Penerapan Naive Bayes Untuk Klasifikasi Penyakit Endokrin Pada Pasien Lansia

Susannda Dwi Yulianti Kusuma[1], Hidayatullah Al Islami[2], dan Perani Rosyani[3]*

[1,2,3]Fakultas Ilmu Komputer Prgram Studi Teknik Informatikan, Universitas Pamulang

** Penulis Korespondensi : , dosen00837@unpam.ac.id*

ABSTRACT

The application of the Naïve Bayes algorithm has shown great potential in the classification of endocrine diseases in elderly patients. This study aims to develop a classification model using the algorithm, utilizing diabetes-related data obtained from public datasets. The process involves data collection, preprocessing, model training, and evaluation using Orange software. The results show that Naïve Bayes' algorithm is able to achieve high accuracy in data classification. The implementation of this model is expected to be a medical decision support system for faster and more accurate diagnosis, as well as improve the efficiency of health services for elderly patients. The advantages of this method lie in its ease of use, time efficiency, and intuitive visualization capabilities, making it an effective tool in medical data analysis. The main advantages of this approach include ease of implementation, time efficiency in data analysis, and the ability to visualize intuitively through a software interface. Thus, this research not only contributes to the development of health technology but also opens up opportunities for further integration with more complex AI-based health information systems. The adoption of this model is expected to be able to encourage the improvement of the quality of health services in the future.

Article History

Received : 10-10-2024
Revised : 10-11-2024
Accepted : 25-12-2024

Keywords

Naive Bayes
Classification
Endocrine Diseases

ABSTRAK

Penerapan algoritma Naïve Bayes telah menunjukkan potensi besar dalam klasifikasi penyakit endokrin pada pasien lanjut usia. Penelitian ini bertujuan untuk mengembangkan model klasifikasi menggunakan algoritma tersebut, memanfaatkan data terkait penyakit diabetes yang diperoleh dari dataset publik. Proses melibatkan pengumpulan data, preprocessing, pelatihan model, dan evaluasi menggunakan perangkat lunak Orange. Hasil menunjukkan bahwa algoritma Naïve Bayes mampu mencapai akurasi tinggi dalam klasifikasi data. Implementasi model ini diharapkan dapat menjadi sistem pendukung keputusan medis untuk diagnosis yang lebih cepat dan akurat, sekaligus meningkatkan efisiensi layanan kesehatan bagi pasien lansia. Keunggulan metode ini terletak pada kemudahan penggunaan, efisiensi waktu, dan kemampuan visualisasi yang intuitif, menjadikannya alat yang efektif dalam analisis data medis. Keunggulan utama dari pendekatan ini mencakup kemudahan implementasi, efisiensi waktu dalam analisis data, serta kemampuan untuk divisualisasikan secara intuitif melalui antarmuka perangkat lunak. Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan teknologi kesehatan tetapi juga membuka peluang untuk integrasi lebih lanjut dengan sistem informasi kesehatan berbasis AI yang lebih kompleks. Adopsi model ini diharapkan mampu mendorong perbaikan kualitas layanan kesehatan di masa depan.

PENDAHULUAN

Sistem endokrin merupakan kumpulan kelenjar yang memiliki peran penting dalam menjaga keseimbangan fungsi tubuh manusia. Kelenjar-kelenjar ini menghasilkan hormon yang disekresikan langsung ke dalam darah dan berperan dalam berbagai aktivitas tubuh, seperti metabolisme, pertumbuhan, perkembangan seksual, dan fungsi reproduksi. Salah satu gangguan utama pada sistem endokrin adalah diabetes melitus, yang prevalensinya semakin meningkat, khususnya di kalangan lansia di Indonesia. Menurut laporan Kementerian Kesehatan tahun 2023, prevalensi diabetes pada usia di atas 15 tahun mencapai 11,7%, dengan tingkat kejadian yang lebih tinggi pada kelompok lansia.

Permasalahan utama dalam penanganan diabetes melitus pada lansia adalah keterlambatan diagnosis serta rendahnya angka kepatuhan terhadap pengobatan. Hanya sekitar 6,06% dari penderita yang menjalani pengobatan, dan kurang dari separuhnya melakukan kunjungan ulang untuk perawatan lanjutan. Untuk mengatasi tantangan ini, pemanfaatan kecerdasan buatan (AI) dalam diagnosis dini dapat menjadi solusi yang efektif. Salah satu metode dalam AI yang dapat digunakan untuk klasifikasi penyakit endokrin adalah Naïve Bayes, yang berbasis probabilistik dan sering diterapkan dalam diagnosis medis.

