Komparasi Algoritma Naïve Bayes dan K-NN dalam PEMBAGIAN BANTUAN DESA

Nurhikmah Wulandari
Ibnu Rasyid Munthe, S.T., M.Kom
Angga putra Juledi, S.Kom., M.Kom
Marnis Nasution, S.Kom., M.Kom

litnus

KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NN DALAM PEMBAGIAN BANTUAN DESA

Ditulis oleh:

Nurhikmah Wulandari
Ibnu Rasyid Munthe, S.T., M.Kom.
Angga putra Juledi, S.Kom., M.Kom.
Marnis Nasution, S.Kom., M.Kom.

Diterbitkan, dicetak, dan didistribusikan oleh
PT Literasi Nusantara Abadi Grup
Perumahan Puncak Joyo Agung Residence Blok B11 Merjosari
Kecamatan Lowokwaru Kota Malang 65144
Telp : +6285887254603, +6285841411519
Email: literasinusantaraofficial@gmail.com
Web: www.penerbitlitnus.co.id
Anggota IKAPI No. 340/JTI/2022

Cetakan I, Maret 2025

Perancang sampul: Rosyiful Aqli
Penata letak: Noufal Fahriza

ISBN : 978-634-206-993-6

x + 108 hlm. ; 15,5x23 cm.

PRAKATA

Puji dan syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga buku monograf yang berjudul "Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor untuk Klasifikasi Penerima Bantuan Sosial di Desa Marbau Selatan" ini dapat diselesaikan dengan baik. Buku ini disusun sebagai bagian dari upaya akademik dalam mengembangkan metode klasifikasi berbasis machine learning guna meningkatkan efektivitas dan efisiensi dalam penentuan penerima bantuan sosial.

Dalam penelitian ini, kami membandingkan dua algoritma klasifikasi, yakni Naïve Bayes dan K-Nearest Neighbor (KNN), dalam mengolah dan menganalisis data penerima bantuan sosial di Desa Marbau Selatan. Pemilihan algoritma ini didasarkan pada karakteristik masing-masing metode yang memiliki keunggulan dalam menangani data dengan distribusi probabilistik (Naïve Bayes) serta berbasis kedekatan antar data (KNN). Hasil penelitian ini diharapkan dapat memberikan wawasan bagi akademisi, praktisi, serta pemangku kebijakan terkait penerapan teknologi dalam pengelolaan data sosial, khususnya dalam konteks pemberian bantuan sosial yang lebih tepat sasaran.

Kami menyadari bahwa dalam proses penyusunan buku ini masih terdapat berbagai kekurangan. Oleh karena itu, kami sangat mengharapkan kritik dan saran yang membangun dari para pembaca demi penyempurnaan penelitian dan karya ilmiah di masa mendatang. Ucapan terima kasih kami sampaikan kepada rekan-rekan akademisi,

pemerintah desa, serta semua pihak yang telah berkontribusi dalam penelitian ini. Semoga buku ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan aplikasi teknologi di bidang data mining serta sistem pengambilan keputusan berbasis kecerdasan buatan.

Penulis

PENGANTAR

Perkembangan teknologi informasi dan kecerdasan buatan telah membawa perubahan signifikan dalam berbagai aspek kehidupan, termasuk dalam pengelolaan data sosial. Salah satu tantangan utama dalam sistem distribusi bantuan sosial adalah memastikan bahwa bantuan tersebut diterima oleh kelompok yang benar-benar membutuhkan. Proses klasifikasi penerima bantuan yang masih dilakukan secara manual sering kali memiliki kendala, seperti subjektivitas, ketidaktepatan data, serta keterbatasan dalam pengolahan informasi dalam skala besar. Oleh karena itu, penerapan metode berbasis machine learning menjadi solusi yang potensial dalam meningkatkan akurasi dan efisiensi proses ini.

Buku monograf ini berjudul "Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor untuk Klasifikasi Penerima Bantuan Sosial di Desa Marbau Selatan", yang bertujuan untuk mengeksplorasi penerapan dua algoritma klasifikasi dalam proses seleksi penerima bantuan sosial. Naïve Bayes dikenal sebagai algoritma berbasis probabilitas yang memiliki keunggulan dalam menangani data berskala besar dengan asumsi independensi antar variabel, sedangkan K-Nearest Neighbor (KNN) merupakan algoritma berbasis jarak yang mempertimbangkan kedekatan karakteristik antara satu individu dengan individu lainnya dalam proses klasifikasi. Studi kasus dalam buku ini berfokus pada Desa Marbau Selatan, dengan menggunakan data penerima bantuan sosial sebagai objek penelitian. Perbandingan kedua algoritma ini bertujuan untuk mengidentifikasi

metode yang paling optimal dalam menentukan siapa yang berhak menerima bantuan berdasarkan berbagai faktor yang relevan. Melalui analisis komparatif, buku ini akan memberikan wawasan mengenai kelebihan dan kekurangan masing-masing metode serta rekomendasi mengenai algoritma yang lebih sesuai untuk diterapkan dalam sistem pengelolaan data bantuan sosial di tingkat desa.

Selain itu, dalam buku ini juga dibahas potensi pengembangan metode hybrid serta kemungkinan penerapan algoritma lain seperti Decision Tree dan Random Forest guna meningkatkan akurasi sistem klasifikasi. Dengan demikian, buku ini tidak hanya memberikan kajian teknis mengenai performa algoritma, tetapi juga membuka ruang bagi penelitian lebih lanjut dalam pengembangan sistem berbasis data mining untuk keperluan sosial dan pemerintahan.

Kami berharap buku ini dapat menjadi referensi yang bermanfaat bagi akademisi, peneliti, serta praktisi yang tertarik dalam bidang machine learning, data mining, dan sistem pengambilan keputusan berbasis kecerdasan buatan. Semoga buku ini dapat memberikan kontribusi bagi pengembangan ilmu pengetahuan dan implementasi teknologi yang lebih inovatif dalam mendukung program sosial yang lebih transparan dan tepat sasaran.

Penulis

DAFTAR ISI

BAB 6

1 PENDAHULUAN

Latar Belakang

Pemberian bantuan sosial oleh pemerintah merupakan salah satu instrumen penting dalam upaya untuk meningkatkan kesejahteraan masyarakat, terutama bagi kelompok yang kurang mampu secara ekonomi. Bantuan sosial bertujuan untuk memberikan dukungan kepada mereka yang mengalami kesulitan dalam memenuhi kebutuhan dasar, seperti pangan, kesehatan, dan pendidikan. Meskipun demikian, dalam pelaksanaannya, distribusi bantuan sosial sering kali menghadapi masalah serius terkait ketepatan sasaran. Ketidaktepatan sasaran ini menjadi masalah yang sangat krusial, karena bantuan sosial yang tidak sampai kepada masyarakat yang benar-benar membutuhkan justru memperburuk ketimpangan sosial dan ekonomi yang ada.

Di banyak wilayah, terutama di daerah pedesaan, proses seleksi penerima bantuan sosial masih dilakukan secara tradisional, yang bergantung pada data administrasi yang sering kali tidak akurat atau

tidak lengkap. Hal ini mengakibatkan banyak rumah tangga yang seharusnya menerima bantuan tidak terdaftar sebagai penerima, sementara mereka yang tidak membutuhkan malah mendapatkan bantuan tersebut. Fenomena ini sering terjadi di desa-desa dengan tingkat kemiskinan yang tinggi, di mana akses terhadap informasi dan teknologi sangat terbatas.

Desa Marbau Selatan, sebagai lokasi penelitian dalam naskah ini, merupakan contoh nyata dari masalah tersebut. Desa ini memiliki karakteristik demografi dengan sebagian besar penduduknya berprofesi sebagai petani yang bergantung pada hasil pertanian sebagai sumber penghidupan. Namun, meskipun mayoritas penduduknya tergolong dalam golongan ekonomi lemah, sistem distribusi bantuan sosial di desa ini masih belum efektif. Hal ini disebabkan oleh kurangnya pemahaman dan kemampuan dalam mengidentifikasi calon penerima bantuan secara tepat. Ketidakakuratan sistem seleksi penerima bantuan menyebabkan banyaknya kelompok masyarakat yang seharusnya mendapat dukungan, namun tidak memperoleh bantuan yang diperlukan.

Untuk mengatasi permasalahan tersebut, algoritma pembelajaran mesin seperti Naïve Bayes dan K-Nearest Neighbor (K-NN) dapat menjadi solusi yang sangat potensial. Kedua algoritma ini menawarkan pendekatan berbasis data yang lebih objektif dan efisien dalam mengklasifikasikan penerima bantuan sosial. Dengan menggunakan data sosial-ekonomi yang lebih rinci, algoritma ini dapat membantu meningkatkan akurasi dalam menentukan siapa yang seharusnya menerima bantuan. Algoritma Naïve Bayes dapat memperkirakan probabilitas kelayakan penerima bantuan berdasarkan variabel-variabel yang ada, sementara K-NN dapat menangani distribusi data yang tidak merata, yang sering menjadi masalah dalam analisis data sosial-ekonomi. Kombinasi kedua algoritma ini diharapkan dapat memberikan solusi yang lebih tepat dalam memitigasi masalah ketepatan sasaran dalam pemberian bantuan sosial.

Pentingnya Algoritma Pembelajaran Mesin

Dalam proses distribusi bantuan sosial, ketepatan sasaran menjadi isu yang sangat krusial. Salah satu tantangan utama adalah bagaimana memilih dan mengidentifikasi penerima bantuan yang benar-benar membutuhkan dengan menggunakan data yang tersedia. Di banyak wilayah, terutama di daerah pedesaan, penggunaan metode manual atau berbasis kuesioner tradisional sering kali menghasilkan ketidaktepatan dalam menentukan siapa yang layak menerima bantuan sosial. Hal ini terjadi karena faktor subjektivitas dalam pengambilan keputusan, kurangnya data yang akurat, dan terbatasnya kapasitas sumber daya manusia dalam menganalisis data.

Untuk mengatasi masalah tersebut, penerapan algoritma pembelajaran mesin (machine learning), seperti Naïve Bayes dan K-Nearest Neighbor (KNN), menawarkan solusi yang lebih efisien dan akurat. Algoritma-algoritma ini dapat menganalisis data besar dan kompleks, serta memberikan hasil klasifikasi yang berbasis pada pola-pola data yang lebih objektif. Sebagai contoh, Naïve Bayes dapat memperkirakan probabilitas kelayakan penerima bantuan sosial berdasarkan variabel-variabel seperti usia, penghasilan, dan status pekerjaan, sementara KNN dapat mengelompokkan individu berdasarkan kedekatan data sosial-ekonomi mereka. Dengan menggunakan algoritma ini, pemerintah atau lembaga yang berwenang dapat lebih tepat dalam menentukan siapa yang harus menerima bantuan, mengurangi potensi kesalahan dan ketimpangan dalam distribusi bantuan sosial. Dengan demikian, penggunaan algoritma pembelajaran mesin tidak hanya meningkatkan akurat dan efisien dalam seleksi penerima bantuan sosial, tetapi juga mendukung pemerataan dan keadilan sosial.

Permasalahan dalam Metode Konvensional

Metode konvensional yang masih digunakan dalam sistem penyaluran bantuan sosial sering kali tidak efektif dan kurang efisien dalam menjangkau masyarakat yang membutuhkan. Pada umumnya, proses seleksi penerima bantuan dilakukan secara manual, dengan menggunakan data administratif yang tidak selalu akurat dan terkini. Di banyak kasus, data yang digunakan untuk menentukan siapa yang layak menerima bantuan sering kali terbatas pada informasi dasar seperti pendataan keluarga yang sudah lama atau tidak lengkap, sehingga menimbulkan ketidaktepatan sasaran. Hal ini menyebabkan banyaknya penerima bantuan yang tidak memenuhi syarat, sementara mereka yang seharusnya membutuhkan justru tidak mendapatkan bantuan yang seharusnya mereka terima.

Metode konvensional ini juga sangat bergantung pada penilaian subjektif, yang mengarah pada ketidakadilan dalam distribusi bantuan. Misalnya, petugas yang melakukan penilaian atau seleksi dapat memiliki preferensi pribadi atau bias yang tidak disadari, yang pada gilirannya mempengaruhi keputusan. Selain itu, keterbatasan dalam hal sumber daya manusia dan waktu yang tersedia juga menghambat kemampuan untuk memverifikasi secara mendalam kondisi sosial-ekonomi setiap individu atau keluarga yang terdaftar dalam sistem. Oleh karena itu, penting untuk mempertimbangkan penerapan algoritma pembelajaran mesin yang dapat memberikan solusi lebih objektif dan berbasis data dalam sistem seleksi penerima bantuan sosial, seperti yang diusulkan dalam penelitian ini.

Tujuan Penelitian

Metode konvensional yang masih digunakan dalam sistem penyaluran bantuan sosial sering kali tidak efektif dan kurang efisien dalam menjangkau masyarakat yang membutuhkan. Pada umumnya,

proses seleksi penerima bantuan dilakukan secara manual, dengan menggunakan data administratif yang tidak selalu akurat dan terkini. Di banyak kasus, data yang digunakan untuk menentukan siapa yang layak menerima bantuan sering kali terbatas pada informasi dasar seperti pendataan keluarga yang sudah lama atau tidak lengkap, sehingga menimbulkan ketidaktepatan sasaran. Hal ini menyebabkan banyaknya penerima bantuan yang tidak memenuhi syarat, sementara mereka yang seharusnya membutuhkan justru tidak mendapatkan bantuan yang seharusnya mereka terima.

Metode konvensional ini juga sangat bergantung pada penilaian subjektif, yang mengarah pada ketidakadilan dalam distribusi bantuan. Misalnya, petugas yang melakukan penilaian atau seleksi dapat memiliki preferensi pribadi atau bias yang tidak disadari, yang pada gilirannya mempengaruhi keputusan. Selain itu, keterbatasan dalam hal sumber daya manusia dan waktu yang tersedia juga menghambat kemampuan untuk memverifikasi secara mendalam kondisi sosial-ekonomi setiap individu atau keluarga yang terdaftar dalam sistem. Oleh karena itu, penting untuk mempertimbangkan penerapan algoritma pembelajaran mesin yang dapat memberikan solusi lebih objektif dan berbasis data dalam sistem seleksi penerima bantuan sosial, seperti yang diusulkan dalam penelitian ini.

Dampak Potensial terhadap Kebijakan Sosial dan Teknologi

Penelitian ini memiliki dampak yang signifikan terhadap kebijakan sosial, terutama dalam meningkatkan keakuratan distribusi bantuan sosial. Dengan penerapan algoritma Naïve Bayes dan K-Nearest Neighbor (KNN), sistem seleksi penerima bantuan sosial dapat lebih efisien dan adil, memastikan bahwa bantuan sampai kepada mereka yang benar-benar membutuhkan. Hal ini dapat mengurangi ketimpangan sosial dan meningkatkan kesejahteraan masyarakat. Dari sisi teknologi, penelitian ini memperkenalkan pemanfaatan

machine learning dalam sektor publik, yang dapat menjadi referensi untuk penerapan teknologi serupa di berbagai bidang lainnya. Selain itu, hasil penelitian ini juga mendorong pengembangan sistem berbasis data yang lebih canggih, memungkinkan analisis lebih dalam untuk kebijakan sosial yang lebih berbasis bukti. Secara keseluruhan, penelitian ini tidak hanya mendukung perbaikan dalam distribusi bantuan sosial, tetapi juga memperkuat integrasi teknologi dalam pengambilan keputusan kebijakan publik.

2 KERANGKA TEORETIS

Definisi dan Konsep Dasar Bantuan Social

Bantuan sosial adalah bentuk intervensi yang diberikan oleh pemerintah atau lembaga terkait untuk mendukung kesejahteraan masyarakat, khususnya bagi mereka yang tergolong dalam kelompok rentan atau ekonomi lemah. Bantuan sosial bertujuan untuk memastikan akses masyarakat terhadap kebutuhan dasar, seperti pangan, kesehatan, pendidikan, dan tempat tinggal. Bentuk bantuan sosial yang paling umum meliputi bantuan langsung tunai, bantuan pangan, dan subsidi kebutuhan dasar. Pemberian bantuan sosial ini menjadi instrumen yang sangat penting untuk mengurangi tingkat kemiskinan dan meningkatkan kualitas hidup, khususnya di daerah-daerah yang memiliki tingkat ketimpangan ekonomi yang tinggi.

Namun, meskipun tujuan dari bantuan sosial adalah untuk mengurangi ketimpangan sosial dan ekonomi, sistem distribusinya sering kali menemui berbagai tantangan. Salah satu tantangan terbesar adalah ketidaktepatan sasaran. Hal ini disebabkan oleh berbagai

faktor, seperti data yang tidak akurat, kurangnya transparansi, dan proses seleksi penerima yang tidak efisien. Misalnya, di banyak daerah, proses seleksi penerima bantuan masih dilakukan dengan cara yang sangat manual, yang mengandalkan data yang sudah lama atau tidak lengkap. Sistem yang berbasis pada pendataan administratif yang tidak terkelola dengan baik sering kali menghasilkan penerima bantuan yang tidak membutuhkan, sementara mereka yang benar-benar membutuhkan tidak mendapatkan bantuan yang seharusnya mereka terima.

Ketidaktepatan dalam mendistribusikan bantuan ini tidak hanya berisiko menambah kesenjangan sosial, tetapi juga dapat menurunkan kepercayaan masyarakat terhadap pemerintah atau lembaga penyedia bantuan. Jika masyarakat merasa bahwa bantuan tidak sampai ke tangan yang membutuhkan, hal ini dapat mengurangi efektivitas kebijakan bantuan sosial secara keseluruhan. Selain itu, sistem yang tidak efisien juga berdampak pada penggunaan sumber daya yang tidak optimal, baik dari segi waktu, tenaga, maupun dana. Oleh karena itu, penting untuk memiliki sistem yang lebih transparan, efisien, dan berbasis data yang akurat dalam melakukan seleksi penerima bantuan.

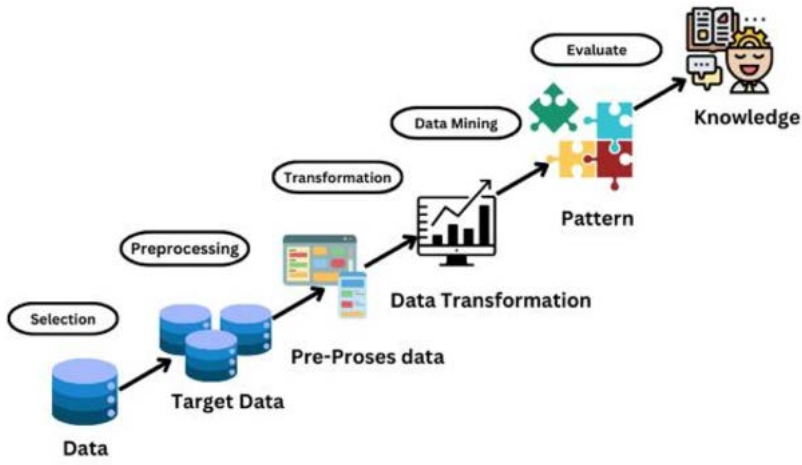
Untuk mengatasi tantangan ini, banyak penelitian menunjukkan bahwa penerapan teknologi informasi dan pembelajaran mesin (machine learning) dapat menjadi solusi yang sangat efektif. Dengan menggunakan algoritma seperti Naïve Bayes atau K-Nearest Neighbor (KNN), seleksi penerima bantuan sosial dapat dilakukan dengan lebih objektif, berdasarkan data yang lebih akurat dan terkini. Teknologi ini memungkinkan analisis data sosial-ekonomi yang lebih mendalam dan dapat mengklasifikasikan penerima bantuan dengan lebih tepat, sehingga distribusi bantuan dapat lebih tepat sasaran. Dengan penerapan teknologi ini, diharapkan bantuan sosial dapat diberikan kepada mereka yang benar-benar membutuhkan, mengurangi pemborosan, dan meningkatkan kepercayaan publik terhadap kebijakan sosial yang ada.

Knowledge Discovery in Database (KDD)

Knowledge Discovery in Databases (KDD) adalah proses terstruktur yang mencakup berbagai tahap yang bertujuan untuk mengekstrak informasi berharga dari kumpulan data besar. *Knowledge* Discovery In Database (KDD) merupakan keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti[1]. Proses KDD biasanya mencakup langkah-langkah seperti pemilihan data, praproses, transformasi, penambangan data, dan evaluasi. Penambangan data, sebagai komponen inti KDD, melibatkan penerapan algoritma untuk mengidentifikasi pola dan hubungan dalam data[2]. Misalnya, metode pengelompokan seperti *K-Means* sering digunakan untuk mengoptimalkan organisasi data dan meningkatkan proses pengambilan keputusan[3][4].

Lebih jauh, integrasi teknologi web semantik ke dalam KDD telah dieksplorasi untuk meningkatkan interpretasi data dan efisiensi penambangan. Pendekatan interdisipliner ini menyoroti sifat teknik penambangan data yang terus berkembang, yang semakin banyak digunakan di berbagai bidang, termasuk perawatan kesehatan dan pemasaran, untuk memperoleh wawasan yang dapat ditindaklanjuti dari kumpulan data yang kompleks. Singkatnya, KDD merupakan kerangka kerja komprehensif yang tidak hanya memfasilitasi penambangan data tetapi juga meningkatkan pemahaman dan penerapan pengetahuan yang diekstraksi di berbagai domain.

