BAB II

LANDASAN TEORI

2.1. Mampu Menyelesaikan Studi

Siswa penerima Kartu Indonesia Pintar (KIP) yang mampu menyelesaikan studi menunjukkan keberhasilan program bantuan pendidikan ini dalam meningkatkan akses dan kualitas pendidikan. Kehadiran siswa yang konsisten merupakan salah satu indikator utama kesuksesan mereka. Siswa-siswa ini cenderung lebih rajin dan disiplin dalam mengikuti pelajaran, karena bantuan KIP mengurangi beban finansial yang sering menjadi penghalang utama untuk hadir secara rutin di sekolah. Dengan berkurangnya hambatan ekonomi, siswa penerima KIP dapat lebih fokus pada pembelajaran mereka, meningkatkan kehadiran dan partisipasi aktif dalam kegiatan sekolah. Selain kehadiran, pengetahuan siswa dan partisipasi orang tua juga memainkan peran penting dalam keberhasilan siswa penerima KIP menyelesaikan studi. Siswa yang menerima KIP umumnya menunjukkan peningkatan dalam pengetahuan dan keterampilan akademis, karena mereka dapat memanfaatkan sumber daya pendidikan yang lebih baik dan mengikuti program pembelajaran tambahan. Partisipasi orang tua juga sangat penting, karena dukungan dan dorongan dari keluarga membantu siswa tetap termotivasi dan berprestasi. Dengan adanya partisipasi aktif dari orang tua dalam pendidikan anak mereka, termasuk dalam mendukung kebutuhan belajar di rumah dan menghadiri pertemuan sekolah, siswa penerima KIP memiliki lingkungan yang kondusif untuk mencapai kesuksesan akademis.

2.2. Machine Learning

Machine Learning merupakan suatu proses penggalian informasi yang tersembunyi dan pola-pola yang dapat digunakan untuk pengambilan keputusan dalam suatu dataset besar [10] [11]. Salah satu aspek penting dari Machine Learning adalah model klasifikasi, di mana tujuannya adalah untuk mengelompokkan atau mengkategorikan data ke dalam kelas atau kelompok yang telah ditentukan. Model klasifikasi menggunakan algoritma pembelajaran mesin untuk mempelajari pola dari data pelatihan yang telah diberikan dan kemudian memprediksi kelas atau label dari data baru.

Proses Machine Learning dimulai dengan pengumpulan data dari berbagai sumber dan kemudian membersihkan, mengintegrasikan, dan mengolah data tersebut untuk mempersiapkannya untuk analisis lebih lanjut [12]. Selanjutnya, proses ini melibatkan pemilihan atribut atau fitur yang relevan untuk digunakan dalam pembuatan model. Setelah data siap, model klasifikasi dapat dibangun dengan menggunakan berbagai algoritma, seperti Naive Bayes, Decision Trees, Support Vector Machines, atau Neural Networks.

Model klasifikasi memainkan peran kunci dalam memprediksi atau mengklasifikasikan data baru ke dalam kategori yang sesuai [13]. Model ini dapat digunakan dalam berbagai bidang, mulai dari keuangan dan pemasaran hingga kesehatan dan ilmu pengetahuan. Sebagai contoh, dalam industri keuangan, model klasifikasi dapat digunakan untuk mendeteksi potensi kecurangan kartu kredit atau untuk menilai risiko kredit.

Penting untuk memvalidasi dan menguji model klasifikasi menggunakan data yang tidak terlibat dalam pembuatan model (data uji) untuk memastikan keandalan dan generalisasi model terhadap data baru. Selain itu, proses iteratif dapat dilakukan untuk meningkatkan kinerja model dengan mengoptimalkan parameter dan memilih fitur yang lebih relevan.

Dengan terus berkembangnya teknologi dan peningkatan ketersediaan data, penggunaan Machine Learning dan model klasifikasi menjadi semakin penting untuk mendapatkan wawasan berharga dari volume data yang terus bertambah [14]. Model klasifikasi ini memberikan alat yang efektif dalam mengambil keputusan dan mengeksplorasi potensi yang terkandung dalam dataset besar untuk mendukung perkembangan di berbagai sektor. Untuk tahapan pelaksanaan pada implementasi Machine Learning yaitu sebagai berikut:

1. Pemahaman Bisnis dan Masalah (Business Understanding)

Pemahaman Bisnis dan Masalah (Business Understanding) adalah tahap awal yang sangat penting dalam proses Machine Learning. Pada tahap ini, penelitian harus mendalam untuk mengidentifikasi tujuan bisnis utama yang ingin dicapai melalui analisis data. Ini mencakup pemahaman menyeluruh tentang kebutuhan dan tantangan bisnis yang dihadapi, serta pertanyaan spesifik yang perlu dijawab dengan menggunakan Machine Learning. Misalnya, dalam konteks bisnis, tujuan bisa mencakup peningkatan retensi pelanggan, identifikasi pola pembelian, atau optimasi proses operasional.

Selain tujuan bisnis, pemahaman yang baik tentang masalah yang ingin dipecahkan menjadi kunci kesuksesan. Ini mencakup pengidentifikasian masalah

spesifik yang dapat diatasi melalui analisis data. Misalnya, mungkin ada tantangan dalam mengidentifikasi tren konsumen atau menganalisis efektivitas strategi pemasaran. Dengan memahami secara mendalam masalah yang dihadapi, penelitian dapat merancang pendekatan analisis yang sesuai dan menghasilkan wawasan yang dapat digunakan untuk menginformasikan keputusan bisnis.

Pada akhirnya, pemahaman bisnis dan masalah membantu penelitian untuk fokus pada aspek-aspek data yang paling relevan dan signifikan. Hal ini memastikan bahwa hasil dari Machine Learning tidak hanya bermanfaat secara teknis, tetapi juga memberikan nilai tambah yang nyata untuk pemecahan masalah dan pencapaian tujuan bisnis yang telah ditetapkan. Tahap ini membentuk dasar yang solid untuk perjalanan analisis data, memastikan bahwa setiap langkah yang diambil memiliki dampak yang positif dan sesuai dengan kebutuhan bisnis yang spesifik.