Naïve Bayes adalah salah satu metode klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dalam pengambilan keputusan. Metode ini mengasumsikan bahwa setiap fitur dalam dataset bersifat independen satu sama lain, sehingga perhitungannya lebih sederhana dibandingkan metode lainnya. Dalam diagnosis penyakit endokrin, Naïve Bayes dapat mengolah data pasien, seperti kadar gula darah, tekanan darah, indeks massa tubuh (BMI), dan faktor risiko lainnya untuk menentukan probabilitas seseorang menderita diabetes atau gangguan endokrin lainnya.[1]

Kelebihan metode ini *c*epat dalam melakukan klasifikasi, bahkan pada dataset besar. Memiliki performa yang baik dalam kondisi fitur yang independen. Tidak memerlukan banyak data pelatihan untuk mendapatkan hasil yang akurat. Mudah diimplementasikan dan memiliki interpretasi hasil yang jelas. [2]

Masalah utama yang dihadapi dalam diagnosis penyakit endokrin, khususnya diabetes pada lansia, adalah keterlambatan deteksi dan kurangnya alat bantu diagnosis yang efisien di fasilitas kesehatan tingkat pertama (FKTP). Selain itu, tantangan dalam penerapan AI dalam dunia medis meliputi kurangnya standar yang jelas, risiko bias algoritma, serta kekhawatiran publik terkait privasi data kesehatan.

Sebagai solusi, penerapan metode Naïve Bayes dalam sistem pakar berbasis AI dapat membantu dokter dalam mengidentifikasi tanda-tanda awal diabetes dan gangguan endokrin lainnya secara lebih cepat dan akurat. Dengan pengolahan data pasien yang lebih sistematis, diharapkan metode ini dapat memberikan rekomendasi yang tepat dalam perencanaan pengobatan dini.

Penelitian ini bertujuan untuk menerapkan metode Naïve Bayes dalam klasifikasi penyakit endokrin pada pasien lansia guna meningkatkan efektivitas diagnosis dini. Dengan adanya penelitian ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai efektivitas Naïve Bayes dalam diagnosis dini penyakit endokrin serta bagaimana teknologi ini dapat diintegrasikan dalam sistem layanan kesehatan untuk meningkatkan kualitas perawatan pasien lansia.

TINJAUAN PUSTAKA

Naive Bayes

Naive Bayes merupakan salah satu metode klasifikasi yang paling banyak digunakan dan dikenal karena memiliki tingkat akurasi yang tinggi. Algoritma ini telah menjadi subjek dari berbagai penelitian dalam bidang klasifikasi.[3] Tidak seperti metode klasifikasi lainnya, seperti logistic regression untuk data ordinal maupun nominal, algoritma Naive Bayes tidak memerlukan pemodelan yang rumit atau pengujian statistik untuk digunakan.[4]

Metode Naive Bayes didasarkan pada prinsip probabilitas sederhana dan dirancang untuk bekerja dengan asumsi bahwa variabel-variabel penjelas bersifat independen satu sama lain. Proses pembelajaran pada algoritma ini lebih menekankan pada penghitungan probabilitas. Salah satu keunggulan utama dari algoritma Naive Bayes adalah kemampuannya untuk menghasilkan tingkat kesalahan yang lebih rendah, terutama ketika digunakan pada dataset berukuran besar. Selain itu, algoritma ini juga memiliki kecepatan dan akurasi yang lebih tinggi dibandingkan metode lain saat diterapkan pada data dengan jumlah besar.

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i2.7312

Pengertian Penyakit Endokrin

Gangguan pada sistem endokrin adalah kondisi kesehatan yang memengaruhi fungsi normal sistem endokrin dalam tubuh. Sistem endokrin sendiri terdiri dari delapan kelenjar utama yang bertugas memproduksi hormon, seperti kelenjar tiroid, kelenjar pituitari, kelenjar adrenal, kelenjar timus, dan pankreas. Sistem ini memiliki peran penting dalam mengatur berbagai proses vital tubuh, termasuk pertumbuhan, perkembangan, metabolisme, fungsi reproduksi, dan bahkan suasana hati.[5]

Gangguan pada sistem ini biasanya terjadi ketika kadar hormon yang dihasilkan oleh kelenjar terlalu tinggi atau terlalu rendah. Ketidakseimbangan ini menjadi indikasi adanya masalah atau penyakit pada sistem endokrin. Selain itu, gangguan endokrin juga dapat muncul ketika tubuh tidak mampu merespons hormon secara normal meskipun kadarnya sudah sesuai.[6]

Beberapa jenis gangguan sistem endokrin cukup sering ditemukan. Salah satu yang paling umum adalah diabetes mellitus, penyakit kronis yang ditandai dengan tingginya kadar gula dalam darah akibat gangguan pada produksi atau penggunaan insulin. Selain diabetes, terdapat gangguan lain seperti hipertiroidisme dan hipotiroidisme, yang disebabkan oleh ketidakseimbangan hormon tiroid akibat masalah pada kelenjar tiroid. Gangguan ini dapat menyebabkan berbagai gejala, mulai dari perubahan berat badan, kelelahan, hingga gangguan fungsi tubuh lainnya.

Gangguan endokrin lainnya meliputi sindrom ovarium polikistik (PCOS), yang sering terjadi pada wanita dan memengaruhi fungsi ovarium, serta kondisi seperti akromegali, yang disebabkan oleh produksi hormon pertumbuhan yang berlebih. Selain itu, ada juga sindrom Cushing, yang diakibatkan oleh tingginya kadar hormon kortisol dalam tubuh.