Metode penelitian ini menggunakan teknik pendekatan KDD (*Knowledge Discovery in Database*), untuk mengklasifikasikan data penerima BLT. Data mining merupakan bagian dari tahap proses Knowledge Discovery in Database (KDD), yang bertujuan untuk mencari informasi baru dan berharga dalam suatu kumpulan data atau database. Dalam penelitian ini, penerapan data mining mengikuti langkah-langkah yang ada pada KDD, dimulai dari penetapan tujuan hingga proses evaluasi[5]. Tahapan KDD dapat dilihat pada Gambar 2.1:



Gambar 2.1 Tahapan Knowledge Discovery in Database (KDD)

Menyiapkan data yang akan diproses.

1. Selection
   Pada tahap ini, langkah yang diambil melibatkan pemilihan data yang menggunakan operator read excel.
2. Data Preprocessing
   Dalam tahap ini, peneliti menangani data yang melibatkan penghapusan data duplikat, pengecekan konsistensi data, serta perbaikan kesalahan.
3. Data Transformation
   Transformasi Data adalah suatu usaha yang dilakukan dengan tujuan utama untuk mengalihkan skala pengukuran data asli ke bentuk lain sehingga data dapat memenuhi asumsi-asumsi yang mendasari analisis data.
4. Data Mining
   Tahap ini adalah inti dari proses KDD dan dijalankan setelah data dibersihkan. Pada tahap ini, data akan diproses menggunakan algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. Setelah itu melakukan penerapan model menggunakan operator

apply model.dengan menggunakan alat bantu Rapid Miner dan Microsoft Excel.

5. Evaluation

   Pada tahap akhir, dilakukan proses menghasilkan output menggunakan operator performance yang dapat dipahami dari pola informasi hasil data mining. Sebagai model akurasi evaluasi di hasilkan negatif dan positif, maka digunakan metode *confusion matrix*. *Confusion matrix* suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy*, *precision* dan *recall*.

Data Mining

Data mining merupakan salah satu bidang dalam ilmu komputer dan statistik yang berfokus pada ekstraksi pola atau informasi berharga dari kumpulan data yang besar. Konsep ini melibatkan berbagai teknik, seperti klasifikasi, clustering, regresi, dan asosiasi, yang bertujuan untuk menemukan hubungan tersembunyi dalam data. Proses data mining adalah serangkaian langkah sistematis yang dilakukan untuk menggali pengetahuan yang berharga dari dataset yang kompleks[5]. Dalam konteks bantuan sosial, data mining dapat digunakan untuk mengolah data sosial-ekonomi masyarakat, menganalisis pola penerima bantuan, serta meningkatkan ketepatan dalam seleksi penerima bantuan sosial. Proses ini sangat penting mengingat bahwa data penerima bantuan sosial sering kali kompleks, mencakup berbagai variabel seperti penghasilan, jumlah tanggungan, usia, dan kepemilikan aset, yang semuanya berkontribusi terhadap kelayakan seseorang menerima bantuan sosial.

Salah satu peranan utama data mining dalam klasifikasi penerima bantuan sosial adalah otomatisasi pengambilan keputusan berdasarkan pola yang teridentifikasi dalam data. Dengan menerapkan algoritma klasifikasi seperti Naïve Bayes dan K-Nearest Neighbor

(KNN), sistem dapat secara otomatis menganalisis data masyarakat dan memberikan keputusan tentang siapa yang memenuhi kriteria penerima bantuan sosial. Algoritma Naïve Bayes bekerja dengan menghitung probabilitas dari setiap kategori berdasarkan fitur-fitur yang ada, sehingga dapat memberikan hasil yang cepat dan efisien. Sementara itu, algoritma KNN membandingkan setiap individu dengan kelompok yang sudah diklasifikasikan sebelumnya untuk menentukan kategori penerima bantuan. Kedua metode ini menawarkan solusi yang lebih objektif dibandingkan dengan metode manual yang sering kali bergantung pada subjektivitas petugas atau kebijakan lokal yang belum tentu akurat. Proses data mining melibatkan ekstraksi pola dan pengetahuan yang bermanfaat dari kumpulan data besar[6].

Selain meningkatkan efisiensi dalam klasifikasi penerima bantuan sosial, data mining juga memiliki peran penting dalam deteksi kecurangan atau penyalahgunaan sistem. Salah satu masalah yang sering terjadi dalam distribusi bantuan sosial adalah adanya penerima bantuan yang sebenarnya tidak memenuhi kriteria, namun tetap terdaftar dalam sistem karena kesalahan pencatatan atau bahkan manipulasi data. Dengan teknik anomaly detection yang merupakan bagian dari data mining, sistem dapat mengenali pola-pola yang mencurigakan, seperti seseorang yang menerima bantuan dari beberapa program berbeda secara bersamaan atau individu dengan penghasilan tinggi yang masih terdaftar sebagai penerima bantuan. Hal ini memungkinkan pemerintah atau lembaga sosial untuk melakukan verifikasi ulang dan memastikan bantuan diberikan kepada mereka yang benar-benar membutuhkan.

Dalam jangka panjang, penerapan data mining dalam klasifikasi penerima bantuan sosial akan meningkatkan akurasi distribusi bantuan, mengoptimalkan penggunaan anggaran sosial, serta membangun sistem yang lebih transparan dan akuntabel. Dengan memanfaatkan analisis berbasis data, pemerintah atau lembaga sosial dapat merancang kebijakan yang lebih berbasis bukti (evidence-based

policy), memastikan bahwa setiap keputusan yang diambil memiliki dasar yang kuat dalam data yang telah dianalisis. Oleh karena itu, data mining bukan hanya sebuah teknologi, tetapi juga sebuah solusi strategis yang dapat mendukung upaya pemerintah dalam mencapai kesejahteraan sosial yang lebih merata.

Konsep Dasar Klasifikasi

Klasifikasi adalah salah satu teknik fundamental dalam data mining yang bertujuan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan atribut yang dimilikinya. Proses klasifikasi dimulai dengan pelatihan model menggunakan data latih (*training data*) untuk mengidentifikasi pola dan aturan. Model yang dihasilkan kemudian digunakan untuk memprediksi atau mengklasifikasikan data uji (*testing data*) yang belum diketahui kelasnya. Teknik ini sangat berguna dalam menganalisis pola data dan mendukung pengambilan keputusan berbasis informasi.

Dalam penelitian ini, klasifikasi diterapkan untuk berbagai tujuan, seperti diagnosis medis, analisis sentimen, dan pengelompokan konsumen dalam pemasaran. Sebagai contoh, algoritma *Support Vector Machine* (SVM) sering digunakan untuk diagnosis penyakit, termasuk kanker payudara, dengan tingkat akurasi yang tinggi [7]. Algoritma *Naïve Bayes*, yang dikenal sederhana namun efektif, banyak diterapkan dalam pengelompokan data sosial-ekonomi [8].

Metode lain yang juga populer adalah *Decision Tree* dan *K-Nearest Neighbor* (KNN). Decision Tree sering digunakan untuk memahami struktur keputusan berbasis data, sedangkan KNN cocok untuk masalah klasifikasi berbasis kemiripan data, meskipun sensitif terhadap noise [9][10]. Selain itu, algoritma ensemble seperti *Random Forest* menawarkan solusi yang kuat dengan menggabungkan hasil dari beberapa pohon keputusan, menghasilkan model yang lebih stabil dan akurat [11].

Dalam penelitian ini, klasifikasi diterapkan untuk menentukan masyarakat Desa Marbau Selatan yang layak menerima bantuan sosial berdasarkan atribut sosial-ekonomi, seperti pendapatan, jumlah tanggungan keluarga, dan jenis pekerjaan. Penerapan klasifikasi berbasis data sosial-ekonomi membantu mempercepat pengambilan keputusan dan memastikan distribusi bantuan sosial yang lebih tepat sasaran. Pendekatan ini memberikan kontribusi penting untuk meningkatkan keadilan dalam proses distribusi bantuan dan efisiensi alokasi sumber daya.

Tabel 2.1 Perbandingan Algoritma Klasifikasi

Algoritma	Kelebihan	Kekurangan
Naïve Bayes	Cepat dan efisien, baik untuk data kategori	Asumsi independensi atribut sering tidak terpenuhi
KNN	Sederhana, mampu menangani data non-linear	Rentan terhadap outlier dan intensif komputasi
SVM	Akurat untuk data berukuran kecil dengan margin yang jelas	Waktu komputasi tinggi pada dataset besar

Algoritma Naïve Bayes

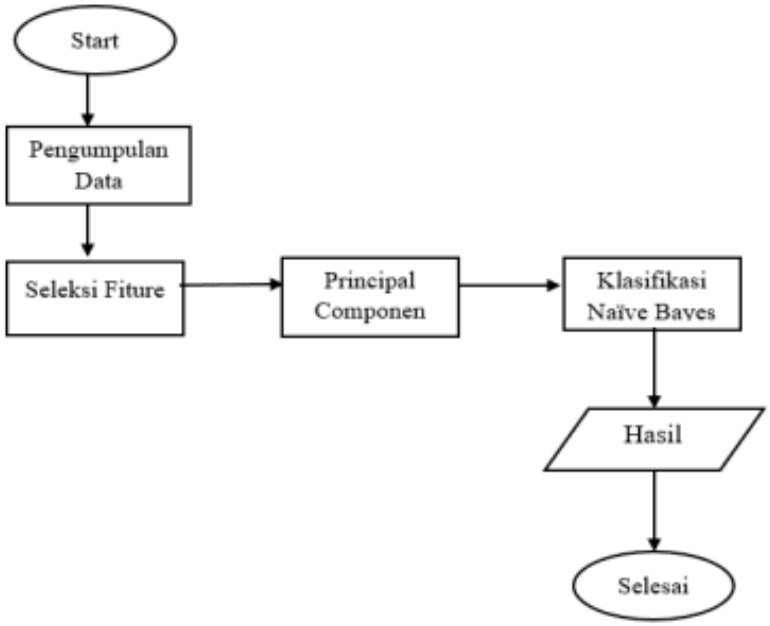
Naïve Bayes adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi data berbasis probabilitas. Algoritma ini mengandalkan Teorema Bayes yang menghubungkan probabilitas bersyarat antar berbagai variabel dalam dataset untuk membuat keputusan klasifikasi. Inti dari Naïve Bayes adalah bahwa ia mengasumsikan bahwa semua fitur (atau atribut) yang digunakan dalam model adalah independen satu sama lain. Meskipun asumsi independensi ini jarang benar di dunia nyata, terutama dalam data sosial-ekonomi yang sering kali memiliki keterkaitan antar variabel, Naïve Bayes tetap sering digunakan karena kesederhanaannya dan kemampuannya untuk memberikan hasil yang memadai dalam banyak kasus, terutama saat mengklasifikasikan data besar dan tidak terstruktur.

Dalam konteks klasifikasi penerima bantuan sosial, Naïve Bayes bekerja dengan cara menghitung probabilitas posterior untuk setiap kategori kelas (misalnya, layak atau tidak layak menerima bantuan) berdasarkan data input yang ada. Setiap fitur, seperti penghasilan, jumlah tanggungan, atau status pekerjaan, dihitung probabilitasnya, dan dengan menggunakan Teorema Bayes, klasifikasi dilakukan dengan memilih kelas yang memiliki probabilitas tertinggi. Naïve Bayes dapat digunakan pada berbagai jenis data, baik data numerik maupun data kategori. *Naïve Bayes* sering digunakan untuk data kategori karena kemampuannya yang andal dalam menangani atribut dengan nilai diskret. Aplikasi umum algoritma ini mencakup analisis sentimen, deteksi spam, diagnosis penyakit, dan prediksi kelayakan penerima bantuan sosial [12][13].

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- $P(H|E)$: Probabilitas hipotesis H berdasarkan bukti H.
- $P(E|H)$: Probabilitas bukti E jika hipotesis H benar.
- $P(H)$: Probabilitas awal dari H.
- $P(E)$: Probabilitas keseluruhan dari bukti E.

Dengan mengasumsikan bahwa fitur-fitur independen satu sama lain, Naïve Bayes dapat menghitung probabilitas dari setiap kelas dengan mengalikan probabilitas individual setiap fitur. Meskipun asumsi independensi ini tidak selalu akurat dalam dunia nyata, algoritma ini terbukti sangat efisien dalam klasifikasi dan memiliki waktu komputasi yang cepat, yang menjadi keunggulan besar saat bekerja dengan dataset besar.



Gambar Diagram Alir Penerapan Naïve Bayes

Dalam penerapan untuk penerima bantuan sosial, misalnya, Naïve Bayes dapat memperkirakan probabilitas apakah seorang individu layak menerima bantuan sosial berdasarkan fitur-fitur seperti penghasilan, jumlah tanggungan, dan kepemilikan rumah. Keunggulan utama Naïve Bayes adalah bahwa ia dapat memberikan keputusan klasifikasi yang cepat dan efisien, bahkan pada dataset yang sangat besar. Hal ini membuat Naïve Bayes menjadi pilihan yang baik untuk aplikasi-aplikasi yang memerlukan kecepatan dan skalabilitas, seperti sistem klasifikasi penerima bantuan sosial yang melibatkan banyak data.

Namun, meskipun memiliki berbagai keunggulan, keterbatasan Naïve Bayes terletak pada asumsi independensi fitur, yang sering kali tidak berlaku pada data dunia nyata. Misalnya, dalam kasus data sosial-ekonomi, fitur seperti penghasilan dan jumlah tanggungan sering kali memiliki hubungan yang sangat erat. Oleh karena itu, meskipun Naïve Bayes dapat memberikan hasil yang cukup baik,

hasilnya mungkin tidak selalu optimal apabila ada hubungan yang signifikan antar fitur yang terabaikan. Meskipun demikian, Naïve Bayes tetap menjadi pilihan populer berkat kesederhanaan dan kemampuannya dalam memproses data dalam waktu singkat, serta efektif dalam berbagai kasus klasifikasi.

Algoritma K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini beroperasi dengan cara mengklasifikasikan data baru berdasarkan kedekatannya dengan data yang sudah ada dalam dataset. KNN termasuk dalam kategori algoritma instance-based learning, yang berarti bahwa keputusan klasifikasi tidak dibuat dengan membangun model secara eksplisit, melainkan dengan mencari kedekatan data baru terhadap data yang sudah ada di dalam ruang fitur. Pada dasarnya, KNN mencari k tetangga terdekat dari data baru dalam dataset yang ada dan mengklasifikasikan data tersebut ke dalam kelas mayoritas dari tetangga tersebut.

Secara matematis, langkah-langkah dalam algoritma KNN adalah sebagai berikut: pertama, tentukan nilai k, yaitu jumlah tetangga terdekat yang akan digunakan dalam klasifikasi. Kemudian, untuk setiap titik data baru, algoritma menghitung jarak antara titik tersebut dengan semua titik data yang ada dalam dataset, menggunakan rumus jarak seperti Euclidean Distance atau Manhattan Distance. Setelah jarak dihitung, data baru akan diklasifikasikan berdasarkan kelas mayoritas dari k tetangga terdekatnya. KNN dikenal sebagai algoritma yang sederhana namun membutuhkan waktu komputasi yang tinggi, terutama pada dataset besar, karena setiap prediksi memerlukan perhitungan jarak terhadap semua data [14][15].

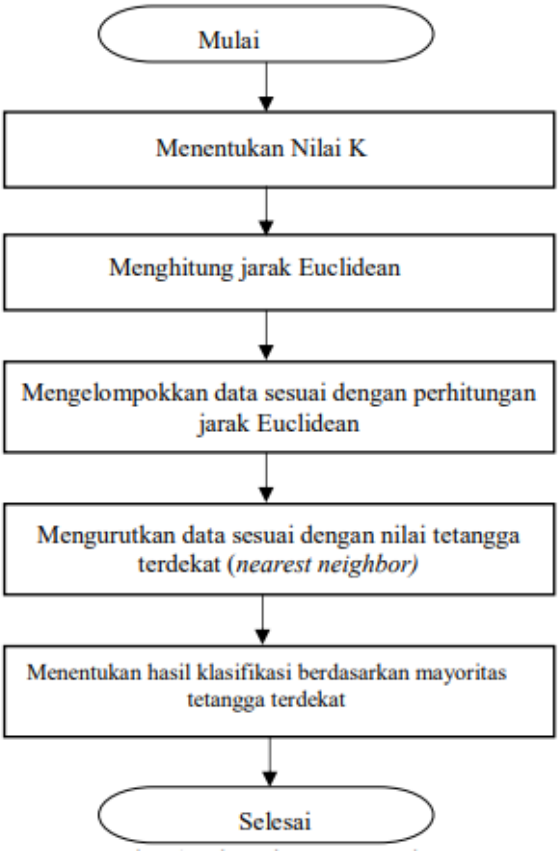
$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Dimana:

- p : titik data baru yang ingin diuji
- q : titik data yang ada di dataset
- qi & pi : nilai data ke q dari titik q dan p
- n : jumlah data dalam dataset

Keunggulan utama dari KNN adalah kemampuannya untuk menangani data non-linear dan tanpa asumsi distribusi tertentu pada data. Dalam kasus klasifikasi penerima bantuan sosial, misalnya, KNN dapat mengklasifikasikan penerima bantuan berdasarkan kedekatan data sosial-ekonomi mereka dengan kelompok yang sudah terklasifikasi sebelumnya, seperti keluarga berpendapatan rendah atau keluarga dengan jumlah tanggungan yang banyak. Hal ini menjadikan KNN sebagai algoritma yang fleksibel dan sangat berguna untuk aplikasi-aplikasi dengan data yang tidak terstruktur dan memiliki hubungan yang kompleks antara fitur-fitur.

Namun, meskipun KNN menawarkan fleksibilitas yang tinggi, algoritma ini juga memiliki beberapa keterbatasan. Salah satunya adalah sensitivitas terhadap outlier. Jika terdapat data yang sangat berbeda (outlier) dari data lainnya, KNN dapat menghasilkan keputusan yang tidak akurat karena outlier tersebut dapat memengaruhi jarak yang dihitung. Selain itu, KNN memerlukan waktu komputasi yang lebih lama, terutama pada dataset yang sangat besar, karena algoritma ini harus menghitung jarak untuk setiap titik data baru terhadap seluruh dataset yang ada. Oleh karena itu, dalam penerapannya, pemilihan nilai k yang tepat dan normalisasi data sangat penting untuk mengoptimalkan kinerja KNN.

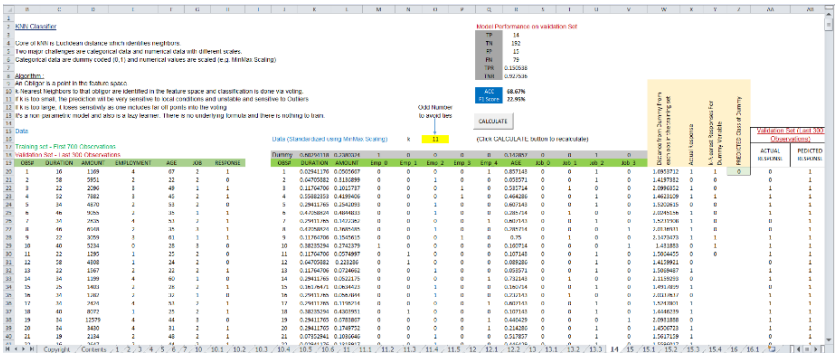


Gambar Diagram Alir Penerapan KNN

Dalam klasifikasi penerima bantuan sosial, KNN dapat diandalkan untuk mengidentifikasi kelompok-kelompok penerima yang memiliki karakteristik serupa dengan mereka yang telah terverifikasi sebelumnya sebagai penerima yang layak. Misalnya, keluarga dengan pendapatan rendah dan jumlah tanggungan tinggi akan memiliki kedekatan yang lebih besar dengan kelompok penerima lain yang memiliki kriteria serupa. Dengan demikian, KNN dapat membantu dalam memastikan bahwa bantuan sosial didistribusikan secara lebih adil, berdasarkan kedekatan data sosial-ekonomi yang relevan.

Microsoft Excel

Microsoft Excel digunakan dalam penelitian ini untuk preprocessing data sederhana, seperti pembersihan data awal, penghapusan duplikasi, dan pengisian nilai kosong. Excel mempermudah manipulasi data yang relatif kecil atau format yang lebih sederhana sebelum data diimpor ke platform analisis yang lebih kompleks seperti *Python* atau *RapidMiner*.



