TESIS

ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA

Disusun oleh:

Nama	**: Almi Yulistia Alwanda**
NIM	**: 22.55.2352**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

TESIS

ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA

COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR AND NAÏVE BAYES ALGORITHMS IN STUDENT CONCENTRATION SELECTION

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Almi Yulistia Alwanda**
NIM	**: 22.55.2352**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

HALAMAN PENGESAHAN

ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA

COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR AND NAÏVE BAYES ALGORITHMS IN STUDENT CONCENTRATION SELECTION

Dipersiapkan dan Disusun oleh

Almi Yulistia Alwanda

22.55.2352

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada Senin, 05 Agustus 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 05 Agustus 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA

COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR AND NAÏVE BAYES ALGORITHMS IN STUDENT CONCENTRATION SELECTION

Dipersiapkan dan Disusun oleh

Almi Yulistia Alwanda

22.55.2352

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 05 Agustus 2024

Pembimbing Utama

Prof. Dr. Ema Utami, S.Si., M.Kom
NIK. 190302037

Anggota Tim Penguji

Dhani Ariatmanto,S.Kom.,M.Kom., Ph.D
NIK. 190302197

Pembimbing Pendamping

Alva Hendi Muhammad, S.T., M.Eng., Ph.D.
NIK. 190302493

Ainul Yaqin, M.Kom.
NIK. 190302255

Prof. Dr. Ema Utami, S.Si., M.Kom
NIK. 190302037

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 05 Agustus 2024
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Almi Yulistia Alwanda
NIM : 22.55.2352
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom
Dosen Pembimbing Pendamping : Ainul Yaqin, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 05 Agustus 2024
Yang Menyatakan,

Almi Yulistia Alwanda

HALAMAN PERSEMBAHAN

Bismillahirrahmanirrahim

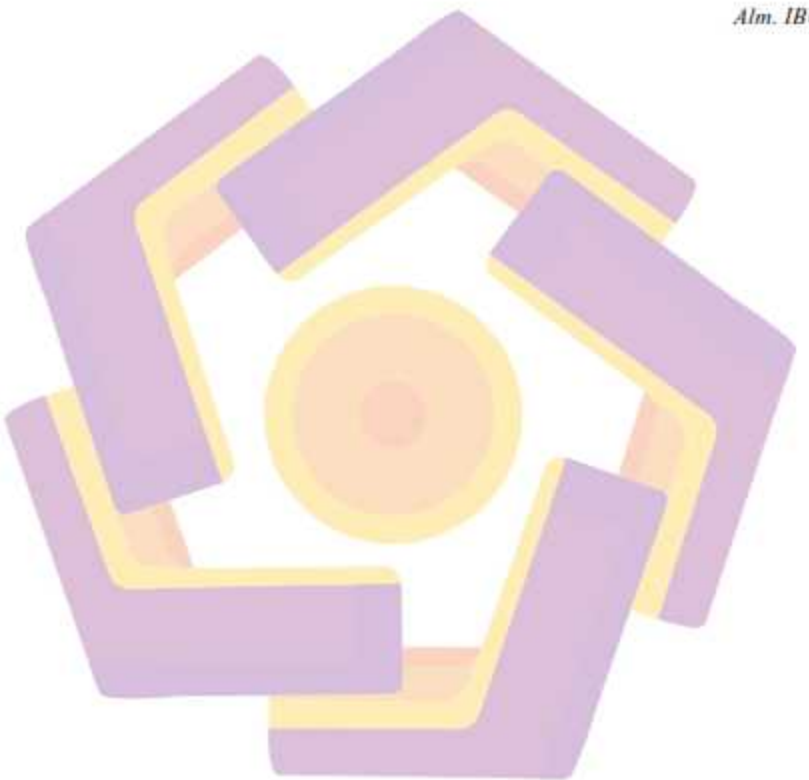
Dengan rasa syukur yang tak terhingga, penelitian tesis ini saya persembahkan sebagai ungkapan terima kasih kepada Allah SWT, Yang Maha Pengasih dan Penyayang, atas segala rahmat, petunjuk, dan karunia-Nya yang melimpah dalam setiap langkah perjalanan hidup. Kepada orang tua tercinta, yang telah memberikan cinta, doa, dan dukungan tak terhingga selama ini. Terima kasih atas pengorbanan, kasih sayang, dan dorongan yang memotivasi untuk terus berjuang dan berkarya.

Tidak lupa kepada Universitas AMIKOM Yogyakarta, atas fasilitas, ilmu pengetahuan, dan pengalaman yang telah diberikan selama proses pendidikan. Terima kasih kepada seluruh dosen, staff, dan rekan-rekan di universitas yang telah memberikan inspirasi, bimbingan, dan bantuan dalam mengejar mimpi dan cita-cita. Semoga segala jerih payah, doa, dan keikhlasan yang telah diberikan menjadi amal jariyah yang bermanfaat bagi kita semua.

HALAMAN MOTTO

"Jangan pernah risau dengan biaya pendidikan karna ilmu itu mahal. Tetaplah menuntut ilmu setinggi yang kamu bisa bagaimanapun kondisinya"

Alm. IBU



KATA PENGANTAR

Puji dan syukur kita panjatkan kehadirat Allah SWT atas rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis ini yang merupakan syarat dalam menyelesaikan jenjang Pendidikan S2 Pascasarjana Teknik Informatika Universitas Amikom Yogyakarta yang berjudul "ANALISIS PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBOR DAN NAÏVE BAYES DALAM PEMILIHAN KONSENTRASI MAHASISWA" dengan baik dan tepat waktu. Pada kesempatan ini Penulis ingin menyampaikan ucapan terimakasih kepada:

1. Kedua Orang Tua saya, adik dan nenek tercinta atas doa serta dukungannya selama ini.
2. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas AMIKOM Yogyakarta.
3. Ibu Prof. Dr. Kusrini, M.Kom., selaku Direktur Pascasarjana Universitas AMIKOM Yogyakarta.
4. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom., selaku Wakil Diektur Pascasarjana Universitas AMIKOM Yogyakarta dan juga sebagai dosen pembimbing utama yang telah banyak membantu memberikan ilmu, masukan dan saran dalam penelitian ini..
5. Bapak Ainul Yaqin, M.Kom selaku dosen pendamping yang telah mengarahkan penulis dan memberikan saran yang membangun dalam penelitian ini.
6. Segenap Dosen dan staff Magister Teknik Informatika Universitas AMIKOM Yogyakarta yang telah memberikan ilmu, wawasan, bantuan

dan pengalaman baru pada penulis selama perkuliahan

7. Rekan-rekan seperjuangan MTI PJJ Angkatan 8 yang memberikan motivasi, pengalaman baru dan juga semangat yang luar biasa semasa perkuliahan dan penyelesaian tesis ini.

Penulis menyadari bahwa dalam penyusunan tesis ini ada kekurangan. Oleh karena itu penulis dengan sangat senang hati menerima kritik dan saran yang membangun dari pembaca. Akhir kata, penulis berharap semoga tesis ini dapat memberikan manfaat bagi yang membacanya.

Yogyakarta, 05 Agustus 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Pada tahap penentuan konsentrasi studi dalam pendidikan tinggi, mahasiswa dihadapkan pada pilihan krusial yang memengaruhi arah karir mereka di masa depan. Universitas Hamzanwadi, Nusa Tenggara Barat, telah menjadi pusat pendidikan tinggi terkemuka dengan Fakultas Tekniknya yang menawarkan program studi seperti Teknik Informatika, Sistem Informasi, Teknik Komputer, dan Teknik Lingkungan. Kenaikan minat mahasiswa dalam program-program ini menciptakan kebutuhan akan penelitian yang mendukung mahasiswa dalam pengambilan keputusan mengenai konsentrasi studi yang sesuai.

Penelitian terdahulu menggunakan algoritma machine learning, seperti Support Vector Machine (SVM), untuk memprediksi konsentrasi studi mahasiswa, namun dengan tingkat akurasi yang belum optimal. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa beberapa algoritma klasifikasi, termasuk SVM, K-Nearest Neighbors (KNN), dan Naive Bayes, dalam meramalkan konsentrasi mahasiswa. Evaluasi terhadap ketiga algoritma tersebut menunjukkan kelebihan dan kekurangan masing-masing, dengan Naive Bayes memberikan hasil terbaik dengan akurasi sekitar 84%.

