ANALISIS PEMANFAATAN *MACHINE LEARNING* GUNA PREDIKSI INDEKS PEMBANGUNAN MANUSIA

Fernando Winantya Atmojo[1], Catarina Ivanda Nurlita[2], Nurchim[3]
[1,2,3] Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Duta Bangsa Surakarta
email: [1]210101014@mhs.udb.ac.id, [2]210101010@mhs.udb.ac.id, [3]nurchim@udb.ac.id*

Penelitian ini bertujuan untuk membandingkan akurasi dari beberapa metode machine learning dalam memprediksi Indeks Pembangunan Manusia (IPM) dengan memanfaatkan tiga indikator utama: Angka Harapan Hidup, Rata-rata Lama Sekolah, Harapan Lama Sekolah, dan Pengeluaran per Kapita. Metode yang dibandingkan meliputi K-Nearest Neighbors (KNN), Random Forest, AdaBoost, dan Support Vector Machine (SVM), yang semakin populer dalam analisis prediktif di bidang ini. Data yang digunakan adalah data sekunder dari Badan Pusat Statistik (BPS) Indonesia, mencakup berbagai kabupaten/kota di seluruh Indonesia. Hasil penelitian menunjukkan bahwa SVM, meskipun memiliki Mean Squared Error (MSE) tertinggi pada data pelatihan (9.283), menghasilkan MSE terendah pada data pengujian (4.419), sehingga memberikan prediksi yang paling akurat. Metode ini diikuti oleh AdaBoost, Random Forest, dan KNN dalam hal akurasi prediksi. Temuan ini menyoroti efektivitas SVM dalam memprediksi IPM dan memberikan wawasan berharga untuk penerapan metode pembelajaran mesin dalam analisis data pembangunan manusia.

***Kata Kunci*— Perbandingan, MSE, Machine Learning, IPM**

I. PENDAHULUAN

Analisis data dan penggunaan teknik machine learning telah menjadi alat yang sangat berguna untuk mengolah dan menganalisis berbagai indikator pembangunan manusia di era komputer saat ini[1]. Untuk memprediksi dan menganalisis IPM, teknik seperti K-Nearest Neighbors (KNN), Random Forest, AdaBoost, dan Support Vector Machine (SVM) semakin populer digunakan dalam penelitian[1][2][3]. Masing-masing dari metode ini memiliki karakteristik dan keunggulan tersendiri dalam hal menangani data dan membuat prediksi yang akurat. KNN adalah teknik sederhana yang berguna untuk klasifikasi dan regresi yang mencari objek dalam data pelatihan yang paling dekat dengan objek dalam data pengujian[4]. Random Forest adalah metode klasifikasi dengan pohon keputusan[5]. AdaBoost atau Boosting merupakan Teknik ensemble menggabungkan kekuatan beberapa model lemah untuk membentuk model yang lebih kuat, konsep kerja AdaBoost dengan membangun kombinasi dari suatu model dalam proses klasifikasi dan prediksi[6]. Support Vector Machine (SVM) adalah metode pengelompokan diskriminatif dengan menggunakan sebuah hyperplane sebagai pemisah antar kelas dengan memaksimalkan margin diantara kelas-kelas tersebut[4][6][7].

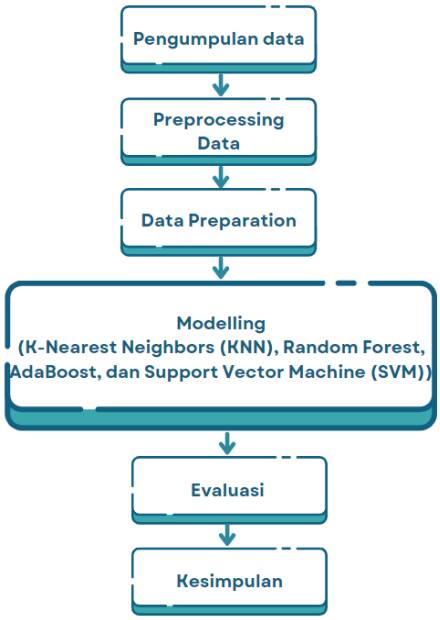
Salah satu alat penting untuk mengukur kesejahteraan penduduk suatu negara atau daerah adalah indeks pembangunan manusia (IPM). Dengan memahami dan mengukur IPM, pemerintah dan pemangku kepentingan dapat merancang kebijakan yang lebih baik guna meningkatkan kualitas hidup masyarakat[8]. IPM mencerminkan capaian dalam tiga dimensi dasar pembangunan manusia yaitu kesehatan yang diukur dengan Angka Harapan Hidup (AHH), pendidikan yang diukur dari Jumlah Rata-rata Lama Sekolah dan Jumlah Harapan Lama Sekolah, standar hidup layak yang diukur melalui Pengeluaran Per Kapita dan data ini diambil dari BPS Indonesia tahun 2023[9].

Beberapa penelitian membandingkan metode dengan berbagai persoalan, artikel ini membandingkan tiga metode ensemble learning yaitu Random Forest, Suport Vector Machine (SVM) dan AdaBoost dalam mengklasifikasikan Indeks Pembangunan Manusia (IPM). data yang digunakan 3 variabel. Dengan metode penelitian preprocessingdata, resampling dan cross-validation, hingga Pembangunan model menggunakan tiga algoritma tersebut. Penelitian ini memperoleh bahwa model Random Forest memiliki performa Terbaik dengan nilai akurasi sebesar 85,23%. Spesifisitas 71,63%, sensitivitas 95,05% dan statistik kappa 0,7698 dibandingkan dengan SVM dan AdaBoost[6].

