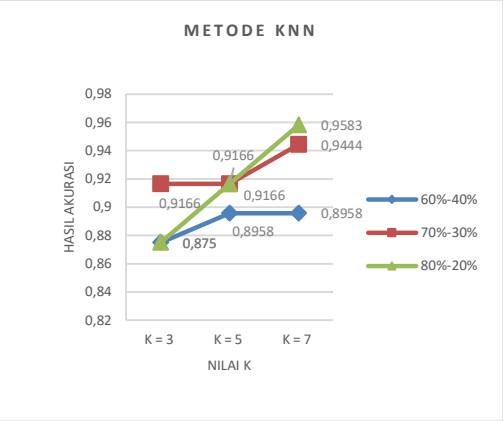
Predicted / *Actual*	**Sedang**	**Tinggi**	**Sangat Tinggi**
Sedang	*True Positive*	*False Negative*	*False Negative*
Tinggi	*False Negative*	*True Positive*	*False Negative*
Sangat Tinggi	*False Negative*	*False Negative*	*True Positive*

Tabel *Confusion Matrix* yang akan ditampilkan pada penelitian ini dengan 3 (tiga) label yang merupakan 3 (tiga) kategori IPM. 3 (tiga) kategori tersebut, yaitu Sedang, Tinggi, dan Sangat Tinggi. Pada baris paling atas merupakan label prediksi, dan pada kolom paling kiri merupakan label asli atau aktual. *True positive* artinya label yang diprediksi sama dengan label asli, sedangkan *False Negative* merupakan label yang diprediksi tidak sama dengan label asli.

4. HASIL DAN PEMBAHASAN

Berdasarkan proses pengujian dan hasil dari penerapan metode KNN dan ANN pada klasifikasi Indeks Pembangunan Manusia (IPM) kabupaten/kota di Pulau Jawa menggunakan tools *Jupyter Notebook*, dengan pembahasan sebagai berikut.

A. Metode KNN



Gambar 3. Grafik Hasil Akurasi KNN

Gambar 3. Merupakan grafik dari hasil akurasi metode KNN. Pada pembagian 60%-40%, nilai *K* dengan hasil akurasi tertinggi adalah *K=5* dan *K=7*. Pada pembagian 70%-30%, nilai K yang memiliki hasil akurasi tertinggi adalah *K=7*, dengan nilai akurasi sebesar 94,44%. Pada pembagian 80%-20%, nilai K yang memiliki hasil akurasi tertinggi adalah *K=7*, dengan nilai akurasi sebesar 95,83%. Secara keseluruhan, pembagian dan nilai K yang memiliki hasil akurasi tertinggi adalah *K=7* pada pembagian 80%-20% dengan akurasi sebesar 95,83%.

Tabel 3. *Mean Cross Validation Score*

Nilai K	***Cross Validation Score***
K=3	0,8485
K=5	0,7981
K=7	0,8224

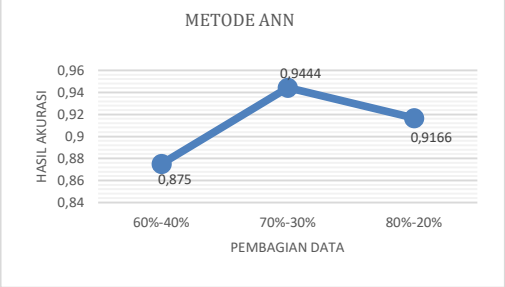
Hasil evaluasi model dengan menggunakan *K-Fold Cross-Validation* menunjukkan bahwa nilai *K* yang memiliki model terbaik adalah *K=3*, dengan nilai *mean Cross Validation Score* 84,85%.

Tabel 4. *Confusion Matrix KNN*

Predicted / *Actual*	**Sedang**	**Tinggi**	**Sangat Tinggi**
Sedang	6	0	0
Tinggi	1	13	0
Sangat Tinggi	0	0	4

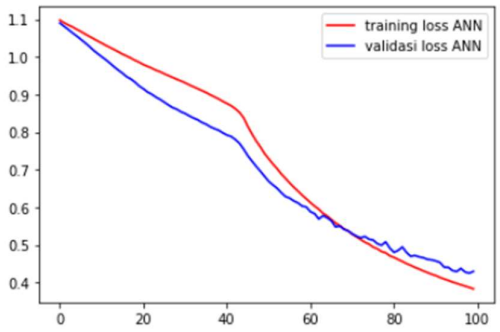
Tabel 4. merupakan tabel *Confusion Matrix* metode KNN dengan nilai *K=7*. Data yang diprediksi benar sebagai label 'Sedang' adalah 6 data. Data yang seharusnya memliki label 'Sedang', namun diprediksi sebagai label 'Tinggi' dan data yang seharusnya memiliki label 'Sedang', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada. Kemudian data yang diprediksi benar sebagai label 'Tinggi' adalah 13 data. Data yang seharusnya memliki label 'Tinggi' namun diprediksi sebagai label 'Sedang' terdapat 1 data. Data yang seharusnya memiliki label 'Tinggi', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada. Terakhir, data yang diprediksi benar sebagai label 'Sangat Tinggi' adalah 6 data. Data yang seharusnya memliki label 'Sedang', namun diprediksi sebagai label 'Tinggi' dan data yang seharusnya memiliki label 'Sedang', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada.