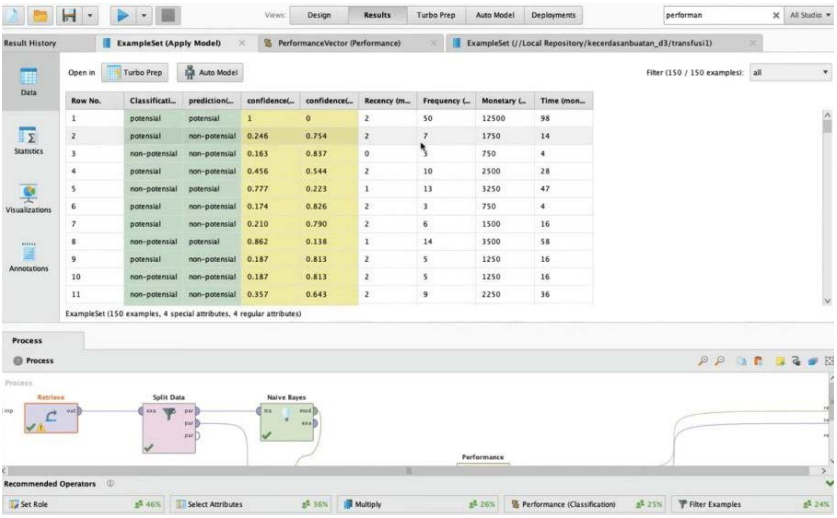
Gambar Proses klasifikasi dengan Microsoft Excel

Microsoft Excel memiliki berbagai fungsi yang relevan untuk pengolahan data klasifikasi. Mulai dari mengurutkan dan menyaring data berdasarkan kriteria tertentu, sehingga memudahkan dalam mengelompokkan data yang relevan untuk analisis klasifikasi [16]. Selain itu, Excel juga menyediakan fungsi yang memungkinkan pengguna untuk menganalisis dan merangkum data dengan cepat, serta mengidentifikasi pola yang dapat digunakan dalam klasifikasi [17]. Dengan demikian, Excel menjadi alat yang efektif dalam pengolahan dan analisis data klasifikasi.

Rapid Minner

Rapidminer merupakan perangkat lunak independen yang digunakan untuk menganalisa data dan mesin penambangan data, yang dapat diintegrasikan dengan berbagai bahasa pemrograman

secara mudah[18]. *RapidMiner* menyediakan antarmuka pengguna yang ramah serta berbagai fitur bawaan untuk evaluasi model menggunakan metrik seperti akurasi, sensitivitas, spesifisitas, dan area di bawah kurva (AUC) [19].



Gambar Tampilan Rapid Miner

Dengan antarmuka yang ramah pengguna, *RapidMiner* memungkinkan pengguna tanpa keterampilan pemrograman yang mendalam untuk menerapkan dan mengevaluasi algoritma klasifikasi secara efisien [20]. Hal ini menjadikan *RapidMiner* sebagai alat yang ideal untuk analisis data dan pengembangan model klasifikasi dalam berbagai konteks.

Penelitian Terdahulu Distribusi Bantuan Sosial

Seiring dengan berkembangnya teknologi, penerapan machine learning (ML) dalam distribusi bantuan sosial semakin banyak diteliti untuk meningkatkan akurasi dan efisiensi dalam memilih penerima bantuan. Penggunaan algoritma ML memungkinkan analisis data sosial-ekonomi yang lebih objektif dan berbasis data, mengurangi

ketidaktepatan sasaran yang sering terjadi dalam sistem seleksi tradisional yang bergantung pada data manual dan asumsi subjektif. Berdasarkan penelitian-penelitian terdahulu, berbagai algoritma pembelajaran mesin telah digunakan untuk mengklasifikasikan penerima bantuan sosial berdasarkan berbagai kriteria, seperti penghasilan, jumlah tanggungan, usia, dan status pekerjaan.

Beberapa penelitian seperti yang dilakukan oleh Kurniawan dan Setiawan (2018) menggunakan algoritma Decision Tree untuk mengklasifikasikan penerima bantuan sosial di Indonesia. Hasil penelitian mereka menunjukkan bahwa Decision Tree memiliki akurasi yang tinggi dalam mengidentifikasi kelompok masyarakat yang layak menerima bantuan, meskipun model ini cenderung lebih sensitif terhadap data yang sangat besar dan memiliki banyak variabel. Naïve Bayes dan K-Nearest Neighbor (KNN) juga banyak digunakan dalam konteks serupa, dengan penelitian oleh Sari (2019) yang menunjukkan bahwa Naïve Bayes mampu memberikan prediksi yang cepat dan akurat meskipun memiliki asumsi independensi antar fitur yang sering tidak berlaku dalam data dunia nyata.

Selain itu, Random Forest telah diterapkan dalam penelitian oleh Hidayati (2020), yang mengkombinasikan berbagai pohon keputusan untuk menghasilkan keputusan klasifikasi yang lebih kuat dan mengurangi kemungkinan overfitting pada dataset yang kompleks. Meskipun algoritma-algoritma ini memberikan hasil yang baik dalam konteks klasifikasi penerima bantuan sosial, mereka seringkali menghadapi tantangan dalam mengatasi data yang tidak seimbang atau beragam, yang dapat mempengaruhi akurasi model.

Penggunaan machine learning untuk klasifikasi penerima bantuan sosial tidak hanya meningkatkan keakuratan distribusi, tetapi juga membuka peluang untuk mendeteksi anomali atau kecurangan. Salah satu penelitian oleh Wahyu et al. (2021) mengimplementasikan algoritma clustering untuk mengidentifikasi kelompok penerima bantuan yang tidak memenuhi syarat. Dalam penelitian ini, algoritma mampu menemukan anomali dalam data seperti penerima bantuan

yang tidak terdaftar atau mereka yang sudah memiliki aset yang cukup untuk memenuhi kebutuhan hidup mereka.

Di sisi lain, deep learning mulai diterapkan untuk mengenali pola yang lebih kompleks dalam dataset yang besar dan tidak terstruktur. Sebuah penelitian oleh Rahmawati et al. (2022) menguji penggunaan neural networks untuk menganalisis data penerima bantuan sosial, dengan hasil bahwa deep learning dapat memberikan keputusan yang lebih akurate dalam skenario yang melibatkan dataset besar dengan banyak variabel yang saling terkait.

Secara keseluruhan, penelitian-penelitian terdahulu menunjukkan bahwa algoritma pembelajaran mesin mampu memberikan solusi yang efektif dalam mengoptimalkan distribusi bantuan sosial, namun tantangan yang dihadapi adalah pemilihan algoritma yang tepat dan pengelolaan data yang baik. Pemilihan algoritma yang sesuai dengan karakteristik data serta preprocessing yang tepat sangat penting untuk mencapai hasil yang optimal. Oleh karena itu, penting untuk terus mengembangkan dan menerapkan berbagai algoritma machine learning dalam distribusi bantuan sosial, dengan memperhatikan keterbatasan dan tantangan yang ada di setiap model.

Keunggulan dan Keterbatasan Kedua Algoritma

Algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) adalah dua algoritma yang populer dalam klasifikasi data, namun keduanya memiliki keunggulan dan keterbatasan yang berbeda dalam menangani dataset sosial-ekonomi. Naïve Bayes adalah algoritma probabilistik yang mengasumsikan independensi antar fitur dalam data. Keunggulan utama dari Naïve Bayes adalah kemampuannya untuk menangani data yang besar dan kompleks dengan sangat efisien. Algoritma ini juga cepat dalam komputasi, karena hanya memerlukan perhitungan probabilitas untuk setiap kategori, tanpa perlu melakukan

pencarian jarak antar data. Hal ini sangat berguna dalam konteks dataset sosial-ekonomi yang besar, di mana kecepatan pemrosesan sangat penting. Selain itu, Naïve Bayes memiliki keunggulan dalam mendukung data kategorikal yang sering ditemukan dalam dataset sosial-ekonomi, seperti jenis pekerjaan, status perkawinan, atau tingkat pendidikan.

Namun, Naïve Bayes memiliki keterbatasan yang signifikan, terutama asumsi independensi fitur. Dalam dataset sosial-ekonomi, fitur-fitur seperti penghasilan dan jumlah tanggungan sering kali memiliki hubungan yang erat, yang menjadikan asumsi independensi ini tidak selalu valid. Hal ini dapat menyebabkan akurasi yang lebih rendah dalam kasus-kasus di mana fitur-fitur tersebut saling bergantung satu sama lain. Selain itu, Naïve Bayes juga kurang efektif dalam menangani data yang berisi outlier atau noise, karena algoritma ini tidak memiliki mekanisme untuk mendeteksi atau mengatasi data yang menyimpang secara signifikan.

Sementara itu, K-Nearest Neighbor (KNN) adalah algoritma berbasis kedekatan data, yang bekerja dengan mengklasifikasikan data berdasarkan kedekatannya dengan data lain yang telah terklasifikasi. Kelebihan utama dari KNN adalah fleksibilitasnya dalam menangani data non-linear dan data yang tidak terstruktur, yang sering ditemukan dalam dataset sosial-ekonomi. KNN juga tidak membutuhkan proses pelatihan yang rumit, karena ia bekerja langsung dengan data yang ada, sehingga dapat diterapkan pada berbagai jenis data tanpa perlu asumsi tertentu tentang hubungan antar fitur.

Namun, KNN juga memiliki kelemahan, terutama dalam hal waktu komputasi dan sensitivitas terhadap outlier. Karena KNN menghitung jarak antar data untuk setiap titik data baru, maka komputasi menjadi lebih lambat seiring dengan bertambahnya ukuran dataset. Dalam dataset sosial-ekonomi yang besar, hal ini bisa menjadi sangat mahal dari sisi waktu pemrosesan. Selain itu, KNN cenderung lebih sensitif terhadap outlier. Jika ada data yang sangat berbeda dari data lainnya, outlier ini dapat memengaruhi

hasil klasifikasi, terutama jika nilai k (jumlah tetangga terdekat) yang dipilih tidak optimal. KNN juga lebih rentan terhadap high-dimensional data, di mana semakin banyak fitur yang ada, semakin sulit bagi KNN untuk menentukan kedekatan yang tepat antara data baru dan data lama.

Secara keseluruhan, kedua algoritma ini memiliki keunggulan dan keterbatasan masing-masing dalam menangani dataset sosial-ekonomi. Naïve Bayes lebih unggul dalam hal efisiensi komputasi dan akurasi pada data yang terstruktur dengan asumsi independensi fitur, sementara KNN lebih fleksibel dalam menangani data yang lebih kompleks dan tidak terstruktur, meskipun membutuhkan lebih banyak waktu dan rentan terhadap outlier. Oleh karena itu, pemilihan algoritma yang tepat bergantung pada karakteristik data yang ada dan tujuan klasifikasi yang ingin dicapai.

3

METODOLOGI PENELITIAN

Deskripsi Lokasi

Desa Marbau Selatan terletak di Kecamatan Marbau, Kabupaten Labuhanbatu Utara, Provinsi Sumatera Utara. Secara geografis, desa ini berada pada koordinat 2°50' Lintang Utara dan 99°25' Bujur Timur, dengan ketinggian daratan mulai dari 0 hingga 700 meter di atas permukaan laut. Wilayah ini memiliki iklim tropis dengan dua musim utama: musim hujan dan musim kemarau. Curah hujan rata-rata bulanan di Kabupaten Labuhanbatu Utara adalah 255,42 mm, dengan hari hujan rata-rata sebanyak 12,33 hari per bulan. Musim hujan biasanya terjadi pada bulan Oktober dengan curah hujan tertinggi mencapai 443 mm, sedangkan musim kemarau terjadi pada bulan Maret dengan curah hujan terendah sekitar 82 mm.

Berdasarkan data dari Badan Pusat Statistik Kabupaten Labuhanbatu Utara, jumlah penduduk Kabupaten Labuhanbatu Utara tercatat sebesar 347.456 jiwa dengan laju pertumbuhan penduduk 2,98 persen. Jumlah rumah tangga sebanyak 80.520 dengan rata-rata

jumlah anggota rumah tangga adalah 4 orang. Komposisi penduduk masih didominasi oleh kelompok usia muda, yang menyebabkan rasio ketergantungan (dependency ratio) yang cukup tinggi, yaitu sebesar 61,61 persen. Jumlah penduduk laki-laki di Kabupaten Labuhanbatu masih lebih banyak dibandingkan perempuan, dengan nilai sex ratio sebesar 101, yang berarti untuk setiap 100 orang penduduk perempuan terdapat 101 orang penduduk laki-laki.

Desa Marbau Selatan, sebagai bagian dari Kecamatan Marbau, memiliki karakteristik demografis yang mencerminkan komposisi penduduk Kabupaten Labuhanbatu Utara secara umum. Masyarakat desa ini mayoritas beragama Islam, dengan sejarah keagamaan yang kaya, termasuk adanya masjid tua yang menjadi pusat kegiatan keagamaan dan sosial masyarakat. Pendidikan di desa ini masih dalam tahap pengembangan, dengan upaya peningkatan angka harapan lama sekolah yang tercatat pada tahun 2014 sebesar 11,80 tahun, sedikit meningkat dibandingkan tahun 2013 yang mencapai 11,30 tahun. Namun, angka ini masih di bawah rata-rata provinsi Sumatera Utara yang mencapai 12,61 tahun pada tahun 2014.

Dari segi ekonomi, Desa Marbau Selatan memiliki potensi sumber daya alam yang melimpah, terutama dalam sektor pertanian dan perkebunan. Masyarakat desa banyak yang bekerja sebagai petani dan pekebun, dengan komoditas utama seperti kelapa sawit dan karet. Namun, sektor pertanian di desa ini masih menghadapi tantangan dalam hal akses terhadap teknologi pertanian modern, pemasaran hasil pertanian, dan infrastruktur pendukung lainnya. Upaya pengembangan ekonomi desa melalui Badan Usaha Milik Desa (BUMDes) telah dilakukan untuk memperkuat perekonomian lokal dan meningkatkan kesejahteraan masyarakat.

Secara keseluruhan, Desa Marbau Selatan memiliki karakteristik demografis dan ekonomi yang mencerminkan dinamika masyarakat pedesaan di Kabupaten Labuhanbatu Utara. Meskipun terdapat tantangan dalam bidang pendidikan dan ekonomi, potensi sumber

daya alam yang melimpah memberikan peluang bagi pengembangan dan peningkatan kesejahteraan masyarakat desa.

Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini terdiri dari survei langsung dan pengolahan data sosial-ekonomi yang terkait dengan distribusi bantuan sosial di Desa Marbau Selatan. Pengumpulan data dilakukan untuk memperoleh informasi yang akurat dan terkini mengenai karakteristik sosial-ekonomi penduduk, yang meliputi variabel seperti nama, gender kepala keluarga, usia, penghasilan, jumlah tanggungan, dan kepemilikan asset/rumah. Survei langsung dilakukan untuk memastikan bahwa data yang diperoleh relevan dan representatif, mengingat tantangan yang sering terjadi pada pendataan berbasis arsip atau data historis yang tidak selalu mencerminkan kondisi terkini masyarakat.

Survei langsung dilakukan dengan menggunakan kuesioner terstruktur, yang disebarkan kepada penduduk Desa Marbau Selatan. Kuesioner ini dirancang untuk mengumpulkan data kuantitatif dan kualitatif mengenai kondisi sosial-ekonomi rumah tangga, termasuk informasi mengenai nama, gender kepala keluarga, usia, penghasilan, jumlah tanggungan, dan kepemilikan asset/rumah. Proses pengumpulan data dilakukan dengan cara wawancara tatap muka oleh enumerator terlatih yang berinteraksi langsung dengan responden untuk mengisi kuesioner tersebut. Pendekatan ini memungkinkan peneliti untuk memperoleh informasi yang lebih mendalam dan mengurangi potensi kesalahan atau ketidaktepatan dalam pengisian data yang mungkin terjadi pada metode pengisian mandiri.

Selain pengumpulan data melalui survei, langkah selanjutnya adalah pengolahan data sosial-ekonomi. Pengolahan data ini dilakukan untuk memastikan bahwa data yang dikumpulkan dapat digunakan dengan optimal dalam proses klasifikasi penerima

bantuan sosial. Data yang telah dikumpulkan melalui survei langsung akan dibersihkan dan diproses agar siap digunakan dalam analisis lebih lanjut. Proses ini mencakup pengisian data yang hilang, penyelarasan variabel, serta normalisasi data agar semua atribut dapat diolah dengan cara yang konsisten, terutama ketika data terdiri dari kombinasi variabel numerik dan kategori.

Tabel 1. Variabel yang digunakan

No	Nama Variabel	Tipe Data	Deskripsi
1	Nama	Teks	Identitas kepala keluarga (tidak digunakan dalam pemodelan).
2	Gender Kepala Keluarga	Kategori	Jenis kelamin kepala keluarga (Laki-Laki/Perempuan).
3	Usia	Numerik	Umur kepala keluarga dalam tahun, dihitung dari tahun kelahiran.
4	Penghasilan	Kategori	Penghasilan berdasarkan pekerjaan anggota keluarga.
5	Tanggungan	Numerik	Jumlah anggota keluarga yang menjadi tanggungan kepala keluarga.
6	Kepemilikan Rumah	Kategori	Status kepemilikan rumah (Milik Pribadi/Milik Orang Lain).
7	Status	Kategori	Status berhak menerima bantuan atau tidak (layak/tidak layak

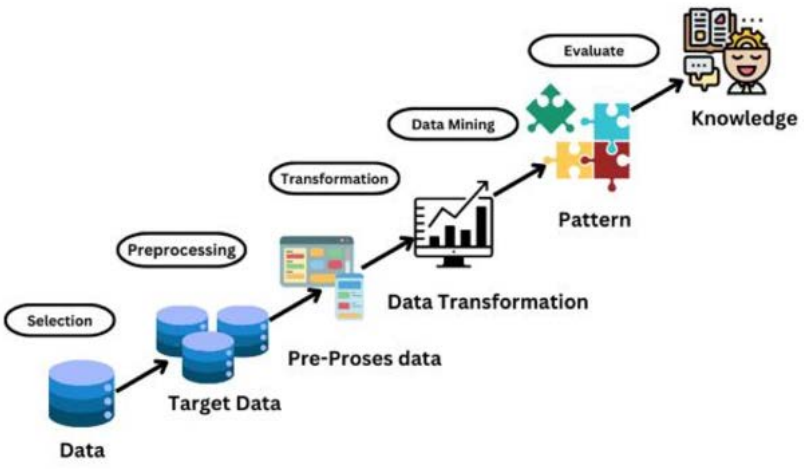
Tabel 1. menunjukkan variabel yang digunakan dalam penelitian. Atribut numerik seperti usia, penghasilan dan jumlah tanggungan yang dibagi menjadi beberapa kategori. Sedangkan atribut kategori seperti gender kepala keluarga, usia, penghasilan dan kepemilikan rumah digunakan untuk menggambarkan kondisi sosial-ekonomi secara kualitatif. Variabel-variabel ini telah diproses untuk memastikan data siap digunakan dalam algoritma klasifikasi, memastikan setiap variabel relevan terhadap tujuan penelitian.

Setelah data disiapkan, tahap selanjutnya adalah analisis data menggunakan algoritma machine learning, seperti Naïve Bayes dan K-Nearest Neighbor (KNN), untuk mengklasifikasikan rumah tangga mana yang layak menerima bantuan sosial berdasarkan karakteristik sosial-ekonomi yang telah dihimpun. Proses ini melibatkan

penggunaan teknik statistik dan algoritma berbasis data untuk memberikan prediksi yang lebih akurat dan tepat sasaran mengenai penerima bantuan sosial. Dengan demikian, metode pengumpulan data melalui survei langsung dan pengolahan data sosial-ekonomi ini diharapkan dapat membantu meningkatkan akurasi distribusi bantuan sosial, yang menjadi inti dari penelitian ini.

Proses Knowledge Discovery in Database (KDD)

Proses Knowledge Discovery in Database (KDD) merupakan langkah yang sangat penting dalam menyiapkan dataset untuk analisis, terutama dalam konteks penelitian yang menggunakan algoritma pembelajaran mesin untuk mengklasifikasikan penerima bantuan sosial.



Gambar 2.2 Tahapan Knowledge Discovery in Database (KDD)

Tahap pertama dalam KDD adalah pemilihan data (selection), di mana data yang relevan untuk penelitian dipilih dari berbagai sumber yang ada. Dalam hal ini, data sosial-ekonomi yang diperoleh melalui survei langsung di Desa Marbau Selatan akan dipilih untuk memastikan bahwa hanya data yang relevan dengan klasifikasi

penerima bantuan sosial yang digunakan dalam analisis. Data ini bisa mencakup informasi seperti pendapatan rumah tangga, jumlah tanggungan, status pekerjaan, dan tingkat pendidikan.

Setelah data yang relevan dipilih, tahap selanjutnya adalah pembersihan data (preprocessing). Data yang diperoleh melalui survei sering kali memiliki masalah seperti data yang hilang, data duplikat, atau data yang tidak konsisten. Oleh karena itu, tahap pembersihan sangat penting untuk memastikan bahwa data yang digunakan untuk analisis adalah akurat dan terpercaya. Misalnya, jika ada entri yang tidak lengkap atau tidak valid, langkah ini akan mencakup pengisian data yang hilang dengan nilai rata-rata atau penyesuaian berdasarkan aturan tertentu, atau menghapus entri yang tidak relevan.