Tujuan dari penelitian ini adalah untuk membandingkan akurasi dari metode KNN, Random Forest, AdaBoost, dan SVM dalam memprediksi IPM berdasarkan tiga indikator utama: Angka Harapan Hidup, Rata-rata Lama Sekolah, Harapan Lama Sekolah dan Pengeluaran per Kapita[6][10][11]. Dengan melakukan perbandingan ini, diharapkan dapat menentukan model terbaik untuk identifikasi indeks pembangunan manusia Selain itu, perbandingan ini juga akan membantu menentukan metode terbaik untuk diterapkan dalam penelitian dan kebijakan di masa mendatang[6].

II. METODE PENELITIAN

Implementasi metode yang digunakan adalah metode data mining menggunakan algoritma K-Nearest Neighbors (KNN), Random Forest, AdaBoost, dan Support Vector Machine (SVM). Secara keseluruhan, tahapan dalam penelitian seperti pada Gambar 1 berikut.



Gambar. 1. Diagram Alur Penelitian

a. Pengumpulan data
Data yang digunakan yaitu Indeks Pembangunan Manusia merupakan data sekunder yang bersumber dari website resmi Badan Pusat Statistika Indonesia Tahun 2023. Data tersebut terdiri 514 objek dengan unit analisis berupa kabupaten/kota di Indonesia.

b. *Preprocessing* Data
Data *Preprocessing* merupakan Langkah awal dalam menyiapkan data dan melakukan transformasi terharap data mentah sesuai dengan format untuk dilakukan analisis selanjutnya[2]. Tahapan-tahapan dalam preprocessing meliputi memasukkan dataset ke dalam dataframe menggunakan pandas, menampilkan informasi dari dataset, menampilkan deskripsi statistik dataset, menemukan dan menangani missing values di dataset, menangani outliers dataset, visualisasi hubungan antar fitur numerik dengan fungsi pairplot.

c. Data *Preparation*
1. *Splitting* Data
Pada tahap ini melibatkan proses pembagian data (data splitting) menjadi dua bagian yaitu data latih (training data) dan data uji (test data) dengan perbandingan 80:20. Langkah ini untuk memastikan bahwa model yang dibangun dapat dievaluasi dengan baik dan memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya.
2. Standarisasi Data
Dalam proses awal pemodelan, teknik transformasi yang paling umum digunakan adalah standarisasi. Studi ini menggunakan library Scikit-learn's StandardScaler, yang menyesuaikan distribusi data dengan mengurangi mean dan membagi dengan standar deviasi untuk melakukan standarisasi fitur. StandardScaler menghasilkan distribusi dengan standar deviasi 1 dan rata-rata 0, dan sekitar 68% dari nilai berada di antara -1 dan 1.

d. *Modelling*
Langkah selanjutnya adalah pemodelan menggunakan beberapa algoritma. Dalam penelitian ini akan digunakan algoritma sebagai berikut:
1. *K-Nearest Neighbors (KNN)*
Dalam data train, algoritma K-Nearest Neighbor (KNN) digunakan untuk menghitung jarak antara dua sampel. Algoritma menemukan sejumlah k tetangga terdekat, di mana k adalah bilangan positif[12]. Algoritma ini dinamakan K-Nearest Neighbor karena fungsinya. Tujuan umum dari KNN adalah sebagai berikut: pertama, menentukan jumlah tetangga terdekat k; kedua, menghitung jarak antara dokumen uji dan dokumen pelatihan; ketiga, mengurutkan data berdasarkan jarak Euclidean terkecil; dan keempat, menentukan kelompok uji berdasarkan label pada k tetangga terdekat[13]. Untuk proyek ini, n_neighbors = 10 tetangga digunakan, dengan catatan bahwa nilai k adalah nilai yang sangat penting dan berdampak pada kinerja model. Untuk mengetahui jarak antara dua titik, metode geometri klasik digunakan. Pada titik ini, data pelatihan dilatih dan data uji disimpan untuk evaluasi, yang akan dibahas di modul Evaluasi Model.
2. *Random Forest*
Prinsip kolaborasi antar model membantu model prediktif yang menggunakan metode bagging multi-model[14]. Konsep grup model ini adalah membentuk kelompok model yang bekerja sama untuk menyelesaikan masalah[15]. Akibatnya, tingkat keberhasilan model ansambel biasanya lebih tinggi daripada model tunggal[15]. Setiap model dalam model grup melakukan prediksi secara mandiri. Kemudian, hasil dari masing-masing model terpisah digabungkan untuk membuat prediksi akhir[14][16].
3. *AdaBoost*
Cara kerja AdaBoost dimulai dengan memberikan bobot yang sama pada semua kasus dalam data pelatihan. Model kemudian memeriksa hasil observasi dan memberikan bobot lebih tinggi pada model yang memiliki kesalahan, sehingga model tersebut akan diikutsertakan dalam tahap berikutnya. Proses ini dilakukan berulang kali hingga model mencapai tingkat akurasi yang diinginkan. Model pertama dibuat dengan dataset pelatihan, dan model kedua dibuat untuk mengoreksi kesalahan yang ditemukan pada model pertama. Setelah proses selesai, kesalahan diminimalkan dan seluruh dataset dapat diprediksi dengan benar. Proyek ini menggunakan learning_rate sebesar 0,05, yang menentukan nilai yang diberikan kepada setiap regressor selama iterasi boosting, dan random_state sebesar 55,

yang mengatur generator angka acak. Kelebihan metode AdaBoost termasuk implementasi yang mudah dan waktu pengujian yang cepat, yang membuatnya cocok untuk kondisi real-time. Namun, untuk memastikan model yang ideal untuk dataset yang digunakan, metode ini membutuhkan hypertuning yang tepat.