Gangguan-gangguan ini dapat berdampak signifikan pada kualitas hidup penderitanya jika tidak dikelola dengan baik. Oleh karena itu, deteksi dini dan pengelolaan yang tepat menjadi langkah penting untuk mencegah komplikasi lebih lanjut. Perawatan gangguan sistem endokrin umumnya melibatkan pendekatan holistik, termasuk pengobatan medis, perubahan gaya hidup, serta pemantauan rutin terhadap kadar hormon dalam tubuh. (Puji, 2022)

Gangguan endokrin merupakan kondisi kesehatan yang berkaitan dengan kelenjar endokrin dalam tubuh. Sistem endokrin sendiri terdiri dari jaringan kelenjar yang bertugas memproduksi hormon, yaitu senyawa kimia yang berfungsi sebagai pengirim sinyal melalui aliran darah. Hormon ini berperan penting dalam mengatur berbagai fungsi tubuh, termasuk pengendalian nafsu makan, proses pernapasan, pertumbuhan tubuh, keseimbangan cairan, pembentukan karakteristik seksual sekunder seperti pembesaran payudara atau testis (feminisasi dan virilisasi), hingga menjaga kestabilan berat badan.

Sistem endokrin memainkan peranan yang sangat vital dalam menjaga keseimbangan internal tubuh, meskipun ada berbagai perubahan eksternal yang terjadi. Namun, ketika terjadi gangguan pada kelenjar endokrin, sistem ini tidak dapat menjalankan fungsinya dengan baik, yang kemudian berdampak pada kesehatan secara menyeluruh.

METODE

Penyakit endokrin merupakan salah satu penyekit kelompok yang sering ditemukan pada pasien lanjut usia (lansia). Penyakit ini beberapa melibatkan gangguan pada sistem endokrin, yang bertugas mengatur hormon dalam di dalam tubuh. Pada artikel ini pembahasan system endokrin yang akan di bahas ialah penyakit Diabetes. Deteksi dini dan klasifikasi penyakit endokrin sangat penting untuk menentukan dan menemukan langkah pengobatan yang tepat. Dalam era digital, penerapan teknologi kecerdasan buatan (Artificial Intelligence/AI) telah membantu proses klasifikasi menjadi cepat dan akurat. Salah satu metode yang sering digunakan adalah algoritma Naïve Bayes, yang dapat diterapkan menggunakan perangkat lunak yaitu Orange.

Pada tahap pengolahan data dalam Data Science, berbagai algoritma diterapkan untuk mendukung proses analisis dan pemrosesan data. Salah satu cabang penting dalam Data Science adalah Data Mining, yang berfokus pada identifikasi pola-pola tertentu dari kumpulan data. Berikut adalah beberapa algoritma utama yang sering digunakan dalam proses tersebut:

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i2.7312

1. NaïveBayes
   Naive Bayes adalah algoritma klasifikasi yang didasarkan pada prinsip probabilitas dan statistik. Algoritma ini memanfaatkan data historis untuk memperkirakan peluang atau kemungkinan kejadian di masa depan. Dengan pendekatan berbasis probabilitas, Naive Bayes sering digunakan dalam berbagai aplikasi seperti analisis sentimen, klasifikasi teks, dan sistem rekomendasi.
2. C4.5(DecisionTree)
   C4.5, yang sering disebut sebagai Decision Tree, adalah algoritma populer untuk membuat model prediksi berbasis pohon keputusan. Prosesnya dimulai dari titik awal atau root node, yang kemudian bercabang menjadi keputusan-keputusan berdasarkan atribut data. Algoritma ini banyak digunakan dalam bahasa pemrograman seperti R untuk membantu pengambilan keputusan dalam berbagai situasi, mulai dari analisis risiko hingga diagnosa medis.
3. K-MeansClustering
   K-Means adalah algoritma pengelompokan data yang bersifat non-hirarki. Algoritma ini membagi data ke dalam beberapa kelompok atau cluster berdasarkan kesamaan karakteristik. Data dengan atribut serupa akan dikelompokkan dalam satu cluster, sementara data dengan atribut yang berbeda akan ditempatkan dalam cluster lainnya. K-Means sering digunakan dalam segmentasi pasar, analisis pola belanja, dan pengelompokan data besar.[7]



Gambar 1. Tampilan Orange

Dengan penerapan algoritma-algoritma ini, Data Science mampu menemukan wawasan dan pola yang sebelumnya tidak terlihat, sehingga mendukung pengambilan keputusan yang lebih efektif dan berbasis data.

Aplikasi yang akan di gunakan dalam mengolah data pada artikel ini ialah aplikasi Orange. Orange adalah perangkat lunak open-source untuk analisis data dan pembelajaran mesin yang dirancang dengan antarmuka berbasis drag and drop. Orange memungkinkan pengguna untuk memproses data, membangun model pembelajaran mesin, dan mengevaluasi performanya tanpa memerlukan keterampilan pemrograman yang mendalam. Artikel ini menjelaskan metode penyelesaian menggunakan Orange dalam penerapan Naïve Bayes untuk klasifikasi penyakit endokrin pada pasien lansia.