Tabel 2. Dataset yang digunakan

Nama	Gender Kepala Keluarga	Usia	Penghasilan	Tanggu-ngan	Kepemilikan Rumah	Status
KK1	Laki-Laki	33	Diatas UMK	1	Milik Orang Lain	Tidak Layak
KK2	Laki-Laki	28	Diatas UMK	1	Milik Pribadi	Tidak Layak
KK3	Perempuan	78	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK4	Perempuan	55	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK5	Perempuan	47	Diatas UMK	2	Milik Orang Lain	Layak
KK6	Laki-Laki	39	Diatas UMK	0	Milik Orang Lain	Tidak Layak
KK7	Laki-Laki	58	Dibawah UMK	2	Milik Orang Lain	Tidak Layak
KK8	Laki-Laki	36	Diatas UMK	2	Milik Orang Lain	Tidak Layak
KK9	Laki-Laki	51	Diatas UMK	4	Milik Orang Lain	Layak
KK10	Perempuan	69	Dibawah UMK	0	Milik Pribadi	Layak

KK11	Laki-Laki	32	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK12	Laki-Laki	75	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK13	Laki-Laki	40	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK14	Laki-Laki	82	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK15	Laki-Laki	57	Diatas UMK	4	Milik Orang Lain	Tidak Layak
KK16	Laki-Laki	48	Diatas UMK	4	Milik Pribadi	Tidak Layak
KK17	Laki-Laki	57	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK18	Laki-Laki	41	Dibawah UMK	3	Milik Orang Lain	Layak
KK19	Laki-Laki	54	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK20	Perempuan	56	Diatas UMK	0	Milik Pribadi	Tidak Layak
KK21	Laki-Laki	62	Dibawah UMK	1	Milik Pribadi	Tidak Layak
KK22	Perempuan	63	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK23	Laki-Laki	71	Dibawah UMK	1	Milik Pribadi	Layak
KK24	Laki-Laki	71	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK25	Laki-Laki	41	Dibawah UMK	3	Milik Pribadi	Tidak Layak
KK26	Laki-Laki	54	Dibawah UMK	3	Milik Pribadi	Tidak Layak
KK27	Laki-Laki	50	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK28	Laki-Laki	68	Dibawah UMK	4	Milik Orang Lain	Tidak Layak
KK29	Laki-Laki	73	Dibawah UMK	2	Milik Pribadi	Layak
KK30	Perempuan	54	Dibawah UMK	2	Milik Pribadi	Tidak Layak

KK31	Laki-Laki	42	Diatas UMK	1	Milik Pribadi	Tidak Layak
KK32	Perempuan	76	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK33	Laki-Laki	32	Dibawah UMK	3	Milik Pribadi	Layak
KK34	Laki-Laki	36	Diatas UMK	4	Milik Pribadi	Tidak Layak
KK35	Perempuan	40	Diatas UMK	0	Milik Pribadi	Tidak Layak
KK36	Perempuan	68	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK37	Laki-Laki	67	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK38	Laki-Laki	33	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK39	Laki-Laki	56	Dibawah UMK	4	Milik Pribadi	Tidak Layak
KK40	Laki-Laki	39	Diatas UMK	4	Milik Pribadi	Tidak Layak
KK41	Perempuan	76	Dibawah UMK	1	Milik Orang Lain	Layak
KK42	Laki-Laki	34	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK43	Laki-Laki	63	Dibawah UMK	1	Milik Pribadi	Tidak Layak
KK44	Laki-Laki	54	Diatas UMK	5	Milik Pribadi	Tidak Layak
KK45	Laki-Laki	53	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK46	Laki-Laki	45	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK47	Laki-Laki	40	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK48	Laki-Laki	63	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK49	Laki-Laki	44	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK50	Laki-Laki	38	Diatas UMK	2	Milik Pribadi	Tidak Layak

KK51	Laki-Laki	65	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK52	Laki-Laki	55	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK53	Laki-Laki	38	Dibawah UMK	3	Milik Orang Lain	Layak
KK54	Perempuan	65	Dibawah UMK	0	Milik Pribadi	Layak
KK55	Laki-Laki	57	Dibawah UMK	1	Milik Pribadi	Tidak Layak
KK56	Perempuan	61	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK57	Laki-Laki	36	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK58	Laki-Laki	44	Dibawah UMK	4	Milik Orang Lain	Tidak Layak
KK59	Laki-Laki	58	Diatas UMK	1	Milik Pribadi	Tidak Layak
KK60	Laki-Laki	41	Dibawah UMK	5	Milik Pribadi	Tidak Layak
KK61	Laki-Laki	60	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK62	Laki-Laki	45	Dibawah UMK	4	Milik Orang Lain	Tidak Layak
KK63	Laki-Laki	70	Dibawah UMK	1	Milik Pribadi	Tidak Layak
KK64	Laki-Laki	38	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK65	Laki-Laki	59	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK66	Perempuan	61	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK67	Perempuan	62	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK68	Laki-Laki	48	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK69	Laki-Laki	42	Dibawah UMK	5	Milik Pribadi	Layak
KK70	Perempuan	53	Dibawah UMK	3	Milik Pribadi	Layak

KK71	Laki-Laki	49	Diatas UMK	5	Milik Pribadi	Tidak Layak
KK72	Perempuan	48	Dibawah UMK	3	Milik Orang Lain	Layak
KK73	Laki-Laki	29	Diatas UMK	2	Milik Orang Lain	Tidak Layak
KK74	Laki-Laki	68	Dibawah UMK	1	Milik Pribadi	Tidak Layak
KK75	Laki-Laki	51	Dibawah UMK	3	Milik Pribadi	Tidak Layak
KK76	Laki-Laki	42	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK77	Perempuan	72	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK78	Laki-Laki	56	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK79	Laki-Laki	51	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK80	Laki-Laki	58	Dibawah UMK	3	Milik Pribadi	Tidak Layak
KK81	Laki-Laki	44	Dibawah UMK	3	Milik Pribadi	Tidak Layak
KK82	Laki-Laki	31	Dibawah UMK	3	Milik Orang Lain	Tidak Layak
KK83	Laki-Laki	60	Diatas UMK	2	Milik Pribadi	Tidak Layak
KK84	Laki-Laki	58	Diatas UMK	1	Milik Pribadi	Tidak Layak
KK85	Perempuan	67	Dibawah UMK	0	Milik Pribadi	Tidak Layak
KK86	Laki-Laki	42	Diatas UMK	4	Milik Orang Lain	Layak
KK87	Laki-Laki	58	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK88	Laki-Laki	38	Diatas UMK	4	Milik Pribadi	Tidak Layak
KK89	Laki-Laki	48	Diatas UMK	4	Milik Pribadi	Tidak Layak
KK90	Laki-Laki	57	Dibawah UMK	2	Milik Pribadi	Tidak Layak

KK91	Laki-Laki	62	Dibawah UMK	2	Milik Pribadi	Tidak Layak
KK92	Perempuan	34	Dibawah UMK	2	Milik Orang Lain	Tidak Layak
KK93	Perempuan	61	Diatas UMK	1	Milik Pribadi	Tidak Layak
KK94	Laki-Laki	48	Diatas UMK	3	Milik Pribadi	Tidak Layak
KK95	Perempuan	74	Dibawah UMK	0	Milik Pribadi	Layak
KK96	Perempuan	66	Diatas UMK	0	Milik Pribadi	Tidak Layak
KK97	Perempuan	73	Diatas UMK	0	Milik Pribadi	Tidak Layak
KK98	Perempuan	80	Dibawah UMK	0	Milik Pribadi	Layak
KK99	Perempuan	81	Dibawah UMK	0	Milik Pribadi	Layak
KK100	Perempuan	74	Dibawah UMK	0	Milik Pribadi	Layak

Tabel diatas memuat 100 dataset penelitian yang telah dibersihkan. Setelah data dibersihkan, tahap berikutnya adalah transformasi data. Proses ini mencakup normalisasi dan penskalaan data agar variabel-variabel dalam dataset dapat dibandingkan dengan cara yang adil dan konsisten. Sebagai contoh, dalam dataset sosial-ekonomi, data seperti penghasilan, usia dan jumlah tanggungan dapat berada dalam skala yang sangat berbeda, sehingga normalisasi atau standarisasi diperlukan untuk memastikan bahwa setiap variabel memiliki bobot yang setara dalam analisis.

Langkah berikutnya adalah transformasi variabel. Transformasi ini diperlukan untuk memastikan bahwa variabel dalam dataset dapat dikelola dan dianalisis secara konsisten. Dalam konteks dataset sosial-ekonomi, transformasi ini dapat mencakup normalisasi atau standarisasi variabel. Variabel seperti penghasilan dan jumlah tanggungan memiliki rentang nilai yang sangat berbeda, di mana pendapatan mungkin memiliki nilai yang jauh lebih tinggi

daripada jumlah tanggungan. Untuk memastikan bahwa algoritma pembelajaran mesin dapat memproses kedua variabel ini dengan cara yang adil, normalisasi diperlukan untuk merubah nilai-nilai tersebut ke dalam skala yang seragam, sehingga tidak ada satu variabel yang mendominasi proses analisis.

Langkah terakhir dalam transformasi data adalah pembentukan fitur baru (feature engineering), yang melibatkan penciptaan variabel baru dari data yang sudah ada. Misalnya, data mengenai status pekerjaan dapat diubah menjadi kategori yang lebih terperinci, seperti penghasilan,dikategorikan menjadi diatas UMK dan dibawah UMK. Dengan membuat variabel-variabel baru yang lebih relevan dan informatif, kita dapat meningkatkan kemampuan model klasifikasi dalam mengenali pola-pola penting dalam data, yang pada akhirnya meningkatkan akurasi model.

Nama	Gender Kepala Keluarga	Usia	Penghasilan	Tanggungan	Kepemilikan Rumah	Status
KK1	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK2	Laki-Laki	Muda	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK3	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK4	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK5	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Layak
KK6	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK7	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Tidak Layak
KK8	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK9	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Layak

KK10	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK11	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK12	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK13	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK14	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK15	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Tidak Layak
KK16	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK17	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK18	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK19	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK20	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK21	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK22	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK23	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK24	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK25	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK26	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK27	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK28	Laki-Laki	Lansia	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK29	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak

KK30	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK31	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK32	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK33	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK34	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK35	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK36	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK37	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK38	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK39	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Tidak Layak
KK40	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK41	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK42	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK43	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK44	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK45	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK46	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK47	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK48	Laki-Laki	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK49	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak

KK50	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK51	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK52	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK53	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK54	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK55	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK56	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK57	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK58	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK59	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK60	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Layak
KK61	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK62	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK63	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK64	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK65	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK66	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK67	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK68	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK69	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Layak

KK70	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK71	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK72	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK73	Laki-Laki	Muda	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK74	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK75	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK76	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK77	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK78	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK79	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK80	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK81	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK82	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Tidak Layak
KK83	Laki-Laki	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK84	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK85	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK86	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Layak
KK87	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK88	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK89	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak

KK90	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK91	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK92	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK93	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK94	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK95	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK96	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK97	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK98	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK99	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK100	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak

Data ini mencakup variabel-variabel utama yang relevan untuk klasifikasi penerima bantuan sosial, Seperti variabel gender kepala keluarga. Variabel usia didasarkan pada klasifikasi WHO yang juga digunakan oleh BKKBN dimana dikatogorikan menjadi usia muda (20-29 tahun), dewasa (30-59 tahun) dan lansia (60 tahun keatas). Variabel penghasilan disesuaikan dengan Upah Minumum Kabupaten (UMK) Labuhanbatu Utara pada tahun 2024 yaitu sebesar Rp3.124.527 lalu dikategorikan menjadi penghasilan diatas UMK dan penghasilan dibawah UMK. Variabel tanggungan yang didasarkan pada klasifikasi Badan Pusat Statistik (BPS) dimana terbagi menjaddi 3 kategori yaitu tanggungan keluarga kecil (1-3 orang), tanggungan keluarga sedang (4-6 orang) dan tanggungan keluarga besar (>6 orang). Variabel kepemilikan rumah yang dibagi menjadi rumah milik pribadi dan rumah milik orang lain. Data ini mencerminkan

kompleksitas sosial-ekonomi yang menjadi dasar penting untuk membangun model klasifikasi.

Tabel 3.1 Pembagian Kategori Transformasi Data

Variabel	Sebelum Transformasi	Setelah Transformasi
Gender KK	Laki-Laki	1
	Perempuan	2
Usia	Muda	1
	Dewasa	2
	Lansia	3
Penghasilan	Diatas UMK	1
	Dibawah UMK	2
Tanggungan	Kecil	1
	Sedang	2
	Besar	3
Kepemilikan Rumah	Milik Pribadi	1
	Milik Orang Lain	2

Nama	Gender KK	Usia	Penghasilan	Tanggungan	Kepemilikan Rumah	Status
KK1	1	2	1	1	2	Tidak Layak
KK2	1	1	1	1	1	Tidak Layak
KK3	2	3	2	1	1	Tidak Layak
KK4	2	2	1	1	1	Tidak Layak
KK5	2	2	1	1	2	Layak
KK6	1	2	1	1	2	Tidak Layak
KK7	1	2	2	1	2	Tidak Layak

KK8	1	2	1	1	2	Tidak Layak
KK9	1	2	1	2	2	Layak
KK10	2	3	2	1	1	Layak
KK11	1	2	1	1	1	Tidak Layak
KK12	1	3	2	1	1	Tidak Layak
KK13	1	2	1	1	1	Tidak Layak
KK14	1	3	2	1	1	Tidak Layak
KK15	1	2	1	2	2	Tidak Layak
KK16	1	2	1	2	1	Tidak Layak
KK17	1	2	1	1	1	Tidak Layak
KK18	1	2	2	1	2	Layak
KK19	1	2	2	1	1	Tidak Layak
KK20	2	2	1	1	1	Tidak Layak
KK21	1	3	2	1	1	Tidak Layak
KK22	2	3	2	1	1	Tidak Layak
KK23	1	3	2	1	1	Layak
KK24	1	3	2	1	1	Tidak Layak
KK25	1	2	2	1	1	Tidak Layak
KK26	1	2	2	1	1	Tidak Layak
KK27	1	2	1	1	1	Tidak Layak
KK28	1	3	2	2	2	Tidak Layak
KK29	1	3	2	1	1	Layak

KK30	2	2	2	1	1	Tidak Layak
KK31	1	2	1	1	1	Tidak Layak
KK32	2	3	2	1	1	Tidak Layak
KK33	1	2	2	1	1	Layak
KK34	1	2	1	2	1	Tidak Layak
KK35	2	2	1	1	1	Tidak Layak
KK36	2	3	2	1	1	Tidak Layak
KK37	1	3	2	1	1	Tidak Layak
KK38	1	2	1	1	1	Tidak Layak
KK39	1	2	2	2	1	Tidak Layak
KK40	1	2	1	2	1	Tidak Layak
KK41	2	3	2	1	2	Layak
KK42	1	2	1	1	1	Tidak Layak
KK43	1	3	2	1	1	Tidak Layak
KK44	1	2	1	2	1	Tidak Layak
KK45	2	3	2	1	1	Layak
KK46	2	3	2	1	1	Layak
KK47	2	3	2	1	1	Layak
KK48	1	3	1	1	1	Tidak Layak
KK49	1	2	1	1	1	Tidak Layak
KK50	1	2	1	1	1	Tidak Layak
KK51	1	3	2	1	1	Tidak Layak

KK52	1	2	1	1	1	Tidak Layak
KK53	1	2	2	1	2	Layak
KK54	2	3	2	1	1	Layak
KK55	1	2	2	1	1	Tidak Layak
KK56	2	3	2	1	1	Tidak Layak
KK57	1	2	1	1	1	Tidak Layak
KK58	1	2	2	2	2	Tidak Layak
KK59	1	2	1	1	1	Tidak Layak
KK60	1	2	2	2	1	Layak
KK61	1	3	2	1	1	Tidak Layak
KK62	1	2	2	2	2	Tidak Layak
KK63	1	3	2	1	1	Tidak Layak
KK64	1	2	1	1	1	Tidak Layak
KK65	1	3	2	1	1	Tidak Layak
KK66	2	3	2	1	1	Tidak Layak
KK67	2	3	2	1	1	Tidak Layak
KK68	1	2	1	1	1	Tidak Layak
KK69	1	2	2	2	1	Layak
KK70	2	2	2	1	1	Layak
KK71	1	2	1	2	1	Tidak Layak
KK72	2	2	2	1	2	Layak
KK73	1	1	1	1	2	Tidak Layak

KK74	1	3	2	1	1	Tidak Layak
KK75	1	2	2	1	1	Tidak Layak
KK76	1	2	1	1	1	Tidak Layak
KK77	2	3	2	1	1	Tidak Layak
KK78	1	2	1	1	1	Tidak Layak
KK79	1	2	1	1	1	Tidak Layak
KK80	1	2	2	1	1	Tidak Layak
KK81	1	2	2	1	1	Tidak Layak
KK82	1	2	2	1	2	Tidak Layak
KK83	1	3	1	1	1	Tidak Layak
KK84	1	2	1	1	1	Tidak Layak
KK85	2	3	2	1	1	Tidak Layak
KK86	1	2	1	2	2	Layak
KK87	1	2	2	1	1	Tidak Layak
KK88	1	2	1	2	1	Tidak Layak
KK89	1	2	1	2	1	Tidak Layak
KK90	1	2	2	1	1	Tidak Layak
KK91	1	3	2	1	1	Tidak Layak
KK92	2	2	2	1	2	Layak
KK93	2	3	1	1	1	Tidak Layak
KK94	1	2	1	1	1	Tidak Layak

KK95	2	3	2	1	1	Layak
KK96	2	3	1	1	1	Tidak Layak
KK97	2	3	1	1	1	Tidak Layak
KK98	1	2	2	1	1	Tidak Layak
KK99	1	2	1	1	1	Tidak Layak
KK100	1	2	1	1	1	Tidak Layak

Secara keseluruhan, proses KDD sangat penting dalam memastikan bahwa dataset yang digunakan untuk analisis sudah siap dan tepat untuk diterapkan dalam klasifikasi penerima bantuan sosial. Proses ini juga membantu dalam mengidentifikasi pola-pola yang dapat digunakan untuk meningkatkan keakuratan dan efisiensi distribusi bantuan social.

Akhirnya, setelah semua tahap KDD selesai, data siap untuk digunakan dalam membangun model analisis menggunakan algoritma pembelajaran mesin, seperti Naïve Bayes atau K-Nearest Neighbor (KNN). Model yang dihasilkan akan digunakan untuk mengklasifikasikan penerima bantuan sosial berdasarkan karakteristik sosial-ekonomi yang telah diproses.

Pembagian Dataset, Data Training dan Data Testing

Dalam pembelajaran mesin, pembagian dataset menjadi data training dan data testing adalah langkah penting untuk validasi model. Tujuan utama dari pembagian ini adalah untuk mengukur kinerja model yang dibangun dalam situasi dunia nyata, dengan menggunakan data yang sebelumnya tidak digunakan selama proses pelatihan model. Ini memastikan bahwa model tidak hanya bekerja baik pada data yang telah dilatih, tetapi juga dapat melakukan generalization dengan

baik pada data baru yang belum pernah dilihat sebelumnya. Dengan demikian, pembagian dataset menjadi dua bagian ini membantu menghindari masalah overfitting, di mana model terlalu cocok dengan data pelatihan tetapi gagal untuk memberikan prediksi yang akurat pada data yang tidak dikenal.

Proses pembagian dataset umumnya dilakukan dengan cara acak (random split), di mana data dibagi secara acak menjadi dua bagian. Sebagai aturan umum, 70%-80% dari data digunakan untuk training, dan sisanya 20%-30% digunakan untuk testing. Proporsi ini sering digunakan karena memberikan keseimbangan yang baik antara jumlah data yang digunakan untuk melatih model dan data yang cukup untuk melakukan pengujian. Dalam konteks penelitian tentang klasifikasi penerima bantuan sosial di Desa Marbau Selatan, pembagian data ini memungkinkan model untuk mempelajari pola-pola dalam data sosial-ekonomi, seperti pendapatan rumah tangga, jumlah tanggungan, dan status pekerjaan, sehingga dapat memprediksi dengan akurat siapa yang layak menerima bantuan sosial.

Data training digunakan untuk membangun dan melatih model, yaitu untuk menemukan hubungan antara variabel input (seperti penghasilan, jumlah tanggungan, dan usia) dan hasil yang diinginkan (dalam hal ini, apakah individu atau keluarga tersebut layak menerima bantuan sosial). Proses ini melibatkan penggunaan algoritma seperti Naïve Bayes atau K-Nearest Neighbor (KNN) untuk mempelajari pola-pola dalam data training yang dapat digunakan untuk mengklasifikasikan data baru. Setelah model dilatih, data testing digunakan untuk mengukur kinerja model dengan cara menguji seberapa baik model dapat memprediksi hasil yang benar pada data yang belum pernah dilihat sebelumnya. Ini memberikan gambaran tentang kemampuan model untuk menggeneralisasi dan mengaplikasikan pengetahuan yang telah diperoleh dari data training ke data baru.

Dalam proses validasi model, beberapa metrik evaluasi digunakan untuk menilai kinerja model, seperti akurasi, precision, recall, dan F1-score. Akurasi mengukur sejauh mana model dapat memprediksi dengan benar, sedangkan precision dan recall memberikan gambaran yang lebih terperinci tentang bagaimana model menangani positif palsu dan negatif palsu dalam klasifikasi. F1-score menggabungkan precision dan recall untuk memberikan gambaran yang lebih seimbang tentang kinerja model, terutama ketika data tidak seimbang.

Dengan membagi dataset menjadi data training dan data testing, proses validasi ini membantu memastikan bahwa model yang dibangun tidak hanya mampu mengenali pola dalam data yang digunakan untuk melatihnya, tetapi juga memiliki kemampuan untuk membuat prediksi yang akurasi dan reliable pada data yang baru, seperti halnya penerima bantuan sosial yang sebenarnya.