4. *Support Vector Machine*
   Model machine learning multifungsi diterapkan untuk menyelesaikan berbagai masalah, termasuk klasifikasi, regresi, dan deteksi outlier[17]. Support Vector Machine (SVM) bertujuan untuk menemukan hyperplane optimal dalam ruang dimensi N, yang berfungsi sebagai garis pemisah yang jelas antara titik-titik data masukan[17].

e. Evaluasi
   Model machine learning bertipe regresi digunakan untuk memprediksi nilai kontinu, dan kinerjanya dapat dievaluasi menggunakan metrik kesalahan seperti Mean Squared Error (MSE)[16]. MSE menghitung selisih kuadrat antara nilai sebenarnya dan nilai prediksi, kemudian mengambil rata-rata dari selisih kuadrat tersebut. Semakin kecil nilai MSE, semakin baik kinerja model dalam melakukan prediksi. MSE dapat dihitung dengan menggunakan rumus[16]:

$$MSE = \sum \frac{(y - y\_pred)^2}{n} \quad (1)$$

Y adalah nilai prediksi, Y adalah nilai aktual, dan n merupakan jumlah data[16].

III. HASIL DAN PEMBAHASAN

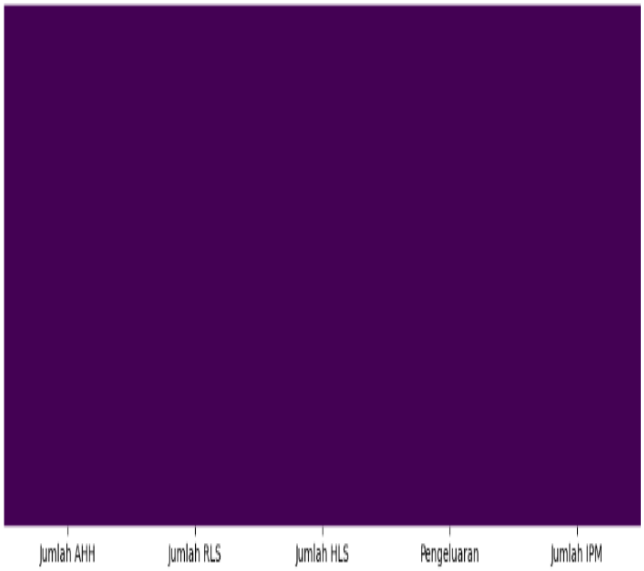
a. Data *Pre-Processing*
   1. Statistik Deskriptif

Tabel 1.
Statistik Deskriptif

No	Statistik	Jumlah AHH	Jumlah RLS	Jumlah HLS	Pengeluaran	Jumlah IPM
1	Minimum	63.880	1.710	4.330	4352.000	35.190
2	Maximum	77.930	13.040	17.930	24975.000	88.280
3	Mean	73.058	8.652	13.149	11015.134	71.269
4	Std.	2.254	1.614	1.300	2779.370	6.342
5	Count	514	514	514	514	514

Pada Tabel 1. Menunjukan Jumlah AHH (Angka Harapan Hidup) menunjukkan rata-rata sebesar 73.06 tahun, dengan nilai minimum sebesar 63.88 tahun di Kabupaten Timor Tengah Selatan, Provinsi Nusa Tenggara Timur, dan nilai maksimum sebesar 77.93 tahun di Kota Yogyakarta, Provinsi Daerah Istimewa Yogyakarta. Jumlah RLS (Rata-rata Lama Sekolah) menunjukkan rata-rata sebesar 8.65 tahun, dengan nilai minimum sebesar 1.71 tahun di Kabupaten Yahukimo, Provinsi Papua, dan nilai maksimum sebesar 13.04 tahun di Kota Yogyakarta, Provinsi Daerah Istimewa Yogyakarta. Jumlah HLS (Harapan Lama Sekolah) menunjukkan rata-rata sebesar 13.15 tahun, dengan nilai minimum sebesar 4.33 tahun di Kabupaten Pegunungan Bintang, Provinsi Papua, dan nilai maksimum sebesar 17.93 tahun di Kota Denpasar, Provinsi Bali. Pengeluaran menunjukkan rata-rata sebesar 11,015.13 unit, dengan nilai minimum sebesar 4,352.00 unit di Kabupaten Maluku Barat Daya, Provinsi Maluku, dan nilai maksimum sebesar 24,975.00 unit di Kota Surabaya, Provinsi Jawa Timur. IPM (Indeks Pembangunan Manusia) rata-rata adalah 71,27, dengan nilai minimum 35.19 di Kabupaten Nduga, Papua, dan nilai maksimum 88.28 di Kota Yogyakarta, Daerah Istimewa Yogyakarta.

