ANALISIS FAKTOR SOSIAL-EKONOMI TERHADAP KETIMPANGAN PENDIDIKAN MENGGUNAKAN MACHINE LEARNING

Mata Kuliah Riset Informatika

Dosen Pengampu: Dr. Basuki Rahmat, S.Si., MT.

Disusun Oleh:

Wanda Gustrifa (21081010083)

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"

JAWA TIMUR

2024

KATA PENGANTAR

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar Belakang

Ketimpangan pendidikan di Provinsi Banten mencerminkan tantangan yang signifikan dalam upaya mencapai pemerataan akses dan kualitas pendidikan di Indonesia. Meskipun Provinsi Banten terletak dekat dengan ibu kota negara, Jakarta, dan memiliki potensi ekonomi yang cukup besar, masih terdapat kesenjangan yang mencolok dalam hal akses pendidikan antara daerah perkotaan dan pedesaan. Data dari Badan Pusat Statistik (BPS) menunjukkan bahwa tingkat partisipasi pendidikan di daerah pedesaan Banten masih rendah, dengan banyak sekolah yang kekurangan fasilitas dan tenaga pengajar yang berkualitas. Hal ini menyebabkan siswa dari daerah tersebut tidak mendapatkan pendidikan yang setara dengan rekan-rekan mereka di kota-kota besar seperti Serang dan Tangerang (Hujaimah et al., 2023).

Faktor sosial ekonomi berperan penting dalam ketimpangan pendidikan di Banten. Banyak keluarga di daerah pedesaan menghadapi tantangan ekonomi yang signifikan, termasuk pendapatan yang rendah dan kurangnya akses terhadap sumber daya pendidikan. Penelitian menunjukkan bahwa siswa dari latar belakang sosial ekonomi rendah sering kali mengalami kesulitan dalam mengakses pendidikan berkualitas, baik karena biaya pendidikan yang tinggi maupun karena kurangnya dukungan keluarga (Surayana & Agustang, 2020). Selain itu, kondisi infrastruktur yang buruk di beberapa wilayah juga menjadi penghalang bagi anak-anak untuk mendapatkan pendidikan yang layak. Keterbatasan ini menciptakan siklus ketidakadilan di mana anak-anak dari keluarga kurang mampu terjebak dalam kondisi yang sulit untuk keluar dari kemiskinan.

Dalam konteks ini, penerapan machine learning dapat memberikan pendekatan baru untuk menganalisis data terkait ketimpangan pendidikan di Provinsi Banten. Dengan memanfaatkan algoritma machine learning untuk menganalisis data besar tentang faktor sosial ekonomi dan hasil pendidikan, peneliti dapat mengidentifikasi

pola-pola yang mungkin tidak terlihat melalui metode analisis tradisional (Kansal et al., 2023). Misalnya, teknik clustering dapat digunakan untuk mengelompokkan daerah berdasarkan karakteristik sosial ekonomi mereka serta tingkat akses pendidikan, sehingga memberikan wawasan lebih mendalam tentang bagaimana faktor-faktor tersebut saling berinteraksi.Penelitian ini bertujuan untuk menganalisis pengaruh faktor sosial ekonomi terhadap ketimpangan pendidikan di Provinsi Banten menggunakan pendekatan machine learning. Dengan menganalisis data dari berbagai sumber, termasuk survei sosial ekonomi dan data pendidikan, diharapkan penelitian ini dapat memberikan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi ketimpangan pendidikan. Hasil penelitian ini tidak hanya akan memperkaya literatur akademis tetapi juga memberikan rekomendasi kebijakan bagi pemerintah daerah dan pemangku kepentingan lainnya dalam upaya mengurangi ketimpangan pendidikan di Banten.

Selain itu, penting untuk menekankan bahwa analisis faktor sosial ekonomi terhadap ketimpangan pendidikan di Banten tidak hanya relevan untuk kebijakan lokal tetapi juga dapat memberikan kontribusi pada diskusi nasional tentang akses pendidikan. Dengan menggunakan pendekatan berbasis data dan algoritma machine learning, penelitian ini berpotensi untuk menghasilkan wawasan yang dapat diterapkan di daerah lain dengan tantangan serupa. Melalui kolaborasi antara pemerintah, lembaga penelitian, dan masyarakat sipil, kita dapat mengembangkan strategi yang lebih efektif untuk meningkatkan akses dan kualitas pendidikan bagi semua anak di Provinsi Banten.

Dengan demikian, penelitian ini bertujuan untuk memberikan kontribusi signifikan dalam memahami dinamika ketimpangan pendidikan di Provinsi Banten serta menawarkan solusi berbasis data untuk meningkatkan akses dan kualitas pendidikan bagi seluruh lapisan masyarakat.

1.2 Rumusan Masalah

1. Apa saja faktor sosial ekonomi yang paling berpengaruh terhadap ketimpangan

akses Pendidikan di Indonesia?

2. Bagaimana penerapan algoritma machine learning dapat membantu dalam menganalisis dan mengidentifikasi pola ketimpangan pendidikan berdasarkan data sosial ekonomi?
3. Bagaimana tingkat ketimpangan akses Pendidikan di Indonesia berdasarkan data social ekonomi?

1.3 Tujuan Penelitian

1. Mengidentifikasi faktor utama yang memengaruhi ketimpangan pendidikan.
2. Mengetahui peran algoritma machine learning dalam menganalisis dan mengidentifikasi pola ketimpangan pendidikan berdasarkan data sosial ekonomi
3. Mengukur tingkat ketimpangan pendidikan di Indonesia

3.1 Manfaat Penelitian

Penelitian ini memberikan manfaat bagi sejumlah pihak. Adapun manfaat penelitian ini bagi berbagai pihak terkait adalah sebagai berikut:

A. Bagi Akademisi dan Peneliti

a) Menambah pengetahuan tentang penerapan algoritma machine learning, seperti regresi dan clustering, dalam analisis ketimpangan pendidikan berdasarkan data sosial ekonomi.
b) Menyediakan referensi untuk penelitian lebih lanjut di bidang pendidikan, analisis data, dan penerapan machine learning dalam konteks sosial ekonomi, khususnya di Indonesia.