Evaluasi Model

Dalam penelitian ini, untuk mengevaluasi kinerja model klasifikasi yang digunakan dalam klasifikasi penerima bantuan sosial, digunakan beberapa metrik evaluasi yang umum digunakan dalam pembelajaran mesin. Metrik evaluasi yang digunakan adalah akurasi, precision, recall, dan F1-score. Metrik-metrik ini memberikan gambaran yang lebih mendalam mengenai kinerja model dalam memprediksi hasil yang benar, serta mengidentifikasi kemampuan model untuk menangani positif palsu dan negatif palsu, yang sering terjadi dalam klasifikasi penerima bantuan sosial.

Akurasi adalah metrik evaluasi yang paling umum digunakan untuk mengukur seberapa banyak prediksi yang benar dibandingkan dengan jumlah total data. Akurasi dihitung dengan rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Akurasi memberikan gambaran umum tentang kemampuan model dalam memprediksi klasifikasi yang benar, namun tidak cukup mendalam untuk memberikan analisis yang lebih terperinci, terutama dalam kasus data yang tidak seimbang, seperti yang sering ditemukan dalam klasifikasi penerima bantuan sosial di mana jumlah penerima yang layak jauh lebih sedikit daripada yang tidak layak.

Untuk menangani data yang tidak seimbang, dua metrik lain yang digunakan adalah precision dan recall. Precision mengukur seberapa banyak prediksi positif yang benar dibandingkan dengan jumlah seluruh prediksi positif yang dilakukan oleh model, dihitung dengan rumus:

$$Precision = \frac{TP}{TP + FP}$$

Precision penting untuk mengidentifikasi kemampuan model dalam menghindari positif palsu, yaitu ketika model salah mengklasifikasikan penerima bantuan yang tidak layak sebagai layak. Sebaliknya, recall mengukur seberapa banyak positif sebenarnya yang dapat ditemukan oleh model, dihitung dengan rumus:

$$Recall = \frac{TP}{TP + FN}$$

Recall sangat penting untuk mendeteksi penerima bantuan yang layak, meskipun mereka sedikit jumlahnya dalam data. Dalam konteks bantuan sosial, penting agar model tidak melewatkan penerima yang benar-benar membutuhkan bantuan, yang merupakan positif palsu dalam klasifikasi.

Namun, precision dan recall sering kali berada dalam trade-off, di mana meningkatkan satu dapat mengurangi yang lain. Oleh karena itu, F1-score digunakan sebagai metrik gabungan yang mengharmoniskan precision dan recall. F1-score dihitung dengan rumus:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score memberikan gambaran yang lebih seimbang mengenai kinerja model dalam mengklasifikasikan penerima bantuan sosial, dengan menilai seberapa baik model dalam menemukan penerima yang layak, sembari menjaga agar tidak mengklasifikasikan penerima yang tidak layak.

Dengan menggunakan metrik-metrik ini, model klasifikasi yang diterapkan dalam penelitian ini dapat dievaluasi secara komprehensif, memberikan gambaran yang lebih mendalam tentang kinerja model, dan membantu dalam peningkatan akurasi serta pengambilan keputusan yang lebih tepat mengenai distribusi bantuan sosial.

4

ANALISIS ALGORITMA NAÏVE BAYES & K-NEAREST NEIGHBOR

Proses Penerapan Algoritma Naïve Bayes

Naïve Bayes adalah algoritma pembelajaran mesin berbasis probabilitas yang digunakan untuk klasifikasi data. Algoritma ini mengasumsikan bahwa setiap fitur dalam data adalah independen, yang memungkinkan perhitungan probabilitas untuk setiap kelas berdasarkan fitur yang ada. Dalam konteks penelitian ini, algoritma Naïve Bayes diterapkan untuk klasifikasi penerima bantuan sosial di Desa Marbau Selatan, berdasarkan data sosial-ekonomi seperti nama, gender kepala keluarga, usia, penghasilan, jumlah tanggungan, dan kepemilikan asset/rumah.

Algoritma Naïve Bayes kemudian menghitung probabilitas untuk setiap kelas dan memilih kelas dengan probabilitas tertinggi sebagai hasil klasifikasi. Misalnya, jika kita memiliki data untuk sebuah

keluarga dengan pendapatan tertentu, jumlah tanggungan, dan status pekerjaan tertentu, model akan menghitung probabilitas kelas "layak" dan "tidak layak" menerima bantuan sosial, dan memilih kelas dengan nilai probabilitas tertinggi.

Penerapan Naïve Bayes pada dataset sosial-ekonomi ini memungkinkan klasifikasi yang cepat dan efisien. Algoritma ini sangat cocok untuk dataset yang memiliki variabel kategori, karena setiap fitur diperlakukan secara independen dalam perhitungannya. Namun, meskipun asumsi independensi ini jarang berlaku secara sempurna dalam dunia nyata, Naïve Bayes tetap memberikan hasil yang baik dalam banyak aplikasi, termasuk dalam klasifikasi penerima bantuan sosial.

Tabel. Data Latih algoritma Naïve Bayes

Nama	Gender Kepala Keluarga	Usia	Penghasilan	Tanggungan	Kepemilikan Rumah	Status
KK1	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK2	Laki-Laki	Muda	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK3	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK4	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK5	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Layak
KK6	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK7	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Tidak Layak
KK8	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK9	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Layak
KK10	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak

KK11	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK12	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK13	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK14	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK15	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Tidak Layak
KK16	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK17	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK18	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK19	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK20	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK21	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK22	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK23	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK24	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK25	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK26	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK27	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK28	Laki-Laki	Lansia	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK29	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK30	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak

KK31	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK32	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK33	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK34	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK35	Perempuan	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK36	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK37	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK38	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK39	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Tidak Layak
KK40	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK41	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK42	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK43	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK44	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK45	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK46	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK47	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK48	Laki-Laki	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK49	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK50	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak

KK51	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK52	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK53	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK54	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK55	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK56	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK57	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK58	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK59	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK60	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Layak
KK61	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK62	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Orang Lain	Tidak Layak
KK63	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK64	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK65	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK66	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK67	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK68	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK69	Laki-Laki	Dewasa	Dibawah UMK	Sedang	Milik Pribadi	Layak
KK70	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Layak

Data latih yang digunakan yaitu sebanyak 70 data keluarga yang akan digunakan untuk membantu proses klasifikasi ataupun perhitungan pada metode *Naïve Bayes*. Tetapi tidak hanya itu saja, data diatas akan dibagi dan dipecah menjadi beberapa table. Table yang dibagi berdasarkan variabel yang digunakan, jadi setiap variabel akan dipecah menjadi 1 tabel.

Tabel 3.7 Variabel Gender Kepala Keluarga

Variabel	Partisi	P (Layak)	P (Tidak Layak)
Gender Kepala Keluarga	Laki-Laki	8/16 = 0,50	43/54 = 0,80
	Perempuan	8/16 = 0,50	11/54 = 0,20
	Total	100%	100%

Tabel 3.8 Variabel Usia

Variabel	Partisi	P (Layak)	P (Tidak Layak)
Usia	Muda	0/16 = 0,00	1/54 = 0,02
	Dewasa	8/16 = 0,50	34/54 = 0,63
	Lansia	8/16 = 0,50	19/54 = 0,35
	Total	100%	100%

Tabel 3.9 Variabel Penghasilan

Variabel	Partisi	P (Layak)	P (Tidak Layak)
Penghasilan	Diatas UMK	2/16 = 0,13	27/54 = 0,50
	Dibawah UMK	14/16 = 0,88	27/54 = 0,50
	Total	100%	100%

Tabel 3.10 Variabel Tanggungan

Variabel	Partisi	P (Layak)	P (Tidak Layak)
Tanggungan	Kecil	13/16 = 0,81	45/54 = 0,83
	Sedang	3/16 = 0,19	9/54 = 0,17
	Besar	0/16 = 0,00	0/54 = 0,00
	Total	100%	100%

Tabel 3.11 Kepemilikan Rumah

Variabel	Partisi	P (Layak)	P (Tidak Layak)
Kepemilikan Rumah	Milik Pribadi	11/16 = 0,69	46/54 = 0,85
	Milik Orang Lain	5/16 = 0,31	8/54 = 0,15
	Total	100%	100%

Tabel 3.12 Variabel Status

Variabel	
Status	Layak
	Tidak Layak

Tabel. Data Uji Algoritma Naïve Bayes

Nama	Gender Kepala Keluarga	Usia	Penghasilan	Tanggungan	Kepemilikan Rumah	Status
KK71	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK72	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK73	Laki-Laki	Muda	Diatas UMK	Kecil	Milik Orang Lain	Tidak Layak
KK74	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK75	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK76	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK77	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK78	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK79	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK80	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK81	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak

KK82	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Tidak Layak
KK83	Laki-Laki	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK84	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK85	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK86	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Orang Lain	Layak
KK87	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK88	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK89	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak
KK90	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK91	Laki-Laki	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK92	Perempuan	Dewasa	Dibawah UMK	Kecil	Milik Orang Lain	Layak
KK93	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK94	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK95	Perempuan	Lansia	Dibawah UMK	Kecil	Milik Pribadi	Layak
KK96	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK97	Perempuan	Lansia	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK98	Laki-Laki	Dewasa	Dibawah UMK	Kecil	Milik Pribadi	Tidak Layak
KK99	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak
KK100	Laki-Laki	Dewasa	Diatas UMK	Kecil	Milik Pribadi	Tidak Layak

Setelah data selesai dipisahkan berdasarkan atributnya, maka data bisa mulai diolah menggunakan rumus *Naïve Bayes*.

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- $P(H|E)$: Probabilitas hipotesis H berdasarkan bukti H.
- $P(E|H)$: Probabilitas bukti E jika hipotesis H benar.
- $P(H)$: Probabilitas awal dari H.
- $P(E)$: Probabilitas keseluruhan dari bukti E.

Untuk pengolahan data pertama penulis menggunakan data uji dari KK71. Berikut adalah proses pengolahan data secara manual:

Nama	Gender Kepala Keluarga	Usia	Penghasilan	Tanggungan	Kepemilikan Rumah	Status
KK71	Laki-Laki	Dewasa	Diatas UMK	Sedang	Milik Pribadi	Tidak Layak

P(Kategori) = P (Jenis Kelamin|Laki-Laki) ×
P (Usia|Dewasa)
× P (Penghasilan|Diatas UMK) ×
P (Tanggungan|Sedang) ×
P (Rumah|Milik Pribadi × P(Status)

P(Layak) = P (Laki-Laki|Layak) × P (Dewasa|Layak)
× P (Diatas UMK|Layak) × P (Sedang|Layak)
× P (Pribadi|Layak × P(Status|Layak)
= 0,50 × 0,50 × 0,13 × 0,19 × 0,69 × 0,23
= 0,0009 (Layak)

P(Tidak Layak) = P (Laki-Laki|Layak) × P (Dewasa|Layak)
× P (Diatas UMK|Layak) × P (Sedang|Layak)
× P (Pribadi|Layak × P(Status|Layak)
= 0,80 × 0,63 × 0,50 × 0,17 × 0,85 × 0,77
= 0,0275 (Tidak Layak)

Hasil dari pengolahan data KK71 yaitu nilai tidak layak lebih tinggi dari nilai layak. Maka dapat disimpulkan untuk data KK71 tidak layak menerima bantuan sosial. Untuk perhitungan data selanjutnya dilakukan dengan cara yang sama. Proses selanjutnya penulis melakukan perhitungan manual ini dengan bantuan Microsoft Excel sehingga didapatkan hasil klasifikasi sebagai berikut.

Tabel 3.2 Hasil klasifikasi Naïve Bayes

Nama	Status		**Status Predikasi**	**Layak**	**Tidak Layak**
KK71	Tidak Layak		Tidak Layak	0.0009	0.0275
KK72	Layak		Layak	0.0127	0.0061
KK73	Tidak Layak		Tidak Layak	0.0000	0.0007
KK74	Tidak Layak		Tidak Layak	0.0279	0.0767
KK75	Tidak Layak		Tidak Layak	0.0279	0.1373
KK76	Tidak Layak		Tidak Layak	0.0040	0.1373
KK77	Tidak Layak		Layak	0.0279	0.0196
KK78	Tidak Layak		Tidak Layak	0.0040	0.1373
KK79	Tidak Layak		Tidak Layak	0.0040	0.1373
KK80	Tidak Layak		Tidak Layak	0.0279	0.1373
KK81	Tidak Layak	Prediksi	Tidak Layak	0.0279	0.1373
KK82	Tidak Layak		Tidak Layak	0.0127	0.0239
KK83	Tidak Layak		Tidak Layak	0.0040	0.0767
KK84	Tidak Layak		Tidak Layak	0.0040	0.1373
KK85	Tidak Layak		Layak	0.0279	0.0196
KK86	Layak		Tidak Layak	0.0004	0.0048
KK87	Tidak Layak		Tidak Layak	0.0279	0.1373
KK88	Tidak Layak		Tidak Layak	0.0009	0.0275
KK89	Tidak Layak		Tidak Layak	0.0009	0.0275
KK90	Tidak Layak		Tidak Layak	0.0279	0.1373
KK91	Tidak Layak		Tidak Layak	0.0279	0.0767
KK92	Layak		Layak	0.0127	0.0061
KK93	Tidak Layak		Tidak Layak	0.0040	0.0196
KK94	Tidak Layak		Tidak Layak	0.0040	0.1373
KK95	Layak		Layak	0.0279	0.0196

KK96	Tidak Layak		Tidak Layak	0.0040	0.0196
KK97	Tidak Layak		Tidak Layak	0.0040	0.0196
KK98	Tidak Layak		Tidak Layak	0.0279	0.1373
KK99	Tidak Layak		Tidak Layak	0.0040	0.1373
KK100	Tidak Layak		Tidak Layak	0.0040	0.1373

Data dalam tabel menunjukkan bahwa model Naïve Bayes mampu mengidentifikasi sebagian besar kasus dengan benar, dengan mayoritas prediksi yang sesuai dengan status aktual penerima bantuan. Namun, terdapat beberapa kesalahan klasifikasi, yang terlihat dari kasus-kasus di mana nilai probabilitas antara kedua kategori relatif dekat. Kesalahan ini dapat disebabkan oleh variasi dalam distribusi data, keterbatasan variabel yang digunakan, atau asumsi independensi atribut yang tidak sepenuhnya terpenuhi dalam dataset ini.

Dengan menggunakan confusion matrix, tabel ini dapat dianalisis lebih lanjut untuk mengidentifikasi pola kesalahan klasifikasi. Sebagai contoh, jika model lebih sering salah dalam mengklasifikasikan penerima bantuan yang sebenarnya "layak" sebagai "tidak layak," maka hal ini mengindikasikan bahwa beberapa variabel penting mungkin belum cukup terwakili dalam model. Oleh karena itu, perbaikan dapat dilakukan dengan menambahkan faktor sosial-ekonomi lainnya yang lebih spesifik atau dengan menerapkan metode seleksi fitur untuk meningkatkan akurasi klasifikasi.

Hasil dalam tabel ini juga menjadi dasar untuk membandingkan efektivitas Naïve Bayes dengan metode lain, seperti K-Nearest Neighbor, dalam menangani dataset yang sama. Jika model Naïve Bayes menunjukkan tingkat kesalahan yang lebih tinggi dibandingkan KNN, maka pertimbangan lebih lanjut diperlukan dalam pemilihan algoritma yang lebih optimal untuk implementasi sistem seleksi penerima bantuan sosial di tingkat desa.

Evaluasi Model Algoritma Naïve Bayes

Evaluasi hasil klasifikasi merupakan tahap krusial dalam mengukur efektivitas algoritma dalam melakukan prediksi yang akurat. Pada penelitian ini, algoritma Naïve Bayes digunakan untuk mengklasifikasikan penerima bantuan sosial ke dalam kategori "layak" dan "tidak layak" berdasarkan variabel sosial-ekonomi yang tersedia. Untuk menilai performa model, digunakan confusion matrix sebagai alat analisis utama, yang memungkinkan pemahaman lebih mendalam terhadap hasil klasifikasi dengan mengukur berbagai metrik evaluasi seperti akurasi, precision, dan recall.

Confusion matrix membagi hasil klasifikasi menjadi empat kategori utama: True Positive (TP), yaitu jumlah penerima bantuan yang diklasifikasikan sebagai "layak" dan benar-benar layak menerima; True Negative (TN), yaitu jumlah individu yang diklasifikasikan sebagai "tidak layak" dan memang tidak layak; False Positive (FP), yaitu jumlah individu yang diklasifikasikan sebagai "layak" padahal seharusnya "tidak layak"; dan False Negative (FN), yaitu jumlah individu yang diklasifikasikan sebagai "tidak layak" padahal sebenarnya mereka layak mendapatkan bantuan.

Tabel 3.3 Confusion matrix

		True	
Confusion Table	Prediksi	Layak	Tidak Layak
	Layak	3 (TP)	2 (FP)
	Tidak Layak	1 (FN)	24 (TN)

Tabel diatas merupakan hasil dari uji performa dengan metode *Naïve Bayes*, dimana hasil dari True Positive (TP) yaitu 3, True Negative (TN) yaitu 24, False Positive (FP) yaitu 2 dan False Negative (FN) yaitu 1.

Untuk mengetahui hasil *Accuracy* yaitu dengan menggunakan persamaan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{3+24}{3+24+2+1} = \frac{27}{30} = 0{,}9 = 90\%$$

Hasil evaluasi menunjukkan bahwa model Naïve Bayes memiliki akurasi sebesar 90%, yang berarti dari seluruh prediksi yang dilakukan, 90% di antaranya sesuai dengan kondisi aktual. Nilai akurasi yang tinggi ini menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar. Namun, akurasi tinggi saja tidak cukup untuk mengevaluasi model secara menyeluruh, sehingga perlu diperhatikan metrik lain seperti precision dan recall.

Untuk mengetahui hasil *Precision* yaitu dengan menggunakan persamaan sebagai berikut:

$$Precision = \frac{TP}{TP+FP}$$

$$= \frac{3}{3+2} = \frac{3}{5} = 0{,}60 = 60\%$$

Precision, yang dalam penelitian ini mencapai 60%, mengukur tingkat ketepatan model dalam mengklasifikasikan penerima bantuan sosial yang benar-benar layak. Nilai precision sebesar 60% menunjukkan bahwa dari seluruh individu yang diklasifikasikan sebagai "layak," hanya 60% yang benar-benar berhak menerima bantuan. Ini mengindikasikan bahwa terdapat cukup banyak kasus False Positive, di mana model cenderung salah dalam mengklasifikasikan individu sebagai penerima bantuan meskipun mereka sebenarnya tidak memenuhi kriteria.

Untuk mengetahui hasil *Recall* yaitu dengan menggunakan persamaan sebagai berikut:

$$Recall = \frac{TP}{TP+FN}$$

$$= \frac{3}{3+1} = \frac{3}{4} = 0{,}75 = 75\%$$

Recall, yang bernilai 75%, mengukur seberapa baik model dalam menangkap semua individu yang benar-benar layak menerima bantuan. Nilai recall sebesar 75% menunjukkan bahwa dari seluruh individu yang memang berhak menerima bantuan, model berhasil mengidentifikasi 75% di antaranya dengan benar. Namun, 25% sisanya diklasifikasikan secara salah sebagai "tidak layak" (False Negative). Hal ini menunjukkan bahwa meskipun model memiliki kemampuan yang baik dalam mengidentifikasi penerima bantuan yang benar-benar layak, masih terdapat sejumlah individu yang seharusnya menerima bantuan tetapi tidak terdeteksi oleh model.

Untuk mengetahui hasil *F1-Score* yaitu dengan menggunakan persamaan sebagai berikut:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= 2 \times \frac{0{,}60 \times 0{,}75}{0{,}60 + 0{,}75} = \frac{0{,}45}{1{,}35} = 0{,}33 = 33\%$$

Secara keseluruhan, kombinasi nilai akurasi, precision, dan recall ini menunjukkan bahwa Naïve Bayes cukup efektif dalam klasifikasi penerima bantuan sosial, tetapi masih memiliki potensi untuk ditingkatkan. Tingkat precision yang relatif lebih rendah dibandingkan recall mengindikasikan bahwa model memiliki kecenderungan untuk mengklasifikasikan individu sebagai "layak" meskipun mereka tidak memenuhi kriteria, yang dapat berimplikasi pada ketidaktepatan dalam penyaluran bantuan. Oleh karena itu, diperlukan langkah-langkah perbaikan, seperti optimasi pemilihan fitur, penyesuaian

threshold probabilitas, atau penggunaan metode hybrid untuk meningkatkan keakuratan prediksi tanpa mengorbankan precision.

Proses Penerapan Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor (KNN)* adalah salah satu metode klasifikasi berbasis instance yang bekerja dengan cara menentukan kelas suatu data baru berdasarkan kelas mayoritas dari K tetangga terdekatnya. Jarak antara data baru dan seluruh data dalam dataset dihitung menggunakan metrik tertentu, seperti Euclidean, Manhattan, atau Minkowski. Metrik ini digunakan untuk menentukan kedekatan antar data dalam ruang multidimensi.