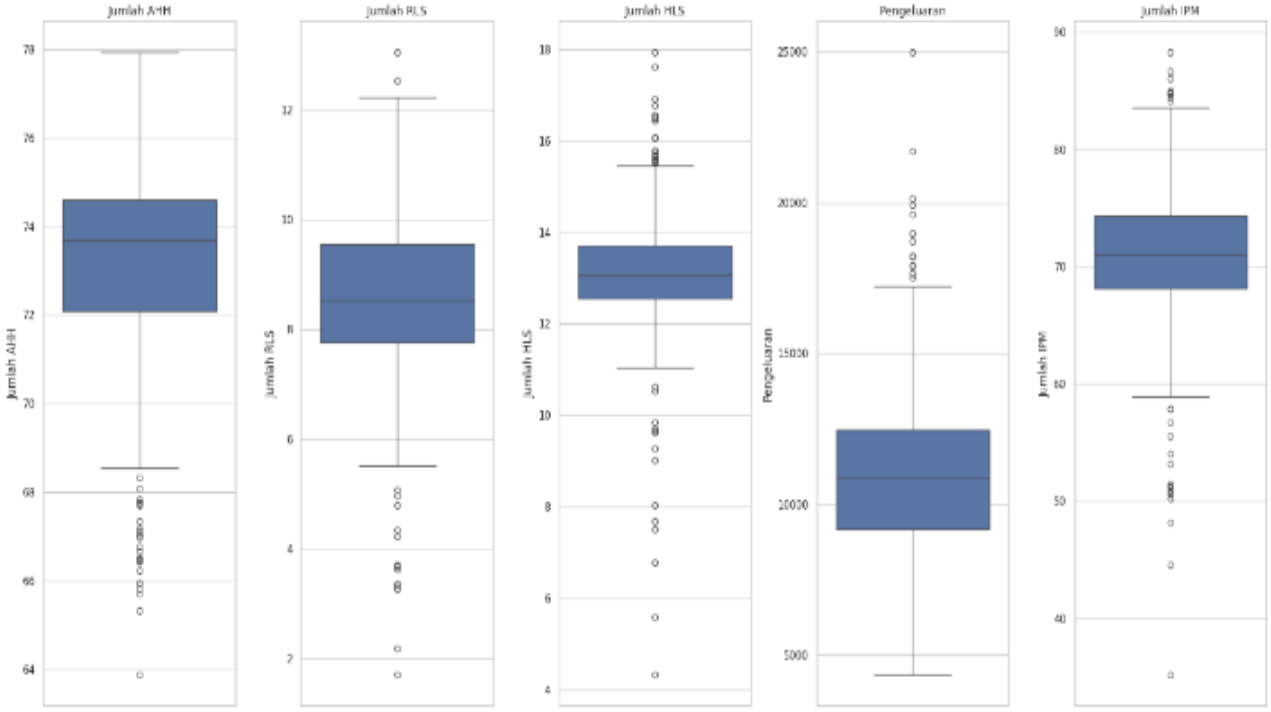
2. Pengecekan *Missing Value*



Gambar 2. Pengecekan Missing Value Menggunakan Heatmap

Pada Gambar 2. menunjukkan bahwa variabel jumlah AHH, RLS, HLS, keluaran, dan IPM tidak memiliki nilai/data yang hilang, penelitian ini bisa di lanjut tanpa mereduksi data.

3. Visualisasi Boxplot Outliner



Gambar 3. Visualisasi Boxplot Outliner

Pada Gambar 3. Jumlah AHH dan RLS menunjukkan distribusi data yang relatif simetris dengan beberapa outliner kecil, sedangkan jumlah HLS menunjukkan pola yang serupa. Sebaliknya, variabel pengeluaran menunjukkan outliner yang sangat tinggi yang menunjukkan perbedaan pengeluaran per kapita antar daerah. Meskipun ada beberapa anomali di bagian bawah, angka IPM menunjukkan distribusi yang hampir normal.

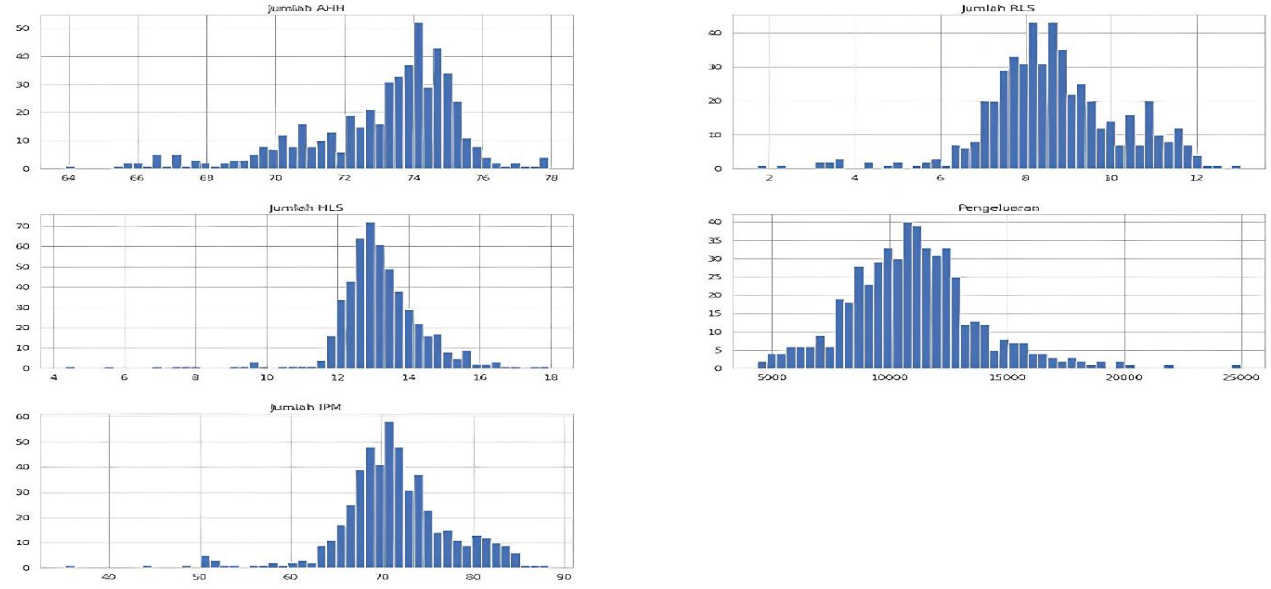
4. Penanganan *Outliner*

Pada Tabel 2. penelitian ini, outliner diidentifikasi untuk mencegah bias akibat nilai ekstrim sehingga mengurangi dampak nilai ekstrim yang dapat mempengaruhi hasil analisis sekaligus menjaga ukuran dataset tetap konstan. Dengan menggunakan Interquartile Range (IQR) Hasil dari penanganan outliner menunjukkan bahwa data menjadi 453 baris dan 5 kolom.