B. Bagi Pemerintah dan Pembuat Kebijakan

a) Memberikan wawasan yang lebih mendalam mengenai faktor-faktor sosial ekonomi yang mempengaruhi ketimpangan pendidikan di Provinsi Banten, sehingga dapat merumuskan kebijakan yang lebih tepat sasaran.
b) Menghasilkan rekomendasi kebijakan berbasis data yang dapat membantu pemerintah daerah dalam merancang program-program pendidikan yang lebih efektif untuk mengurangi ketimpangan.

C. Bagi Lembaga Pendidikan

a) Meningkatkan pemahaman lembaga pendidikan tentang tantangan yang dihadapi oleh siswa dari latar belakang sosial ekonomi rendah, sehingga dapat mengembangkan program intervensi yang lebih sesuai.
b) Memfasilitasi pengembangan kurikulum dan metode pembelajaran yang lebih inklusif, dengan mempertimbangkan kebutuhan siswa dari berbagai latar belakang.

D. Bagi Peneliti Selanjutnya

a) Menjadi sumber referensi bagi peneliti lain yang ingin mengeksplorasi topik serupa atau menerapkan algoritma machine learning dalam konteks pendidikan dan sosial ekonomi.
b) Membuka peluang untuk penelitian lanjutan yang dapat memperdalam pemahaman tentang ketimpangan pendidikan di daerah lain di Indonesia atau negara lain dengan karakteristik serupa.

3.2 Batasan Masalah

Dalam penelitian ini, untuk menjaga fokus dan kejelasan, beberapa batasan masalah telah ditetapkan sebagai berikut:

Ruang Lingkup Geografis: Penelitian ini hanya akan dilakukan di Provinsi Banten, sehingga temuan dan rekomendasi yang dihasilkan tidak dapat digeneralisasi untuk daerah lain di Indonesia tanpa analisis tambahan.

Fokus pada Faktor Sosial Ekonomi: Penelitian ini akan membatasi analisis pada faktor-faktor sosial ekonomi yang mempengaruhi ketimpangan pendidikan, seperti pendapatan keluarga, tingkat pendidikan orang tua, status pekerjaan, dan kondisi sosial ekonomi lainnya. Faktor-faktor lain yang mungkin berkontribusi, seperti kebijakan pendidikan dan budaya lokal, tidak akan dianalisis secara mendalam.

Penggunaan Data Sekunder: Penelitian ini akan menggunakan data sekunder yang diperoleh dari survei sosial ekonomi, data pendidikan, dan sumber-sumber resmi

lainnya. Keterbatasan dalam kualitas dan ketersediaan data dapat memengaruhi hasil analisis.

Metode Analisis: Penelitian ini akan menggunakan algoritma machine learning tertentu, seperti random forest dan clustering untuk menganalisis data. Algoritma lain yang mungkin relevan tetapi tidak digunakan dalam penelitian ini, seperti neural networks atau deep learning, tidak akan dianalisis.

Periode Waktu: Penelitian ini akan membatasi analisis pada data yang tersedia dalam rentang waktu tertentu, misalnya data pendidikan dan sosial ekonomi dari tahun 2019 hingga 2024. Temuan penelitian mungkin tidak mencerminkan perubahan yang terjadi sebelum atau setelah periode tersebut.

Keterbatasan Variabel: Penelitian ini akan fokus pada variabel-variabel yang dapat diukur secara kuantitatif dan dapat diakses melalui data yang tersedia. Variabel kualitatif yang mungkin mempengaruhi ketimpangan pendidikan, seperti sikap masyarakat terhadap pendidikan atau motivasi siswa, tidak akan dianalisis secara mendalam.

BAB II

TINJAUAN PUSTAKA

2.1. Definisi Pendidikan

Dalam UU SISDIKNAS No. 20 tahun 2003, pendidikan disebutkan sebagai usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya dan masyarakat (Abd Majid, 2014). Pendidikan menjadi hal yang sangat krusial bagian setiap orang. Pemerintah bahkan telah mencanangkan berbagai program wajib sekolah. Bahkan pentingnya pendidikan juga terlihat dari besarnya APBN yang disediakan oleh pemerintah untuk bidang pendidikan sebesar 20%. Menurut Modouw menyebutkan bahwa pada prinsipnya terdapat tiga aspek di dalam istilah pendidikan yang saling mengisi yaitu usaha sadar dan terencana, memengaruhi atau menciptakan lingkungan yang menunjang pembelajaran, serta perubahan dan kemampuan diri (Modouw, 2013).

Pendidikan menjadi indikator utama dalam pembangunan sumber daya manusia (SDM) yang berimplikasi pada kualitas sumber daya manusia. Pendidikan memiliki posisi yang strategis dalam pembangunan daerah dan nasional. Pendidikan juga merupakan salah satu indikator kemajuan suatu bangsa karena berdampak pada peningkatan kualitas hidup dan kesejahteraan masyarakat untuk mewujudkan masyarakat yang makmur dan sejahtera (Pribadi, 2015). Sukirno menjelaskan bahwa pendidikan merupakan satu investasi yang sangat berguna untuk pembangunan ekonomi. Di satu pihak untuk memperoleh pendidikan diperlukan waktu dan uang. Pada masa selanjutnya setelah pendidikan diperoleh, masyarakat dan individu akan memperoleh manfaat. Individu yang memperoleh pendidikan tinggi cenderung memperoleh pendapatan yang lebih tinggi dibandingkan dengan tidak berpendidikan. Semakin tinggi pendidikan, semakin tinggi pula pendapatan yang diperoleh.