2. Pemahaman Data (Data Understanding)

Pemahaman Data (Data Understanding) dalam Machine Learning merupakan tahap kritis yang membuka pintu ke dalam karakteristik dan potensi data yang akan dianalisis. Pada tahap ini, penelitian memusatkan perhatian pada eksplorasi data untuk mendapatkan wawasan yang mendalam mengenai sifat dan struktur dataset. Hal ini melibatkan pemahaman distribusi variabel, identifikasi potensi nilai-nilai yang hilang atau outlier, serta pengeksplorasian tren atau pola awal yang mungkin muncul. Analisis kualitas data menjadi fokus, termasuk pemeriksaan kebersihan data, konsistensi format, dan penanganan nilai yang hilang.

Selain itu, tahap Pemahaman Data juga melibatkan penentuan relevansi variabel terhadap tujuan analisis, memastikan bahwa variabel yang dipilih memiliki kontribusi signifikan dalam menjawab pertanyaan atau masalah yang diajukan. Misalnya, dalam penelitian tentang analisis minat masyarakat menggunakan media sosial, pemahaman data akan mencakup pemahaman terhadap jenis data yang terkumpul, seperti postingan, like, atau komentar, serta bagaimana variabel-variabel tersebut dapat mencerminkan minat masyarakat.

Keberhasilan tahap Pemahaman Data memberikan landasan yang kokoh untuk pemrosesan data selanjutnya, memastikan bahwa data yang digunakan dalam analisis adalah representatif dan bersih. Pemahaman yang mendalam ini juga membantu mengidentifikasi potensi kompleksitas dalam dataset, memandu pemilihan metode analisis yang sesuai, dan mengarahkan perhatian pada aspek-aspek kritis yang akan dibahas dalam tahap-tahap selanjutnya dari proses Machine Learning.

3. Pemrosesan Data (Data Preparation)

Pemrosesan Data (Data Preparation) dalam Machine Learning memegang peranan kunci dalam menyediakan fondasi yang solid untuk analisis data yang efektif. Pada tahap ini, data yang telah dikumpulkan diolah untuk memastikan kebersihan, konsistensi, dan keterwakilan yang optimal. Hal ini mencakup langkah-langkah pembersihan data, seperti penanganan nilai yang hilang dan outlier, serta transformasi variabel yang diperlukan untuk memenuhi format atau skala yang diinginkan. Proses normalisasi atau standarisasi dapat diterapkan untuk memastikan variabel-variabel memiliki pengaruh yang seimbang dalam analisis.

Selain itu, pemrosesan data juga melibatkan pemilihan dan ekstraksi fitur yang relevan agar data yang dianalisis sesuai dengan tujuan penelitian. Dengan memastikan bahwa data telah dipersiapkan dengan baik, tahap ini membuka jalan untuk pengembangan model dan analisis lebih lanjut, memastikan bahwa data yang digunakan dalam proses Machine Learning adalah yang terbaik dalam memberikan wawasan yang berarti dan mendukung pengambilan keputusan yang informasional.

4. Pemodelan (Modeling)

Pemodelan (Modeling) dalam Machine Learning adalah tahap di mana model atau algoritma diterapkan ke data yang telah dipersiapkan untuk mengekstrak pola atau hubungan yang signifikan. Pada tahap ini, dipilihnya algoritma atau metode yang sesuai dengan tujuan analisis menjadi kunci. Misalnya, dalam konteks analisis minat masyarakat menggunakan media sosial dengan Metode Naive Bayes dan metode Naïve Bayes, tahap pemodelan melibatkan penerapan kedua teknik tersebut untuk mengidentifikasi pola kompleks dan memprediksi minat berdasarkan data yang ada. Setelah memilih model, dilakukan proses pelatihan menggunakan data yang telah dipersiapkan sebelumnya. Melalui iterasi dan penyesuaian, model ditingkatkan untuk mencapai kinerja yang optimal. Pemodelan menjadi langkah penting dalam menghasilkan wawasan yang dapat mendukung pengambilan keputusan, dan evaluasi yang cermat pada tahap ini memastikan bahwa model yang dikembangkan dapat memahami dengan baik struktur dan kompleksitas data yang dihadapi.

5. Evaluasi Model (Model Evaluation)

Evaluasi Model (Model Evaluation) merupakan tahap penting dalam proses Machine Learning yang bertujuan untuk mengukur kinerja dan efektivitas model yang telah dikembangkan. Pada tahap ini, model yang telah dilatih dievaluasi menggunakan set data yang tidak pernah dilihat sebelumnya untuk menghindari overfitting. Metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan kurva ROC digunakan untuk mengukur sejauh mana model mampu memprediksi dengan benar dan mengklasifikasikan data. Evaluasi model yang cermat memberikan wawasan tentang kemampuan model untuk memahami pola atau tren dalam data yang mungkin belum pernah dihadapi sebelumnya. Apabila model tidak memenuhi standar kinerja yang diinginkan, tahap ini memungkinkan penyesuaian dan optimalisasi model, sehingga hasil akhir yang diperoleh memiliki validitas dan relevansi yang tinggi. Evaluasi model menjadi langkah kritis dalam memastikan bahwa model yang dikembangkan dapat memberikan wawasan yang dapat diandalkan dan dapat diaplikasikan dalam konteks bisnis atau penelitian yang diinginkan.