Tabel 2.
Penanganan Outliner Menggunakan metode Interquartile Range

Jumlah Baris	Jumlah Kolom
453	5

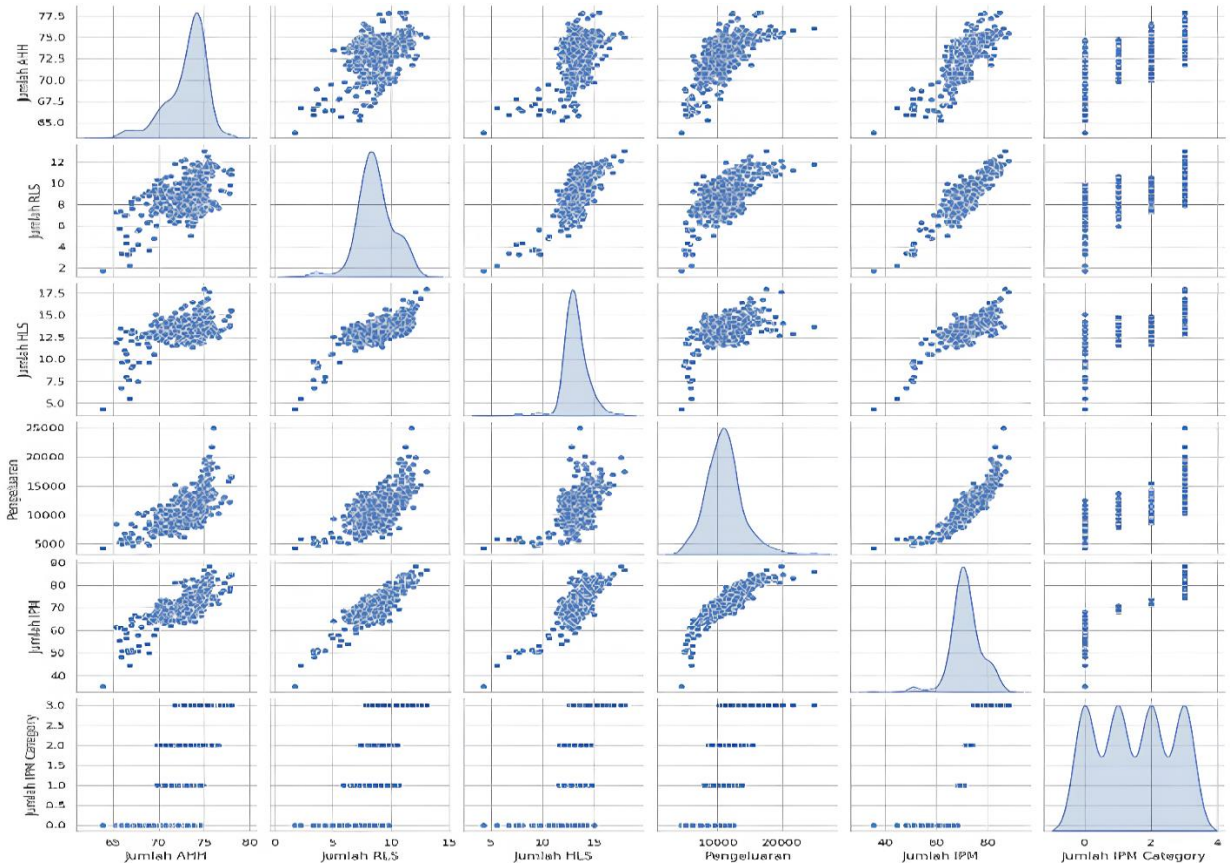
5. Visualisasi Histogram



Gambar 4. Visualisasi Histogram Setiap Variabel

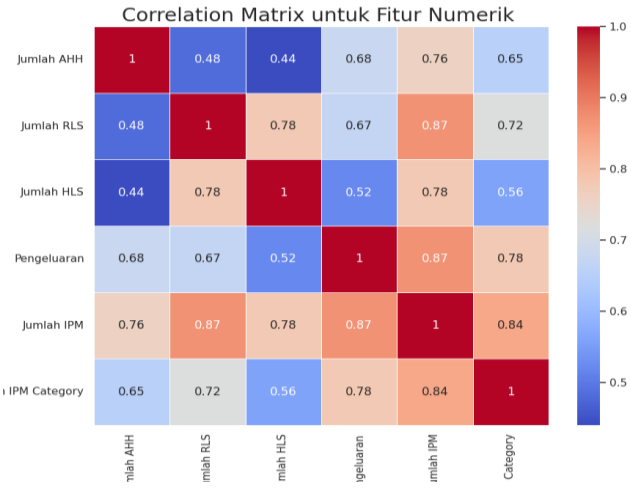
Pada Gambar 4. menunjukkan histogram untuk beberapa variabel penting dalam data Indeks Pembangunan Manusia (IPM), yaitu Jumlah AHH (Angka Harapan Hidup), Jumlah RLS (Rata-rata Lama Sekolah), Jumlah HLS (Harapan Lama Sekolah), Pengeluaran, dan Jumlah IPM. Histogram menunjukkan distribusi Angka Harapan Hidup yang sebagian besar terkonsentrasi pada nilai tertentu. Histogram Harapan Lama Sekolah menunjukkan bahwa kebanyakan daerah memiliki harapan lama sekolah yang relatif tinggi, dengan sedikit daerah yang memiliki harapan lama sekolah yang rendah[18]. Distribusi Rata-rata Lama Sekolah terlihat memiliki variasi yang cukup besar, yang menunjukkan adanya kesenjangan pendidikan di berbagai daerah[18]. Distribusi pengeluaran menunjukkan rentang yang luas, yang menunjukkan perbedaan besar dalam tingkat pengeluaran antar daerah; beberapa daerah memiliki pengeluaran yang sangat tinggi, sedangkan yang lain memiliki pengeluaran yang sangat rendah, yang menunjukkan perbedaan ekonomi yang signifikan. Distribusi indeks pembangunan manusia (IPM) menunjukkan bahwa sebagian besar daerah memiliki nilai IPM sedang hingga tinggi, tetapi ada juga beberapa daerah dengan IPM rendah, menunjukkan bahwa meskipun banyak daerah memiliki tingkat pembangunan manusia yang baik, ada beberapa daerah yang masih membutuhkan perhatian.