Peningkatan dalam pendidikan memberi beberapa manfaat dalam mengurangi tingkat kemiskinan dan sekaligus dapat mempercepat pertumbuhan ekonomi (Sukirno, 2004).

2.2. Ketimpangan Pendidikan

Ketimpangan pendidikan merupakan kondisi ketidakmerataan lulusan pendidikan dari penduduk di suatu daerah. Ukuran ketimpangan pendidikan adalah indeks Gini pendidikan yang mengukur rasio rata-rata capain tahun sekolah dari semua penduduk. Indeks Gini Pendidikan memiliki koefisien berkisar antara 0 hingga 1. Semakin rendah indeks koefisien, semakin baik tingkat kemerataan capaian pendidikan, dan semakin tinggi indeks koefisien, menunjukkan terjadinya ketidakmerataan atau ketimpangan pendidikan. Kategori ketimpangan sesuai dengan Indeks Gini Pendidikan (Todaro & Smith, 2006) yaitu (1) indeks 0,71 ke atas adalah wilayah dengan ketimpangan sangat tinggi, (2) indeks 0,5-0,70 adalah wilayah dengan ketimpangan tinggi, (3) indeks 0,36-0,49 adalah wilayah dengan ketim-pangan sedang, (4) indeks 0,21-0,35 adalah wilayah dengan ketimpangan rendah, dan (5) indeks 0,20 ke bawah adalah wilayah dengan ketimpangan sangat rendah (Sholikhah et al., 2014).

2.3. Sosial Ekonomi

Sosial menurut KBBI adalah hal-hal yang berkenaan dengan masyarakat atau sifat-sifat kemasyarakatan yang memperhatikan umum. Jadi sosial bisa dikakatan sebuah perilaku manusia yang berhubungan ataupun bekerja sama satu sama lain dalam kehidupan bermasyarakatnya, dengan tujuan untuk memenuhi kebutuhan dan keiginanan didalam hidupnya masing-masing baik kebutuhan sandang, papan dan juga pangan. Sedangkan ekonomi dapat diartikan sebagai prilaku manusia dalam mencari alat pemuas kebutuhan untuk mencapai kesejahteraan dan kebahagiaan di dalam kehidupannya.

Sosial ekonomi menurut Soerjono Soekanto (2007:89) adalah posisi seseorang dalam masyarakat berkaitan dengan orang lain dalam arti lingkungan pergaulan, prestasinya, dan hak-hak serta kewajibannya dalam berhubungan dengan sumber daya.

Menurut Soekanto (2001:237) menyatakan bahwa komponen pokok kedudukan sosial ekonomi meliputi ukuran kekayaan, ukuran kekuasaan, ukuran kehormatan, ukuran ilmu pengetahuan. Kondisi ekonomi berperan penting dalam pendidikan seorang anak. Menurut Gerungan (2009: 196), peranan kondisi ekonomi dalam pendidikan anak memegang satu posisi yang sangat penting. Dengan adanya perekonomian yang cukup memadai, lingkungan material yang dihadapi anak dalam keluarganya jelas 11 lebih luas, maka ia akan mendapat kesempatan yang lebih luas juga untuk mengembangkan kecakapan yang tidak dapat ia kembangkan tanpa adanya sarana dan prasarana itu.

Dapat ditarik kesimpulkan kondisi sosial ekonomi yaitu suatau posisi, kedudukan, jabatan, kepemilikan yang dimiliki seorang individu ataupun kelompok yang berkaitan dengan tingkat pendidikan, tingkat pendapatan, kepemilikan aset rumah tangga, dan pemenuhan kebutuhan keluarga dan pekerjaan yang dimiliki yang akan sangat mempengaruhi status sosial seseorang, kelompok ataupun keluarga di lingkungan masayarakatnya.

Berikut ini beberapa faktor sosial orang tua yang dapat mempengaruhi perkembangan anak menurut Gerungan (2009:199): 1) Keutuhan keluarga Yang dimaksud dengan keutuhan keluarga adalah keutuhan dalam struktur keluarga, yaitu bahwa keluarga terdiri dari ayah, ibu, dan anak. Apabila salah satu unsur keluarga diatas tidak ada, maka struktur keluarga tidak utuh. Ketidakutuhan keluarga berpengaruh negatif terhadap perkembangan sosial anak. Pengaruh negatif itu bisa mempengaruhi kecakapan-kecakapan anak disekolah. Dalam penilaian kaum psikologi, anak-anak dari keluarga utuh memperoleh nilai psikologis yang lebih baik dari pada anak-anak dari keluarag utuh dalam hal fleksibilitas, penyesuaian diri, pengertian akan orang-orang dan situasi diluarnya, dan dalam hal pengendalian diri. 2) Sikap dan kebiasaan orang tua Umumnya sikap mendidik yang otoriter, overprotective, sikap penolakan orang tua terhadap anak-anak dapat menjadi suatu kendala bagi perkembangan sosial anak. 3) Status anak Yang dimaksud dengan status anak adalah status anak sebagai anak sulung, anak bungsu atau anak tunggal. Selain itu status anak sebagai anak tiri juga mempengaruhi interaksi sosial keluarga