6. Optimasi dan Tuning (Optimization and Tuning)

Optimasi dan Tuning (Optimization and Tuning) adalah tahap yang strategis dalam proses Machine Learning yang bertujuan untuk meningkatkan kinerja dan akurasi model yang telah dikembangkan. Pada tahap ini, dilakukan penyesuaian terhadap parameter-model, serta eksplorasi berbagai konfigurasi yang memungkinkan peningkatan efisiensi dan ketepatan model. Proses ini dapat melibatkan eksperimen dengan berbagai nilai parameter, pemilihan fitur yang

lebih optimal, atau penggunaan teknik ensemble untuk menggabungkan kekuatan beberapa model. Melalui proses iteratif ini, tujuannya adalah mencapai model yang dapat menghasilkan prediksi yang lebih akurat dan relevan dengan konteks analisis. Optimasi dan tuning juga memainkan peran penting dalam mencegah overfitting atau underfitting, memastikan bahwa model dapat diterapkan dengan baik pada data baru dan menghasilkan hasil yang generalis. Dengan fokus pada peningkatan kinerja model, tahap ini menegaskan kualitas dan kehandalan model Machine Learning yang digunakan dalam menghadapi tantangan dan kompleksitas data yang sebenarnya.

7. Implementasi (Deployment)

Implementasi (Deployment) dalam konteks Machine Learning mencakup penerapan hasil analisis ke dalam lingkungan praktis atau keputusan bisnis sehari-hari. Setelah model Machine Learning dioptimalkan dan divalidasi dengan baik, tahap implementasi menjadi kunci untuk mengubah wawasan yang diperoleh menjadi aksi yang nyata. Misalnya, dalam penelitian analisis minat masyarakat menggunakan media sosial dengan Metode Naive Bayes dan metode Naïve Bayes, tahap ini melibatkan penerapan temuan terkait minat masyarakat ke dalam strategi pemasaran, pengelolaan konten media sosial, atau pengembangan kampanye yang lebih terarah. Proses implementasi juga mempertimbangkan integrasi model Machine Learning dengan sistem atau proses yang ada, memastikan bahwa wawasan yang dihasilkan dapat diterapkan secara efektif dalam konteks operasional sehari-hari. Kesuksesan implementasi menciptakan dampak yang nyata dan berkelanjutan, membuktikan nilai analisis Machine

Learning dalam mendukung pengambilan keputusan yang lebih baik dan responsif terhadap dinamika masyarakat atau bisnis yang terus berkembang.

8. Monitoring dan Pemeliharaan (Monitoring and Maintenance)

Monitoring dan Pemeliharaan (Monitoring and Maintenance) merupakan tahap penting dalam siklus Machine Learning yang menekankan pada keberlanjutan dan relevansi model yang telah diimplementasikan. Setelah fase implementasi, tahap ini melibatkan pemantauan terus-menerus terhadap kinerja model dan responsnya terhadap perubahan data. Proses ini memastikan bahwa model tetap relevan seiring berjalannya waktu dan dapat mengatasi perubahan dalam pola atau tren yang mungkin terjadi. Monitoring juga mencakup identifikasi indikasi overfitting atau penurunan kinerja, memungkinkan untuk melakukan penyesuaian atau optimalisasi tambahan jika diperlukan. Pemeliharaan melibatkan pembaruan terhadap model, integrasi dengan data baru, dan penyesuaian terhadap perubahan dalam kebutuhan bisnis atau lingkungan eksternal. Dengan pendekatan yang proaktif terhadap pemantauan dan pemeliharaan, perusahaan atau peneliti dapat memastikan bahwa model Machine Learning tetap efektif dan memberikan nilai tambah yang berkelanjutan sepanjang waktu.

Penelitian yang dilakukan oleh [15] bahwassanya penerapan Machine Learning dalam konteks meningkatkan akurasi prediksi absensi melibatkan strategi holistik yang berfokus pada pencarian kombinasi optimal antara metode pemilihan fitur dan teknik pengklasifikasian. Pemilihan fitur merupakan langkah awal yang kritis, di mana metode seperti Information Gain dan Recursive Feature

Elimination digunakan untuk mengidentifikasi atribut-atribut kunci yang memiliki dampak signifikan terhadap prediksi absensi. Selanjutnya, pendekatan pengklasifikasian bagging diterapkan untuk membangun ensemble model yang berdasarkan sampel data yang diambil secara acak dengan penggantian. Hasil dari kombinasi metode ini menunjukkan peningkatan signifikan dalam akurasi prediksi absensi hingga mencapai 92%, menghasilkan model yang lebih tahan terhadap overfitting dan lebih mampu menggeneralisasi pola dari dataset. Implementasi ini memberikan dampak positif pada efisiensi manajemen sumber daya manusia, memungkinkan organisasi untuk melakukan perencanaan operasional yang lebih efektif dan mengoptimalkan alokasi tenaga kerja dengan lebih tepat. Keberhasilan penerapan ini tidak hanya meningkatkan produktivitas organisasi tetapi juga memberikan kontribusi pada kemajuan metodologi Machine Learning yang dapat diaplikasikan pada tantangan prediksi di berbagai konteks bisnis.

Menurut [16], penerapan Machine Learning dalam konteks prediksi nilai suatu ukuran untuk fakta baru dengan menggunakan model klaster memberikan hasil yang sangat akurat. Machine Learning, khususnya dengan pendekatan klastering, memungkinkan analis untuk mengungkap pola-pola tersembunyi dan hubungan kompleks dalam dataset. Dengan membentuk kelompok atau klaster berdasarkan kesamaan karakteristik, model klaster dapat memberikan landasan yang kokoh untuk memprediksi nilai suatu ukuran untuk fakta baru yang serupa. Indah meyakini bahwa tingkat akurasi yang tinggi dalam prediksi ini memiliki implikasi positif dalam mendukung pengambilan keputusan yang lebih informasional dan perencanaan strategis yang lebih efektif. Machine Learning

menjadi instrumen kuat yang membantu meningkatkan kecerdasan analisis, memastikan hasil yang diperoleh bermanfaat dan dapat diandalkan dalam konteks pengembangan strategi bisnis atau pengelolaan sumber daya.

2.3. Model Klasifikasi

Klasifikasi dalam Machine Learning merupakan proses esensial yang bertujuan untuk mengelompokkan atau mengkategorikan data ke dalam kelas atau kategori yang telah ditentukan. Dalam konteks ini, tujuan utama adalah untuk mengembangkan model prediktif yang dapat memahami pola-pola yang ada dalam dataset dan memprediksi kelas atau label dari data baru yang diberikan. Klasifikasi membuka pintu bagi berbagai aplikasi, seperti deteksi spam email, identifikasi penyakit berdasarkan gejala, atau prediksi keberhasilan pelanggan dalam suatu bisnis.