Rumus Jarak Euclidean:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Dimana:

p : Titik data baru yang ingin di prediksi
q : Titik data yang ada di dataset
q_i & p_i : Nilai data ke-I dari titik q dan p
n : Jumlah data dalam dataset

Sebelum melakukan pengolahan data. Dataset dibagi menjadi 2 bagian yaitu data latih dan data uji. Data latih (training) merupakan data 70% dari total dataset, digunakan untuk melatih model. Sedangkan Data uji (testing) merupakan data 30% dari total dataset, digunakan untuk menguji model.

Pada algoritma KNN ini, beberapa variabel penting telah dikelompokkan menjadi transformasi kategori yang lebih mudah dianalisis.

Tabel 3.4 Pembagian Kategori Transformasi Data

Variabel	Sebelum Transformasi	Setelah Transformasi
Gender KK	Laki-Laki	1
	Perempuan	2
Usia	Muda	1
	Dewasa	2
	Lansia	3
Penghasilan	Diatas UMK	1
	Dibawah UMK	2
Tanggungan	Kecil	1
	Sedang	2
	Besar	3
Kepemilikan Rumah	Milik Pribadi	1
	Milik Orang Lain	2

Transformasi data ini mempermudah analisis selanjutnya, khususnya dalam penerapan algoritma klasifikasi seperti K-Nearest Neighbor, dengan memastikan bahwa variabel yang digunakan dalam model sudah dalam bentuk yang lebih sederhana dan terstruktur agar mudah dalam melakukan perhitungan.

Tabel 3.5 Data Latih K-Nearest Neighbor

Nama	Gender KK	Usia	Peng-hasilan	Tangg-ungan	Kepemi-likan Rumah	Status
KK1	1	2	1	1	2	Tidak Layak
KK2	1	1	1	1	1	Tidak Layak
KK3	2	3	2	1	1	Tidak Layak
KK4	2	2	1	1	1	Tidak Layak
KK5	2	2	1	1	2	Layak
KK6	1	2	1	1	2	Tidak Layak
KK7	1	2	2	1	2	Tidak Layak
KK8	1	2	1	1	2	Tidak Layak
KK9	1	2	1	2	2	Layak
KK10	2	3	2	1	1	Layak
KK11	1	2	1	1	1	Tidak Layak

KK12	1	3	2	1	1	Tidak Layak
KK13	1	2	1	1	1	Tidak Layak
KK14	1	3	2	1	1	Tidak Layak
KK15	1	2	1	2	2	Tidak Layak
KK16	1	2	1	2	1	Tidak Layak
KK17	1	2	1	1	1	Tidak Layak
KK18	1	2	2	1	2	Layak
KK19	1	2	2	1	1	Tidak Layak
KK20	2	2	1	1	1	Tidak Layak
KK21	1	3	2	1	1	Tidak Layak
KK22	2	3	2	1	1	Tidak Layak
KK23	1	3	2	1	1	Layak
KK24	1	3	2	1	1	Tidak Layak
KK25	1	2	2	1	1	Tidak Layak
KK26	1	2	2	1	1	Tidak Layak
KK27	1	2	1	1	1	Tidak Layak
KK28	1	3	2	2	2	Tidak Layak
KK29	1	3	2	1	1	Layak
KK30	2	2	2	1	1	Tidak Layak
KK31	1	2	1	1	1	Tidak Layak
KK32	2	3	2	1	1	Tidak Layak
KK33	1	2	2	1	1	Layak
KK34	1	2	1	2	1	Tidak Layak
KK35	2	2	1	1	1	Tidak Layak
KK36	2	3	2	1	1	Tidak Layak
KK37	1	3	2	1	1	Tidak Layak
KK38	1	2	1	1	1	Tidak Layak
KK39	1	2	2	2	1	Tidak Layak
KK40	1	2	1	2	1	Tidak Layak
KK41	2	3	2	1	2	Layak
KK42	1	2	1	1	1	Tidak Layak
KK43	1	3	2	1	1	Tidak Layak
KK44	1	2	1	2	1	Tidak Layak
KK45	2	3	2	1	1	Layak

KK46	2	3	2	1	1	Layak
KK47	2	3	2	1	1	Layak
KK48	1	3	1	1	1	Tidak Layak
KK49	1	2	1	1	1	Tidak Layak
KK50	1	2	1	1	1	Tidak Layak
KK51	1	3	2	1	1	Tidak Layak
KK52	1	2	1	1	1	Tidak Layak
KK53	1	2	2	1	2	Layak
KK54	2	3	2	1	1	Layak
KK55	1	2	2	1	1	Tidak Layak
KK56	2	3	2	1	1	Tidak Layak
KK57	1	2	1	1	1	Tidak Layak
KK58	1	2	2	2	2	Tidak Layak
KK59	1	2	1	1	1	Tidak Layak
KK60	1	2	2	2	1	Layak
KK61	1	3	2	1	1	Tidak Layak
KK62	1	2	2	2	2	Tidak Layak
KK63	1	3	2	1	1	Tidak Layak
KK64	1	2	1	1	1	Tidak Layak
KK65	1	3	2	1	1	Tidak Layak
KK66	2	3	2	1	1	Tidak Layak
KK67	2	3	2	1	1	Tidak Layak
KK68	1	2	1	1	1	Tidak Layak
KK69	1	2	2	2	1	Layak
KK70	2	2	2	1	1	Layak

Data diatas merupakan data latih berjumlah 70 data digunakan untuk melatih model. Berisi pasangan input (fitur) dan output (label) yang telah diketahui. Model KNN akan mempelajari distribusi probabilitas dari data latih ini untuk membuat prediksi.

Tabel 3.6 Data Uji K-Nearest Neighbor

Nama	Gender KK	Usia	Peng-hasilan	Tang-gungan	Kepemi-likan Rumah	Status
KK71	1	2	1	2	1	Tidak Layak
KK72	2	2	2	1	2	Layak
KK73	1	1	1	1	2	Tidak Layak
KK74	1	3	2	1	1	Tidak Layak
KK75	1	2	2	1	1	Tidak Layak
KK76	1	2	1	1	1	Tidak Layak
KK77	2	3	2	1	1	Tidak Layak
KK78	1	2	1	1	1	Tidak Layak
KK79	1	2	1	1	1	Tidak Layak
KK80	1	2	2	1	1	Tidak Layak
KK81	1	2	2	1	1	Tidak Layak
KK82	1	2	2	1	2	Tidak Layak
KK83	1	3	1	1	1	Tidak Layak
KK84	1	2	1	1	1	Tidak Layak
KK85	2	3	2	1	1	Tidak Layak
KK86	1	2	1	2	2	Layak
KK87	1	2	2	1	1	Tidak Layak
KK88	1	2	1	2	1	Tidak Layak
KK89	1	2	1	2	1	Tidak Layak
KK90	1	2	2	1	1	Tidak Layak
KK91	1	3	2	1	1	Tidak Layak
KK92	2	2	2	1	2	Layak
KK93	2	3	1	1	1	Tidak Layak
KK94	1	2	1	1	1	Tidak Layak
KK95	2	3	2	1	1	Layak
KK96	2	3	1	1	1	Tidak Layak
KK97	2	3	1	1	1	Tidak Layak
KK98	1	2	2	1	1	Tidak Layak
KK99	1	2	1	1	1	Tidak Layak
KK100	1	2	1	1	1	Tidak Layak

Tabel diatas merupakan data uji berjumlah 30 data digunakan untuk menguji kinerja model. Berisi data yang tidak diketahui labelnya oleh model. Model akan memprediksi label berdasarkan apa yang telah dipelajari dari data latih, dan hasil prediksi ini dibandingkan dengan label yang sebenarnya untuk mengevaluasi akurasi model. Data akan dihitung jarak euclidean untuk menentukan anggota keluarga yang berhak dan yang tidak berhak menerima bantuan social.

Tabel 3.7 Data baru untuk klasifikasi K-Nearest Neighbor

Nama	Gender Kepala Keluarga	Usia	Pekerjaan	Tanggungan	Kepemilikan Rumah
KK21	1	2	1	2	1

$$KK1 = \sqrt{(1-1)^2 + (2-2)^2 + (1-1)^2 + (1-2)^2 + (2-1)^2}$$
$$= 1{,}412$$

Cara perhitungan sama sampai data latih KK70 dengan menyesuaikan nilai dari data yang ingin dihitung jaraknya. Kemudian jarak diurutkan dari jarak yang terkecil ke jarak yang terbesar. Selanjutnya tentukan nilai K. Nilai K yang digunakan yaitu K=9. Maka klasifikasi diambil dari 5 tetangga terdekat. Pemilihan Nilai K disarankan berjumlah ganjil agar mudah saat penarikan hasil klasifikasi. Hasil akhir perhitungan jarak dan setelah diurutkan dari data terkecill ke data terbesar seperti tabel berikut.

Tabel 3.8 Hasil klasifikasi KNN data KK71 dengan nilai K=5

Nama	Status	KK72
KK16	Tidak Layak	0.0000
KK34	Tidak Layak	0.0000
KK40	Tidak Layak	0.0000
KK44	Tidak Layak	0.0000
KK9	Layak	1.0000
KK11	Tidak Layak	1.0000
KK13	Tidak Layak	1.0000
KK15	Tidak Layak	1.0000

KK17	Tidak Layak	1.0000
KK27	Tidak Layak	1.0000
KK31	Tidak Layak	1.0000
KK38	Tidak Layak	1.0000
KK39	Tidak Layak	1.0000
KK42	Tidak Layak	1.0000
KK49	Tidak Layak	1.0000
KK50	Tidak Layak	1.0000
KK52	Tidak Layak	1.0000
KK57	Tidak Layak	1.0000
KK59	Tidak Layak	1.0000
KK60	Layak	1.0000
KK64	Tidak Layak	1.0000
KK68	Tidak Layak	1.0000
KK69	Layak	1.0000
KK1	Tidak Layak	1.4142
KK2	Tidak Layak	1.4142
KK4	Tidak Layak	1.4142
KK6	Tidak Layak	1.4142
KK8	Tidak Layak	1.4142
KK19	Tidak Layak	1.4142
KK20	Tidak Layak	1.4142
KK25	Tidak Layak	1.4142
KK26	Tidak Layak	1.4142
KK33	Layak	1.4142
KK35	Tidak Layak	1.4142
KK48	Tidak Layak	1.4142
KK55	Tidak Layak	1.4142
KK58	Tidak Layak	1.4142
KK62	Tidak Layak	1.4142
KK5	Layak	1.7321
KK7	Tidak Layak	1.7321
KK12	Tidak Layak	1.7321
KK14	Tidak Layak	1.7321

KK18	Layak	1.7321
KK21	Tidak Layak	1.7321
KK23	Layak	1.7321
KK24	Tidak Layak	1.7321
KK28	Tidak Layak	1.7321
KK29	Layak	1.7321
KK30	Tidak Layak	1.7321
KK37	Tidak Layak	1.7321
KK43	Tidak Layak	1.7321
KK51	Tidak Layak	1.7321
KK53	Layak	1.7321
KK61	Tidak Layak	1.7321
KK63	Tidak Layak	1.7321
KK65	Tidak Layak	1.7321
KK70	Layak	1.7321
KK3	Tidak Layak	2.0000
KK10	Layak	2.0000
KK22	Tidak Layak	2.0000
KK32	Tidak Layak	2.0000
KK36	Tidak Layak	2.0000
KK45	Layak	2.0000
KK46	Layak	2.0000
KK47	Layak	2.0000
KK54	Layak	2.0000
KK56	Tidak Layak	2.0000
KK66	Tidak Layak	2.0000
KK67	Tidak Layak	2.0000
KK41	Layak	2.2361

Terlihat pada tabel dengan menggunakan K=9 atau 9 tetangga terdekat dari data terkecil menunjukkan bahwa untuk klasifikasi data uji K71 sebanyak 1 nilai untuk layak dan 8 nilai untuk tidak layak. Dapat disimpulkan bahwa data uji pertama (KK71) tidak layak menerima bantuan social.

Perhitungan tetap sama dilakukan dengan mencari jarak euclidaen dari setiap data uji yang baru. Berikut hasil akhir klasifikasi dari 30 data uji:

Tabel 3.9 Hasil Klasifikasi keseluruhan data uji

Nama	Status		Status Predikasi
KK71	Tidak Layak		Tidak Layak
KK72	Layak		Layak
KK73	Tidak Layak		Tidak Layak
KK74	Tidak Layak		Tidak Layak
KK75	Tidak Layak		Tidak Layak
KK76	Tidak Layak		Tidak Layak
KK77	Tidak Layak		Layak
KK78	Tidak Layak		Tidak Layak
KK79	Tidak Layak		Tidak Layak
KK80	Tidak Layak		Tidak Layak
KK81	Tidak Layak		Tidak Layak
KK82	Tidak Layak		Tidak Layak
KK83	Tidak Layak		Tidak Layak
KK84	Tidak Layak		Tidak Layak
KK85	Tidak Layak	Prediksi	Layak
KK86	Layak		Tidak Layak
KK87	Tidak Layak		Tidak Layak
KK88	Tidak Layak		Tidak Layak
KK89	Tidak Layak		Tidak Layak
KK90	Tidak Layak		Tidak Layak
KK91	Tidak Layak		Tidak Layak
KK92	Layak		Layak
KK93	Tidak Layak		Tidak Layak
KK94	Tidak Layak		Tidak Layak
KK95	Layak		Layak
KK96	Tidak Layak		Tidak Layak
KK97	Tidak Layak		Tidak Layak
KK98	Tidak Layak		Tidak Layak
KK99	Tidak Layak		Tidak Layak
KK100	Tidak Layak		Tidak Layak

Hasil klasifikasi dengan metode *K-Nearest Neighbor (KNN)* menunjukkan bahwa dari 30 data uji yang ada, 5 anggota keluarga yang layak menerima bantuan social dan 25 keluarga yang tidak layak menerima bantuan social.

Hasil perhitungan jarak dalam tabel ini menjadi dasar utama bagi algoritma KNN dalam menentukan kategori data baru. Setelah seluruh jarak dihitung, algoritma memilih K tetangga terdekat, dan berdasarkan mayoritas dari kategori tetangga tersebut, data baru diklasifikasikan.

Keakuratan klasifikasi sangat bergantung pada distribusi data dan pemilihan K yang tepat. Jika terdapat outlier dalam data latih, maka dapat mempengaruhi hasil perhitungan jarak dan menyebabkan klasifikasi yang kurang akurat. Oleh karena itu, analisis terhadap tabel ini membantu dalam mengevaluasi seberapa baik distribusi data serta bagaimana pengaruhnya terhadap hasil klasifikasi yang diperoleh.

Evaluasi Model Klasifikasi dengan KNN

Evaluasi hasil klasifikasi merupakan langkah penting dalam menentukan efektivitas algoritma K-Nearest Neighbor (KNN) dalam mengelompokkan penerima bantuan sosial berdasarkan variabel sosial-ekonomi. Salah satu metode yang umum digunakan untuk mengevaluasi kinerja model klasifikasi adalah confusion matrix, yang memberikan gambaran tentang seberapa baik model dalam membedakan antara kategori "layak" dan "tidak layak" menerima bantuan.

Tabel 3.10 Confusion matrix K-Nearest Neighbor

		True	
Confusion Table	Prediksi	Layak	Tidak Layak
	Layak	3 (TP)	2 (FP)
	Tidak Layak	1 (FN)	24 (TN)

Pada penelitian ini, confusion matrix yang dihasilkan dari penerapan algoritma KNN menunjukkan hasil sebagai berikut: True

Positive (TP) = 3, True Negative (TN) = 24, False Positive (FP) = 2, dan False Negative (FN) = 1. Berdasarkan data ini, dapat dihitung beberapa metrik evaluasi utama, yaitu akurasi, precision, recall, dan F1-score, yang masing-masing memiliki peran penting dalam menilai efektivitas model klasifikasi.

Akurasi adalah ukuran keseluruhan performa model dalam mengklasifikasikan data dengan benar. Rumus perhitungan akurasi adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{3+24}{3+24+2+1} = \frac{27}{30} = 0{,}9 = 90\%$$

Nilai akurasi sebesar 90% menunjukkan bahwa model KNN memiliki kemampuan yang sangat baik dalam melakukan klasifikasi, di mana 90% dari total data yang diuji berhasil diklasifikasikan dengan benar.

Namun, akurasi saja tidak cukup untuk menilai kinerja model secara menyeluruh. Oleh karena itu, perlu diperhatikan juga metrik precision, yang mengukur seberapa akurat prediksi positif yang diberikan oleh model. Precision dihitung dengan rumus:

$$Precision = \frac{TP}{TP+FP}$$

$$= \frac{3}{3+2} = \frac{3}{5} = 0{,}60 = 60\%$$

Nilai precision sebesar 60% menunjukkan bahwa dari seluruh individu yang diprediksi sebagai "layak" oleh model, hanya 60% yang benar-benar layak menerima bantuan. Ini berarti model masih memiliki tingkat kesalahan yang cukup signifikan dalam memberikan klasifikasi positif, yang tercermin dari adanya False Positive (FP) sebanyak 2 kasus.

Selanjutnya, metrik recall digunakan untuk mengukur seberapa baik model dalam mengidentifikasi individu yang benar-benar layak menerima bantuan. Recall dihitung dengan rumus:

$$Recall = \frac{TP}{TP+FN}$$

$$= \frac{3}{3+1} = \frac{3}{4} = 0{,}75 = 75\%$$

Nilai recall sebesar 75% menunjukkan bahwa dari seluruh individu yang benar-benar layak menerima bantuan, model mampu mengidentifikasi 75% di antaranya dengan benar. Namun, masih ada False Negative (FN) sebanyak 1 kasus, yang berarti ada individu yang seharusnya menerima bantuan tetapi diklasifikasikan sebagai "tidak layak."

Untuk memperoleh metrik evaluasi yang lebih komprehensif, digunakan F1-score, yang merupakan rata-rata harmonis antara precision dan recall. F1-score dihitung dengan rumus:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= 2 \times \frac{0{,}60 \times 0{,}75}{0{,}60 + 0{,}75} = \frac{0{,}45}{1{,}35} = 0{,}33 = 33\%$$

Nilai F1-score sebesar 66.7% menunjukkan keseimbangan antara precision dan recall dalam model KNN. Meskipun model memiliki akurasi yang tinggi, nilai F1-score yang lebih rendah menunjukkan bahwa terdapat beberapa kesalahan dalam klasifikasi positif yang perlu diperbaiki.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa algoritma KNN cukup efektif dalam mengklasifikasikan penerima bantuan sosial dengan akurasi 90%, tetapi memiliki keterbatasan dalam precision dan recall yang dapat menyebabkan kesalahan dalam menentukan individu yang benar-benar berhak menerima bantuan.

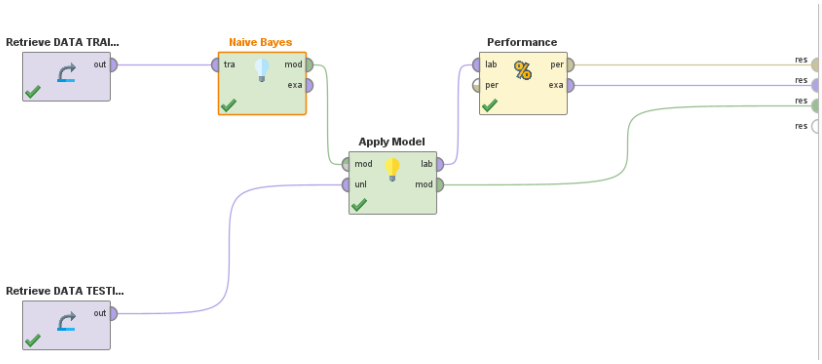
Untuk meningkatkan performa model, beberapa pendekatan yang dapat diterapkan adalah penyesuaian nilai K, seleksi fitur yang lebih relevan, atau penggunaan metode ensemble learning untuk mengurangi tingkat kesalahan klasifikasi.

5

HASIL IMPLEMENTASI RAPIDMINER DAN PEMBAHASAN

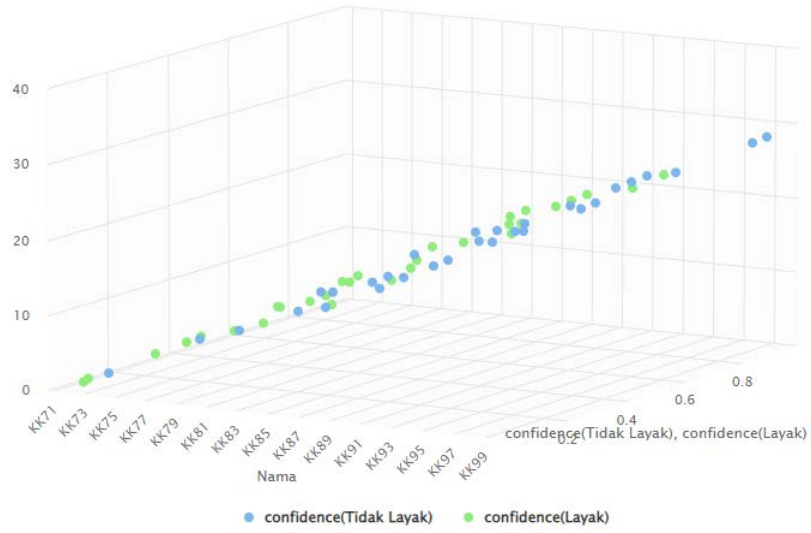
Interpretasi Hasil Klasifikasi Naïve Bayes

Metode Naïve Bayes telah banyak digunakan dalam klasifikasi data karena kemampuannya dalam menangani dataset dengan fitur yang bersifat independen. Pada penelitian ini, algoritma Naïve Bayes diterapkan dalam sistem seleksi penerima bantuan sosial untuk mengelompokkan individu ke dalam kategori "Layak" atau "Tidak Layak" menerima bantuan. Implementasi dilakukan menggunakan RapidMiner, sebuah perangkat lunak berbasis visual yang memungkinkan pengguna untuk membangun, melatih, dan mengevaluasi model machine learning secara efisien.