Tahapan Penyelesaian Menggunakan Orange:

1. Pengumpulan Data

Tahap pertama adalah pengumpulan data yang relevan. Data yang digunakan untuk penelitian ini dapat berasal dari rekam medis rumah sakit atau dataset publik yang tersedia

yang di ambil dari Kaggale.com terkait penyakit Diabetes. Dataset tersebut mencakup atribut seperti:

- Usia pasien.
- Jenis kelamin.
- Gejala klinis.

2. Preprocessing Data

Setelah data terkumpul, langkah selanjutnya adalah melakukan preprocessing menggunakan widget *File* di Orange. Tahap ini bertujuan untuk memastikan data bersih dan siap digunakan. Langkah-langkahnya meliputi:

- Mengatasi Nilai Kosong: Nilai yang hilang (*missing values*) diatasi dengan metode imputasi, seperti mengganti dengan rata-rata atau median untuk data numerik.
- Normalisasi Data: Data dinormalisasi untuk memastikan semua atribut berada pada skala yang sama.
- Pemisahan Atribut: Atribut fitur dipisahkan dari atribut target untuk memastikan proses pelatihan model berjalan dengan benar.

3. . Pemilihan Algoritma Naïve Bayes

Naïve Bayes adalah algoritma yang pembelajaran mesin sangat sederhana namun efektif, terutama untuk masalah klasifikasi. Algoritma ini didasarkan pada Teorema Bayes dengan asumsi bahwa semua fitur bersifat independen. Dalam Orange, widget *Naïve Bayes* digunakan untuk memilih algoritma ini.

4. Pelatihan Model

Pada tahap ini, dataset dibagi menjadi data pelatihan dan data pengujian. Pembagian ini dilakukan menggunakan widget *Data Sampler* atau *Test & Score* di Orange. Data pelatihan digunakan untuk membangun model prediktif berdasarkan algoritma Naïve Bayes. Proses ini melibatkan perhitungan probabilitas untuk setiap kategori dalam atribut target berdasarkan fitur yang diberikan.

5. Evaluasi Model

Setelah model dilatih, evaluasi dilakukan untuk mengukur kinerjanya. Beberapa metrik yang digunakan antara lain:

- Akurasi: Persentase prediksi yang benar dibandingkan dengan total data.
- Presisi: Ketepatan model dalam mengidentifikasi kasus positif.
- Recall: Kemampuan model untuk menemukan semua kasus positif.
- F1-Score: Rata-rata harmonis antara presisi dan recall.

Widget *Confusion Matrix* digunakan untuk melihat distribusi prediksi model, sedangkan *ROC Analysis* digunakan untuk memvisualisasikan kurva ROC dan menghitung nilai Area Under Curve (AUC).

6. Visualisasi dan Interpretasi Hasil

Hasil klasifikasi ini divisualisasikan dalam bentuk grafik dan tabel menggunakan fitur widget seperti *Scatter Plot*, *Box Plot*, atau *Data Table*. Visualisasi ini membantu pengguna memahami pola data dan hubungan antar atribut.

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i2.7312

HASIL DAN PEMBAHASAN

1. Pengumpulan data

Pada pengumpulan data yang kami lakukan dengan cara mengambil dataset dari kaggel.com , terkait data diabetes, berikut datasetnya.

Tabel 1. Pengumpulan dataset

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1

Tahap awal dari proyek ini melibatkan pengumpulan data, yang merupakan langkah kunci dalam membangun model prediktif yang berkualitas. Dataset yang digunakan diperoleh dari Kaggle, sebuah platform komunitas yang menyediakan berbagai dataset publik untuk penelitian dan pengembangan. Kaggle dikenal luas sebagai sumber terpercaya bagi para peneliti dan data scientist, menawarkan dataset berkualitas dengan topik yang beragam. Dalam proyek ini, dataset yang dipilih berisi informasi terkait diabetes, meliputi berbagai parameter kesehatan, seperti kadar gula darah, tekanan darah, usia, indeks massa tubuh (BMI), jumlah kehamilan (khusus untuk wanita), serta riwayat keluarga terkait diabetes.

Pengumpulan dataset ini tidak hanya sekadar mengunduh data, tetapi juga memerlukan pemahaman menyeluruh terhadap struktur data yang tersedia. Sebelum data digunakan lebih lanjut, analisis eksploratif dilakukan untuk memastikan relevansi dataset dengan tujuan proyek. Langkah ini melibatkan pemeriksaan atribut, analisis distribusi data, serta identifikasi potensi masalah seperti nilai yang hilang (missing values) atau outlier yang dapat memengaruhi hasil model. Selain itu, ukuran dataset juga menjadi faktor penting; dataset yang besar dan bervariasi memberikan peluang lebih besar untuk menghasilkan model yang akurat dan dapat diterapkan pada populasi yang lebih luas.