Salah satu metode klasifikasi yang umum digunakan adalah Decision Trees (pohon keputusan). Decision Trees membangun struktur pohon yang menggambarkan serangkaian keputusan berdasarkan fitur-fitur tertentu dari dataset. Proses ini memungkinkan interpretasi yang mudah dan memberikan keputusan yang dapat dijelaskan secara visual. Meskipun rentan terhadap overfitting, teknik tuning parameter dan pruning dapat membantu meningkatkan kinerja model Decision Trees.

Metode Naive Bayes juga populer dalam klasifikasi, terutama dalam konteks teks. Naive Bayes didasarkan pada teorema Bayes dan menghitung probabilitas kelas berdasarkan kondisi atau atribut yang diamati. Meskipun melakukan asumsi independensi antar atribut, Naive Bayes efisien dan sering

memberikan hasil yang baik, terutama pada dataset dengan dimensi tinggi atau ketidakseimbangan kelas.

K-Nearest Neighbors (k-NN) adalah metode klasifikasi yang bekerja dengan mengidentifikasi k entitas terdekat dalam ruang atribut dan menentukan label mayoritas dari entitas tersebut. K-NN cocok untuk masalah klasifikasi yang melibatkan pola spasial dalam data. Walaupun sederhana, k-NN dapat memberikan hasil yang baik pada data yang sesuai dengan asumsi dasarnya.

Support Vector Machines (SVM) adalah model klasifikasi yang kuat untuk masalah klasifikasi biner. SVM berusaha menemukan hyperplane terbaik yang memisahkan kelas-kelas dalam ruang atribut. Meskipun mungkin memerlukan penalaan parameter yang cermat, SVM efektif dalam menangani dataset kompleks dan memiliki toleransi tinggi terhadap overfitting.

Klasifikasi dalam Machine Learning memainkan peran kritis dalam membantu organisasi membuat keputusan berdasarkan pola-pola yang ada dalam data mereka. Pemilihan metode klasifikasi harus mempertimbangkan karakteristik data, sifat masalah, dan tujuan akhir dari analisis tersebut untuk memastikan pengembangan model yang efektif dan relevan.

2.4. Metode Naïve Bayes

Metode Naive Bayes merupakan salah satu pendekatan yang populer dalam klasifikasi dan analisis teks. Metode ini didasarkan pada teorema Bayes yang mencoba menghitung probabilitas kejadian berdasarkan informasi yang ada. Keunikan dari Naive Bayes terletak pada asumsi bahwa semua variabel yang

digunakan dalam analisis adalah independen satu sama lain, walaupun dalam konteks dunia nyata hal ini sering kali tidak sepenuhnya akurat.

Salah satu aplikasi paling umum dari metode Naive Bayes adalah dalam klasifikasi teks, seperti deteksi spam email atau kategorisasi dokumen. Dalam klasifikasi teks, Naive Bayes bekerja dengan menganalisis kemunculan kata-kata tertentu dalam dokumen untuk memprediksi kelas atau kategori dokumen tersebut. Meskipun asumsi tentang independensi antara kata-kata tidak selalu terpenuhi, metode Naive Bayes tetap efektif dan efisien, terutama ketika dihadapkan pada dataset teks besar.

$$P\ (A|B) = \frac{P\ (B|A)\ .P\ (A)}{P\ (B)}$$

P (A|B) = Probabilitas A bersyarat yang diberikan oleh B

P (B|A) = Probabilitas B bersyarat yang diberikan oleh A

P (A) = Probabilitas kejadian A

P (B) = Probabilitas kejadian B

Langkah-langkah dalam menggunakan metode Naive Bayes melibatkan pembelajaran dari data pelatihan untuk menghitung probabilitas munculnya setiap kata dalam setiap kelas, yang kemudian digunakan untuk mengklasifikasikan dokumen baru. Naive Bayes juga dapat diterapkan dalam konteks klasifikasi yang lebih umum, seperti prediksi penyakit berdasarkan gejala atau pengklasifikasian gambar berdasarkan fitur-fitur tertentu.

Kelebihan utama dari metode Naive Bayes melibatkan kemudahan implementasi dan kinerja yang baik, terutama ketika berhadapan dengan data teks. Meskipun asumsi independensi seringkali tidak terpenuhi secara sempurna, Naive

Bayes tetap memberikan hasil yang memuaskan dalam banyak kasus. Meskipun sederhana, metode ini tetap menjadi salah satu pilihan yang populer dalam analisis data, terutama ketika dihadapkan pada masalah klasifikasi dengan fitur yang cukup kompleks.

Model klasifikasi dalam Machine Learning menjadi salah satu elemen sentral dalam pengembangan strategi pengambilan keputusan berbasis data. Tujuan utamanya adalah untuk memprediksi kategori atau kelas dari suatu entitas berdasarkan pola-pola yang ditemukan dalam data pelatihan. Dalam model klasifikasi, terdapat beberapa metode yang telah terbukti efektif, dan ini mencakup Decision Trees, Naive Bayes, k-Nearest Neighbors, dan Support Vector Machines.

Salah satu model klasifikasi yang umum digunakan adalah Decision Trees. Decision Trees membangun pohon keputusan dengan membagi dataset berdasarkan fitur-fitur tertentu hingga mencapai keputusan atau label tertentu. Kelebihan Decision Trees meliputi kemampuan interpretasi yang baik dan kemampuan menangani kumpulan data yang kompleks. Namun, mereka rentan terhadap overfitting, di mana model menjadi terlalu rumit dan memfitting dengan baik pada data pelatihan tetapi tidak umum pada data baru.