B. Metode *Artificial Neural Network*



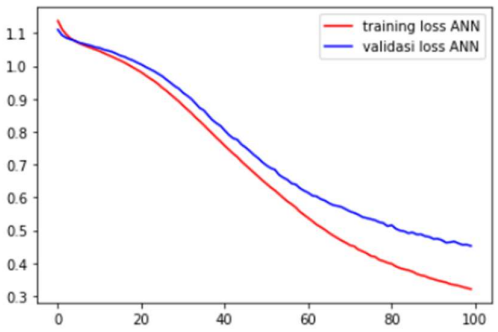
Gambar 4. Grafik Hasil Akurasi ANN

Gambar 4. merupakan grafik hasil akurasi metode ANN. Pada pembagian 60%-40%, nilai akurasi yang dihasilkan adalah 87,50%, sedangkan pada pembagian 70%-30%, nilai akurasinya adalah sebesar 94,44%, dan pada pembagian 80%-20%, nilai akurasinya adalah 91,67%. Dari ketiga pembagian tersebut, yang memiliki hasil akurasi tertinggi untuk penerapan metode ANN adalah pada pembagian 70%-30% dengan hasil akurasi sebesar 94,44%.



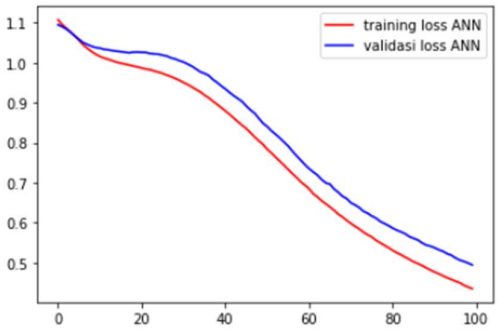
Gambar 5. Grafik *Training Loss* dan *Validasi Loss* 60%-40%

Gambar 5. merupakan grafik *training loss* dan *validasi loss* dari pembagian data *traning* dan data *testing* 60%-40%. Pada Gambar 5. menunjukkan bahwa terjadi *underfitting* pada model ANN yang telah dilatih, meskipun garis grafik *training loss* berada lebih tinggi dibandingkan garis grafik *validasi loss* dan memiliki bentuk garis grafik yang hampir sama, tetapi pada akhir *Epoch*, garis grafik *training loss* mengalami penurunan, sedangan garis grafik *validasi loss* mengalami sebaliknya, sehingga hal itu menyebabkan terjadinya *underfitting* pada model ANN tersebut.



Gambar 6. Grafik *Training Loss* dan *Validasi Loss* 70%-30%

Gambar 6. merupakan grafik *training loss* dan *validasi loss* dari pembagian data *traning* dan data *testing* 60%-40%. Garis berwarna biru merupakan garis yang menunjukkan grafik *validasi loss,* sedangkan garis berwarna merah merupakan grafik *training loss* dengan sumbu x menunjukkan jumlah *Epoch* dan sumbu y berupa nilai *loss*. Pada Gambar 6. menunjukkan bahwa terjadi *overfitting* pada model ANN yang telah dilatih, karena garis grafik *training loss* berada lebih tinggi dibandingkan garis grafik *validasi loss*, tetapi pada akhir *Epoch* dan juga garis grafik *training loss* dan *validasi loss* memiliki perbedaan kurva pada akhir *Epoch*.



Gambar 7. Grafik *Training Loss* dan *Validasi Loss* 80%-20%

Gambar 7. menunjukkan bahwa terjadi *overfitting* pada model ANN pembagian data training dan testing, 80%-20%. Hal itu dapat dilihat pada perbedaan kurva garis grafik antara *training loss* dan *validasi loss* yang memiliki gap perbedaan yang cukup besar dan juga garis grafik *validasi loss* yang lebih tinggi dibandingkan garis grafik *training loss*.