Pemilihan dataset dari Kaggle dilakukan dengan mempertimbangkan beberapa kriteria penting. Dataset yang dipilih harus mencakup atribut yang relevan untuk prediksi diabetes, didukung oleh dokumentasi yang jelas, dan berasal dari sumber yang dapat dipercaya. Dokumentasi ini sering kali mencakup deskripsi atribut, metode pengumpulan data, dan referensi studi sebelumnya, yang membantu memahami konteks dataset. Sebagai contoh, atribut seperti kadar gula darah dan riwayat keluarga memberikan informasi penting yang dapat meningkatkan akurasi prediksi diabetes.

Selain memilih dataset yang relevan, aspek legalitas dan etika juga menjadi perhatian utama dalam pengumpulan data. Dataset dari Kaggle biasanya memiliki lisensi yang mengatur

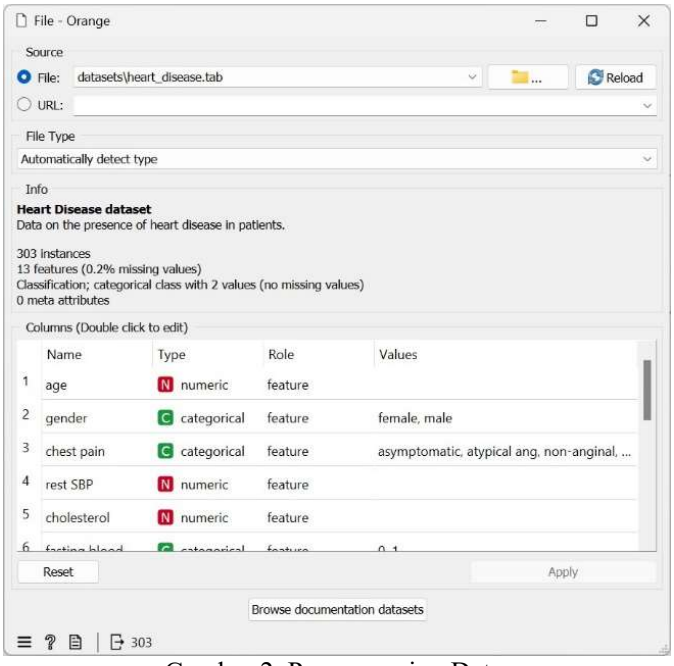
penggunaannya, baik untuk keperluan akademis maupun komersial. Oleh karena itu, penting untuk memastikan bahwa penggunaannya sesuai dengan ketentuan lisensi yang berlaku. Mengingat data yang digunakan mencakup informasi kesehatan individu, perlindungan privasi juga sangat penting. Meskipun data dari Kaggle umumnya sudah dianonimkan, verifikasi tambahan dilakukan untuk memastikan tidak ada informasi yang dapat mengidentifikasi individu secara langsung.

Setelah dataset dikumpulkan, langkah berikutnya adalah melakukan validasi awal untuk memastikan data siap digunakan. Validasi ini mencakup pemeriksaan konsistensi data dan memastikan bahwa semua atribut berada dalam format yang sesuai. Misalnya, atribut kadar gula darah harus dinyatakan dalam satuan yang seragam, seperti mg/dL, dan nilainya harus masuk akal. Data yang tidak memenuhi kriteria ini, seperti nilai yang tidak realistis atau atribut yang tidak relevan, perlu dihapus atau diperbaiki agar analisis berikutnya tidak terganggu.

Secara keseluruhan, proses pengumpulan data ini merupakan fondasi yang sangat penting untuk keberhasilan proyek. Pemilihan dataset dari Kaggle memberikan keuntungan berupa akses ke data yang relevan dan berkualitas tinggi, sehingga model prediktif yang dihasilkan dapat lebih andal. Dengan memprioritaskan kualitas data sejak awal, risiko kesalahan dalam analisis dapat diminimalkan. Langkah ini tidak hanya mendukung tahap preprocessing dan pelatihan model, tetapi juga memastikan hasil akhir dari analisis memiliki akurasi yang tinggi dan relevansi yang signifikan dalam konteks prediksi diabetes.

2. Preprocessing Data

Setelah data terkumpul selanjutnya masuk ke aplikasi orange untuk melakukan pembersihan data agar siap digunakan.