Untuk penelitian selanjutnya, disarankan untuk melakukan eksperimen lebih lanjut dengan mempertimbangkan faktor-faktor lain yang dapat memengaruhi kinerja model klasifikasi. Dengan demikian, dapat diperbaiki performa model klasifikasi untuk memprediksi konsentrasi studi mahasiswa secara lebih akurat, sehingga membantu dalam pengambilan keputusan yang lebih cerdas bagi mahasiswa dan staf akademis.

Kata kunci: Konsentrasi, SVM, KNN, Naïve Bayes

ABSTRACT

In the stage of determining study concentrations in higher education, students are faced with crucial decisions that affect their future career paths. Hamzanwadi University, located in West Nusa Tenggara, has established itself as a leading center of higher education with its Faculty of Engineering offering programs such as Informatics Engineering, Information Systems, Computer Engineering, and Environmental Engineering. The increased interest of students in these programs has created a need for research to support students in making informed decisions regarding their study concentrations.

Previous research has utilized machine learning algorithms, such as Support Vector Machine (SVM), to predict students' study concentrations, albeit with suboptimal accuracy rates. Therefore, this study aims to compare the performance of several classification algorithms, including SVM, K-Nearest Neighbors (KNN), and Naive Bayes, in forecasting student concentrations. Evaluation of these three algorithms revealed their respective strengths and weaknesses, with Naive Bayes yielding the best results with an accuracy rate of approximately 84%.

For future research, it is recommended to conduct further experiments considering other factors that may influence the performance of classification models. By doing so, the performance of classification models in predicting student study concentrations can be improved, thereby aiding in making smarter decisions for both students and academic staff.

Keyword: concentrations, SVM, KNN, Naïve Bayes

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam perjalanan pendidikan tinggi, mahasiswa dihadapkan pada tahap penentuan konsentrasi studi atau peminatan yang menjadi fokus utama dalam pengembangan keterampilan dan pengetahuan mereka. Keputusan ini menjadi langkah krusial yang memengaruhi arah dan relevansi karir di masa depan. Sejalan dengan perkembangan pesat dalam bidang teknologi dan tuntutan pasar kerja yang dinamis, mahasiswa diharapkan mampu membuat pilihan konsentrasi yang cerdas dan sesuai dengan minat serta potensi mereka.

Universitas Hamzanwadi, yang terletak di Nusa Tenggara Barat, telah membuktikan dirinya sebagai salah satu Perguruan Tinggi terkemuka di wilayah tersebut. Dengan enam fakultas yang mencakup berbagai disiplin ilmu, Fakultas Teknik menjadi salah satu pilar penting dalam menyediakan pendidikan tinggi yang berkualitas. Fakultas Teknik memiliki peran yang signifikan dengan menyelenggarakan empat program studi, yaitu Teknik Informatika, Sistem Informasi, Teknik Komputer, dan Teknik Lingkungan.

Keberhasilan Fakultas Teknik, terutama dalam program-program studi tersebut, tercermin dari pertumbuhan tahunan jumlah mahasiswa baru. Prospek kerja yang semakin besar dalam bidang-bidang tersebut menjadi alasan utama di balik peningkatan signifikan ini. Dengan demikian, kenaikan minat mahasiswa untuk bergabung dengan program studi di Fakultas Teknik menciptakan kebutuhan

akan penelitian yang dapat membantu mahasiswa dalam pengambilan keputusan mengenai konsentrasi studi atau peminatan yang paling sesuai dengan minat dan potensi mereka.

Namun, proses pengambilan keputusan mengenai konsentrasi studi seringkali merupakan tantangan yang kompleks bagi mahasiswa. Faktor-faktor seperti minat pribadi, potensi akademis, dan tuntutan industri menjadi pertimbangan yang perlu dipertimbangkan. Ketepatan dalam pemilihan konsentrasi akan memengaruhi peluang pekerjaan yang diminati oleh mahasiswa setelah menyelesaikan studi. Pemilihan bidang konsentrasi sangat penting karena berkaitan erat dengan kompetensi dan durasi studi yang akan dimiliki oleh mahasiswa. Jika mahasiswa salah memilih bidang konsentrasi, hal ini dapat merugikan mereka dalam hal waktu, tenaga, biaya, dan bahkan dapat menurunkan kualitas prestasi akademik. Selain itu, setelah mahasiswa menentukan konsentrasi, hal ini akan memengaruhi pemilihan topik untuk tugas akhir yang akan mereka ambil di semester akhir. Oleh karena itu, pemilihan konsentrasi studi yang tepat sangat penting bagi mahasiswa, karena keputusan ini tidak hanya menentukan kesesuaian dengan minat dan bakat mereka, tetapi juga berperan besar dalam menentukan keberhasilan karier di masa depan.

Sebagai respons terhadap kebutuhan ini, institusi pendidikan tinggi, khususnya Fakultas Teknik Universitas Hamzanwadi menyediakan beragam opsi konsentrasi studi atau peminatan yang dapat dipilih oleh mahasiswa. Perkembangan teknologi data mining saat ini merupakan sebuah alat yang sangat efisien dalam ranah teknologi informasi, terutama di dalam konteks bisnis yang penuh persaingan, terutama ketika masyarakat memasuki era Big Data. Dalam beberapa tahun

belakangan, penerapan data mining dalam sektor pendidikan telah menjadi semakin umum di perguruan tinggi, di mana teknologi ini digunakan untuk melakukan prediksi terkait mahasiswa, termasuk evaluasi kinerja mereka dalam proses pembelajaran, analisis kelulusan tepat waktu, mengidentifikasi potensi mahasiswa yang berisiko drop out, mengelola penerimaan mahasiswa baru, menentukan minat mahasiswa, dan berbagai aspek lainnya.

Dalam konteks ini, penelitian sebelumnya telah mencoba memanfaatkan algoritma machine learning, khususnya Support Vector Machine (SVM), untuk memprediksi konsentrasi studi mahasiswa. Meskipun demikian, hasil penelitian sebelumnya menunjukkan tingkat akurasi yang belum memuaskan, sekitar 66%, terutama saat menggunakan metode validasi silang (cross-validation) dengan 5 fold.

Ketidakoptimalan ini menggugah kebutuhan akan penyelidikan lebih lanjut untuk meningkatkan performa model klasifikasi. Oleh karena itu, penelitian ini bertujuan untuk memperdalam pemahaman dengan melakukan perbandingan antara beberapa algoritma klasifikasi yang umum digunakan, seperti Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes.

Dengan menyelidiki dan membandingkan performa ketiga algoritma ini, penelitian ini bertujuan untuk mengidentifikasi model klasifikasi yang paling optimal dalam meramalkan konsentrasi mahasiswa. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi positif dalam meningkatkan akurasi dan efektivitas sistem pendukung keputusan, yang pada gilirannya akan membantu

mahasiswa dan staf akademis dalam proses pengambilan keputusan mengenai konsentrasi studi yang sesuai.

Dalam rangka meningkatkan akurasi dan efektivitas prediksi, penelitian ini akan melakukan perbandingan antara tiga algoritma klasifikasi yang umum digunakan, yaitu Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes. Dengan melibatkan algoritma-algoritma ini, diharapkan dapat ditemukan model klasifikasi yang lebih optimal untuk meramalkan konsentrasi mahasiswa dengan tingkat akurasi yang lebih tinggi. Langkah ini diambil guna memberikan kontribusi lebih lanjut dalam pengembangan sistem pendukung keputusan yang dapat membantu mahasiswa dan staf akademis dalam proses pengambilan keputusan mengenai konsentrasi studi.

Banyak implementasi dan penelitian mengenai klasifikasi dan prediksi berbagai macam topic dengan menggunakan algoritma Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes. Salah satunya adalah penelitian yang dilakukan oleh Fradenly Keminhard Wattimury dkk. Pada penelitiannya dilakukan pengklasifikasian minat skripsi mahasiswa akhir dengan menggunakan algoritma Support Vector Machine (SVM) yang didasarkan pada penilaian tes minat Holland yaitu 6 Fitur (Realistis, Investigatif, Artistik, Sosial, Enterprising dan Conventional). Algoritma SVM menghasilkan tingkat akurasi 66% dengan menggunakan 5-Fold Cross Validation (Fradenly Keminhard et al., 2019).