2.4. Data Mining

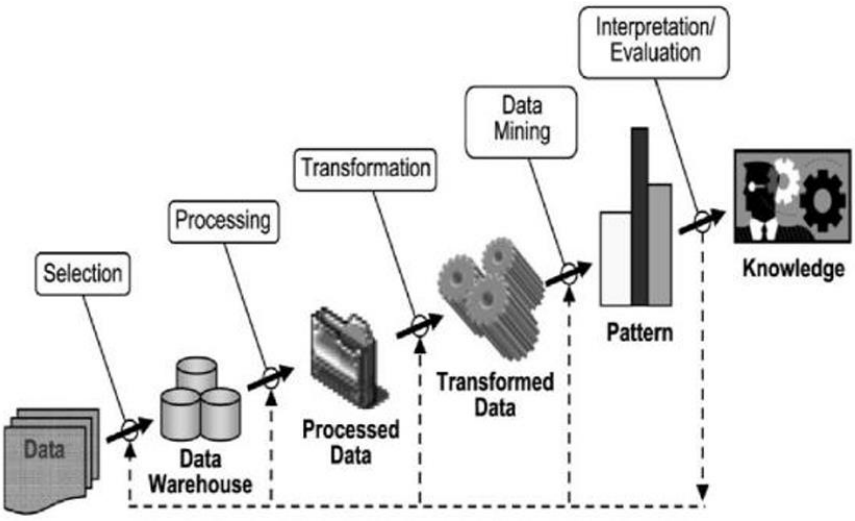
Data mining adalah proses mengidentifikasi pola atau informasi menarik dalam data yang dipilih dengan memanfaatkan berbagai teknik atau metode tertentu. Dalam praktiknya, terdapat beragam metode, teknik, atau algoritma yang dapat digunakan dalam data mining, dan pemilihan metode yang tepat sangat bergantung pada tujuan serta tahap-tahap dalam proses Knowledge Discovery in Database (KDD) (Yuli Mardi, 2019). Data mining mulai dikenal sejak tahun 1990-an sebagai metode yang efektif untuk mengekstraksi pola dan informasi yang dapat mengungkap hubungan antar data. Teknik ini digunakan untuk mengelompokkan data ke dalam satu atau lebih cluster, sehingga objek-objek dalam satu cluster memiliki tingkat kesamaan yang tinggi satu sama lain (Zahra et al., 2024).

Menurut (Zahra et al., 2024), data mining memiliki beberapa keunggulan sebagai alat analisis, di antaranya:

a) Data mining dapat mengelola data dalam jumlah besar dan kompleks.
b) Data mining juga mampu menangani data dengan berbagai jenis atribut.
c) Data mining memiliki kemampuan untuk mencari dan memproses data secara otomatis yang disebut semi otomatis karena beberapa teknik dalam data mining membutuhkan input parameter dari pengguna secara manual.
d) Data mining dapat memanfaatkan pengalaman dan kesalahan sebelumnya untuk meningkatkan kualitas analisis, menghasilkan output terbaik.

Namun, data mining juga memiliki kekurangan. Dalam pencariannya, data mining tidak bekerja pada data secara individu tetapi sebagai satu set individu atau kumpulan yang memenuhi kriteria tertentu.

Terdapat beberapa tahapan dalam melakukan proses data mining. Rincian setiap tahap dalam proses Knowledge Discovery in Database (KDD) dapat dilihat pada Gambar 2.1.



Gambar 2.1 Tahap Data Mining

(Sumber: Suliman, 2021)

Menurut (Suliman, 2021), rincian setiap tahap dalam proses Knowledge Discovery dapat dijelaskan sebagai berikut:

Data Selection

Data dari sekumpulan data operasional perlu diseleksi terlebih dahulu sebelum memulai tahap eksplorasi informasi dalam Knowledge Discovery in Database (KDD). Data hasil seleksi ini nantinya akan digunakan dalam proses data mining dan disimpan dalam berkas terpisah dari basis data operasional.

Pre-processing/Cleaning

Sebelum proses data mining dilaksanakan, perlu dilakukan pembersihan pada data yang menjadi fokus Knowledge Discovery in Database (KDD). Proses ini meliputi penghapusan data duplikat, pengecekan konsistensi data, dan perbaikan kesalahan, seperti kesalahan penulisan. Selain itu, dilakukan juga proses enrichment, yaitu memperkaya data yang sudah ada dengan informasi tambahan yang relevan, seperti data eksternal yang dibutuhkan untuk Knowledge Discovery in Database (KDD).

Transformation

Transformasi data atau coding, dilakukan untuk mengubah data yang telah dipilih agar sesuai dengan kebutuhan proses data mining. Proses ini bersifat kreatif dalam Knowledge Discovery in Database (KDD) dan sangat bergantung pada jenis atau pola informasi yang ingin ditemukan dalam basis data.

Data Mining

Data mining adalah tahap menemukan pola atau informasi yang bermanfaat dalam data yang telah dipilih, menggunakan teknik atau metode tertentu. Teknik dan metode dalam data mining beragam, dan pemilihan metode yang tepat tergantung pada tujuan serta keseluruhan proses Knowledge Discovery in Database (KDD).

Interpretation/Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu disajikan dalam format yang mudah dipahami oleh pihak terkait. Tahap ini merupakan bagian dari proses Knowledge Discovery in Database (KDD) yang disebut interpretasi. Proses ini mencakup evaluasi untuk memastikan apakah pola atau informasi yang ditemukan sesuai atau bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Langkah akhir dalam Knowledge Discovery in Database (KDD) adalah menyampaikan pengetahuan dalam bentuk yang mudah dimengerti oleh pengguna.

Data mining dapat dikategorikan ke dalam beberapa kelompok berdasarkan tugas yang dapat dilaksanakan (Yuli Mardi, 2019), yaitu:

Description (Deskripsi)

Para peneliti dan analis sering berusaha mencari cara untuk menggambarkan pola dan trend yang tidak terlihat dalam data.

Estimation (Estimasi)

Estimasi serupa dengan klasifikasi, namun lebih memusatkan perhatian pada variabel target yang berbentuk numerik bukan kategori. Model dikembangkan dengan

memanfaatkan record lengkap yang menyediakan nilai variabel target sebagai nilai prediksi. Selanjutnya, dalam evaluasi berikutnya estimasi nilai dari variabel target ditentukan berdasarkan nilai variabel yang diprediksi.

Prediction (Prediksi)

Prediksi memiliki kesamaan dengan klasifikasi dan estimasi, tetapi yang membedakannya adalah nilai yang akan muncul di masa mendatang. Beberapa algoritma dan teknik yang digunakan dalam klasifikasi dan estimasi juga bisa digunakan untuk prediksi dalam kondisi yang tepat.