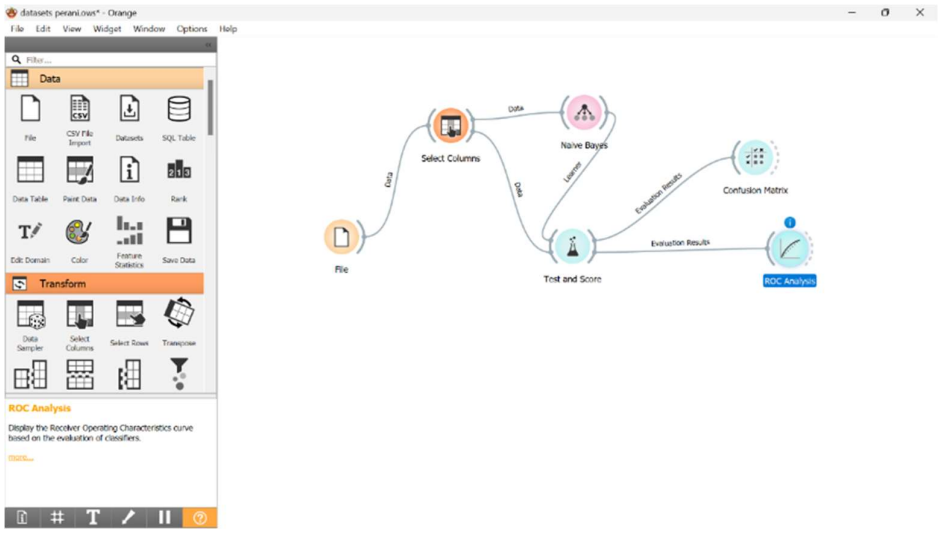
Gambar 2. Preprovessing Data

Tahap berikutnya setelah pengumpulan data adalah preprocessing, yang bertujuan untuk memastikan data dalam kondisi optimal sebelum digunakan dalam analisis. Tahap ini mencakup langkah-langkah untuk membersihkan, menyusun, dan menyiapkan data agar kompatibel dengan algoritma pembelajaran mesin. Dalam proyek ini, data yang telah dikumpulkan diproses menggunakan aplikasi Orange, alat analitik berbasis visual yang memungkinkan langkah-langkah preprocessing dilakukan dengan lebih mudah dan efisien.

Secara keseluruhan, preprocessing adalah kunci untuk menghasilkan model yang andal dan akurat. Dengan mengatasi potensi masalah seperti nilai kosong, data tidak valid, dan perbedaan skala, kualitas data dapat ditingkatkan, sehingga model pembelajaran mesin dapat memanfaatkan informasi tersebut secara maksimal. Hasilnya, analisis menjadi lebih bermakna dan relevan dalam mendukung pengambilan keputusan.

3. Pemilihan Algoritma Naïve Bayes

Algoritma ini didasarkan pada Teorema Bayes dengan asumsi bahwa semua fitur bersifat independen. Dalam Orange, widget *Naïve Bayes* digunakan untuk memilih algoritma.

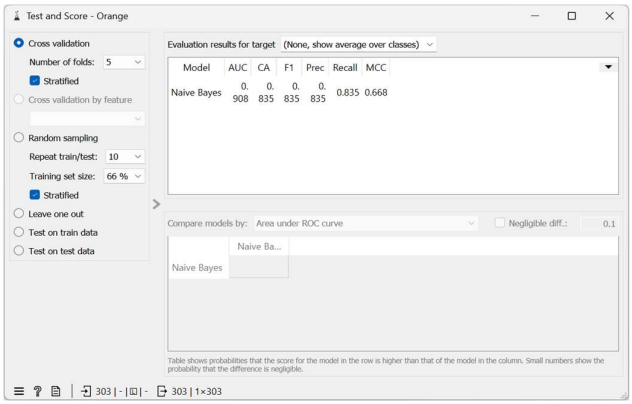


Gambar 3. widget *Naïve Bayes*

Setelah data selesai melalui tahap preprocessing, langkah selanjutnya adalah menentukan algoritma yang paling sesuai untuk membangun model prediktif. Pemilihan algoritma menjadi salah satu langkah penting dalam pipeline analisis, karena algoritma yang dipilih akan sangat memengaruhi kinerja dan akurasi model. Pada proyek ini, algoritma Naïve Bayes dipilih karena karakteristiknya yang sederhana namun efektif untuk menangani dataset dengan ukuran besar dan kompleksitas tinggi. Meskipun algoritma ini mengasumsikan bahwa setiap fitur dalam dataset bersifat independen—yang dalam kenyataannya jarang terjadi—Naïve Bayes tetap mampu memberikan hasil yang cukup baik dalam berbagai skenario praktis.

4. Pelatihan Model

Pada tahap ini, dataset dibagi menjadi data pelatihan dan data pengujian. Pembagian ini dilakukan menggunakan *Test & Score* di Orange. Data pelatihan digunakan untuk membangun model prediktif berdasarkan algoritma Naïve Bayes. Proses ini melibatkan perhitungan probabilitas untuk setiap kategori dalam atribut target berdasarkan fitur yang diberikan. Berikut hasil Test & Scor :



Gambar 4. Test & Score

Tahap pelatihan model adalah bagian penting dari proses pembelajaran mesin, di mana dataset dibagi menjadi dua bagian utama: data pelatihan dan data pengujian. Pembagian ini dilakukan untuk memastikan model dapat belajar dari pola yang terdapat dalam data yang telah diketahui sekaligus mengukur kemampuan prediksinya pada data baru yang belum pernah dilihat sebelumnya. Dalam analisis ini, pembagian dataset dilakukan menggunakan fitur **Test & Score** di aplikasi Orange. Fitur ini memungkinkan pembagian dataset dilakukan secara otomatis dengan proporsi tertentu, seperti 70% data pelatihan dan 30% data pengujian, sesuai dengan kebutuhan proyek.