Algoritma SVM juga digunakan untuk menentukan metode klasifikasi yang paling cocok untuk mengekstraksi fitur linear (seperti jalan, tepi bangunan dan

pembatas jalan) dalam pemetaan bahan permukaan perkotaan yang rinci dan skala kecil. Penelitian ini dilakukan oleh Siti Aekbal Salleh dkk dengan judul "Support Vector Machine (SVM) and Object Based Classification in Earth Linear Features Extraction : A Comparison". asil klasifikasi SVM mengandung permukaan impermabel yang salah diklasifikasikan dan fitur perkotaan lainnya, serta objek yang bercampur. Klasifikasi ini mencapai akurasi keseluruhan 75.1%.

Eleni Tsalera dkk melakukan penelitian dengan judul "Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm". penelitian ini berfokus pada pengklasifikasian kebisingan lingkungan ke dalam kategori yang berbeda menggunakan fitur suara yang diukur, dan memanfaatkan algoritma KNN dengan berbagai konfigurasi untuk mencapai hasil yang baik dalam hal akurasi klasifikasi. Algoritma ini telah dikonfigurasi untuk mempertimbangkan 1 hingga 3 tetangga, sementara tiga metrik jarak (Euclidean, Chebyshev, dan kosinus) digunakan untuk membuat 9 model yang mencapai kinerja antara 70% dan 85%.

Penelitian dengan judul "Traffic incident prediction and classification system using naïve bayes algorithm" yang dilakukan oleh Michael Libnao dkk. penelitian ini bertujuan untuk mengembangkan sistem prediksi dan klasifikasi insiden lalu lintas yang efektif menggunakan algoritma Naïve Bayes, dengan harapan meningkatkan manajemen insiden dan aliran lalu lintas, serta memberikan manfaat bagi pihak-pihak terkait. Hasilnya Sistem TIPCS yang menggunakan Algoritma Naïve Bayes mencapai akurasi 70,03% dalam memprediksi dan mengklasifikasikan insiden lalu lintas

Penelitian yang dilakukan oleh Mustafa dkk yang berjudul "Accuracy assessment of RFerns, NB, SVM and KNN machine learning classifiers in acualture". Penelitian tersebut bertujuan untuk membandingkan keakuratan dan efektivitas algoritma SVM, NB, RFerns dan KNN untuk mengidentifikasi kondisi yang menyebabkan penyakit pada ikan. Hasilnya menunjukkan bahwa klasifikasi SVM dan RFerns menghasilkan hasil yang akurat (100% untuk keduanya) selama fase pengujian, sementara klasifikasi kNN dan NB mencapai akurasi yang lebih rendah (91,3% untuk keduanya). Namun, penelitian saya berbeda dengan penelitian yang dilakukan oleh Mustafa dkk dalam beberapa aspek penting. Penelitian Mustafa dkk berfokus pada identifikasi kondisi yang menyebabkan penyakit pada ikan menggunakan algoritma SVM, Naive Bayes, RFerns, dan KNN, di mana hasil akurasi menunjukkan bahwa SVM dan RFerns mencapai akurasi sempurna (100%). Di sisi lain, penelitian saya bertujuan untuk membandingkan keakuratan algoritma SVM, KNN, dan Naive Bayes dalam mengklasifikasikan pemilihan konsentrasi mahasiswa berdasarkan atribut akademis.

Budiman dkk melakukan penelitian untuk membandingkan algoritma klasifikasi data mining untuk penelusuran minat calon mahasiswa baru. Teknik data mining yang digunakan dalam klasifikasi ini di antaranya Naïve Bayes, Decision Tree J48, dan K-Nearest Neighbor. Dataset yang digunakan sebanyak 5934 record, kemudian memberikan hasil pengujian terhadap ketiga model klasifikasi maka nilai akurasi tertinggi diperoleh pada klasifikasi Decision Tree J48 yang memperoleh nilai 90,3%, sedangkan klasifikasi K-Nearest Neighbor memiliki akurasi yang lebih

rendah yaitu 87,52% dan klasifikasi naïve bayes memiliki akurasi yaitu 87,24% (Budiman et al., 2021).

Penelitian yang dilakukan oleh Khaerul Anam dkk untuk peminatan program studi dengan menerapkan algoritma pada model klasifikasi. Algortima yang digunakan sebagai komparasi yaitu algoritma Decision Tree (C4.5), Naive Bayes, k-Nearest Neighbour dan Support Vector Machine (SVM). . Hasil dari pengujian akurasi algoritma didapatkan algoritma SVM memiliki akurasi terbaik dengan nilai 80,76%. Sedangkan algoritma dengan akurasi paling rendah adalah Naive Bayes dengan nilai 74,64%. Sedangkan dua algoritma lainnya memiliki tingkat akurasi berurutan yaitu 80,47% untuk Decision Tree dan 76,09% untuk k-NN (Khaerul Anam et al., 2022).

Budiman dkk melakukan penelitian untuk penelusuran minat calon mahasiswa dengan menggunakan algoritma Naïve Bayes (NB), Decision Tree J48 (J48), K-Nearest Neighbor (K-NN), Random Forest (RF) dan Support Vector Machine (SVM). Pengujian kelima algoritma klasifikasi bertujuan untuk menganalisis kinerja masing-masing algoritma, sehingga memberikan informasi dan pengetahuan data set sebagai rekomendasi dalam strategi promosi dimasa yang akan datang. Hasil pengujian menunjukkan tingkat akurasi tertinggi pada algoritma klasifikasi J48 sebesar 90,34% diikuti RF sebesar 89,04%, SVM sebesar 88,43%, K-NN sebesar 87,53%, SVM sebesar 87,25% (Budiman et al., 2022).

Penelitian selanjutnya dilakukan oleh Nora Lizarti dkk untuk menentukan peminatan studi STMIK Amik Riau dengan menggunakan algoritma K-Nearest Neighbor (KNN). Pengujian menggunakan tools RapidMiner untuk mengukur

performa algoritma. Hasil pengujian yang dilakukan terhadap 183 data latih dan 100 data uji menyatakan algoritma K-NN memiliki performa dengan hasil Acuracy, Recall, Precision, F Measure, dan Clasificassion Error dengan nilai 98%, 100%, 100%, 91.67%, dan 2% (Nora Lizarty et al., 2019).

Indah Hidayanti dkk melakukan penelitian untuk menentukan konsentrasi jurusan mahasiswa dengan menggunakan algoritma C4.5 dan Naïve Bayes dengan menggunakan aplikasi Rapid Miner sebagai alat bantu untuk mengklasifikasikan penjurusan mahasiswa. Pada penelitian ini diketahui algoritma C4.5 memiliki tinggkat akurasi 48,06 % dan naïve bayes 42,79% (Indah Hidayanti et al., 2020).

1.2. Rumusan Masalah

Adapun rumusan masalah pada penelitian ini yaitu :

a. Bagaimana tingkat akurasi dan efektivitas model klasifikasi Support Vector Machine (SVM) dalam meramalkan konsentrasi mahasiswa di Fakultas Teknik Universitas Hamzanwadi?
b. Sejauh mana perbandingan performa antara Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes dalam konteks klasifikasi konsentrasi mahasiswa dapat memberikan wawasan terkait keunggulan dan kelemahan masing-masing algoritma?
c. Apakah Penggunaan model klasifikasi Naïve Bayes dapat memberikan hasil yang lebih optimal dibandingkan dengan Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) dalam meramalkan konsentrasi mahasiswa?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini antara lain :

a. Data yang digunakan adalah data mahasiswa di program studi Sistem Informasi Fakultas Teknik Universitas Hamzanwadi.
b. Fokus perbandingan dalam penelitian ini terbatas pada tiga algoritma klasifikasi yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes.
c. Penelitian ini membatasi analisis pada fitur atau variabel tertentu yang digunakan untuk meramalkan konsentrasi mahasiswa. Beberapa faktor seperti variabel tingkat partisipasi dalam kegiatan ekstrakulikuler atau variabel lainnya yang mungkin mempengaruhi konsentrasi mahasiswa tidak akan diperhitungkan dalam penelitian ini.
d. Pengumpulan data dilaksanakan pada bulan November 2023 dan akan diolah menggunakan Google Colab.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah :

a. Menilai tingkat akurasi model klasifikasi Support Vector Machine (SVM) dalam meramalkan konsentrasi mahasiswa di Fakultas Teknik Universitas Hamzanwadi.
b. Membandingkan performa algoritma Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naïve Bayes dalam konteks klasifikasi konsentrasi mahasiswa.

c. Menilai apakah model klasifikasi Naïve Bayes memberikan hasil yang lebih optimal dalam meramalkan konsentrasi mahasiswa dibandingkan dengan model Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN).