Metode Naive Bayes, sementara sederhana, juga merupakan pilihan yang populer. Berdasarkan teorema Bayes, metode ini menghitung probabilitas kelas tertentu berdasarkan kondisi atau atribut yang diamati. Keunggulan utama Naive Bayes adalah efisiensinya, terutama dalam pengolahan data teks, dan kinerjanya yang baik pada dataset yang besar. Meskipun asumsi independensi antar atribut

(yang sering kali tidak terpenuhi), Naive Bayes tetap menjadi pilihan yang kuat dalam klasifikasi.

Selanjutnya, k-Nearest Neighbors (k-NN) bekerja dengan mengidentifikasi k entitas terdekat dalam ruang atribut dan memberikan label mayoritas dari entitas tersebut. Meskipun sederhana, k-NN efektif pada dataset yang memiliki pola spasial yang jelas, tetapi dapat menjadi komputasi secara intensif pada dataset yang besar.

Support Vector Machines (SVM) adalah model klasifikasi lain yang kuat, terutama ketika menangani masalah klasifikasi biner. SVM mencari pemisah terbaik (hyperplane) antara kelas-kelas yang ada. Meskipun dapat memerlukan tuning parameter yang cermat, SVM efektif dalam menangani kumpulan data yang kompleks dan memiliki toleransi tinggi terhadap overfitting.

Dalam keseluruhan, pilihan model klasifikasi dalam Machine Learning bergantung pada karakteristik dataset dan sifat masalah klasifikasi yang dihadapi. Pengembang harus mempertimbangkan trade-off antara interpretabilitas, kecepatan komputasi, dan ketepatan prediksi untuk memilih model yang paling sesuai dengan kebutuhan spesifiknya.

2.5. Aplikasi Orange

Aplikasi Orange adalah platform analisis data visual yang berfokus pada pengembangan model prediktif, eksplorasi data, dan analisis statistik. Dikembangkan dengan antarmuka pengguna yang bersahabat, Orange memungkinkan pengguna dari berbagai latar belakang, termasuk peneliti, analis

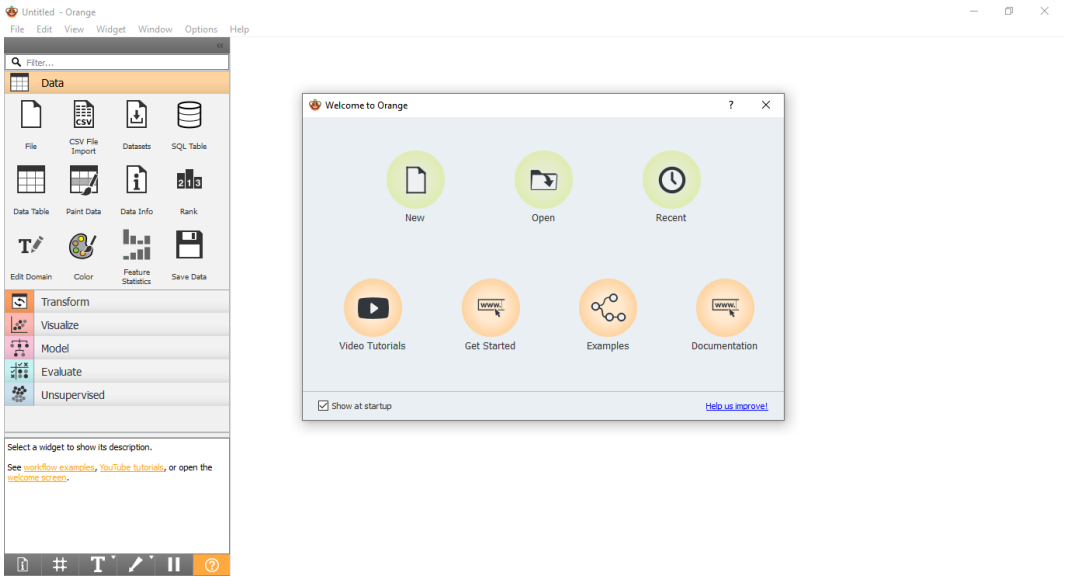
data, dan pemula dalam ilmu data, untuk menjalankan analisis data kompleks tanpa perlu pengetahuan pemrograman mendalam.

Salah satu keunggulan utama Orange adalah fokusnya pada pendekatan visual dalam pembangunan model. Pengguna dapat membangun alur kerja analisis data dengan menarik dan menjatuhkan komponen visual (widget) yang mewakili berbagai langkah dalam proses analisis. Ini membuatnya sangat cocok untuk mereka yang lebih suka berinteraksi dengan data mereka secara visual, memahami setiap langkah secara intuitif.

Orange memiliki sejumlah widget yang kuat, termasuk widget untuk pre-processing data, visualisasi data, pembangunan model, dan evaluasi model. Misalnya, dengan menggunakan widget "Data Table," pengguna dapat dengan cepat melihat dan memahami struktur data mereka, sementara widget "Scatter Plot" memungkinkan visualisasi pola-pola dalam data.

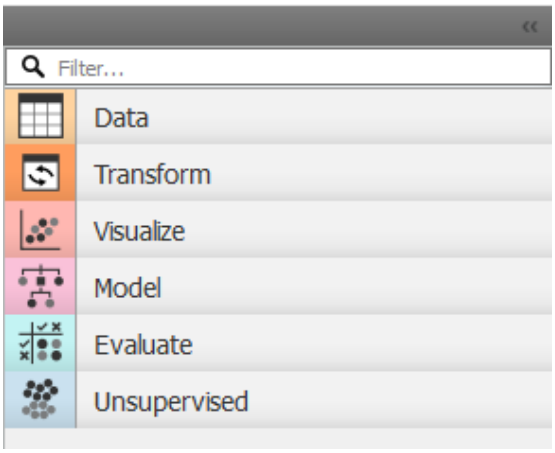
Dalam konteks pembangunan model prediktif, Orange menawarkan widget seperti "Classification Tree" untuk Decision Trees, "Naive Bayes" untuk model Naive Bayes, dan widget "SVM" untuk Support Vector Machines. Selain itu, platform ini mendukung algoritma machine learning yang lebih canggih seperti Random Forest dan Gradient Boosting melalui widget yang sesuai. Orange tidak hanya terbatas pada analisis data konvensional. Platform ini juga menyediakan widget untuk tugas analisis teks, seperti "Text File" untuk membaca data teks, "Text Processing" untuk memproses teks, dan "Topic Modelling" untuk mengeksplorasi tema dalam dokumen.