Tabel 5. *Confusion Matrix* ANN

Predicted / *Actual*	Sedang	Tinggi	Sangat Tinggi
Sedang	11	0	0
Tinggi	3	17	0
Sangat Tinggi	0	0	5

Pada Tabel 5. merupakan tabel *Confusion Matrix* pada pembagian 70%-30%, dengan jumlah data pada tabel *Confusion Matrix* sama dengan jumlah data pada dataframe y_test yaitu sebesar 36 data. Data yang diprediksi benar sebagai label Sedang adalah 11 data. Data yang seharusnya memliki label 'Sedang', namun diprediksi sebagai label 'Tinggi' dan data yang seharusnya memiliki label 'Sedang', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada. Kemudian data yang diprediksi benar sebagai label 'Tinggi' adalah 17 data. Data yang seharusnya memliki label 'Tinggi', namun diprediksi sebagai label 'Sedang' terdapat 3 data. Data yang seharusnya memiliki label 'Tinggi', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada. Terakhir, data yang diprediksi benar sebagai label 'Sangat Tinggi' adalah 5 data. Data yang seharusnya memliki label 'Sedang', namun diprediksi sebagai label 'Tinggi' dan data yang seharusnya memiliki label 'Sedang', namun diprediksi sebagai label 'Sangat Tinggi' terdapat 0 data atau tidak ada.

C. Perbandingan Hasil Akurasi

Proses perhitungan dari penerapan metode KNN dan ANN pada objek yang sama menghasilkan nilai akurasi yang berbeda. Pada Gambar 4.28 menunjukkan nilai akurasi tertinggi KNN pada *K=7* dan pembagian 80%-20% sebesar 0,9583 atau 95,83%. Sedangkan pada Gambar 4.29 menunjukkan bahwa nilai akurasi tertinggi ANN pada pembagian data 70%-30% sebesar 0,9444 atau 94,44%, sehingga secara akurasi metode KNN memberikan nilai yang lebih tinggi yaitu 95,83%, dibandingkan dengan metode ANN yaitu 94,44%.

5. KESIMPULAN

Hasil dari penelitian ini menunjukkan bahwa pada objek penelitian ini, model KNN dan ANN tidak lebih unggul dari satu sama lain.

- Metode KNN, hasil akurasi tertinggi (K=7(80%-20%) : 95,83%) bukan merupakan model terbaik.
- Metode ANN, hasil akurasi tertinggi (70%-30% : 94,44%) merupakan model overfitting.

6. REFERENSI

[1] A. Nuryani, A. Julia and Y. Sandaya, "Proyeksi Ketercapaian Bonus Demografi di Indonesia Tahun 2035," *Bandung Conference Series: Economics Studies,* vol. 2, pp. 264-272, 2022.

[2] Tenaga Ahli Madya Kedeputian, "Opini: Publikasi: Sekretaris Kabinet Republik Indonesia," 18 April 2017. [Online]. *Available*: https://setkab.go.id/memacu-infrastruktur-mempercepat-pemerataan-pembangunan/#:~:text=Pemerataan%20pembangunan%20merupakan%20jawaban%20terhadap,baru%2C%20meningkatnya%20pendapatan%20dan%20kesejahteraanmasyarakat.

[3] P. Z. A. Fakrulloh, "https://dukcapil.kemendagri.go.id/berita/baca/809/distribusi-penduduk-indonesia-per-juni-2021-jabar-terbanyak-kaltara-paling-sedikit," 2021. [Online].

[4] Badan Pusat Statistik, "https://www.bps.go.id/subject/26/indeks-pembangunan-manusia.html#:~:text=IPM%20dibentuk%20oleh%203%20(tiga,Standar%20hidup%20layak," 2014. [Online].

[5] Badan Pusat Statistik, "https://sirusa.bps.go.id/sirusa/index.php/indikator/1873," 2014. [Online].

[6] M. Fathurrahman and N. Qisthi, "Klasifikasi Indeks Pembangunan Manusia (IPM) Di Pulau Sumatera Pada Dataset *Multi-Class* Dengan Metode *Artificial Neural Network* (ANN)," *Prosiding Seminar Nasional Fisika,* vol. 7, pp. 378-384, 2021.

[7] F. Fauzi, "*K-Nearset Neighbor* (K-NN) dan *Support Vector Machine* (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah," *Jurnal MIPA,* pp. 118-124, 2017.

[8] Y. Astriningtias and R. Mardhiyah, "Aplikasi Data Mining untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa," *Jurnal Informatika,* vol. 8, pp. 837-848, 2014.

[9] H. Susanto and Sudiyanto, "Data Mining untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan dan Prestasi Masa Lalu," *Jurnal Pendidikan Vokasi,* vol. 4, pp. 222-231, 2014.