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah :

a. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengelolaan akademik di Fakultas Teknik Universitas Hamzanwadi dengan memberikan informasi tentang tingkat akurasi dan efektifitas model klasifikasi. Informasi ini dapat digunakan untuk mengoptimalkan perencanaan dan pengembangan program akademik yang sesuai dengan kebutuhan dan minat mahasiswa.
b. Perbandingan performa antara Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naïve Bayes dapat membantu dalam pemilihan algoritma yang paling sesuai untuk tugas klasifikasi konsentrasi mahasiswa. Ini akan memberikan wawasan kepada pengambil keputusan terkait keunggulan dan kelemahan masing-masing algoritma, sehingga dapat dipilih metode yang paling efektif dan efisien.
c. Penelitian ini dapat memberikan pandangan yang lebih mendalam tentang apakah penggunaan model klasifikasi Naïve Bayes lebih optimal dibandingkan dengan SVM dan KNN dalam meramalkan konsentrasi mahasiswa. Jika Naïve Bayes terbukti lebih efisien, hasil penelitian ini dapat menjadi dasar untuk mengoptimalkan sistem ramalan pemilihan konsentrasi mahasiswa di masa depan, sehingga respon terhadap perubahan dapat lebih cepat dan akurat.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian sebelumnya yang terkait dengan metode dan algoritma yang digunakan pada penelitian ini. Penelitian yang dilakukan oleh Iis Istianah dkk untuk pengukuran pemilihan mata kuliah peminatan. Faktor utama yang mempengaruhi pada pemilihan peminatan antara lain memahami mata kuliah inti pada program studi tertentu. Algoritma yang digunakan adalah naïve bayes dan data yang digunakan adalah data mahasiswa tahun angkatan 2013, 2014, 2015 & 2016. Dengan memunculkan 2 peminatan yaitu software engineering dan networking. Hasil pengujian yang dilakukan adalah tingkat akurasi naïve bayes sebesar 98,06% (Iis Istianah et al., 2021).

Aditya Elanda Gumanti dkk melakukan penelitian untuk klasifikasi topik skripsi mahasiswa di Fakultas Ilmu Komputer dengan menggunakan algoritma K-Nearest Neighbor. Fakultas iLmu Komputer Universitas Lancang Kuning memberikan beberapa pilihan topuk yang dapat dipilih oleh mahasiswa. Kesimpulan yang didapat dari penelitian ini optimasi nilai K menggunakan k-fold cross validation menghasilkan tingkat akurasi 56,67% dengan nilai k-fold cross validation = 2 dan nilai K-5. Dari hasil klasifikasi menggunakan algoritma KNN hasilnya adalah sebanyak 73 mahasiswa berminat mengambil topik skripsi Rekayasa Perangkat Lunak (RPL), 48 mahasiswa berminat mengambil topik skripsi

kecerdasan buatan (AI) dan 0 atau tidak ada mahasiswa yang berminat mengambil topic skripsi jaringan.

Penelitian selanjutnya dilakukan oleh Nanang untuk membandingkan algoritma Decision Tree dan Naïve Bayes dalam penjurusan siswa MAN 1 Kota Tangerang Selatan. Dalam penelitian ini, data yang digunakan adalah data hasil ujian masuk dan data nilai dari hasil psikotes yang diselenggarakan di sekolah untuk siswa baru yang telah dinyatakan lulus seleksi masuk MAN 1 Kota Tangerang Selatan. Penelitian ini menggunakan dua algoritma klasifikasi data mining yaitu Naïve Bayesdan Decision Tree, dalam penelitian ini dapat dibuktikan Naïve Bayesmenghasilkan akurasi lebih tinggi dibandingkan Decision Tree, dengan nilai akurasi 95,92 % untuk algoritma Naïve Bayesdan 91,84 % untuk algoritma Decision Tree (Nanang., 2022).

Muhammad Zainuri dkk melakukan penelitian untuk komparasi metode klasifikasi algoritma C5.0 dan Naïve Bayes untuk menentukan jurusan siswa. Berdasarkan hasil perbandingan pengujian yang telah dilakukan melalui berbagai skenario terhadap kedua metode tersebut, pengujian 10-fold cross validation yang kemudian dicatat dalam confusion matrixmenghasilkan nilai akurasi yaitu sebesar 60,87% untuk algoritmaC5.0sedangkan untuk naïve bayessebesar 56,52%. Dari hasil yang diperoleh algoritmaC5.0merupakan metode paling baik dibanding naïve bayesyang dibuktikan dengan nilai tingkat akurasi yang didapatkan lebih tinggi.

Penelitian selanjutnya dilakukan oleh Mulya Cahya Ramadanty dkk untuk klasifikasi peminatan program studi di Perguruan Tinggi berdasarkan nilai rapor

dengan menggunakan algoritma C4.5 dan K-Nearest Neighbor. Tujuan dari penelitian ini adalah Mengimplementasikan algoritma C4.5 dan K-Nearest Neighbor (K-NN) dalam klasifikasi peminatan program studi. Algoritma yang digunakan pada penelitian ini yaitu C4.5 dan K-NN. Data yang digunakan adalah nilai rapor Matematika dan mata pelajaran produktif siswa kelas XII jurusan Teknik Komputer Jaringan (TKJ), Teknik Elektronika Industri (TEI), dan Rekayasa Perangkat Lunak (RPL) Sekolah Menengah Kejuruan Negeri (SMKN) 1 Karawang. Hasil yang didapat dari pengujian menggunakan tool RapidMiner sebesar 98,04% untuk algoritma K-NN dan 100% untuk algoritma C4.5. Pada tahap implementasi algoritma K-NN ke program diperoleh hasil sebesar 98% (Mulya Cahya et al., 2021).

Dewi Marini Umi dkk melakukan penelitian untuk menentukan peminatan mahasiswa menggunakan Vector Space Model dan Metode K-Nearest Neighbor (KNN). Tujuan dari penelitian ini yaitu memberikan sebuah penunjang keputusan bagi pihak Jurusan agar setiap judul skripsi yang diajukan oleh mahasiswa sesuai dengan peminatan. Berdasarkan hasil penelitian yang telah dilakukan, model yang dibangun menggunakan algoritma KNN menghasilkan tingkat akurasi yang lebih tinggi jika dibandingkan dengan model yang dibangun menggunakan algoritma VSM. Nilai akurasi tertinggi berdasarkan hasil pengujian pada penelitian ini adalah sebasar 96,85%.

Muhammad Farid Satrio dkk melakukan penelitian untuk pemilihan konsentrasi mahasiswa menggunakan algoritma Naïve Bayes. Hasil pengujian yang telah dilakukan terhadap sample dataset sebanyak 1534 data menggunakan

algoritma naïve bayes, diperoleh bahwa hasil prediksi untuk menentukan konsentrasi memiliki nilai akurasi sebesar 84,27% (Muhammad Farid et al., 2022).

Penelitian yang dilakukan oleh Lingga Kurnia Ramadhani dkk untuk perbandingan metode klasifikasi Naïve Bayes dan Support Vector Machine (SVM) pada predikat kelulusan mahasiswa. Pada penelitian-penelitin sebelumnya Naïve Bayes cenderung menghasilkan performa dn akurasinya yang lebih baik dibandingkan dengan Support Vector Machine. Kemudian setelah dilakukan pengujian dengan menggunakan data kelulusan mahasiswa ternyat Naïve Bayes memiliki akurasi lebih besar yaitu 96,52% dengn tingkat eror 0,03% sedangkan metode Support Vector Machine menghasilkan akurasi 86,93% dengan tingkat eror 0,13% (Lingga Kurnia et al., 2022).

Anang Prayogo dkk melakukan penelitian untuk klasifikasi judul artikel pada jurnal ilmih dengan menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor (KNN). Dataset yang digunakan berupa judul artikel jurnal sejumlah 200 data yang kemudian dipisahkan menjadi data uji dan data latih. Hasil akurasi yang didapat dengan menggunakan K-NN sebesar 85,00% dengan data uji 10% dan k=3. Hasil akurasi menggunakan NB adalah sebesar 81,66% dengan menggunakan data uji 30%. Akurasi tidak terlalu berbeda jauh namun NB menciptakan akurasi yang sedikit lebih besar dibandingkan K-NN (Anang Prayogo, 2023).