Dengan ketersediaan sejumlah besar widget dan fungsionalitas yang dapat dikembangkan, Orange memberikan fleksibilitas dan daya jelajah yang besar bagi pengguna yang ingin menjalankan analisis data yang kompleks tanpa harus menjadi ahli pemrograman. Oleh karena itu, Orange adalah alat yang berharga untuk berbagai kebutuhan analisis data, mulai dari eksplorasi data hingga pembangunan model prediktif yang kompleks. Aplikasi Orange memiliki berbagai fungsi yang membuatnya menjadi platform analisis data yang populer dan bermanfaat di kalangan berbagai pengguna, dari pemula hingga ahli analisis data. Fungsi utama aplikasi ini melibatkan pembangunan model prediktif, eksplorasi data, analisis statistik, dan visualisasi informasi.



Gambar 2.4. 1. *Tampilan Awal Aplikas Orange*

Gambar atas merupakan tampilan awal ketika sudah membuka aplikasi orange. Terdapat beberapa menu yang dapat digunakan yaitu membuat projek baru dan membuka projek.



Gambar 2.4. 2. *Menu pada Aplikasi Orange*

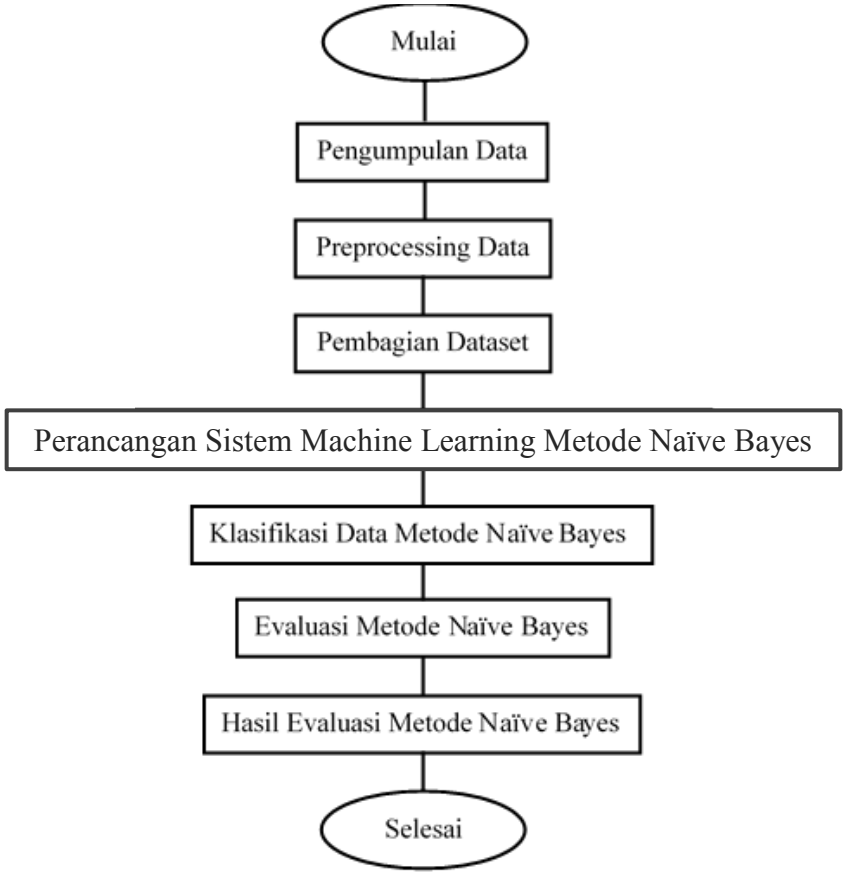
Gambar diatas merupakan menu yang dapat digunakan untuk melakukan analisis data pada aplikasi orange.

Salah satu fungsi utama Orange adalah kemampuannya dalam membangun model prediktif dengan mudah dan intuitif. Dengan berbagai widget dan algoritma machine learning yang disediakan, pengguna dapat membuat model klasifikasi, regresi, atau klastering dengan hanya menggunakan antarmuka grafis tanpa perlu pengetahuan pemrograman mendalam. Ini memudahkan pengguna dari berbagai latar belakang untuk memahami dan mengimplementasikan model prediktif. Selain itu, Orange menawarkan fungsi eksplorasi data yang kuat. Widget visualisasi seperti scatter plot, histogram, dan data table memungkinkan pengguna untuk dengan cepat memahami struktur data dan menemukan pola-pola yang mungkin tersembunyi. Proses eksplorasi data ini mendukung pengambilan keputusan yang informasional dan mendalam.

Manfaat utama Orange terletak pada kemudahan penggunaan dan antarmuka yang intuitif. Fungsi drag-and-drop untuk membangun alur kerja analisis data menjadikan aplikasi ini sangat ramah pengguna. Ini memungkinkan

pengguna dengan berbagai tingkat keahlian untuk melakukan analisis data kompleks tanpa harus menguasai keterampilan pemrograman atau statistik yang tinggi. Kegunaan Orange sangat luas dan beragam. Platform ini dapat digunakan dalam berbagai konteks, seperti riset akademis, analisis data bisnis, atau bahkan pembelajaran dan pengajaran di bidang ilmu data. Oleh karena itu, aplikasi ini dapat membantu pengguna mencapai tujuan analisis data mereka tanpa harus mengatasi kompleksitas teknis yang terkait.