[10] S. B. Imandoust and M. Bolandraftar, "*Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*," *Int. Journal of Engineering Research and Applications,* vol. 3, no. 5, pp. 605-610, 2013.

[11] K. Khamar, "*Short Text Classification Using kNN Based on Distance Function*," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 2, no. 4, 2013.

[12] F. Khademi, S. M. Jamal, N. Deshpande and S. Londe, "*Predicting strength of recycled aggregate concrete using Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System and Multiple Linear Regression*," *International Journal of Sustainable Built Environment,* vol. 5, pp. 355-369, 2016.

[13] R. Dharma, "*Data Preprocessing : Accurate*," 22 Agustus 2022. [Online]. *Available*: https://accurate.id/teknologi/data-preprocessing/.

[14] S. Hussein, "*Data Science: Category: Geospasialis*," 27 10 2021. [Online]. Available: https://geospasialis.com/k-nearest-neighbor/#:~:text=Pada%20metode%20K%20Nearest%20Neighbor,muncul%20pada%20proses%20KNN%20dijalankan..

[15] H. Azis, Purnawansyah, F. Fattah and I. P. Putri, "Performa Klasifikasi K-NN dan *Cross-validation* pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah,* vol. 12, no. 2, pp. 81-86, 2020.

[16] S. S. Haykin, *Neural Networks and Learning Machines*, vol. 3, Pearson, 2009.

[17] Kuncahyo Setyo Nugroho, 13 11 2019. [Online]. Available: https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f.

[18] A. A. Kurniawan and M. Mustikasari, "Evaluasi Kinerja *Mllib Apache Spark* Pada Klasifikasi Berita Palsu Dalam Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK),* vol. 9, no. 3, pp. 489-500, 2022.

[19] L. Setiyani, M. Wahidin, D. Awaludin and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining *Naive Bayes : Systematic Review*," vol. 13, pp. 35-43, 2020.

[20] Kusrini, Konsep dan Aplikasi Sistem Pendukung Keputusan, Yogyakarta: Andi, 2017.

[21] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma *K-Nearest Neighbor*," *Jurnal SIMETRIS,* vol. 10, no. 2, p. 423, 2019.

[22] P. Perez, A. Trier and J. Reyes, "*Prediction of PM2.5 Concentrations Several Hours in Advance Using Neural Networks in Santiago*," vol. 34, pp. 1189-1196, 2000.

[23] M. H. J. Rahanra, Kusrini and E. T. Luthfi, "Analisa Kelulusan Mahasiswa Teknik Informatika Tepat Waktu Menggunakan Algoritma *Artificial Neural Network* (ANN)," *JURNAL FATEKSA: Jurnal Teknologi dan Rekayasa,* vol. 7, pp. 1-11, 2022.

[24] F. Tempola, M. Muhammad and A. Khairan, "Perbandingan Klasifikasi Antara KNN Dan *Naive Bayes* Pada Penentuan Status Gunung Berapi Dengan *K-Fold Cross Validation*," *Jurnal Teknologi Informasi dan Ilmu Teknologi,* vol. 5, no. 5, pp. 577-584, 2018.

[25] P. I. Nainggolan, D. S. Prasvita and D. S. Bukit, "Klasifikasi Informasi Kesehatan Pada Data Media Sosial Menggunakan *Support Vector Machine* Dan *K-Fold Cross Validation*," *Malikussaleh Journal of Mechanical Science and Technology,* vol. 5, pp. 34-38, 2021.

[26] Y. N. Fuadah, I. D. Ubaidullah, N. Ibrahim, F. F. Taliningsing, N. K. SY and M. A. Pramuditho, "Optimasi *Convolutional Neural Network* dan *K-Fold Cross Validation* pada Sistem Klasifikasi Glaukoma," *ELKOMIKA : Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika,* vol. 10, no. 3, pp. 728-741, 2022.

[27] F. Rahmad, Y. Suryanto and K. Ramli, "*Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification*," *IOP Conf. Series: Materials Science and Engineering,* 2020.

[28] Alvana Noor Fariza, "*Data Cleaning*," 17 2 2022. [Online]. Available: https://www.sekawanmedia.co.id/blog/data-cleaning/.

[29] J. Han and M. Kauffman, "*Data Mining Concept and Techniques*," 2006.

[30] Trivusi, "*Data Science*," 16 9 2022. [Online]. Available: https://www.trivusi.web.id/2022/09/normalisasi-data.html.

[31] T. T. Hanifa, Adiwijaya and S. Al-Faraby, "Analisis *Churn Prediction* pada Data Pelanggan PT. Telekomunikasi dengan *Logistic Regression* dan *Underbagging*," *e-Proceeding of Engineering,* vol. 4, no. 2, pp. 3210-3225, 2017.