Data pelatihan digunakan untuk membangun model dengan mengenalkan pola-pola yang ada di dalam dataset kepada algoritma. Pada analisis ini, algoritma Naïve Bayes bekerja dengan menghitung probabilitas masing-masing kategori berdasarkan nilai atribut dalam data pelatihan. Misalnya, algoritma mempelajari hubungan antara kadar gula darah, tekanan darah, indeks massa tubuh, dan atribut lainnya dengan target prediksi, yaitu risiko diabetes (positif atau negatif). Selama proses ini, algoritma terus menyesuaikan parameter internalnya agar dapat memberikan hasil prediksi yang paling akurat berdasarkan data pelatihan.

Sebaliknya, data pengujian digunakan untuk menilai kemampuan model dalam memprediksi data baru yang tidak pernah dilihat sebelumnya. Data ini berfungsi untuk mengevaluasi seberapa baik model dapat diandalkan di luar lingkungan pelatihannya. Proses ini penting untuk menghindari masalah overfitting, di mana model terlalu terfokus pada data pelatihan sehingga gagal memberikan hasil yang baik pada data lain. Evaluasi dengan data pengujian membantu memastikan bahwa model dapat diterapkan secara efektif dalam situasi dunia nyata.

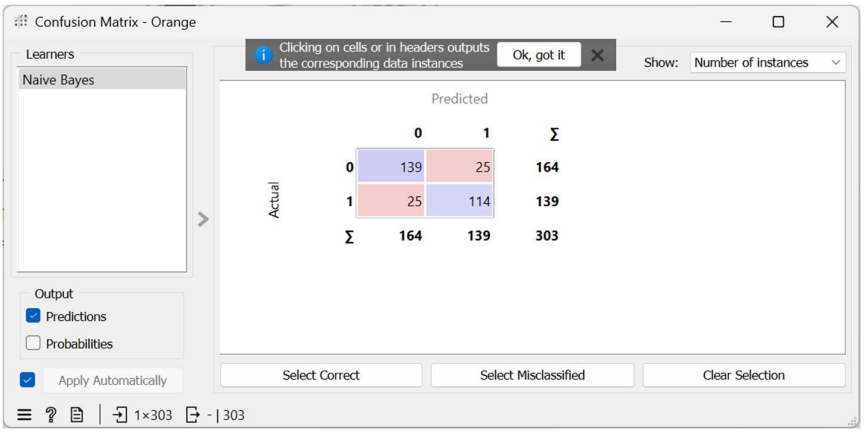
Fitur **Test & Score** tidak hanya membagi dataset menjadi data pelatihan dan pengujian, tetapi juga memberikan berbagai metrik evaluasi untuk menilai kinerja model. Metrik seperti akurasi, presisi, recall, dan F1-score digunakan untuk memberikan gambaran menyeluruh tentang kinerja model. Akurasi menunjukkan persentase prediksi yang benar, sementara presisi mengukur seberapa tepat model dalam memprediksi kategori positif. Recall mengevaluasi kemampuan model dalam menemukan semua kategori positif, dan F1-score memberikan kombinasi seimbang antara presisi dan recall, yang sangat berguna ketika data tidak seimbang.

5. Evaluasi Model

Setelah proses pelatihan model selesai, langkah berikutnya adalah melakukan evaluasi untuk menentukan seberapa baik model tersebut dalam menjalankan tugas yang diberikan. Evaluasi ini dilakukan dengan menggunakan berbagai metrik yang memberikan gambaran menyeluruh tentang kinerja model. Evaluasi kinerja dilakukan untuk memastikan keandalannya. Proses evaluasi ini

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i2.7312

mencakup pengukuran beberapa metrik, seperti akurasi, presisi, recall, dan F1-score. Akurasi menunjukkan seberapa sering model memberikan prediksi yang benar, sedangkan presisi mengukur ketepatan prediksi positif. Recall menilai sejauh mana model dapat menemukan semua kasus positif, dan F1-score, sebagai kombinasi presisi dan recall, memberikan gambaran menyeluruh tentang kinerja model. Untuk memberikan visualisasi lebih lanjut, digunakan widget Confusion Matrix di Orange, yang menunjukkan distribusi prediksi model terhadap kategori target, baik yang benar maupun salah.



Gambar 5. Hasil Widget *Confusion Matrix*

Hasil evaluasi kinerja model pada gambar 5 menunjukkan dengan pendekatan ini memberikan wawasan penting tentang kemampuan model dalam memprediksi risiko diabetes secara efektif. Setiap metrik memiliki kontribusi spesifik dalam mengukur keandalan model, memastikan performa yang optimal dalam mendukung pengambilan keputusan berbasis data.

1. Akurasi

Akurasi memberikan gambaran keseluruhan tentang performa model. Misalnya, jika model memiliki akurasi sebesar 87%, ini berarti 87 dari setiap 100 prediksi sesuai dengan nilai sebenarnya. Namun, akurasi tidak selalu mencerminkan performa model yang sesungguhnya jika dataset memiliki distribusi yang tidak seimbang (misalnya, lebih banyak kategori negatif dibandingkan positif). Oleh karena itu, akurasi harus didukung oleh analisis metrik lainnya seperti presisi, recall, dan F1-score.