2.6. Kerangka Kerja Penelitian



1. Pengumpulan Data

Dalam konteks Machine Learning, pengumpulan data menjadi tahapan awal yang esensial untuk membangun model analisis yang efektif. Proses ini

melibatkan akuisisi dataset yang mencakup informasi yang relevan untuk tujuan analisis atau prediksi yang diinginkan. Machine Learning dapat melibatkan pengumpulan data dari berbagai sumber, termasuk database perusahaan, file teks, data sensor, atau sumber data lainnya. Pemilihan variabel atau fitur yang akan digunakan dalam analisis juga merupakan bagian kunci dari pengumpulan data dalam Machine Learning. Kualitas dataset dan representativitasnya dapat secara signifikan memengaruhi hasil dari model Machine Learning yang dibangun. Oleh karena itu, peneliti harus memastikan bahwa data yang dikumpulkan memadai, bersih, dan sesuai dengan tujuan analisis, serta mungkin melakukan tahapan pre-processing untuk mengatasi masalah seperti nilai yang hilang atau outlier sebelum melakukan proses mining itu sendiri. Keseluruhan, pengumpulan data dalam konteks Machine Learning menjadi fondasi penting untuk memastikan keberhasilan proses analisis dan pengambilan keputusan yang akurat.

2. Preprocessing Data

Preprocessing data merupakan tahap kritis dalam proses Machine Learning yang bertujuan untuk mempersiapkan dan membersihkan data mentah sehingga dapat diolah oleh model analisis dengan lebih efektif dan akurat. Langkah-langkah preprocessing mencakup penanganan nilai-nilai yang hilang, deteksi dan penanganan outlier, normalisasi data, serta pemilihan dan transformasi fitur yang relevan. Tujuan utamanya adalah menghilangkan noise atau gangguan yang dapat memengaruhi kualitas hasil analisis, meningkatkan keakuratan model, dan mengoptimalkan performa algoritma Machine Learning. Proses preprocessing juga dapat melibatkan penggabungan atau pengelompokan kategori data untuk

meningkatkan interpretasi dan generalisasi model. Dengan memastikan bahwa data telah diproses dengan benar sebelum dilibatkan dalam proses mining, hasil analisis dapat menjadi lebih dapat diandalkan, membantu pengambilan keputusan yang lebih baik, dan meningkatkan pemahaman terhadap pola atau tren yang tersembunyi dalam dataset.

3. Pembagian Dataset

Pembagian dataset menjadi dua subset utama, yaitu data training dan data testing, merupakan praktik yang umum dalam Machine Learning untuk menguji dan mengevaluasi performa model yang dibangun. Data training digunakan untuk melatih model dan menyesuaikannya dengan pola atau hubungan dalam data, sementara data testing digunakan untuk menguji sejauh mana model mampu mengeneralisasi dan membuat prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Pembagian dataset ini bertujuan untuk menghindari overfitting, di mana model terlalu beradaptasi dengan data pelatihan dan kurang mampu melakukan generalisasi pada data baru. Proporsi pembagian antara data training dan data testing dapat bervariasi tergantung pada karakteristik dan ukuran dataset, namun, pembagian yang umum adalah sekitar 70-80% untuk data training dan 20-30% untuk data testing. Dengan adanya pembagian dataset ini, pengguna dapat memastikan bahwa model yang dihasilkan memiliki kemampuan prediktif yang baik dan dapat diandalkan pada data yang belum pernah dilihat sebelumnya.

4. Perancangan Sistem Machine Learning Metode Naïve Bayes

Perancangan sistem Machine Learning menggunakan metode Naive Bayes melibatkan beberapa tahap penting untuk membangun model klasifikasi yang

efektif. Pertama, dataset perlu dikumpulkan dan dipersiapkan dengan melakukan preprocessing data, termasuk penanganan nilai yang hilang, normalisasi, dan pemilihan fitur yang relevan. Selanjutnya, dataset dibagi menjadi dua bagian, yaitu data training untuk melatih model dan data testing untuk mengevaluasi kinerja model. Selama tahap pelatihan, probabilitas prior dan likelihood dihitung berdasarkan data training, dan model Naive Bayes dikonstruksi dengan asumsi independensi antara fitur-fitur. Setelah model terlatih, langkah klasifikasi dapat dilakukan pada data testing, di mana probabilitas posterior dihitung dan kelas dengan probabilitas tertinggi dipilih sebagai prediksi. Penting untuk melakukan evaluasi terhadap model, menggunakan metrik-metrik seperti akurasi, presisi, dan recall, untuk memastikan keandalan dan keefektifan model Naive Bayes yang telah dibangun. Keseluruhan, perancangan sistem Machine Learning dengan metode Naive Bayes membutuhkan pemahaman yang baik terhadap dataset, pemilihan fitur yang tepat, dan evaluasi yang cermat terhadap kinerja model.

5. Klasifikasi Data Metode Naïve Bayes

Klasifikasi data dengan metode Naive Bayes adalah proses yang didasarkan pada teorema probabilitas Bayes, di mana model berusaha untuk mengklasifikasikan instance data ke dalam kategori atau kelas yang sesuai. Metode Naive Bayes mengasumsikan independensi kondisional antara fitur-fitur, meskipun asumsi ini bersifat "naive" dan sederhana, namun seringkali cukup efektif dalam berbagai kasus. Selama tahap pelatihan, model Naive Bayes menghitung probabilitas prior dari setiap kelas berdasarkan frekuensi kemunculan kelas dalam data training, serta probabilitas likelihood dari setiap nilai fitur dalam

kelas tersebut. Selanjutnya, saat melakukan klasifikasi pada data baru, model menghitung probabilitas posterior untuk setiap kelas berdasarkan fitur-fitur yang diamati. Kelas dengan probabilitas posterior tertinggi kemudian dianggap sebagai prediksi untuk instance data tersebut. Metode Naive Bayes sering digunakan dalam klasifikasi teks, deteksi spam email, dan aplikasi lainnya, karena sederhana, cepat, dan sering memberikan hasil yang memuaskan terutama ketika independensi antar-fitur dapat dianggap cukup valid.