6. Visualisasi *Pairplot*



Gambar 5. Visualisasi Pairplot Fitur Numerik

Pada Gambar 5. Analisis scatter plot matriks menunjukkan hubungan positif yang cukup kuat antara Jumlah IPM dan beberapa fitur, termasuk Jumlah AHH (Angka Harapan Hidup), Jumlah RLS (Rata-rata Lama Sekolah), Jumlah HLS (Harapan Lama Sekolah), dan Pengeluaran (Pengeluaran Per Kapita). Ketika AHH, RLS, HLS, dan pengeluaran meningkat, IPM juga cenderung meningkat Korelasi yang jelas ini menunjukkan bahwa investasi harus dilakukan dalam bidang-bidang tersebut untuk meningkatkan IPM. Selain itu, kategori IPM menunjukkan bahwa daerah dengan kategori yang lebih tinggi memiliki nilai IPM yang lebih tinggi, sesuai dengan perkiraan. Secara keseluruhan, ciri-ciri yang ditampilkan merupakan ukuran penting untuk menentukan IPM.



Gambar 6. Kolerasi Matrik Fitur Numerik

Pada Gambar 6. Hubungan positif yang kuat antara "Jumlah IPM" dan beberapa fitur numerik, seperti "Jumlah AHH" (0.76), "Jumlah RLS" (0.87), "Jumlah HLS" (0.78), dan "Pengeluaran" (0.87), ditunjukkan dalam Gambar 6. matriks korelasi. Faktor-faktor ini menunjukkan korelasi yang signifikan, menunjukkan bahwa peningkatan angka harapan hidup, rata-rata lama sekolah, harapan lama sekolah, dan pengeluaran umumnya dikaitkan dengan peningkatan IPM[11]. Selain itu, kategori "Jumlah IPM" juga memiliki korelasi yang signifikan dengan "Jumlah IPM" (0.84), serta dengan fitur lainnya, menunjukkan bahwa faktor-faktor kesehatan, pendidikan, dan ekonomi sangat penting dalam menentukan kualitas hidup yang diukur oleh IPM.

b. Data *Preparation*

1. *Split Dataset*

Tabel 3.
Split Dataset

Total # of sample in whole dataset	514
Total # of sample in train dataset	411
Total # of sample in test dataset	103

Pada Tabel 3. Hasil split dataset menunjukkan dari total 514 sampel dalam dataset secara keseluruhan, 411 sampel (sekitar 80%) digunakan untuk dataset pelatihan, dan 103 sampel (sekitar 20%) digunakan untuk dataset pengujian. Untuk memastikan bahwa model dapat dilatih dengan sebagian besar data dan kemudian dievaluasi kinerjanya pada data yang belum pernah dilihat sebelumnya, pembagian ini adalah praktik umum dalam machine learning[19]. Ini memberikan gambaran yang lebih akurat tentang bagaimana model akan berfungsi di dunia nyata.

2. Standarisasi

Tabel 4.
Standarisasi X_train

	Jumlah AHH	Jumlah RLS	Jumlah HLS	Pengeluaran
485	-0.743154	0.400304	0.993999	-0.062970
402	0.187004	-0.573108	-0.479159	0.114161
384	0.733902	1.061490	-0.174111	0.190846
156	1.022424	2.010414	0.688951	2.804614
351	0.553038	-0.181294	-0.397317	0.573911

Hasil Tabel 4. yang menunjukkan beberapa sampel dari dataset pelatihan (X_train) setelah standarisasi menunjukkan bahwa setiap fitur telah dimodifikasi sehingga memiliki mean 0 dan standar deviasi 1. Nilai-nilai yang dihasilkan mencerminkan deviasi standar masing-masing sampel dari rata-rata fitur awal. Misalnya, fitur "Jumlah AHH" sampel dengan indeks 485 menunjukkan nilai AHH sekitar 0.74 standar deviasi di bawah rata-rata, dan fitur "Jumlah RLS" sampel dengan indeks 156 menunjukkan nilai RLS sekitar 2 standar deviasi di atas rata-rata. Proses standarisasi memastikan bahwa setiap fitur berada dalam skala yang konsisten, yang meningkatkan kinerja model pembelajaran mesin dengan mencegah fitur tertentu mendominasi hanya karena skala yang berbeda.

Tabel 5.
Standarisasi X_train Describe

	Jumlah AHH	Jumlah RLS	Jumlah HLS	Pengeluaran
count	411.0000	411.0000	411.0000	411.0000
mean	0.0000	-0.0000	0.0000	0.0000
std	1.0012	1.0012	1.0012	1.0012
min	-3.9255	-4.2158	-6.5504	-2.3653
25%	-0.3750	-0.5149	-0.4494	-0.6431
50%	0.2990	-0.0772	-0.0625	-0.0396
75%	0.6715	0.5227	0.4397	0.5037
max	2.1248	2.7206	3.5683	3.8836

Pada Tabel 5. menunjukkan ringkasan statistik dari dataset pelatihan (X_train) setelah standarisasi, yang memiliki jumlah sampel 411 untuk setiap fitur dan rata-rata (mean) sebesar 0.0000, yang menunjukkan bahwa data telah diatur ulang sehingga rata-rata setiap fitur adalah nol. Nilai minimum dan maksimum fitur berbeda, dengan nilai minimum "Jumlah AHH" sebesar -3.9255 dan nilai maksimum "Jumlah Pengeluaran" sebesar 3.8836, masing-masing menunjukkan deviasi dari rata-rata. Setiap fitur memiliki standar deviasi sekitar 1.0012, yang menunjukkan penyebaran data dalam skala standar deviasi satu. Kuartil pertama (25%), median (50%), dan kuartil ketiga (75%) untuk setiap fitur berada dalam rentang yang mencerminkan distribusi standar. Ini memastikan bahwa semua fitur berada dalam skala yang konsisten dan siap untuk digunakan dalam model pembelajaran mesin.

c. *Modelling*

Penelitian ini menggunakan empat algoritma yang berbeda dari model pembelajaran mesin: K-Nearest Neighbors, Support Vector Machine, Random Forest, dan Ada Bossting. Parameter default yang disediakan oleh library scikit-learn digunakan untuk setiap model.