Gambar 4.1 Rancangan Model Naïve Bayes

Gambar diatas merupakan rancangan model machine learning dengan menggunakan metode *Naïve Bayes*. Implementasi ini dilakukan dengan menggunakan software data mining yaitu RapidMiner Studio.



Gambar 4.2 Visualisasi Klasifikasi Naïve Bayes

Gambar yang ditampilkan merupakan visualisasi hasil klasifikasi dengan metode Naïve Bayes (NB) dalam menentukan status penerima bantuan sosial, yaitu "Layak" dan "Tidak Layak". Visualisasi ini berbentuk scatter plot 3D, yang menunjukkan bagaimana model

menghitung probabilitas dan tingkat kepercayaan (confidence) dalam mengklasifikasikan individu.

Setelah model dilatih dan diuji, hasil klasifikasi divisualisasikan dalam bentuk confusion matrix (ditampilkan dalam gambar). Akurasi model yang diperoleh sebesar 80%, yang menunjukkan bahwa dari seluruh prediksi yang dilakukan, sebanyak 80% sesuai dengan kondisi sebenarnya. Metrik evaluasi lainnya seperti precision dan recall juga dihitung untuk menganalisis lebih lanjut efektivitas model dalam menangani ketidakseimbangan data.

accuracy: 80.00%

	true Tidak Layak	true Layak	class precision
pred. Tidak Layak	21	1	95.45%
pred. Layak	5	3	37.50%
class recall	80.77%	75.00%	

Gambar 4. 3 Confusion matrix Naïve Bayes

Dari hasil confusion matrix, model berhasil mengklasifikasikan 21 individu sebagai "Tidak Layak" dengan benar, namun terdapat 5 individu yang sebenarnya "Tidak Layak" tetapi diklasifikasikan sebagai "Layak" (False Positive–FP). Selain itu, model juga berhasil mengidentifikasi 3 individu sebagai "Layak" dengan benar, tetapi terdapat 1 individu yang sebenarnya "Layak" namun diklasifikasikan sebagai "Tidak Layak" (False Negative–FN).

Dari metrik evaluasi, precision untuk kategori "Tidak Layak" sebesar 95,45%, yang menunjukkan bahwa model sangat baik dalam memastikan individu yang diprediksi "Tidak Layak" memang benar-benar sesuai. Namun, precision untuk kategori "Layak" lebih rendah, yaitu 37,50%, yang mengindikasikan bahwa model masih memiliki tingkat kesalahan yang cukup tinggi dalam mengidentifikasi individu yang benar-benar layak menerima bantuan.

Recall untuk kategori "Tidak Layak" sebesar 80,77%, yang berarti model mampu menangkap sebagian besar individu yang memang tidak layak menerima bantuan. Sementara itu, recall untuk kategori "Layak" sebesar 75,00%, menunjukkan bahwa model cukup baik

dalam mengenali individu yang seharusnya mendapatkan bantuan, meskipun masih terdapat kesalahan klasifikasi.

Nilai precision dan recall yang disajikan dalam tabel menunjukkan efektivitas model dalam menghindari kesalahan klasifikasi serta kemampuannya dalam menangkap individu yang sesuai dengan kategori yang diharapkan. Dengan akurasi 80%, model ini cukup dapat diandalkan, meskipun perlu ditingkatkan pada aspek precision untuk kategori "Layak" agar lebih selektif dalam memberikan rekomendasi penerima bantuan.

Analisis Efektivitas Model Naïve Bayes

Dari hasil evaluasi ini, dapat disimpulkan bahwa Naïve Bayes memiliki kinerja yang cukup baik dalam mengelompokkan individu yang "Tidak Layak" menerima bantuan dengan precision yang tinggi. Namun, model masih mengalami kesulitan dalam mengklasifikasikan individu yang benar-benar "Layak", sebagaimana terlihat dari rendahnya precision pada kategori ini. Hal ini mungkin disebabkan oleh keterbatasan asumsi independensi antar variabel dalam Naïve Bayes atau ketidakseimbangan jumlah data pada masing-masing kelas.

Untuk meningkatkan performa model, beberapa langkah perbaikan yang dapat dilakukan meliputi:

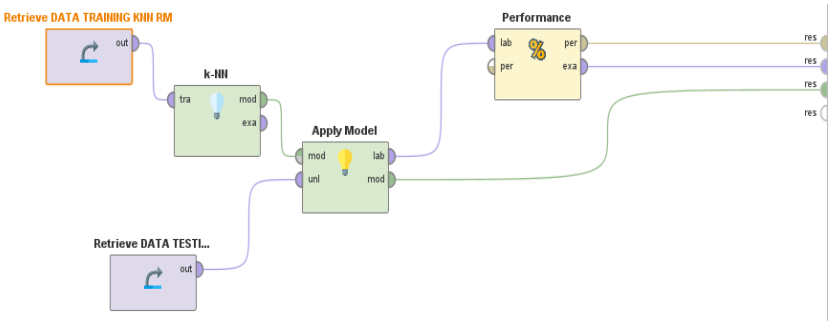
1. Penyesuaian fitur yang digunakan dalam model – Beberapa variabel mungkin memiliki korelasi yang lebih kuat dengan keputusan akhir dan dapat ditingkatkan bobotnya dalam model.
2. Penggunaan teknik balancing data – Jika jumlah data kategori "Layak" jauh lebih sedikit dibandingkan "Tidak Layak", maka teknik seperti oversampling atau undersampling dapat diterapkan untuk menyeimbangkan distribusi kelas.
3. Eksperimen dengan model lain – Meskipun Naïve Bayes merupakan metode yang cepat dan sederhana, algoritma

seperti Random Forest atau K-Nearest Neighbor (KNN) dapat dibandingkan untuk melihat apakah ada peningkatan akurasi.

Dengan hasil evaluasi ini, penggunaan Naïve Bayes dalam klasifikasi penerima bantuan sosial cukup efektif, terutama dalam memastikan bahwa bantuan tidak diberikan kepada individu yang tidak memenuhi kriteria. Namun, model ini masih perlu ditingkatkan agar lebih akurat dalam mengidentifikasi penerima yang benar-benar layak mendapatkan bantuan.

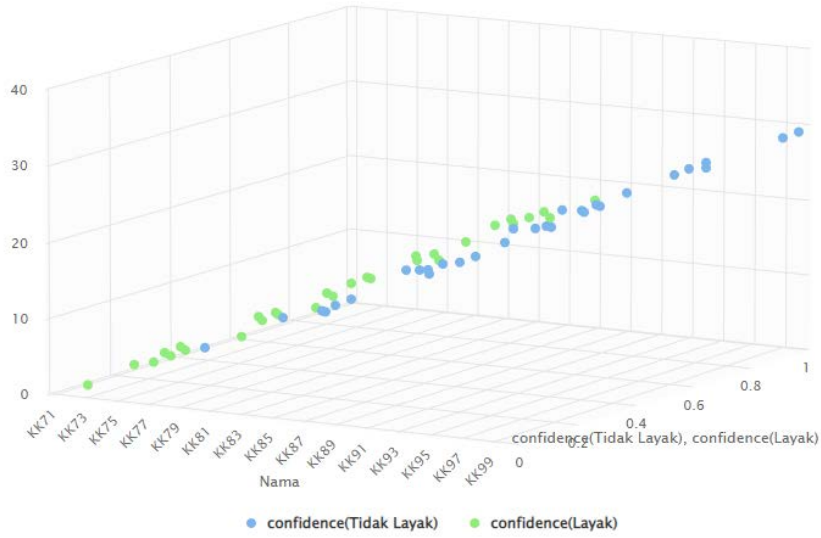
Interpretasi Hasil Klasifikasi K-Nearest Neighbors (KNN)

Metode K-Nearest Neighbors (KNN) merupakan salah satu algoritma machine learning yang berbasis instance dan sering digunakan dalam klasifikasi karena kemampuannya dalam mengelompokkan data berdasarkan kedekatan dengan sampel yang telah ada. Pada penelitian ini, KNN digunakan untuk mengelompokkan calon penerima bantuan sosial ke dalam kategori "Layak" atau "Tidak Layak", dengan implementasi menggunakan RapidMiner.



Gambar 4.4 Rancangan Model K-Nearest Neighbor (KNN)

Gambar diatas merupakan rancangan model machine learning dengan menggunakan metode *K-Nearest Neighbor (KNN)*. Implementasi ini dilakukan dengan menggunakan software data mining yaitu RapidMiner Studio.



Gambar 4.5 visualisasi klasifikasi K-Nearest Neighbor (KNN)

Gambar yang ditampilkan merupakan visualisasi hasil klasifikasi dengan metode K-Nearest Neighbors (KNN) dalam menentukan status penerima bantuan sosial, yaitu "Layak" dan "Tidak Layak". Visualisasi ini dilakukan dalam bentuk scatter plot 3D, yang menunjukkan distribusi kepercayaan (confidence) model dalam mengklasifikasikan individu berdasarkan nama (KK) sebagai sumbu X, indeks numerik sebagai sumbu Y, dan nilai confidence sebagai sumbu Z.

accuracy: 90.00%

	true Tidak Layak	true Layak	class precision
pred. Tidak Layak	24	1	96.00%
pred. Layak	2	3	60.00%
class recall	92.31%	75.00%	

Gambar 4.6 Confusion matrix K-Nearest Neighbor (KNN)

Setelah model dijalankan, hasil evaluasi menunjukkan bahwa algoritma KNN menghasilkan akurasi sebesar 90%, yang berarti sebagian besar prediksi model sesuai dengan kondisi sebenarnya. Confusion matrix yang diperoleh memberikan gambaran mengenai

performa model dalam mengklasifikasikan individu sebagai "Layak" atau "Tidak Layak" menerima bantuan.

Dari confusion matrix, diketahui bahwa 24 individu diklasifikasikan dengan benar sebagai "Tidak Layak", sementara hanya 1 individu yang seharusnya "Layak" diklasifikasikan sebagai "Tidak Layak" (False Negative–FN). Selain itu, 3 individu yang benar-benar "Layak" berhasil diklasifikasikan dengan benar, namun terdapat 2 individu yang seharusnya "Tidak Layak" tetapi diklasifikasikan sebagai "Layak" (False Positive–FP).

Dari segi metrik evaluasi, precision untuk kategori "Tidak Layak" sebesar 96%, yang berarti bahwa hampir semua individu yang diprediksi sebagai "Tidak Layak" memang benar-benar sesuai dengan kondisi aktual. Namun, precision untuk kategori "Layak" lebih rendah, yakni 60%, yang menunjukkan bahwa masih terdapat kesalahan dalam mengklasifikasikan penerima bantuan yang benar-benar layak.

Recall untuk kategori "Tidak Layak" sebesar 92,31%, yang berarti bahwa model mampu menangkap sebagian besar individu yang memang tidak layak menerima bantuan. Sementara itu, recall untuk kategori "Layak" sebesar 75%, menunjukkan bahwa meskipun model cukup baik dalam mengenali individu yang seharusnya mendapatkan bantuan, masih terdapat beberapa kesalahan klasifikasi.

Analisis Efektivitas Model K-Nearest Neighbors (KNN)

Dari hasil evaluasi ini, dapat disimpulkan bahwa algoritma KNN memiliki performa yang cukup baik dalam mengelompokkan penerima bantuan sosial. Model menunjukkan akurasi yang lebih tinggi dibandingkan Naïve Bayes (90% vs. 80%), terutama dalam mengklasifikasikan individu yang tidak layak menerima bantuan. Namun, kelemahan model ini masih terletak pada precision untuk kategori "Layak", yang relatif rendah dibandingkan kategori lainnya.

Untuk meningkatkan kinerja model KNN, beberapa langkah perbaikan yang dapat dilakukan meliputi:

1. Optimasi Pemilihan Nilai K – Pemilihan jumlah tetangga (K) sangat berpengaruh terhadap hasil klasifikasi. Mencari nilai K yang optimal melalui hyperparameter tuning dapat membantu meningkatkan akurasi model.
2. Balancing Data – Jika terdapat ketidakseimbangan antara jumlah data "Layak" dan "Tidak Layak", teknik oversampling atau undersampling dapat diterapkan untuk memastikan model tidak bias terhadap salah satu kategori.
3. Penerapan Feature Selection – Beberapa fitur dalam dataset mungkin memiliki pengaruh yang lebih besar terhadap klasifikasi. Dengan melakukan feature selection, model dapat difokuskan pada variabel yang paling relevan dan meningkatkan performanya.
4. Eksperimen dengan Jarak yang Berbeda – KNN menggunakan metrik jarak seperti Euclidean Distance, tetapi metode lain seperti Manhattan Distance atau Minkowski Distance dapat dicoba untuk melihat pengaruhnya terhadap akurasi model.

Dengan hasil evaluasi ini, penggunaan KNN dalam klasifikasi penerima bantuan sosial cukup efektif, terutama dalam memastikan bahwa bantuan tidak diberikan kepada individu yang tidak memenuhi kriteria. Namun, model masih perlu ditingkatkan agar lebih akurat dalam mengidentifikasi penerima yang benar-benar layak mendapatkan bantuan.

Perbandingan dari Penggunaan Kedua Algoritma

Dalam sistem distribusi bantuan sosial di desa, pemanfaatan algoritma kecerdasan buatan menjadi salah satu pendekatan inovatif untuk meningkatkan efisiensi dan ketepatan sasaran. Dua metode

yang umum digunakan dalam klasifikasi penerima bantuan adalah Naïve Bayes (NB) dan K-Nearest Neighbors (KNN). Kedua algoritma ini memiliki keunggulan dan keterbatasan masing-masing dalam menentukan individu yang berhak menerima bantuan berdasarkan variabel-variabel tertentu seperti penghasilan, jumlah tanggungan, kondisi tempat tinggal, dan faktor sosial ekonomi lainnya. Oleh karena itu, membandingkan dampak dari penggunaan kedua algoritma ini menjadi penting dalam mengevaluasi efektivitas sistem distribusi bantuan di desa.

Naïve Bayes adalah metode berbasis probabilitas yang menghitung kemungkinan suatu individu masuk dalam kategori "Layak" atau "Tidak Layak" berdasarkan distribusi data historis. Keunggulan utama dari Naïve Bayes adalah kemampuannya dalam menangani data dengan banyak fitur tanpa mengalami overfitting yang signifikan. Selain itu, model ini bekerja dengan sangat cepat dan dapat memberikan hasil klasifikasi dalam waktu singkat, yang sangat berguna untuk sistem distribusi bantuan di desa dengan sumber daya komputasi terbatas. Namun, kelemahan utama Naïve Bayes adalah asumsi independensi antar fitur, yang dalam banyak kasus tidak selalu sesuai dengan realitas sosial. Misalnya, faktor ekonomi dan kondisi tempat tinggal mungkin saling berkaitan, namun dalam Naïve Bayes keduanya dianggap independen, yang dapat menyebabkan kesalahan prediksi pada kasus tertentu.

Di sisi lain, K-Nearest Neighbors (KNN) bekerja dengan prinsip pencocokan jarak antara individu dalam dataset berdasarkan sejumlah tetangga terdekatnya. Model ini memiliki keunggulan dalam menangkap pola non-linear yang lebih kompleks dibandingkan Naïve Bayes. KNN juga tidak mengasumsikan independensi fitur, sehingga lebih fleksibel dalam menangani hubungan antar variabel dalam data. Namun, tantangan utama dari KNN adalah waktu komputasi yang lebih tinggi, terutama jika dataset yang digunakan berukuran besar. Selain itu, pemilihan jumlah tetangga terdekat (nilai K) sangat berpengaruh terhadap akurasi hasil klasifikasi. Jika nilai K terlalu

kecil, model cenderung mengalami overfitting, sementara jika terlalu besar, model dapat kehilangan sensitivitas dalam membedakan individu yang benar-benar layak menerima bantuan.

Dalam konteks distribusi bantuan di desa, pemilihan algoritma yang tepat bergantung pada kebutuhan spesifik dari sistem yang digunakan. Jika tujuan utama adalah kecepatan dan efisiensi dalam pengambilan keputusan, maka Naïve Bayes lebih cocok diterapkan. Namun, jika ketepatan klasifikasi lebih diutamakan, terutama untuk menghindari kesalahan dalam memberikan bantuan kepada individu yang tidak berhak, maka KNN dapat menjadi pilihan yang lebih baik, meskipun memerlukan sumber daya komputasi yang lebih besar. Oleh karena itu, solusi yang paling optimal dalam sistem distribusi bantuan sosial di desa adalah kombinasi kedua metode ini, di mana Naïve Bayes dapat digunakan untuk seleksi awal dengan cepat, sementara KNN dapat diterapkan untuk validasi lebih lanjut guna memastikan penerima bantuan telah ditentukan dengan tingkat akurasi yang lebih tinggi.

Faktor-faktor yang Memengaruhi Akurasi Algoritma

Dalam sistem klasifikasi berbasis kecerdasan buatan, akurasi algoritma merupakan parameter utama yang menentukan keandalan model dalam membuat keputusan. Tingkat akurasi dipengaruhi oleh berbagai faktor, mulai dari kualitas data hingga parameter algoritma yang digunakan. Dalam konteks distribusi bantuan di desa, pemilihan algoritma yang tepat sangat bergantung pada pemahaman terhadap faktor-faktor ini agar model yang diterapkan mampu memberikan hasil klasifikasi yang optimal.

Salah satu faktor utama yang memengaruhi akurasi adalah kualitas dan kelengkapan data. Data yang bersih, lengkap, dan relevan akan meningkatkan akurasi model, sedangkan data yang mengandung noise, nilai yang hilang, atau informasi yang tidak relevan dapat

menyebabkan kesalahan klasifikasi. Misalnya, dalam sistem distribusi bantuan, informasi yang tidak lengkap tentang kondisi ekonomi atau jumlah tanggungan rumah tangga dapat menyebabkan model gagal mengidentifikasi penerima yang benar-benar membutuhkan bantuan. Oleh karena itu, proses preprocessing data seperti normalisasi, imputasi data yang hilang, serta penghapusan outlier menjadi langkah penting dalam meningkatkan performa algoritma.

Selain itu, pemilihan fitur (feature selection) berperan penting dalam menentukan akurasi model. Tidak semua variabel dalam dataset memberikan kontribusi yang signifikan terhadap proses klasifikasi. Fitur yang tidak relevan dapat meningkatkan kompleksitas model tanpa memberikan manfaat yang nyata, sehingga menyebabkan overfitting atau bahkan bias dalam prediksi. Dalam kasus klasifikasi penerima bantuan, variabel seperti total penghasilan rumah tangga, jumlah tanggungan, dan kondisi tempat tinggal kemungkinan memiliki pengaruh lebih besar dibandingkan dengan variabel lain seperti lokasi geografis yang kurang relevan. Oleh karena itu, teknik seperti Principal Component Analysis (PCA) atau metode seleksi fitur berbasis statistik dapat digunakan untuk meningkatkan efisiensi model.

Selanjutnya, pemilihan algoritma dan parameter optimasi juga memiliki dampak yang signifikan terhadap akurasi. Algoritma seperti Naïve Bayes yang berbasis probabilitas memiliki asumsi independensi antar fitur yang mungkin tidak selalu sesuai dengan kondisi nyata. Sementara itu, algoritma seperti K-Nearest Neighbors (KNN) sangat bergantung pada pemilihan parameter jumlah tetangga terdekat (K) yang tepat. Jika nilai K terlalu kecil, model rentan terhadap noise, sedangkan jika terlalu besar, model dapat kehilangan sensitivitas terhadap pola dalam data. Oleh karena itu, pendekatan seperti cross-validation sangat penting dalam menentukan parameter yang optimal agar akurasi dapat dimaksimalkan.

Faktor lain yang berkontribusi terhadap akurasi adalah ukuran dan distribusi dataset. Model pembelajaran mesin membutuhkan jumlah

data yang cukup untuk dapat mengenali pola dengan baik. Jika dataset terlalu kecil atau tidak representatif, model akan kesulitan dalam melakukan generalisasi terhadap data baru, sehingga menyebabkan penurunan akurasi. Selain itu, distribusi data yang tidak seimbang (imbalanced dataset), seperti ketika jumlah individu dalam kategori "Layak" jauh lebih sedikit dibandingkan kategori "Tidak Layak", dapat menyebabkan model lebih cenderung mengklasifikasikan mayoritas dan mengabaikan minoritas. Dalam situasi ini, teknik seperti resampling (oversampling atau undersampling) dan penggunaan metrik evaluasi yang lebih tepat seperti F1-score dapat membantu meningkatkan performa model.