6. Evaluasi Metode Naïve Bayes

Evaluasi metode Naive Bayes merupakan langkah penting untuk mengukur kinerja dan keandalan model klasifikasi yang telah dibangun. Metrik evaluasi yang umum digunakan mencakup akurasi, presisi, recall, dan F1-score. Akurasi mengukur sejauh mana model berhasil mengklasifikasikan instance-data dengan benar, sementara presisi menilai ketepatan model dalam mengidentifikasi instance-data positif. Recall, di sisi lain, mengukur sejauh mana model mampu mendeteksi semua instance-data positif yang seharusnya diidentifikasi. F1-score adalah metrik yang menggabungkan presisi dan recall untuk memberikan ukuran holistik terhadap keseimbangan antara keduanya. Selain itu, matriks kebingungan (confusion matrix) dapat memberikan wawasan lebih detail tentang kinerja model, terutama dalam konteks false positives dan false negatives. Evaluasi Naive Bayes juga dapat mencakup analisis ROC (Receiver Operating Characteristic) dan AUC (Area Under the Curve) untuk mengukur tingkat ketahanan model terhadap variasi threshold. Keseluruhan, evaluasi metode Naive Bayes adalah tahap krusial untuk

memastikan bahwa model memberikan hasil klasifikasi yang akurat dan dapat diandalkan pada dataset yang beragam.

7. ***Hasil Evaluasi Metode Naïve Bayes***

Hasil evaluasi metode Naive Bayes memberikan wawasan mendalam terhadap kinerja model klasifikasi yang telah dibangun. Akurasi, presisi, recall, dan F1-score adalah metrik-metrik utama yang digunakan untuk menilai kemampuan model dalam mengklasifikasikan instance-data. Akurasi mengukur persentase instance-data yang diklasifikasikan dengan benar, sementara presisi dan recall memberikan informasi tentang ketepatan dan kelengkapan model dalam mengidentifikasi instance-data positif. Sebuah nilai F1-score yang tinggi menunjukkan keseimbangan yang baik antara presisi dan recall. Analisis matriks kebingungan memberikan gambaran yang lebih rinci tentang kinerja model, termasuk jumlah false positives dan false negatives. Selain itu, kurva ROC dan nilai AUC dapat memberikan informasi tentang sejauh mana model Naive Bayes dapat menangani variasi threshold pengklasifikasian. Hasil evaluasi ini membantu mengidentifikasi kelemahan model dan memandu pengoptimalkan, memastikan bahwa model memberikan prediksi yang konsisten dan dapat diandalkan dalam kasus pengklasifikasian data yang lebih luas.

2.7. Peneliti Terdahulu

Referensi Penelitian	1
Judul	Analysis of the Naïve Bayes Method for Determining Social Assistance Eligibility Public

Nama	Adinda Pratiwi Siregar1)* , Deci Irmayani2) , Mila Nirmala Sari3)
Tahun	2023
Hasil	Metode Naive Bayes merupakan algoritma klasifikasi yang sangat efektif untuk berbagai jenis data, termasuk dalam konteks penentuan kelayakan penerima bantuan sosial. Dalam penelitian "Analisis Metode Naïve Bayes untuk Menentukan Kelayakan Bantuan Sosial", metode ini diterapkan untuk mengklasifikasikan data penerima bantuan berdasarkan fitur-fitur tertentu, seperti pendapatan, jumlah anggota keluarga, dan kondisi sosial-ekonomi lainnya. Metode Naive Bayes bekerja dengan prinsip probabilistik yang sederhana namun kuat, menghitung kemungkinan suatu data termasuk dalam kategori tertentu berdasarkan fitur-fitur yang dimilikinya. Dalam penelitian tersebut, metode Naive Bayes berhasil mencapai akurasi 100%, menunjukkan kemampuannya yang luar biasa dalam menentukan kelayakan penerima bantuan sosial dengan akurasi dan efisiensi tinggi. Hal ini menegaskan bahwa Naive Bayes adalah pilihan yang tepat untuk klasifikasi dalam

	aplikasi praktis seperti ini, karena mampu menangani data dengan kompleksitas tinggi secara efektif [8].
Referensi Penelitian	2
Judul	Implementation of the Naïve Bayes Method to determine the Level of Consumer Satisfaction
Nama	Fitri Febriyani Hasibuan1)*, Muhammad Halmi Dar2), Gomal Juni Yanris3)
Tahun	2023
Hasil	Metode Naive Bayes adalah algoritma klasifikasi yang sangat efektif dan efisien untuk digunakan dalam berbagai aplikasi, termasuk untuk mengklasifikasikan tingkat keMampuan mahasiswa. Dalam penelitian tertentu, metode ini telah berhasil mencapai akurasi sebesar 97%, menunjukkan kemampuannya dalam mengelola data dengan baik dan memberikan prediksi yang akurat. Algoritma ini bekerja dengan menghitung probabilitas dari setiap kategori berdasarkan fitur-fitur yang ada, sehingga memungkinkan pengklasifikasian data dengan cepat dan tepat. Keunggulan ini menjadikan

	Naive Bayes sebagai pilihan yang sangat bagus untuk analisis tingkat keMampuan mahasiswa [17].
Referensi Penelitian	3
Judul	Implementation of the Naïve Bayes Method to Determine Student Interest in Gaming Laptops
Nama	Rico Fadly Nasution[1)*], Muhammad Halmi Dar[2)], Fitri Aini Nasution[3)]
Tahun	2023
Hasil	Metode Naive Bayes adalah algoritma klasifikasi yang sangat efektif dalam menganalisis dan mengelompokkan minat mahasiswa pada laptop gaming. Dalam penelitian tertentu, metode ini telah berhasil mencapai akurasi sebesar 99%, menunjukkan kemampuannya dalam memberikan prediksi yang sangat akurat. Naive Bayes bekerja dengan menghitung probabilitas dari setiap kategori berdasarkan fitur-fitur yang ada, seperti spesifikasi laptop, preferensi mahasiswa, dan faktor harga. Keunggulan ini menjadikan Naive Bayes sebagai pilihan yang sangat bagus untuk melakukan klasifikasi terkait minat mahasiswa pada laptop gaming, memungkinkan pengambilan

	keputusan yang tepat berdasarkan analisis data yang andal [7].