Classification (Klasifikasi)

Dalam klasifikasi variabel, tujuan bersifat kategorik. Sebagai contoh pengklasifikasian persediaan dalam tiga kelas, yaitu persediaan tinggi, persediaan sedang dan persediaan rendah.

Clustering (Pengklusteran)

Clustering merupakan teknik pengelompokan record data, pengamatan atau kasus dalam kelas yang memiliki kemiripan. Cluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record lain dalam cluster.

Association (Asosiasi)

Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu. Dalam dunia bisnis lebih umum disebut sebagai analisis keranjang belanja.

2.5. Machine Learning

Machine learning adalah bagian dari kecerdasan buatan yang sering digunakan untuk menyelesaikan berbagai jenis masalah. Machine learning dapat diartikan sebagai penerapan algoritma komputer dan matematika yang memungkinkan sistem untuk belajar dari data dan membuat prediksi untuk masa depan. Proses pembelajaran ini

melibatkan usaha untuk mencapai kecerdasan melalui dua tahap, yaitu pelatihan (training) dan pengujian (testing) (Roihan et al., 2020).

Menurut (Wijoyo A et al., 2024), algoritma machine learning dibagi menjadi tiga jenis, yaitu:

1. Supervised Learning

Supervised Learning adalah metode machine learning yang menggunakan data berlabel untuk melatih algoritma dalam membuat estimasi terbaik terhadap output (Y) dari input (X). Proses ini efektif untuk tugas klasifikasi dan regresi, dengan contoh algoritma seperti K-NN, Naive Bayes, Decision Tree, dan Support Vector Machine. Meski sederhana dan mudah dipahami, supervised learning membutuhkan data pelatihan yang akurat dan komputasi yang lebih lama dibanding unsupervised learning karena memerlukan pelabelan pada setiap input.

2. Unsupervised Learning

Unsupervised learning adalah algoritma yang bekerja tanpa memerlukan data berlabel. Dalam pendekatan ini, algoritma tidak membutuhkan data pelatihan, melainkan digunakan untuk menemukan pola atau membuat model deskriptif tanpa perlu kategori atau output yang sudah ditentukan. Algoritma unsupervised learning banyak diterapkan dalam clustering dan asosiasi aturan. Keunggulan utamanya adalah fleksibilitasnya dalam mendeteksi pola yang sebelumnya mungkin tidak dikenali. Namun, kekurangannya adalah sulitnya mengidentifikasi informasi spesifik di dalam data karena tidak ada label, sehingga sulit juga membandingkan output dengan input.

3. Semi Supervised dan Reinforcement Learning

Semi-supervised learning adalah algoritma yang menggabungkan pendekatan supervised dan unsupervised, dengan bekerja pada data besar yang sebagian berlabel dan sebagian tidak. Keunggulannya adalah lebih hemat biaya karena hanya sebagian data yang memerlukan pelabelan dan tidak membutuhkan tenaga ahli untuk pemrosesan. Sementara itu, reinforcement learning bertujuan memaksimalkan hasil dan mengurangi risiko dengan mengamati interaksi agen dengan lingkungannya.

Algoritma ini belajar secara berulang, di mana agen mengamati data input, mengambil tindakan, dan menerima "reward" atau umpan balik dari lingkungan. Dengan mengamati input ulang dan mendapatkan umpan balik tambahan, agen memperbaiki keputusannya secara bertahap untuk hasil yang lebih akurat.

2.6. K-Means

K-Means merupakan salah satu algoritma dalam data mining yang bisa digunakan untuk melakukan pengelompokan/clustering suatu data. Ada banyak pendekatan untuk membuat cluster, diantaranya adalah membuat aturan yang mendikte keanggotaan dalam group yang sama berdasarkan tingkat persamaan diantara anggota-anggotanya. Pendekatan lainnya adalah dengan membuat sekumpulan fungsi yang mengukur beberapa properti dari pengelompokan tersebut sebagai fungsi dari beberapa parameter dari sebuah clustering. Metode K-Means adalah metode yang termasuk dalam algoritma clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numerik. Pengelompokan data dengan metode KMeans dilakukan dengan algoritma:

1. Tentukan jumlah kelompok
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat kelompok (centroid/rata-rata) dari data yang ada di masing-masing kelompok. Lokasi centroid setiap kelompok diambil dari rata-rata (mean) semua nilai data pada setiap fiturnya. Jika M menyatakan jumlah data dalam sebuah kelompok, i menyatakan fitur ke-i dalam sebuah kelompok, dan p menyatakan dimensi data, maka persamaan untuk menghitung centroid fitur ke-i digunakan persamaan 1.

$$ci = \frac{1}{m}\sum_{j=1}^{m} xj$$

   persamaan 1 dilakukan sebanyak p dimensi dari i=1 sampai dengan i=p.

4. Alokasikan masing-masing data ke centroid/rata-rata terdekat. Ada beberapa cara

yang dapat dilakukan untuk mengukur jarak data ke pusat kelompok, diantaranya adalah Euclidean [9]. Pengukuran jarak pada ruang jarak (distance space) Euclidean dapat dicari menggunakan persamaan 2.

$$d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Pengalokasian kembali data ke dalam masing-masing kelompok dalam metode K-Means didasarkan pada perbandingan jarak antara data dengan centroid setiap kelompok yang ada. Data dialokasikan ulang secara tegas ke kelompok yang mempunyai cetroid dengan jarak terdekat dari data tersebut. Pengalokasian data ini menurut MacQueen (1967) dapat ditentukan menggunakan persamaan 3.

2.7. Random forest

Random Forest merupakan algoritma machine learning yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan menentukan metode yang lebih efektif dalam memproses. Salah satu keunggulan dari Random Forest adalah kemampuannya untuk menangani dataset besar dengan beragam fitur, serta mengatasi masalah overfitting yang sering muncul pada pohon keputusan tunggal. Selain itu, algoritma ini dapat mempertahankan kinerja yang tinggi dan stabil (Ary Prandika Siregar et al., 2023).