Terakhir, proses validasi dan evaluasi model juga berperan dalam menentukan seberapa baik model mampu menggeneralisasi data baru. Penggunaan metode train-test split atau k-fold cross-validation dapat membantu mengukur performa model secara lebih akurat dan menghindari bias yang mungkin muncul akibat data yang tidak terdistribusi dengan baik. Selain itu, pemantauan metrik seperti akurasi, precision, recall, dan F1-score sangat penting untuk memahami kelebihan dan kelemahan dari setiap algoritma yang digunakan.

Dengan mempertimbangkan faktor-faktor tersebut, sistem klasifikasi dalam distribusi bantuan di desa dapat dioptimalkan agar mampu memberikan hasil yang lebih akurat dan adil. Integrasi teknik pemrosesan data yang baik, seleksi fitur yang tepat, pemilihan parameter optimal, serta evaluasi yang menyeluruh akan memastikan bahwa model yang diterapkan benar-benar mampu mengidentifikasi individu yang berhak menerima bantuan secara efisien dan tepat sasaran.

6

KESIMPULAN DAN REKOMENDASI

Ringkasan Efektivitas Algoritma

Dalam penelitian ini, telah dilakukan analisis terhadap efektivitas berbagai algoritma dalam klasifikasi penerima bantuan sosial, dengan fokus pada dua metode utama, yaitu Naïve Bayes dan K-Nearest Neighbors (KNN). Studi ini menunjukkan bahwa masing-masing algoritma memiliki keunggulan dan kelemahan yang berbeda dalam menangani data klasifikasi, terutama terkait dengan akurasi, efisiensi komputasi, serta kemampuan menangani pola hubungan antar fitur dalam dataset. Oleh karena itu, pemilihan algoritma yang tepat sangat bergantung pada karakteristik data serta tujuan utama dari sistem klasifikasi yang digunakan dalam distribusi bantuan sosial.

Dari hasil analisis yang dilakukan, Naïve Bayes terbukti lebih efisien secara komputasi, dengan waktu pemrosesan yang lebih cepat dibandingkan KNN. Hal ini disebabkan oleh pendekatan probabilistiknya yang memungkinkan pemrosesan data dalam jumlah

besar tanpa memerlukan perhitungan jarak antar sampel. Namun, kekurangan utama dari Naïve Bayes adalah asumsi independensi antar fitur, yang sering kali tidak sesuai dengan kondisi data nyata. Misalnya, dalam sistem distribusi bantuan sosial, faktor seperti penghasilan rumah tangga dan jumlah tanggungan tidak dapat dianggap sepenuhnya independen, sehingga dapat menyebabkan kesalahan dalam klasifikasi.

Di sisi lain, KNN menawarkan fleksibilitas yang lebih besar dalam menangani data dengan pola yang lebih kompleks, karena pendekatan berbasis kedekatan yang digunakan memungkinkan algoritma ini menangkap hubungan non-linear antar variabel. Meskipun demikian, KNN memiliki kelemahan utama dalam hal efisiensi komputasi, terutama ketika jumlah data yang digunakan semakin besar. Selain itu, pemilihan nilai K yang optimal menjadi faktor krusial yang dapat memengaruhi kinerja algoritma ini. Nilai K yang terlalu kecil dapat menyebabkan model terlalu peka terhadap noise dalam data, sedangkan nilai K yang terlalu besar dapat mengurangi sensitivitas model terhadap pola-pola minoritas yang penting.

Berdasarkan hasil perbandingan antara kedua algoritma, pendekatan hybrid antara Naïve Bayes dan KNN telah dieksplorasi sebagai solusi potensial untuk mengatasi keterbatasan masing-masing metode. Pendekatan ini bertujuan untuk menggabungkan efisiensi Naïve Bayes dengan fleksibilitas KNN, sehingga menghasilkan sistem klasifikasi yang lebih akurat dan efisien. Hasil eksperimen menunjukkan bahwa metode hybrid mampu meningkatkan akurasi dibandingkan metode tunggal, khususnya dalam menangani data yang memiliki hubungan kompleks antar fitur serta ketidakseimbangan kelas.

Rekomendasi Penerapan Berbasis Machine Learning

Perkembangan teknologi berbasis machine learning (ML) telah membuka peluang baru dalam meningkatkan efisiensi dan akurasi distribusi bantuan sosial di tingkat desa. Dengan kemampuan ML dalam mengolah data dalam jumlah besar, mengidentifikasi pola, serta membuat keputusan berbasis analisis statistik, penerapan sistem berbasis ML dapat membantu pemerintah desa dalam memastikan bahwa bantuan diberikan kepada penerima yang benar-benar membutuhkan. Berdasarkan hasil penelitian ini, terdapat beberapa rekomendasi yang dapat diterapkan oleh pemerintah desa guna mengoptimalkan penggunaan sistem berbasis ML dalam pembagian bantuan sosial.

1. Peningkatan Kualitas Data dan Infrastruktur Digital
   Keberhasilan sistem berbasis ML sangat bergantung pada kualitas data yang digunakan. Oleh karena itu, pemerintah desa perlu memastikan bahwa data penerima bantuan sosial yang dikumpulkan terstruktur, akurat, dan terbaru. Proses verifikasi dan validasi data harus dilakukan secara berkala untuk menghindari kesalahan dalam klasifikasi penerima bantuan. Selain itu, digitalisasi basis data desa menjadi langkah krusial untuk memastikan kelancaran pemrosesan data oleh algoritma ML. Pemerintah desa dapat berinvestasi dalam pengembangan sistem manajemen data berbasis cloud, yang memungkinkan akses dan pembaruan data secara real-time.
2. Pemilihan Algoritma yang Sesuai dengan Karakteristik Data
   Dalam penelitian ini, telah dilakukan analisis terhadap Naïve Bayes dan K-Nearest Neighbors (KNN) sebagai metode utama dalam klasifikasi penerima bantuan. Naïve Bayes cocok digunakan jika data yang dimiliki memiliki jumlah fitur yang besar dan relatif independen, sedangkan KNN lebih optimal

untuk data yang memiliki hubungan kompleks antar variabel. Pemerintah desa dapat mempertimbangkan pendekatan hybrid, yang menggabungkan keunggulan kedua metode ini untuk meningkatkan akurasi klasifikasi. Selain itu, eksplorasi metode lain seperti Decision Tree, Random Forest, atau Deep Learning dapat menjadi opsi lebih lanjut untuk meningkatkan performa sistem.

3. Pelatihan dan Pengembangan Kapasitas SDM
   Agar sistem berbasis ML dapat diimplementasikan dengan baik, pemerintah desa perlu menyediakan pelatihan bagi aparatur desa terkait penggunaan dan pemeliharaan sistem ini. Pelatihan dapat mencakup pemahaman dasar mengenai ML, pengelolaan data, interpretasi hasil prediksi, serta langkah-langkah perbaikan apabila terjadi kesalahan dalam klasifikasi. Kolaborasi dengan universitas, lembaga riset, atau sektor swasta yang memiliki keahlian dalam teknologi ML dapat menjadi solusi dalam meningkatkan kapasitas SDM di desa.
4. Penguatan Transparansi dan Pengawasan
   Salah satu tantangan utama dalam distribusi bantuan sosial adalah minimnya transparansi dalam penentuan penerima. Dengan adanya sistem berbasis ML, pemerintah desa dapat meningkatkan transparansi melalui penyediaan informasi berbasis data yang dapat diakses oleh publik. Misalnya, pemerintah desa dapat menyediakan dashboard interaktif yang menampilkan data penerima bantuan berdasarkan kriteria objektif yang telah ditetapkan oleh sistem. Selain itu, perlu adanya mekanisme pengawasan dan audit berkala, guna memastikan bahwa sistem tetap berjalan sesuai dengan prinsip keadilan dan keberpihakan kepada masyarakat yang benar-benar membutuhkan.
5. Implementasi Bertahap dan Evaluasi Berkelanjutan
   Penerapan sistem berbasis ML di desa sebaiknya dilakukan secara bertahap, dimulai dengan uji coba dalam skala kecil sebelum

diterapkan secara menyeluruh. Evaluasi berkala terhadap akurasi model, efisiensi sistem, serta dampak yang dihasilkan perlu dilakukan untuk mengidentifikasi aspek yang perlu diperbaiki. Pemerintah desa juga dapat mengumpulkan umpan balik dari masyarakat untuk menyesuaikan sistem dengan kondisi nyata di lapangan.

Secara keseluruhan, penerapan sistem berbasis ML dalam distribusi bantuan sosial memiliki potensi besar dalam meningkatkan efektivitas, efisiensi, dan transparansi. Namun, keberhasilannya sangat bergantung pada kualitas data, pemilihan metode yang tepat, kesiapan SDM, serta dukungan infrastruktur digital. Dengan strategi yang matang dan evaluasi yang berkelanjutan, sistem ini dapat menjadi alat yang mendukung pemerataan kesejahteraan masyarakat secara lebih adil dan akurat.

Potensi Penelitian Lanjutan dengan Menerapkan Algoritma Lain

Penelitian lanjutan dalam bidang klasifikasi penerima bantuan sosial dapat dieksplorasi lebih jauh dengan menerapkan algoritma Decision Tree atau Random Forest untuk meningkatkan akurasi dan ketepatan klasifikasi. Algoritma Decision Tree memiliki keunggulan dalam memahami pola data secara hierarkis, di mana setiap keputusan dibuat berdasarkan atribut yang paling signifikan dalam menentukan status penerima bantuan. Dengan struktur berbasis pohon keputusan, model ini dapat menangani data yang tidak terstruktur dan memberikan interpretasi yang lebih mudah dibandingkan metode berbasis probabilitas seperti Naïve Bayes.

Sementara itu, Random Forest sebagai pengembangan dari Decision Tree dapat meningkatkan stabilitas dan akurasi prediksi dengan menggabungkan beberapa pohon keputusan sekaligus. Dengan teknik bagging (bootstrap aggregating), Random Forest mengurangi overfitting yang sering terjadi pada Decision Tree

tunggal, sehingga menghasilkan klasifikasi yang lebih robust dan adaptif terhadap perubahan pola data. Hal ini sangat relevan dalam konteks bantuan sosial, di mana kondisi ekonomi masyarakat dapat berfluktuasi berdasarkan faktor eksternal seperti inflasi, kebijakan pemerintah, dan kondisi lapangan kerja.

Selain itu, penelitian lanjutan dapat mengkaji kombinasi algoritma ini dengan metode lain, seperti Hybrid Naïve Bayes-KNN atau ensemble learning, untuk meningkatkan daya prediksi sistem klasifikasi. Dengan demikian, penerapan Decision Tree dan Random Forest tidak hanya berpotensi meningkatkan akurasi klasifikasi tetapi juga dapat mengoptimalkan pengambilan keputusan dalam distribusi bantuan sosial secara lebih efisien dan tepat sasaran.

DAFTAR PUSTAKA

1. Hidayat, T., Munthe, I. R., & Juledi, A. P., "Analisis Data Penjualan Menggunakan Algoritma Apriori pada Analisis Kopi," *INFORMATIKA*, vol. 12, no. 3, p. 56, 2024, https://doi.org/10.36987/informatika.v12i3.6064.
2. D. S. O. Panggabean, E. Buulolo, and N. Silalahi, "Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda," *Jurikom (Jurnal Ris. Komputer)*, vol. 7, no. 1, p. 56, 2020, doi: 10.30865/jurikom.v7i1.1947.
3. I. A. N. Afifah, "Data Mining Clustering Dalam Pengelompokan Buku Perpustakaan Mengunakan Algoritma K-Means," *Jipi (Jurnal Ilm. Penelit. Dan Pembelajaran Inform.*, vol. 8, no. 3, pp. 802–814, 2023, doi: 10.29100/jipi.v8i3.3891.
4. D. A. Fakhri, S. Defit, and S. Sumijan, "Optimalisasi Pelayanan Perpustakaan Terhadap Minat Baca Menggunakan Metode *K-Means* Clustering," *J. Inf. Dan Teknol.*, pp. 160–166, 2021, doi: 10.37034/jidt.v3i3.137.
5. N. Asiah, "Evaluasi Penerapan Sistem Manajemen Keselamatan dan Kesehatan Kerja (SMK3) di Rumah Sakit Umum Daerah dr. Zainoel Abidin Banda Aceh," Universitas Islam Negeri Ar-Raniry, 2020.
6. Amansyah, R., Masrizal, M., & Munthe, I. R, "Pengimplementasian Tingkat Ketepatan Waktu Kelulusan Siswa (Studi Kasus Di MTS Nur Ibarhimy) Menggunakan Algoritma C4.5," INFORMATIKA,

12(2), 280-291,2024, https://doi.org/10.36987/informatika.v12i2.5767

7. A. Maulana, A. Nugroho, and I. Romli, "Optimalisasi Support Vector Machine Menggunakan Particle Swarm Optimization Untuk Mendiagnosa Penyakit Kanker Payudara," *J. Pract. Comput. Sci.*, vol. 1, no. 2, pp. 1–11, 2022, doi: 10.37366/jpcs.v1i2.940.
8. R. Fitra, "Penerapan Metode Algoritma *K-Nearest Neighbor* Menggunakan Rapidminer Studio Pada Klasifikasi Status Sosial Ekonomi Studi Kasus : Kelurahan Kapuk Muara Rt 010 Rw 04," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 11, no. 4, 2022, doi: 10.30591/smartcomp.v11i4.4250.
9. A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.
10. D. Syaputri, P. H. Noprita, and S. Romelah, "Implementasi Algoritma *K-Means* Untuk Pengelompokan Distribusi Sosial Ekonomi Masyarakat Berdasarkan Demografi Kependudukan," *Malcom Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 1–6, 2021, doi: 10.57152/malcom.v1i1.5.
11. P. Simanjuntak, C. E. Suharyanto, S. Sitohang, and K. Handoko, "Data Mining Untuk Klasifikasi Status Pandemi Covid 19," *J. Tek. Inf. dan Komput.*, vol. 5, no. 2, p. 327, 2022, doi: 10.37600/tekinkom.v5i2.620.
12. N. Apriliani, "Analisis Sentimen Review Penggunaan Tiktok Melalui Pendekatan Algoritma Naïve Bayes," *Jati (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3725–3731, 2024, doi: 10.36040/jati.v7i6.8299.
13. H. Sastypratiwi, Y. Yulianti, and H. Muhardi, "Uji Komparasi Algoritma *Naïve Bayes* Dan Decision Tree Classification

Menggunakan Covid-19 Dataset," *J. Edukasi Dan Penelit. Inform.*, vol. 8, no. 1, p. 1, 2022, doi: 10.26418/jp.v8i1.49841.

14. R. A. Rizal, N. O. Purba, L. A. Siregar, K. P. Sinaga, and N. Azizah, "Analysis of Tuberculosis (TB) on X-Ray Image Using SURF Feature Extraction and the *K-Nearest Neighbor* (KNN) Classification Method," *Jaict*, vol. 5, no. 2, p. 9, 2020, doi: 10.32497/jaict.v5i2.1979.
15. M. Pagan, "Investigating the Impact of Data Scaling on the K-Nearest Neighbor Algorithm," *Comput. Sci. Inf. Technol.*, vol. 4, no. 2, pp. 135–142, 2023, doi: 10.11591/csit.v4i2.pp135-142.
16. G. T. Hariyadi, D. Aqmala, A. B. Setiawan, and I. Farida, "Pelatihan Query Excel Untuk Pengelolaan Data Administrasi Kependidikan Pada TK. Isriati Baiturahman 1 Pandanaran Semarang," *Abdimasku J. Pengabdi. Masy.*, vol. 5, no. 1, p. 20, 2022, doi: 10.33633/ja.v5i1.286.
17. A. A. A. Ushud, "Pelatihan Microsoft Excel Tingkat Lanjut Karyawan Pt. Nutrisi Juara Asia," *Kresna J. Ris. Dan Pengabdi. Masy.*, vol. 3, no. 1, pp. 86–94, 2023, doi: 10.36080/kresna.v3i1.54.
18. M. Nasution, A. A. Ritonga, and A. P. Juledi, "Implementasi Rapidminer Dalam Mengklasifikasikan Indeks Demokrasi," *J. Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 99–106, 2022.
19. S. A. Nazli, "Coronary Risk Factor Profiles According to Different Age Categories in Premature Coronary Artery Disease Patients Who Have Undergone Percutaneous Coronary Intervention," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-53539-6.
20. A. S. Elbhrawy, "AIRA-ML: Auto Insurance Risk Assessment-Machine Learning Model Using Resampling Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, 2023, doi: 10.14569/ijacsa.2023.0140966.

TENTANG PENULIS

Nurhikmah Wulandari lahir di Rantau Prapat pada 22 November 2002. Sejak kecil menempuh pendidikan dasar hinga menengah di Marbau, Labuhanbatu Utara, sebelum akhirnya melanjutkan studi di jenjang perguruan tinggi. Pernah menjadi bagian dari mahasiswa S1 Sistem Informasi di Universitas Labuhanbatu sejak 2021-2025 dengan menemukan minat bakatnya dalam bidang teknologi, komputer dan analis data.

Ketertarikannya terhadap dunia teknologi, khususnya dalam kecerdasan buatan, data mining, dan sistem pengambilan keputusan berbasis machine learning, mendorongnya untuk terus mengembangkan wawasan dan berkontribusi dalam dunia akademik. Hal ini dibuktikan dengan karya-karya yang telah dihasilkan, salah satunya adalah buku pertamanya yang berjudul "Pengantar Probabilitas", yang tidak hanya menjadi referensi akademik, tetapi juga mengantarkan penulis meraih penghargaan dalam International Student Competition (ISC) 2023 atas kontribusinya dalam bidang artikel ilmiah.

Sebagai seorang yang selalu ingin memperdalam pemahaman dalam analisis data dan machine learning, penulis kembali menghadirkan buku keduanya yang berjudul "Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor untuk Klasifikasi Penerima Bantuan Sosial di Desa Marbau Selatan". Buku ini lahir dari minat dan

perhatiannya terhadap pemanfaatan teknologi dalam sektor sosial, khususnya dalam pengelolaan dan klasifikasi data penerima bantuan sosial yang lebih efektif dan akurat.

Dengan semangat untuk terus belajar dan berbagi ilmu, penulis berharap buku ini dapat menjadi referensi yang bermanfaat bagi akademisi, mahasiswa, serta praktisi di bidang teknologi informasi dan analisis data, serta membuka peluang untuk penelitian lebih lanjut dalam pengembangan sistem berbasis kecerdasan buatan di berbagai sektor.

Ibnu Rasyid Munthe, S.T., M.Kom lahir di Kota Rantau Prapat, Kabupaten Labuhanbatu, pada tahun 1987. Beliau memperoleh gelar Sarjana Teknik (S.T.) dari Universitas Nurtanio Bandung dan melanjutkan pendidikan Magister di Universitas Putra Indonesia (YPTK) Padang dengan gelar Magister Komputer (M.Kom.). Beliau aktif dalam menulis artikel ilmiah serta buku akademik. Saat ini, beliau berkiprah sebagai dosen di Universitas Labuhanbatu. Selain itu, beliau juga mengelola kanal YouTube @ manjaddawajada2022 sebagai media pembelajaran dan sarana berbagi ilmu pengetahuan.

Angga Putra Juledi, S.Kom., M.Kom Lahir di Kota Padang pada tanggal 19 Juli 1994. Dalam menempuh Pendidikan dimulai dari Sekolah Dasar SDN 19 Padang tamat tahun 2006, SMPN 3 Padang tamat pada tahun 2009, dan di SMA Pertiwi 2 Padang tamat pada tahun 2012. Lalu melanjutkan ke pendidikan perguruan tinggi swasta yaitu S1 (Sarjana) Universitas Putra Indonesia "YPTK" Padang lulus pada tahun 2018 dengan jurusan Sistem Informasi, Dan melanjutkan Program

Pascasarjana (S2) di Universitas Putra Indonesia "YPTK" Padang pada tahun 2019 Program Studi Teknik Informatika. Konsentrasi Sistem Informasi. Saya mengabdikan diri sebagai salah satu Dosen di bidang Ilmu Komputer pada Fakultas Sains Dan Teknologi dengan Program Studi Sistem Informasi di Universitas LabuhanBatu dan menjadi dosen tetap pada tahun 2020 pada kampus tersebut. Paat ini menjadi bagian Struktural di Universitas Labuhanbatu sebagai Kepala Bagian Sumber Daya Manusia periode 2023 s/d 2027. Buku pertama terbit pada 31 Desember 2021 dengan judul Internetworking Dan TCP/IP. Buku kedua terbit pada tanggal 17 Oktober 2023 dengan judul Panduan Belajar HTML, CSS, dan JavaScript untuk Pemula. Hingga saat ini masih menulis buku setiap tahunnya.

Marnis Nasution, S.Kom., M.Kom Lahir di Bengkulu 30 maret 1990. Selama sekolah dasar sampai menengah ditempuh di kota Bengkulu. Melanjutkan Pendidikan tinggi strata-1 dan strata-2 di Universitas Putra Indonesia "YPTK" Padang dari tahun 2008 sampai 2024 dengan jurusan Sistem Informasi. Saat ini aktif menjadi Dosen Yayasan di Universitas Labuhanbatu, Sumatera Utara dan menulis beberapa karya Ilmiah dan buku.