Penelitian selanjutnya dilakukan oleh Aulia Putri dkk untuk prediksi kelulusan mahasiswa tingkat akhir dengan menggunakan algoritma K-Nearest Neighbor (KNN), Naïve Bayes dan Support Vector Machine (SVM). Dengan 379 orang mahasiswa tahap akhir sebagai responden. Berdasarkan perbandingan ketiga

algoritma tersebut dengan teknik splitting data sehingga didapatkan bahwa algoritma K-nearest Neighbor (KNN) memiliki rata-rata lebih tinggi dibandingkan dengan Naïve Bayes dan Support Vector Machine (SVM) untuk prediksi kelulusan mahasiswa tingkat akhir dengan akurasi 87,8% dan recall 84% (Aulia Putri et al., 2023).

Arisa Dwi Cahyo melakukan penelitian untuk klasifikasi masa studi sarjana dengan menggunakan metode Naïve Bayes. Pemilihan algoritma ini Naive Bayes dapat digunakan untuk melakukan prediksi probabilitas keanggotaan suatu class. Dengan adanya prediksi kelulusan diharapkan mampu memantu memprediksi tingkat kelulusan mahasiswa. Fitur –fitur yang digunakan dalam prediksikelulusan mahasiswa yaiitu Jenis Kelamin, Status Pernikahan,Status Pekerjaan, dan IPS1,IPS2,IPS3,IPS4.Prediksi kelulusan menggunakan data alumni yang sudah lulus. Berdasarkan perhitungan akurasi Naive Bayes menggunkan k-fold cross validation dengan hasilperhitungan rata-rata k=5 adalah sebesar 95%, dan perhitungan rata-rata k=10 sebesar 94% Berdasarkan hasil tersebut maka algoritma Naive Bayes bisa digunakan untuk prediksi kelulusan mahasiswa.

2.2. Keaslian Penelitian

Tabel 2. 1. Matriks Literature Review

Optimasi Model Klasifikasi : Support Vector Machine, K-Nerast Neighbor dan Naive Bayes untuk Konsentrasi Mahasiswa

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Evaluation of Support Vector Machine, Naïve Bayes, Decision Tree and Gradient Boosting Algorithms for Sentiment Analysis on ChatGPT Twitter Dataset.	Salsabila Rabbani, Dea Safitri, Farida Try Puspa Siregar, Rahmaddeni, Lusiana Efrizoni (Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)), 2023.	Untuk klasifikasi sentiment terhadap ChatGPT terhadap kesulitan dalam menentukan apakah respons yang diberikan oleh pengguna bersifat positif, negative atau netral terhadap penggunaan ChatGPT.	Melalui hasil penelitian, ditemukan bahwa respons pengguna Twitter cenderung negatif terhadap ChatGPT, dan algoritma Support Vector Machine dengan pembagian data 90:10 memiliki nilai akurasi tertinggi dengan pencapaian 80%, dibandingkan dengan algoritma lainnya.	Penambahan dataset dalam jumlah besar dan juga variable lainnya untuk menambahkan hasil akurasi dari metode yang digunakan.	Penelitian yang dilakukan olch salsabila rabbani dkk untuk klasifikasi sentiment terhadap ChatGPT membandingkan algoritma Support Vector Machine, Naïve Bayes, Decision Tree and Gradient Boosting.
2	Comparison Of K-Nearest Neighbor and Naïve Bayes Methods Fo ClassificationOf News Content.	Andi Tejawati, Anindita Septiarini, Rondongalo Rismawati, Novianti Puspitasari, JUTIF (Jurnal Teknik Informatika), 2023	melakukan perbandingan metode K-Nearest Neighbor (KNN) dan Naïve Bayes untuk melakukan klasifikasi konten berita sehinga diperoleh metode yang terbaik.	metode KNN mampu menghasilkan nilai akurasi yang lebih tinggi dibandingkan Naïve Bayes yaitu mencapai 86% dari 51% dengan data uji sejumlah 100 artikel berita.	Untuk mendapatkan hasil akurasi lebih baik lagi, penelitian selanjutnya dapat menggunakan metode lainnya sebagai bahan perbandingan	Penelitian yang dilakukan olch Andi Tejawati untuk melakukan perbandingan metode K-Nearest Neighbor (KNN) dan Naïve Bayes.

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun.	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Comparison of SVM, NBC and KNN Classification Methods in Determining Students Majors at SMK N02 Manokwari.	Siska Howay and Suhirman (Journal of Computer Science and Technology Studies), 2023.	Untuk penentuan jurusan bagi calon siswa SMK.	Disimpulkan bawah Untuk metode KNN, diperoleh akurasi sebesar 54,56%, untuk metode NBC sebesar 74,78%, dan untuk metode SVM sebesar 58,70%. Dengan demikian, dapat disimpulkan bahwa ketiga metode, berdasarkan atribut yang digunakan oleh metode NBC, memiliki akurasi yang tinggi, yaitu sebesar 74,78%.	selanjutnya dapat menggunakan metode lainnya sebagai bahan perbandingan.	algoritma klasifikasi Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), dan K-Nearest Neighbors (KNN). Perbandingan metode NBC, KNN, dan SVM diukur menggunakan akurasi
4	C4.5, k-Nearest Neighbor, Naïve Bayes and Random Forest Algorithms Comparison to Predict Students On Time Graduation.	Gunawan, Hanes and Catherine(Indonesian Journal Of Artificial Intelligence and Data Mining), 2021.	Untuk memprediksi kelulusan tepat waktu.	Disimpulkan bahwa Hasil dari proses klasifikasi dievaluasi menggunakan validasi silang dan matriks kebingungan untuk menentukan algoritma klasifikasi data mining yang paling akurat untuk memprediksi kelulusan mahasiswa	Penelitian selanjutnya dapat dilakukan di sebuah perusahaan atau Lembaga agar mendapatkan dataset yang real dan hasil yang bervariasi dengan algoritma yang lebih bervariasi lagi	Algoritma yang akan dibandingkan dalam penelitian ini adalah C4.5, K-Nearest Neighbor, Naive Bayes, dan Random Forest.

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				tepat waktu, di mana algoritma K-Nearest Neighbor dan Random Forest memiliki akurasi tertinggi sebesar 72,651%, diikuti oleh algoritma C4.5 sebesar 72,453%, dan algoritma Naive Bayes sebesar 71,86%.		
5	Comparative Study of Supervised Algorithms for Prediction of Students Performance	Madhuri T. Sathe and Amol C. Adamuthe (IJ Modern Education and Computer Science), 2021.	Untuk memprediksi kinerja akademis siswa.	Kinerja teknik-teknik ini diuji pada tiga set data yang berbeda. Hasil menunjukkan bahwa kinerja Random Forest dan C5.0 lebih baik daripada J48, CART, NB, KNN, dan SVM.	Dataset yang dimiliki dapat digunakan untuk penelitian selanjutnya dengan menggunakan beberapa metode yang lain guna mendapatkan hasil akurasi atau target nilai yang lebih tinggi lagi dari penelitian sebelumnya.	Comparative Study of Supervised Algorithms for Prediction of Students Performance

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6	Prediction of student exam performance using data mining classification algorithms	Dalia Khairy1 · Nouf Alharbi2 · Mohamed A. Amasha1 · Marwa F. Areed4 · Salem Alkhalaf3 · Rania A. Abougalala1, Education and Information Technologies https://doi.org/10.1007/s10639-024-12619-w	Penelitian ini bertujuan untuk menilai keberhasilan akademik mahasiswa dengan memanfaatkan teknik data mining pendidikan dan metrik statistik untuk menentukan apakah diperlukan promosi tambahan berdasarkan kinerja mereka	Kesimpulannya adalah bahwa algoritma machine learning (ML) dapat secara efektif digunakan untuk memprediksi kinerja akademis mahasiswa. Penelitian ini menganalisis data mahasiswa jurusan Ilmu Komputer dari tahun 2016 hingga 2021 menggunakan lima algoritma ML yaitu Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Neural Network (NN), dan K-Nearest Neighbor (KNN). Hasil penelitian menunjukkan bahwa algoritma RF dan DT memiliki kinerja terbaik dengan akurasi 98.7%, sedangkan NN mencapai akurasi 96.4%, NB 94%, dan KNN 89.6%.	Mengumpulkan data yang lebih besar, menggunakan algoritma yang beragam dan mempertimbangkan fitur tambahan	Algoritma yang dibandingkan adalah Random Forest, Decision Tree, Neural Network, Naïve Bayes dan KNN.