Sebagai pengembangan dari metode Decision Tree, Random Forest terdiri dari beberapa pohon keputusan yang dilatih menggunakan sampel yang berbeda. Setiap pohon memecah atribut berdasarkan subset acak. Random Forest memiliki beberapa manfaat, termasuk meningkatkan akurasi ketika ada data yang hilang, mampu mengatasi outlier, dan efisien dalam penyimpanan data. Selain itu, algoritma ini juga melakukan seleksi fitur yang efektif, sehingga dapat mengidentifikasi atribut terbaik dan meningkatkan performa model klasifikasi. Dengan adanya seleksi fitur, Random Forest dapat beroperasi dengan baik pada data besar dengan parameter yang kompleks (Marlina Haiza et al., 2023).

Menurut (Suci Amaliah et al., 2022), Metode Random Forest pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001 dan termasuk dalam kategori

ensemble learning, yang berarti memanfaatkan kombinasi beberapa model untuk meningkatkan kinerja. Random Forest adalah teknik yang dapat meningkatkan akurasi dengan menghasilkan atribut acak untuk setiap node. Pohon keputusan dibangun dengan menentukan node akar dan diakhiri dengan beberapa node daun untuk mencapai hasil akhir. Proses pembentukan pohon keputusan dalam Random Forest serupa dengan proses pada Classification and Regression Tree (CART), tetapi Random Forest tidak melakukan pemangkasan (pruning). Indeks Gini digunakan untuk memilih fitur di setiap simpul internal dari pohon keputusan, dengan nilai Indeks Gini dapat dihitung melalui rumus berikut:

$$Gini(Si) = 1 - \sum_{i=0}^{c-1} pi^2$$

Di mana:

Si adalah subset data pada node i.

pi adalah frekuensi relatif dari kelas i di dalam subset Si.

c adalah jumlah kelas yang ada.

Untuk menghitung kualitas pemisahan (split) pada fitur k ke dalam subset Si, rumusnya adalah:

$$Ginisplit = 1 - \sum_{i=0}^{c-1} \left(\frac{ni}{n}\right) Gini(si)$$

Di mana:

ni adalah jumlah sampel dalam subset Si setelah pemisahan.

n adalah jumlah total sampel di node yang diberikan.

Fungsi margin untuk Random Forest, yang mengukur seberapa baik model dapat mengklasifikasikan kelas tertentu, dinyatakan sebagai:

$mr(\mathrm{X}, Y) = P_{\Theta}(h(\mathrm{X}, \Theta) = Y) - {}^{\max j} P_{\Theta}(h(\mathrm{X}, \Theta) = oo)$

Di mana:

X adalah fitur input.

Y adalah kelas yang benar.

$P\Theta$ adalah probabilitas yang diprediksi oleh model. Kekuatan himpunan pengklasifikasi dapat diukur dengan:

$s = E\mathrm{x,y}\mathrm{F}mr(\mathrm{X}, Y)$

Batas atas kesalahan generalisasi dapat diturunkan sebagai berikut:

$$Pe = \frac{p(1 - s^2)}{s^2}$$

Di mana:

$\bar{p}$ adalah rata-rata korelasi antara pengklasifikasi.

Rumus-rumus ini memberikan dasar matematis untuk cara Random Forest membangun model dan mengevaluasi kualitas prediksi yang dihasilkan.

BAB III

ANALISIS DESAIN SISTEM

3.1 Tahapan Penelitian

1. Identifikasi Masalah dan Tujuan Penelitian

Tahap awal dalam penelitian ini adalah mengidentifikasi masalah yang ingin diteliti, yaitu ketimpangan pendidikan di Provinsi Banten, serta merumuskan tujuan penelitian. Tujuan utama dari penelitian ini adalah untuk menganalisis pengaruh faktor sosial ekonomi terhadap ketimpangan pendidikan dan menggunakan algoritma machine learning untuk mengidentifikasi pola-pola dalam data.

2. Pengumpulan Data

Pengumpulan data merupakan langkah penting dalam penelitian ini. Data yang akan digunakan mencakup informasi mengenai latar belakang sosial ekonomi keluarga, akses pendidikan, dan hasil belajar siswa. Data akan dikumpulkan dari berbagai sumber, termasuk:

a) Survei sosial ekonomi yang dilakukan oleh pemerintah daerah.
b) Data pendidikan dari Dinas Pendidikan Provinsi Banten.
c) Data sekunder dari Badan Pusat Statistik (BPS) dan lembaga internasional seperti World Bank.

3. Pra-Proses Data

Setelah data terkumpul, langkah selanjutnya adalah pra-proses data. Proses ini meliputi:

Pembersihan Data: Mengidentifikasi dan menangani nilai yang hilang, duplikasi, dan kesalahan entri data.

Transformasi Data: Melakukan normalisasi atau standardisasi variabel numerik agar memiliki skala yang sama.

Pemilihan Fitur: Mengidentifikasi variabel mana yang paling relevan untuk analisis ketimpangan pendidikan.

4. Eksplorasi Data (Data Exploration)

Tahap eksplorasi data dilakukan untuk memahami karakteristik dataset secara keseluruhan. Ini meliputi analisis deskriptif untuk menghitung rata-rata, median, dan distribusi dari setiap variabel. Visualisasi data juga akan dilakukan untuk mendapatkan wawasan lebih lanjut tentang pola-pola dalam data.

5. Penerapan Algoritma Machine Learning

Setelah pra-proses dan eksplorasi data selesai, langkah berikutnya adalah menerapkan algoritma machine learning. Dalam penelitian ini, algoritma seperti K-Means clustering dan regresi akan digunakan untuk menganalisis data. Proses ini meliputi:

a) Menentukan jumlah kluster yang optimal untuk K-Means.
b) Melatih model menggunakan dataset pelatihan.
c) Menguji model dengan dataset pengujian untuk mengevaluasi kinerjanya.