2. Presisi

Presisi mengukur keakuratan prediksi positif model, yaitu seberapa banyak dari prediksi positif yang benar-benar sesuai dengan nilai aktual. Presisi sangat penting dalam situasi di mana meminimalkan false positives menjadi prioritas utama, misalnya dalam merujuk pasien untuk pemeriksaan lebih lanjut. Tingkat presisi yang tinggi akan memastikan bahwa sebagian besar pasien yang dirujuk benar-benar berisiko, sehingga dapat mengurangi biaya tambahan akibat kesalahan prediksi.

3. Recall (Sensitivitas)

Recall mengukur kemampuan model dalam mendeteksi semua kasus positif. Dalam konteks diabetes, recall yang tinggi sangat penting untuk memastikan bahwa sebagian besar individu yang berisiko tidak terlewatkan. Misalnya, jika recall model sebesar 85%, ini berarti 85 dari setiap 100 kasus positif berhasil terdeteksi. Recall yang rendah dapat mengakibatkan konsekuensi serius, seperti keterlambatan diagnosis yang berdampak pada kesehatan pasien.

4. F1-Score

F1-Score adalah rata-rata harmonis antara presisi dan recall, yang memberikan gambaran seimbang tentang performa model. F1-Score sangat relevan ketika data yang digunakan memiliki ketidakseimbangan antara kategori positif dan negatif. Jika F1-Score tinggi, model dapat dikatakan memiliki performa prediksi yang andal tanpa mengorbankan salah satu metrik. Contohnya, jika presisi adalah 80% dan recall adalah 85%, maka F1-Score model adalah sekitar 82.8%.

6. Visualisasi dengan Confusion Matrix

Confusion Matrix memberikan rincian yang lebih spesifik tentang distribusi prediksi model terhadap kategori target:

a) True Positives (TP): Jumlah kasus positif yang diprediksi dengan benar.
b) True Negatives (TN): Jumlah kasus negatif yang diprediksi dengan benar.
c) False Positives (FP): Jumlah kasus negatif yang salah diprediksi sebagai positif.
d) False Negatives (FN): Jumlah kasus positif yang salah diprediksi sebagai negatif.

Sebagai ilustrasi, jika Confusion Matrix menunjukkan:
TP: 120, TN: 100, FP: 30, FN: 20,

Maka metrik evaluasi dihitung sebagai berikut:
Akurasi = (TP + TN) / (TP + TN + FP + FN) = (120 + 100) / (120 + 100 + 30 + 20) = 84%.
Presisi = TP / (TP + FP) = 120 / (120 + 30) = 80%.
Recall = TP / (TP + FN) = 120 / (120 + 20) = 85.7%.
F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall) = 82.8%.

KESIMPULAN

Pendekatan klasifikasi menggunakan algoritma Naïve Bayes juga telah diterapkan untuk mendeteksi penyakit endokrin pada lansia. Proses ini melibatkan pengumpulan dataset dari Kaggle, preprocessing data dengan aplikasi Orange, serta pelatihan dan evaluasi model menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Evaluasi model menunjukkan bahwa algoritma Naïve Bayes mampu memberikan hasil yang cukup baik dalam memprediksi risiko diabetes, dengan akurasi mencapai 84%.

DAFTAR PUSTAKA

[1] G. Kalaivani and P. Mayilvahanan, "Air Quality Prediction and Monitoring using Machine Learning Algorithm based IoT sensor- A researcher's perspective," *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, 2021, doi: 10.1109/ICCES51350.2021.9489153.

[2] S. Abidin, "Deteksi Wajah Menggunakan Metode Haar Cascade Classifier Berbasis Webcam Pada Matlab," *J. Teknol. Elekterika*, vol. 15, no. 1, p. 21, 2018, doi: 10.31963/elekterika.v15i1.2102.

[3] P. Rosyani and O. Hariansyah, "Pengenalan Citra Bunga Menggunakan Segmentasi Otsu Threshold dan Naïve Bayes," pp. 1–7, 2020, doi: 10.30864/jsi.v15i1.304.

[4] K. Sulastri, "Klasifikasi Naïve Bayes pada Analisis Sentimen atas Penolakan Dibukanya Larangan Ekspor Benih Lobster," *KERNEL J. Ris. Inov. Bid. Inform. dan Pendidik. Inform.*, vol. 1, no. 2, pp. 68–75, 2020, doi: 10.31284/j.kernel.2020.v1i2.1501.

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i2.7312

[5] F. Fitriani and R. Fadilla, "Pengaruh Senam Diabetes Terhadap Penurunan Kadar," *J. Kesehat. dan Pengemb.*, vol. 10, no. 19, pp. 114–122, 2020.

[6] Didin Wahyu Utomo, Suprapto, and Nurul Hidayat, "Pemodelan Sistem Pakar Diagnosis Penyakit pada Sistem Endokrin Manusia dengan Metode Dempster-Shafer," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 893–903, 2019.

[7] Perani Rosyani and Fabian Syawali, "Application of Advanced Class Determination System Using K-Means Clustering Method (Case Study: SMK Al-Badar Balaraja)," *Int. J. Integr. Sci.*, vol. 2, no. 10, pp. 1557–1570, 2023, doi: 10.55927/ijis.v2i10.6347.