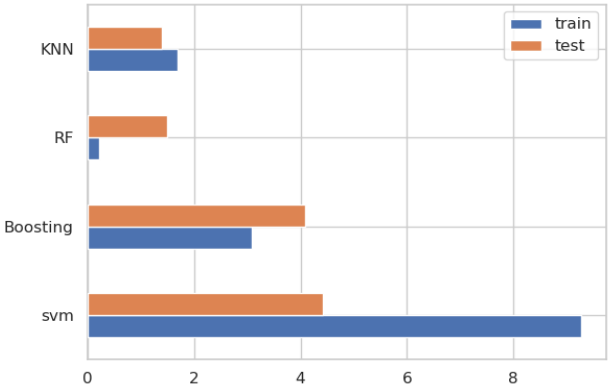
d. Evaluasi

Tabel 6.
Hasil Perhitungan MSE 4 Algoritma

	Train	Test
KNN	1.681049	1.399685
RF	0.224898	1.491243
Boosting	3.086492	4.094086
svm	9.283898	4.418809

Pada Tabel 6. Hasil kinerja masing-masing model pada dataset pelatihan dan pengujian ditunjukkan oleh perhitungan Mean Squared Error (MSE) untuk keempat algoritma. K-Nearest Neighbors (KNN) sangat baik dengan MSE sebesar 0.225 pada data pelatihan dan 1.491 pada data pengujian, menunjukkan bahwa mereka cukup konsisten antara pelatihan dan pengujian. Random Forest (RF) juga sangat baik, dengan MSE sebesar 1.681 pada data pelatihan dan 1.399 pada data pengujian, meskipun terdapat sedikit peningkatan error pada pengujian, menunjukkan bahwa ada model overfitting pada data pelatihan. Seperti yang ditunjukkan oleh MSE yang tinggi (3.086 pada data pelatihan dan 4.094 pada data pengujian), Algoritma Boosting menunjukkan bahwa itu kurang efisien dalam menangkap pola dari data. Dengan

MSE 9.283 pada data pelatihan dan 4.419 pada data pengujian, Support Vector Machine (SVM) memiliki yang paling tinggi dari semua algoritma.



Gambar 7. Visualisasi Hasil Mean Squared Error (MSE)

Pada Gambar 7. Visualisasi hasil Mean Squared Error (MSE) dari empat algoritma machine learning menunjukkan kinerja model pada data pelatihan dan pengujian dapat dilihat dengan melihat hasil Mean Squared Error (MSE) dari empat algoritma pengajaran mesin. K-Nearest Neighbors (KNN) memiliki MSE yang relatif rendah dan seimbang antara pelatihan dan pengujian; mereka memiliki nilai sedikit lebih tinggi pada pengujian, menunjukkan performa yang cukup stabil. Random Forest (RF) memiliki MSE yang sangat rendah pada data pelatihan tetapi meningkat pada data pengujian, menunjukkan perbedaan dalam performa antara data pelatihan dan pengujian. Baik pada pelatihan maupun pengujian, Algoritma Boosting memiliki MSE yang lebih tinggi. Nilai yang lebih tinggi ditunjukkan dalam pengujian, yang menunjukkan bahwa algoritma tersebut tidak memiliki kinerja yang optimal dalam menangkap pola data. Di antara semua algoritma, Support Vector Machine (SVM) memiliki MSE tertinggi.

Tabel 7 .
Hasil Prediksi MSE dari 4 Algoritma

	y_true	Pred.KNN	Pred.RF	Pred.Boosting	Pred.svm
4	76.52	75.7	75.7	75.1	76.1

Tabel 7. menunjukkan hasil perkiraan, dengan nilai sebenarnya 76.52. Prediksi dengan KNN menghasilkan nilai 75,7, Random Forest juga menghasilkan nilai 75,7, Boosting menghasilkan nilai 75,1, dan SVM menghasilkan nilai 76,1. Prediksi SVM tampaknya memberikan hasil yang paling dekat dengan nilai sebenarnya. Dengan demikian, algoritma SVM memiliki kinerja terbaik dengan nilai prediksi yang paling mendekati nilai sebenarnya.

IV. KESIMPULAN

Data pre-processing dalam penelitian ini mencakup sejumlah prosedur penting, termasuk menampilkan deskripsi statistik dataset, mengidentifikasi dan menangani outliers, menangani nilai yang tidak ada, dan menampilkan hubungan antar fitur numerik melalui penggunaan pairplot. Menurut statistik deskriptif, Angka Harapan Hidup (AHH) rata-rata adalah 73,06 tahun, Rata-rata Lama Sekolah (RLS) adalah 8,65 tahun, Harapan Lama Sekolah (HLS) adalah 13,15 tahun, dan Pengeluaran rata-rata adalah 11,015,13 unit. Tidak ada nilai yang hilang pada variabel yang dievaluasi. Ada anomali dalam distribusi data, terutama pada variabel Pengeluaran, yang menunjukkan ketidakmerataan pengeluaran per kapita di berbagai wilayah. Setelah penanganan outliners, dataset yang digunakan untuk analisis berjumlah 453 baris dari 514 baris. Standarisasi dilakukan untuk memastikan bahwa setiap fitur memiliki mean 0 dan standar deviasi 1. Ini meningkatkan konsistensi skala fitur dan kinerja model pembelajaran mesin.