6. Analisis Hasil

Setelah penerapan algoritma machine learning, tahap selanjutnya adalah menganalisis hasil yang diperoleh dari model. Ini termasuk:

a) Menginterpretasikan pola-pola yang teridentifikasi oleh model.
b) Membandingkan hasil dengan literatur yang ada untuk memahami relevansi temuan.

7. Penyusunan Rekomendasi Kebijakan

Berdasarkan analisis hasil, penelitian ini akan menyusun rekomendasi kebijakan untuk pemerintah daerah dan pemangku kepentingan lainnya dalam upaya mengurangi ketimpangan pendidikan di Provinsi Banten. Rekomendasi ini diharapkan dapat membantu dalam merumuskan strategi pendidikan yang lebih efektif.

8. Penulisan Laporan Penelitian

Tahap akhir adalah penulisan laporan penelitian yang mencakup semua aspek dari penelitian ini, mulai dari latar belakang, metodologi, hasil analisis, hingga rekomendasi kebijakan. Laporan ini akan disusun dengan format akademis yang sesuai

3.2 Studi Literature

Dalam tahap awal penelitian ini, peneliti mengumpulkan sumber literatur untuk memahami, mencari, dan mempelajari permasalahan yang diangkat. Tahap ini sangat penting karena memberikan pemahaman mendalam tentang konsep dan definisi metode yang digunakan, termasuk istilah teknis yang relevan. Dengan memanfaatkan berbagai sumber, seperti jurnal, artikel, buku, dan penelitian terdahulu, peneliti memperoleh landasan teoretis yang kuat mengenai metode analisis ketimpangan pendidikan, khususnya dalam konteks penggunaan algoritma machine learning.

3.3 Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan data sekunder. Pengumpulan data dilakukan dengan membaca literature terkait data-data yang digunakan serta mengakses melalui website instansi terkait. Data diperoleh melalui website Badan Pusat Statistik (BPS) Provinsi Banten dan penyedia data Pendidikan. Data yang dikumpulkan terkait social-ekonomi dan Pendidikan dengan rentang 2019-2024.

3.4 Praproses Data

Praproses data yang baik akan meningkatkan kualitas data yang digunakan dalam model machine learning, sehingga dapat menghasilkan prediksi yang lebih akurat.

Praproses data merupakan langkah penting dalam mempersiapkan dataset untuk analisis dan pemodelan. Proses dimulai dengan pengumpulan data dengan sumber file CSV, yang dapat mencakup informasi tentang ketimpangan pendidikan. Selanjutnya,

dilakukan pemeriksaan data untuk memastikan kualitas dan konsistensi data yang dikumpulkan, dengan mengidentifikasi nilai yang hilang, duplikat, dan outlier. Setelah pemeriksaan, tahap berikutnya adalah pembersihan data, di mana masalah yang teridentifikasi diatasi, seperti menghapus duplikat dan mengisi nilai yang hilang. Kemudian, dilakukan konversi tipe data untuk memastikan kolom atau fitur berada dalam format yang sesuai, seperti mengubah kolom tanggal menjadi format tanggal dan mengonversi kolom kategorikal menjadi format numerik. Selanjutnya, dalam tahap feature engineering, fitur baru dibuat dari data yang ada untuk meningkatkan model, seperti menghitung total penjualan per bulan. Proses ini diikuti dengan agregasi data, di mana data dari beberapa sumber digabungkan atau dikelompokkan berdasarkan kategori tertentu untuk analisis lebih lanjut

Tahap terakhir dalam praproses data adalah normalisasi atau standarisasi, di mana skala fitur diubah agar semua fitur memiliki rentang yang sama. Normalisasi digunakan untuk mengubah data ke dalam rentang 0-1, sedangkan standarisasi menghitung nilai z-score. Akhirnya, dilakukan pemisahan data, di mana dataset dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data) dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Proses ini penting untuk memastikan kinerja model dapat diuji setelah dilatih.

3.5 Implementasi Model

Dalam penelitian ini, implementasi model menggunakan algoritma K-Means dan Random Forest bertujuan untuk menganalisis ketimpangan pendidikan di Provinsi Banten dengan pendekatan machine learning. Berikut adalah penjelasan rinci mengenai implementasi kedua model tersebut:

1. Implementasi K-Means

K-Means adalah algoritma clustering yang digunakan untuk mengelompokkan data berdasarkan kesamaan fitur. Dalam konteks penelitian ini, K-Means digunakan untuk mengidentifikasi kelompok-kelompok siswa atau daerah berdasarkan

karakteristik sosial ekonomi mereka. Berikut adalah langkah-langkah implementasi K-Means:

Pemilihan Jumlah Kluster (K): Langkah pertama dalam K-Means adalah menentukan jumlah kluster yang optimal. Metode elbow dapat digunakan untuk membantu memilih nilai K yang tepat dengan memplot varians dalam kluster terhadap jumlah kluster dan mencari titik di mana penurunan varians mulai melambat.

Inisialisasi Centroid: Setelah menentukan nilai K, centroid untuk setiap kluster diinisialisasi secara acak dari data.

Pengelompokan Data: Data kemudian dikelompokkan ke dalam kluster berdasarkan jarak terdekat ke centroid. Proses ini dilakukan dengan menghitung jarak Euclidean antara setiap titik data dan centroid.

Pembaruan Centroid: Setelah semua data dikelompokkan, centroid diperbarui dengan menghitung rata-rata dari semua titik data dalam kluster.

Iterasi: Proses pengelompokan dan pembaruan centroid diulang hingga tidak ada perubahan signifikan dalam posisi centroid atau jumlah iterasi maksimum tercapai.

Analisis Hasil: Setelah model K-Means diterapkan, hasilnya dianalisis untuk mengidentifikasi pola-pola dalam data yang berkaitan dengan akses pendidikan. Kluster yang terbentuk dapat memberikan wawasan tentang kelompok-kelompok siswa atau daerah dengan karakteristik sosial ekonomi serupa.