2.3. Landasan Teori

Data mining adalah disiplin ilmu yang digunakan untuk mengolah data dalam jumlah yang besar untuk menghasilkan informasi. Data mining menggunakan teknik statistic, matematika, kecerdasarn buatan untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai basis data besar. Data mining adalah salah stau bidang yang berkemabng pesat karena besarnya jebutuhan akan nilai tambah dari database skala besar yang makin banyak terakumulasi sejalan dengan pertumbuhan teknologi informasi. Data mining memiliki bebrbagai macam variasi teknik-teknik, metode dan algoritma yang tepat untuk dapat dipadukan dalam sebuah data mining agar bisa mengeluarkan hasil yang akurat.

Dalam klasifikasi terdapat variabel target yang bersifat kategoris yang dibagi menjadi kelas yang sudah ditentukan seperti kelas nasabah yang bermasalah atau tidak, hewan yang masuk ke dalam klasifikasi reptile, amfibi, mamalia, burung atau ikan. Klasifikasi proses penentuan model pola dengan menggambarkan perbedaan kelas data sehingga bisa digunakan untuk mencapai tujuan prediksi kelas yang belum diketahui (Annur Haditsah, 2018). Setiap algoritma bisa menghasilkan klasifikasi yang berbeda. Algoritma terbaik dapat dilihat dari data yang diklasifikasikan secara benar oleh model dengan data sebenarnya atau seberapa akurat model dapat memprediksi kelas klasifikasi. Teknik pada klasifikasi yang sering digunakan adalah seperti Naïve Bayes, K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM).

Pada Support Vetor Machine (SVM), kemampuannya mencakup pengklasifikasian data baik yang bersifat linier maupun non-linier. Data input merupakan nilai variabel-variabel prediktor, sedangkan outputnya adalah variabel target yang saling bergantung. Tujuan utama SVM adalah menemukan fungsi klasifikasi optimal untuk memisahkan anggota dari dua kelas dalam dataset pelatihan. Konsep fungsi klasifikasi "terbaik" dapat diinterpretasikan secara geometris, khususnya dalam konteks dataset terpisah secara linier. Fungsi klasifikasi linier terkait dengan hyperplane pemisah f(x) yang melintasi tengah dua kelas dan berhasil memisahkan keduanya (Neelamegam & Ramaraj, 2013). Rumus SVM yang sering dipakai adalah sebagai berikut :

a.) Polynomial Kernel

Proses Polynomial Kernel merupakan langkah penting dalam mengklasifikasikan dataset training yang telah dinormalisasi. Proses ini dapat diterapkan menggunakan rumus yang tercantum pada persamaan 1 berikut.

$$K(x_i,x) = (y.x^{T}x + r)^{p}, y > 0 \qquad (1)$$

b.) Radial Bias Function (RBF)

RBF adalah salah satu fungsi dalam SVM yang digunakan untuk mengklasifikasikan dataset yang tidak dapat dipisahkan secara linier. Proses ini memiliki keunggulan dalam memberikan akurasi yang sangat baik baik pada tahap training maupun prediksi. Rumus ini dapat diterapkan menggunakan persamaan 2 berikut.

$$K(x_i,x) = \exp(-y|x_i-x|^{2}), y > 0 \qquad (2)$$

c.) Sigmoid Kernel

Sigmoid adalah sebuah proses yang dikembangkan dari jaringan saraf tiruan. Rumus ini dapat diterapkan menggunakan persamaan 3 berikut.

$$K(x_i,x) = \tanh(yx^{T}x + r) \qquad (3)$$

K-NN, singkatan dari K-Nearest Neighbors, merupakan sebuah teknik pengklasifikasian data di mana data tersebut telah sebelumnya diklasifikasikan. Tingkat akurasi algoritma K-NN sangat dipengaruhi oleh berbagai karakteristik, terutama ketika nilai fitur tidak sesuai dengan perkiraan yang ada. Beberapa penelitian yang menggunakan algoritma K-NN secara khusus fokus pada pemilihan fitur dan pembobotan untuk meningkatkan efisiensi algoritma dalam proses klasifikasi.

Algoritma K-NN termasuk dalam kategori instance-based learning, di mana data training disimpan untuk memungkinkan pencarian klasifikasi untuk rekaman baru yang belum terklasifikasi. Hal ini dilakukan dengan membandingkan sebagian besar persamaan atau kesamaan dengan data training yang ada. Peran utama K-NN adalah menemukan jarak terdekat antar data, diikuti dengan mengevaluasi tetangga terdekat sebanyak K pada data latih. Data training direpresentasikan dalam ruang multidimensi, masing-masing mencerminkan karakteristik data.

Berikut adalah langkah-langkah perhitungan dalam algoritma KNN :

a. Menentukan Nilai K yang berupa bilangan bulat positif
b. Menghitung jarak antar data baru dan dengan data training

$$d = \sqrt{(x1 - y1)^2 + (x2 - y2)^2} \qquad (4)$$

Diketahui :

d : jangkauan

x : data sampel

y : data pengujian

c. Menentukan urutan jarak (jangkauan) paling dekat dengan jarak min pada K

d. Label tetangga terdekat digunakan untuk memprediksi data baru sedangkan persamanaan KNN dalam menghitung nilai prediksi.

Algoritma K-NN memiliki beberapa kelebihan, di antaranya keefektifan dalam menghadapi data berjumlah banyak dan kemampuan untuk menghasilkan prediksi dengan tingkat akurasi yang tinggi. Meskipun demikian, terdapat juga beberapa kekurangan yang perlu diperhatikan. Salah satu kekurangan utama dari algoritma K-NN adalah kebutuhan untuk menentukan nilai k yang optimal. Pemilihan nilai k yang tidak tepat dapat mempengaruhi performa algoritma dan menghasilkan prediksi yang kurang akurat. Oleh karena itu, proses pemilihan nilai k menjadi kritis dalam implementasi K-NN.

Selain itu, algoritma K-NN juga memerlukan biaya komputasi yang tinggi, terutama dalam menghitung jarak antara instance pada query dengan data latih. Proses ini dapat menjadi memakan waktu, terutama jika datasetnya sangat besar atau memiliki dimensi yang tinggi. Oleh karena itu, perlu mempertimbangkan secara hati-hati trade-off antara akurasi dan biaya komputasi saat menggunakan algoritma K-NN.

Naive Bayes Classifier (NBC) merupakan sebuah algoritma yang digunakan untuk memprediksi probabilitas suatu kejadian berdasarkan penerapan Teorema

Bayes. Algoritma ini mengasumsikan bahwa setiap fitur pada data adalah independen satu sama lain, meskipun dalam kenyataannya fitur-fitur tersebut mungkin saling terkait. NBC termasuk dalam kategori pengklasifikasi yang sederhana dan efisien.

Metode Naive Bayes ini bekerja dengan menghitung probabilitas kelas suatu instance berdasarkan probabilitas fitur-fitur yang ada pada instance tersebut. Walaupun termasuk dalam klasifikasi sederhana, NBC memiliki kemampuan memproses data dengan cepat, terutama saat jumlah data terbatas atau tidak terlalu banyak.

$$(\mathrm{H}|\mathrm{X}) = \frac{P(X|H)P(H)}{P(X)} \qquad (5)$$

Diketahui:

X	: data (label)
H	: hipotesa X adalah label
P(H)	: peluang H
P(X)	: peluang X
P(X|H)	: peluang X pada H
P(H|X)	: peluang H pada X

Terdapat beberapa tahapan pada algoritma Naive Bayes diantaranya yaitu :

a. Menghitung jumlah label yang ada P(H)
b. Menghitung jumlah kasus dari masing-masing label ((P(X|H)
c. Mengalikan semua variabel label (P(H|X) . P(H))

Performance, dalam konteks penelitian atau pengembangan model dan algoritma, merujuk pada ukuran atau metrik yang digunakan untuk mengevaluasi kualitas atau keakuratan suatu model. Penilaian performa ini memberikan gambaran seberapa baik model tersebut dapat memenuhi tujuan dan konsep penelitian yang telah ditetapkan.