Berdasarkan hasil pembahasan, dapat disimpulkan bahwa model SVM memiliki nilai prediksi paling rendah 76,1, menjadikannya model yang paling dekat dengan nilai sebenarnya. Sebaliknya, model KNN dan Random Forest, masing-masing dengan nilai prediksi 75,7, menunjukkan performa yang sedikit kurang akurat. Model Boosting, dengan nilai prediksi 75,1, memiliki error yang lebih besar daripada KNN dan Random Forest, tetapi tetap lebih besar daripada model SVM.

DAFTAR PUSTAKA

[1] I. Oosterlaken, “Technology and human development,” *Technol. Hum. Dev.*, no. May, pp. 1–147, 2015, doi: 10.4324/9781315770604.

[2] G. Dudek, “A Comprehensive Study of Random Forest for Short-Term Load Forecasting,” *Energies*, vol. 15, no. 20, 2022, doi: 10.3390/en15207547.

[3] L. Bi, O. Tsimhoni, and Y. Liu, “Using the support vector regression approach to model human performance,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 41, no. 3, pp. 410–417, 2011, doi: 10.1109/TSMCA.2010.2078501.

[4] I. A. A. S. Pratiwi and A. W. Wijayanto, “Klasifikasi Indeks Pembangunan Manusia dengan Metode K-Nearest Neighbor dan Support Vector Machine di Pulau Jawa,” *J. Ilmu Komput.*, vol. 15, no. 1, pp. 8–21, 2022.

[5] A. Fauzi, E. Utami, and A. D. Hartanto, “DDoS Penerapan Random Forest dan Adaboost untuk Klasifikasi Serangan DDoS,” *J. Educ.*, vol. 5, no. 3, pp. 7925–7937, 2023, doi: 10.31004/joe.v5i3.1920.

[6] R. I. Arumnisaa and A. W. Wijayanto, “Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI),” *Sistemasi*, vol. 12, no. 1, p. 206, 2023, doi: 10.32520/stmsi.v12i1.2501.

[7] C. A. Pamungkas and W. W. Widiyanto, “Klasifikasi Indeks Pembangunan Manusia Di Indonesia Tahun 2022 Dengan Support Vector Machine,” *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 139–145, 2023, doi: 10.55606/juisik.v3i1.407.

[8] M. B. Setiawan and A. Hakim, “Indeks Pembangunan Manusia Manusia,” *J. Econ. 9(1), 18-26*, vol. 9(1), pp. 18–26, 2008, [Online]. Available: Uny.ac.id

[9] A. Syahrani, “Analisis Pengaruh Kemiskinan, Kesehatan dan Pendidikan Terhadap Indeks Pembangunan Manusia Dalam Perspektif Ekonomi Islam,” *Stud. Kasus di Kabupaten Pesawaran*, pp. 1–147, 2018, [Online]. Available: http://repository.radenintan.ac.id/4442/

[10] A. H. Saptadi, “Perbandingan Akurasi Pengukuran Suhu dan Kelembaban Antara Sensor DHT11 dan DHT22,” *J. INFOTEL - Inform. Telekomun. Elektron.*, vol. 6, no. 2, p. 49, 2014, doi: 10.20895/infotel.v6i2.16.

[11] A. S. Wicaksono and A. M. Yolanda, “Pengelompokkan Kabupaten / Kota di Provinsi Nusa Tenggara Timur Berdasarkan Indikator Indeks Pembangunan Manusia Menggunakan K-Medoids Clustering Penyedia Data Statistik Berkualitas untuk Indonesia Maju Pengelompokkan Kabupaten / Kota di Provinsi Nusa Ten,” *Stat. Terap.*, vol. 1, no. 1, pp. 79–90, 2021.

[12] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, 2016, doi:

10.21037/atm.2016.03.37.

[13] M. A. Miftakhurahmat, N. Safitri, P. A. Kusnadi, and C. Rozikin, "Klasifikasi Pengguna Hashtag Pada Aplikasi Tiktok Menggunakan Perbandingan Metode K-Nearest Neighbors Dan Naïve Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, pp. 427–433, 2023, doi: 10.23960/jitet.v11i3.3150.

[14] L. Britanthia Christina Tanuwijaya, B. Susanto, and A. Saragih, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 68–78, 2020.

[15] G. Awliya Muhammad Ashfania, T. Prahasto, A. Widodo, and T. Warsokusumo, "Penggunaan Algoritma Random Forest untuk Klasifikasi berbasis Kinerja Efisiensi Energi pada Sistem Pembangkit Daya," *Rotasi*, vol. 24, no. 3, pp. 14–21, 2023.

[16] Anggista Oktavia Praneswara, "Perbandingan K-Nearest Neighbors, Support Vector Dan Random Forest Pada Prediksi Medical Cost," *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2035–2048, 2023, doi: 10.33022/ijcs.v12i4.3298.

[17] M. D. F. Tino, Herliyani Hasanah, and Tri Djoko Santosa, "Perbandingan Algoritma Support Vector Machines (Svm) Dan Neural Network Untuk Klasifikasi Penyakit Jantung," *INFOTECH J.*, vol. 9, no. 1, pp. 232–235, 2023, doi: 10.31949/infotech.v9i1.5432.

[18] A. Tanadjaja, "Pemodelan Angka Harapan Hidup (AHH) di Provinsi Papua menggunakan metode Geographically Weighted Regression (GWR)," *J. Sains dan Teknol. Indones.*, vol. 19, no. 2, pp. 91–99, 2017.

[19] I. Setiawan, R. Fina Antika Cahyani, and I. Sadida, "Exploring Complex Decision Trees: Unveiling Data Patterns and Optimal Predictive Power," *J. Innov. Futur. Technol.*, vol. 5, no. 2, pp. 112–123, 2023, doi: 10.47080/iftech.v5i2.2829.