2. Implementasi Random Forest

Random Forest adalah algoritma machine learning berbasis ensemble yang digunakan untuk klasifikasi dan regresi. Dalam penelitian ini, Random Forest digunakan untuk memprediksi hasil pendidikan berdasarkan faktor-faktor sosial ekonomi. Berikut adalah langkah-langkah implementasi Random Forest:

Persiapan Data: Data yang telah dibersihkan dan dipra-proses dibagi menjadi dua set: data pelatihan (training set) dan data pengujian (test set). Umumnya, 70% dari data digunakan untuk pelatihan dan 30% untuk pengujian.

Pelatihan Model: Algoritma Random Forest dilatih menggunakan data pelatihan. Model ini terdiri dari sejumlah pohon keputusan (decision trees) yang dibangun secara acak dari subset data dan fitur. Setiap pohon memberikan suara untuk prediksi akhir.

Pengujian Model: Setelah model dilatih, kinerjanya diuji menggunakan data pengujian. Akurasi model dievaluasi dengan membandingkan prediksi model dengan nilai sebenarnya.

Evaluasi Kinerja: Beberapa metrik evaluasi seperti akurasi, presisi, recall, dan F1-score digunakan untuk menilai kinerja model Random Forest. Selain itu, penting juga untuk memeriksa pentingnya fitur (feature importance) untuk memahami faktor-faktor mana yang paling berkontribusi terhadap prediksi hasil pendidikan.

3.6 Evaluasi Model

Evaluasi model merupakan langkah penting dalam pengembangan sistem machine learning yang efektif, terutama dalam konteks analisis data penjualan di Infinity Jar Surabaya. Hasil evaluasi memberikan gambaran yang jelas tentang performa model dalam memprediksi produk paling laris dan tidak laris berdasarkan data transaksi. Dalam penelitian ini, model dievaluasi menggunakan berbagai metrik, termasuk akurasi, presisi, recall, dan F1-score, yang masing-masing memiliki peran penting dalam menilai kualitas prediksi.

1. **Metrik Evaluasi**

   Akurasi memberikan informasi umum tentang seberapa banyak prediksi yang benar dibandingkan dengan total prediksi yang dibuat. Dalam konteks ini, akurasi tinggi menunjukkan bahwa model mampu mengenali pola penjualan dengan baik.

Presisi dan Recall berfokus pada kinerja model dalam mengidentifikasi produk yang benar-benar laris. Hal ini sangat penting, mengingat kesalahan dalam memprediksi produk yang laris atau tidak laris dapat berdampak signifikan pada strategi pemasaran dan pengelolaan inventaris.

F1-score, sebagai gabungan dari presisi dan recall, memberikan pandangan yang lebih seimbang tentang kinerja model, terutama ketika terdapat ketidakseimbangan antara produk yang laris dan tidak laris.

Confusion Matrix juga diimplementasikan untuk menganalisis prediksi model pada setiap kelas, memberikan wawasan mendalam tentang jenis kesalahan yang terjadi, seperti false positives (produk yang diprediksi laris tetapi sebenarnya tidak) dan false negatives (produk yang diprediksi tidak laris tetapi sebenarnya laris). Informasi ini sangat berharga untuk perbaikan model di masa mendatang.

2. **Proses Evaluasi**

   Proses evaluasi dilakukan secara menyeluruh pada validation set dan test set untuk memastikan bahwa model tidak hanya berfungsi dengan baik pada data yang telah dilatih, tetapi juga mampu menggeneralisasi dengan baik pada data yang tidak terlihat sebelumnya. Hal ini penting untuk menghindari masalah overfitting, di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga kehilangan kemampuannya untuk melakukan prediksi yang akurat pada data baru. Dengan menggunakan validation set, peneliti dapat melakukan penyesuaian parameter dan memilih model yang paling sesuai sebelum menguji kinerjanya pada test set.Pengujian pada test set memberikan gambaran akhir tentang kemampuan model dalam situasi dunia nyata. Dengan demikian, evaluasi yang komprehensif tidak hanya meningkatkan keandalan model tetapi juga memberikan dasar yang kuat untuk pengambilan keputusan yang lebih baik dalam penerapan model di berbagai bidang.

3. **Evaluasi K-Means**

Selain evaluasi model prediksi seperti Random Forest, evaluasi algoritma K-Means juga penting dalam konteks analisis data ini. K-Means digunakan untuk mengelompokkan produk berdasarkan pola penjualannya. Berikut adalah langkah-langkah evaluasi untuk K-Means:

Silhouette Score: Salah satu metrik untuk mengevaluasi hasil clustering adalah silhouette score, yang mengukur seberapa baik setiap titik data terpisah dari kluster lain. Nilai silhouette berkisar antara -1 hingga 1; nilai mendekati 1 menunjukkan bahwa titik data berada jauh dari kluster lain dan dekat dengan kluster sendiri.

Elbow Method: Untuk menentukan jumlah kluster (K) yang optimal, metode elbow digunakan dengan memplot nilai inertia (jumlah jarak kuadrat dari titik ke centroid kluster) terhadap jumlah kluster. Titik di mana penurunan inertia mulai melambat menunjukkan jumlah kluster optimal.

Visualisasi Kluster: Visualisasi hasil clustering melalui plot grafis dapat memberikan wawasan tambahan mengenai distribusi produk dalam kluster-kluster tertentu. Ini membantu memahami pola penjualan dan segmentasi pasar.

Analisis Karakteristik Kluster: Setelah kluster terbentuk, analisis karakteristik setiap kluster dilakukan untuk memahami atribut produk dalam setiap kelompok. Misalnya, kluster dapat menunjukkan produk-produk dengan penjualan tinggi atau rendah berdasarkan faktor-faktor tertentu seperti harga atau kategori produk.