Salah satu cara umum untuk mengukur performance adalah dengan menggunakan confusion matrix, yang membandingkan hasil prediksi model dengan kelas sebenarnya dari data. Dari confusion matrix, beberapa metrik evaluasi umum seperti akurasi, presisi, recall, dan lainnya dapat dihitung.

Tabel 2. 2. Confusion Matrix

Prediksi	Aktual	
	Prediksi	Aktual
Prediksi	TP	FP
Aktual	FN	TN

Dengan:
TP = True Positif
FP = False Positif
FN = False Negatif
TN = True Negatif

- Akurasi, Mengukur sejauh mana model dapat memberikan prediksi yang benar secara keseluruhan.
- Precision, Menunjukkan seberapa tepat model dalam memprediksi kelas positif. Rasio prediksi benar positif terhadap total prediksi positif.
- Recall, Menunjukkan sejauh mana model dapat menangkap semua instance yang seharusnya termasuk dalam kelas positif. Rasio prediksi benar positif terhadap total instance yang sebenarnya positif.

Rumus confusion matrix untuk menghitung accuracy dilihat pada persamaan 6, untuk precision pada persamaan 7 dan untuk perhitungan recall pada persamaan 8.

$$Accuracy = \frac{TP+TN}{Total} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Evaluasi adalah tahap kritis dalam pengembangan model atau algoritma, di mana kualitas dan keakuratan dari hasil yang dihasilkan dievaluasi secara menyeluruh. Proses evaluasi ini dapat melibatkan berbagai metode, dan salah satu pendekatannya adalah dengan melakukan uji coba menggunakan data uji. Penting untuk memilih metode evaluasi yang sesuai dengan hipotesis, tujuan, dan kerangka penelitian yang telah ditentukan sebelumnya. Dengan merinci tujuan penelitian, kita dapat memilih metrik evaluasi yang paling relevan dan sesuai dengan konteks masalah yang dihadapi.

Metode evaluasi yang umum digunakan melibatkan penggunaan data uji yang tidak digunakan dalam proses pelatihan model. Hal ini bertujuan untuk mengukur sejauh mana model dapat menggeneralisasi dan memberikan prediksi yang baik pada data yang belum pernah dilihat sebelumnya. Beberapa metode evaluasi yang umum digunakan melibatkan pembagian dataset menjadi subset pelatihan dan uji, penggunaan teknik validasi silang (cross-validation), atau bahkan pengujian eksternal dengan dataset yang sepenuhnya baru.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini dapat diklasifikasikan sebagai penelitian komparatif atau perbandingan. Fokus utama penelitian ini adalah membandingkan kinerja tiga model klasifikasi, yaitu Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes, dalam konteks konsentrasi mahasiswa.

Penelitian ini bersifat eksplanatif komparatif. Eksplanatif karena penelitian ini bertujuan untuk menjelaskan dan memahami perbandingan kinerja model klasifikasi. Komparatif karena penelitian ini mengevaluasi dan membandingkan tiga model klasifikasi yang berbeda.

Pendekatan penelitian yang digunakan dalam judul ini adalah pendekatan kuantitatif. Penelitian ini akan melibatkan analisis data numerik untuk mengevaluasi dan membandingkan performa SVM, KNN, dan Naïve Bayes. Pendekatan kuantitatif akan memberikan pemahaman yang jelas tentang perbandingan kinerja model klasifikasi untuk konsentrasi mahasiswa.

3.2. Metode Pengumpulan Data

Data didapat dari Fakultas Teknik Universitas Hamzanwadi yang termasuk juga ke dalam data private yang belum digunakan dalam penelitian-penelitian sebelumnya. Data diperoleh dengan cara observasi dan wawancara.

1. Observasi

   Salah satu teknik pengumpulan data yang dilakukan dengan pengamatan langsung ke lokasi penelitian yaitu ke Fakultas Teknik Universita Hamzanwadi.

2. Wawancara

   Wawancara dilakukan di lokasi penelitian dengan staff bagian akademik di Fakultas Teknik Universitas Hamzanwadi.

3.3. Metode Analisis Data

Metode analisis data untuk penelitian ini adalah kuantitatif, metode ini digunakan karena sesuai dengan penelitian yang dilakukan. Data yang diperoleh dari bagian akademik Fakultas Teknik Universitas Hamzanwadi dan instrument penelitian yang digunakan adalah data dan wawancara yang dilakukan di lokasi penelitian.

Support Vector Machine (SVM) merupakan pilihan yang tepat dalam penelitian ini karena SVM mampu menangani baik data linier maupun non-linier dengan baik. Dalam konteks klasifikasi konsentrasi mahasiswa, SVM dapat secara efektif membangun hyperplane yang memisahkan antara kelas konsentrasi yang berbeda. Kemampuan SVM untuk menangani data kompleks dan menemukan batas keputusan yang optimal membuatnya menjadi pilihan yang solid untuk penelitian ini.

K-Nearest Neighbors (KNN) menjadi pilihan yang relevan karena pendekatan ini berfokus pada kesamaan antar data. Dalam konteks konsentrasi

mahasiswa, KNN dapat mengidentifikasi pola dan hubungan antara mahasiswa yang memiliki konsentrasi serupa berdasarkan atribut tertentu seperti minat atau nilai akademis. KNN memberikan fleksibilitas dan dapat memberikan wawasan tentang bagaimana preferensi konsentrasi mahasiswa berkumpul dalam ruang atribut.

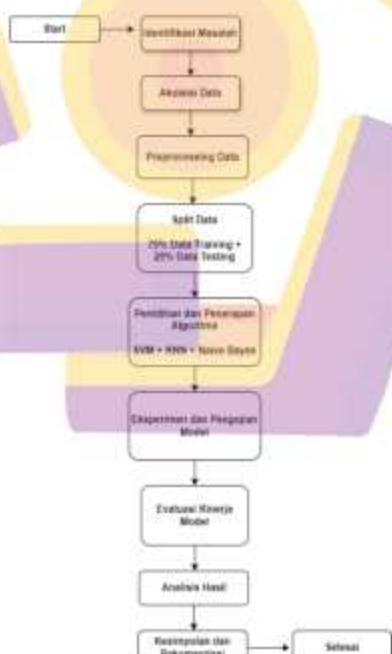
Naïve Bayes adalah pilihan yang cocok untuk penelitian ini karena kemampuannya dalam mengatasi masalah klasifikasi dengan dataset yang besar dan beragam. Dalam penelitian tentang konsentrasi mahasiswa, Naïve Bayes dapat menghasilkan prediksi dengan mempertimbangkan probabilitas atribut yang ada. Keunggulan Naïve Bayes terletak pada kesederhanaan modelnya, serta kemampuannya menangani dataset yang kompleks dan bervariasi, yang seringkali terdapat dalam konteks preferensi konsentrasi mahasiswa.

Perbandingan antara Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naïve Bayes dalam penelitian ini dijustifikasi oleh keunikan dan keunggulan masing-masing algoritma dalam menangani masalah klasifikasi konsentrasi mahasiswa. SVM dikenal efektif dalam menangani kompleksitas data dan memisahkan kelas yang kompleks, sementara KNN menonjol dalam mengidentifikasi pola berdasarkan kedekatan antar data, dan Naïve Bayes memanfaatkan probabilitas atribut untuk klasifikasi. Melalui perbandingan ini, penelitian dapat mengidentifikasi di mana setiap algoritma berkinerja optimal dan di mana masing-masing memiliki batasan, memberikan wawasan yang lebih mendalam terkait pilihan algoritma terbaik dalam mendukung keputusan akademis. Dengan mengeksplorasi diversitas metodologi ketiganya, penelitian dapat

memberikan pemahaman holistik sejauh mana model klasifikasi dapat digeneralisasi untuk memprediksi konsentrasi mahasiswa. Hasil perbandingan juga dapat membantu dalam pemilihan model yang lebih optimal dan berkinerja tinggi dalam meramalkan preferensi konsentrasi mahasiswa di masa depan. Keseluruhan, perbandingan ini diarahkan untuk memberikan landasan yang kuat dalam memahami dan mengoptimalkan penggunaan algoritma klasifikasi dalam konteks penelitian ini.

3.4. Alur Penelitian

Adapun alur penelitian dapat dilihat pada Gambar 3.1.



Gambar 3. 1. Alur Penelitian

Tahapan Alur Penelitian

1. Identifikasi Masalah

Pada tahap ini, peneliti mengidentifikasi masalah yang akan dipecahkan dan menyusun tujuan penelitian. Langkah ini melibatkan kajian literatur untuk memahami konteks dan relevansi penelitian, serta menjelaskan mengapa topik ini penting untuk dieksplorasi.

2. Akuisisi Data

Akuisisi data merupakan tahap pengumpulan data yang akan diolah. Data dapat diperoleh melalui berbagai metode seperti wawancara langsung, kuisioner, observasi, atau ekstraksi data dari sumber yang ada. Metode pengumpulan data harus sesuai dengan tujuan penelitian dan asumsi yang telah ditetapkan sebelumnya. Data juga dapat diperoleh dari sumbersumber seperti internet, lembaga pemerintah, rumah sakit, dan lain-lain. Dataset yang digunakan dalam penelitian ini diperoleh langsung dari staf akademik Fakultas Teknik Universitas Hamzanwadi. Dataset ini mencakup 8 atribut, dengan deskripsi atribut-atribut tersebut dapat dilihat pada Tabel 3.1.

Tabel 3. 1. Deskripsi Atribut Data

No	Atribut	Deskripsi
1.	NAMA	Nama Mahasiswa
2.	JENIS_KELAMIN	Jenis Kelamin Mahasiswa
3.	NIM	Nomor Induk Mahasiswa
4.	IPK	Indeks Prestasi Kumulatif
5.	DATA_SCIENCE	Nilai Mata Kuliah Data Science
6.	RPL	Nilai Mata Kuliah RPL
7.	MULTIMEDIA	Nilai Mata Kuliah Multimedia
8.	KONSENTRASI	Minat Mahasiswa

Pada Tabel 3.1 dapat dilihat beberapa atribut yang digunakan dalam penelitian ini beserta deskripsinya. Penelitian ini menggunakan 8 atribut, yaitu: nama, jenis kelamin mahasiswa, NIM, IPK, nilai mata kuliah yang berpengaruh seperti Data Science, RPL (Rekayasa Perangkat Lunak), dan Multimedia, serta konsentrasi mahasiswa.

3. Preprocessing

Pra-pemrosesan (preprocessing) adalah tahap yang digunakan untuk mempersiapkan data sebelum dilakukan analisis. Tahap ini merupakan langkah awal sebelum dilakukannya pengujian model algoritma yang digunakan yang meliputi pembersihan data (cleaning data) dan pemilihan atribut (feature selection). Dataset yang ada akan diubah menjadi data siap uji, dengan tujuan agar sifat data dapat diuji terhadap model algoritma klasifikasi. Langkah pembersihan data melibatkan penghapusan data kosong dan data yang tidak relevan, sedangkan transformasi data melibatkan perubahan struktur karakteristik data untuk memenuhi kebutuhan penelitian.

4. Splitting data

Data yang telah dipreproses kemudian dibagi menjadi set latih dan set uji. Pembagian ini biasanya dilakukan dengan rasio tertentu, misalnya 75% untuk data latih dan 25% untuk data uji. Pembagian data ini penting untuk memastikan bahwa model dapat dievaluasi secara objektif dan tidak overfit terhadap data latih.

5. Pemilihan dan Penerapan Algoritma

Pada tahap ini, peneliti memilih algoritma pembelajaran mesin yang akan digunakan dalam penelitian, seperti SVM, KNN, dan Naive Bayes. Setiap algoritma diterapkan pada data latih untuk membangun model prediksi. Parameter dan konfigurasi algoritma juga ditentukan dan dioptimalkan selama proses ini.

6. Eksperimen dan Pengujian Model

Setelah model dibangun, eksperimen dilakukan dengan berbagai parameter dan konfigurasi untuk menentukan kinerja terbaik dari setiap algoritma. Model kemudian diuji menggunakan data uji untuk menilai kemampuannya dalam melakukan prediksi. Hasil dari setiap eksperimen dicatat untuk analisis lebih lanjut. Model seperti Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naïve Bayes dapat digunakan untuk mengklasifikasikan konsentrasi mahasiswa berdasarkan data yang relevan seperti IPK, distribusi nilai mata kuliah dan konsentrasi yang diambil.

7. Evaluasi Kinerja Model

Kinerja model dievaluasi menggunakan berbagai metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan confusion matrix. Metrik ini memberikan gambaran tentang seberapa baik model melakukan prediksi dan seberapa efektif algoritma yang digunakan. Evaluasi ini juga membantu dalam membandingkan kinerja antara algoritma yang berbeda.

8. Analisis Hasil

Hasil evaluasi kinerja model dianalisis untuk mengidentifikasi kekuatan dan kelemahan setiap algoritma. Peneliti menjelaskan temuan utama dari hasil eksperimen, membandingkan kinerja algoritma, dan mengeksplorasi faktor-faktor yang mungkin mempengaruhi hasil. Analisis ini memberikan wawasan yang mendalam tentang efektivitas algoritma yang digunakan.

9. Kesimpulan dan Rekomendasi

Pada tahap ini, peneliti menarik kesimpulan berdasarkan hasil penelitian. Peneliti juga memberikan rekomendasi untuk penelitian di masa depan dan mengidentifikasi keterbatasan penelitian yang perlu diperhatikan.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Untuk menentukan hasil pada penelitian ini digunakan tiga algoritma klasifikasi yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes. Ketiga algoritma klasifikasi ini digunakan untuk mengkomparasi nilai akurasi, presisi, recall tertinggi agar dapat digunakan untuk memprediksi dataset yang digunakan dalam penelitian ini. Dataset yang digunakan adalah data mahasiswa sistem informasi Fakultas Teknik Universitas Hamzanwadi.

4.1. Identifikasi Masalah

Tahapan ini melibatkan kajian literatur yang dikumpulkan dari berbagai sumber, termasuk artikel, jurnal, dan buku yang berkaitan dengan topik penelitian. Selanjutnya, dilakukan identifikasi masalah dengan menetapkan konteks penelitian, yakni mahasiswa yang harus menentukan konsentrasi studi atau peminatan, yang menjadi aspek utama dalam pengembangan kompetensi dan pengetahuan mereka, serta adanya peningkatan signifikan dalam jumlah mahasiswa baru di Fakultas Teknik. Penelitian ini bertujuan untuk memprediksi konsentrasi mahasiswa di Fakultas Teknik Universitas Hamzanwadi dengan mengolah dataset menjadi informasi yang lebih bermanfaat. Langkah selanjutnya mencakup perumusan masalah serta penentuan manfaat dan tujuan penelitian.

4.2. Akuisisi data

Dataset yang digunakan dalam penelitian ini bersifat orisinal dan belum pernah dianalisis atau dimanfaatkan dalam penelitian sebelumnya Pengumpulan data dilakukan dengan observasi, wawancara dan pengisian kuesioner ke mahasiswa di lokasi penelitian. Dengan jumlah data mahasiswa sebanyak 160 data mahasiswa. Untuk sample dari dataset dapat dilihat pada tabel 4.1.

Tabel 4. 1. Sample Data Mahasiswa

NIM	Jenis_Kelamin	IPK	Data_Science	RPL	Multimedia	Class
200602001	Perempuan	3.40	B	A	B	1
200602002	Perempuan	3.64	A	B	A	1
200602003	Laki	3.59	B	A	A	0
200602004	Laki	3.24	A	B	B	2
200602005	Perempuan	3.32	B	B	B	1
200602006	Perempuan	3.36	A	A	A	1
200602007	Perempuan	3.21	A	A	B	1
200602008	Laki	3.08	B	B	B	0
200602009	Perempuan	3.48	A	A	A	1
200602010	Perempuan	3.65	A	A	B	1

4.3. Preprocessing Data

Sebelum melewati tahap pra-pemrosesan, data yang akan digunakan itu berjumlah 160 data dan 11 atribut yaitu Nama, NIM, IPK, Jenis Kelamin, Nilai Mata Kuliah Data Science, Nilai Mata Kuliah RPL, Nilai Mata Kuliah Multimedia, Beasiswa, Tanggal Lahir, Alamat, dan Konsentrasi. Pada tahapan ini peneliti

melakukan beberapa langkah penting untuk memastikan kualitas dataset yang digunakan dalam penelitian.

Langkah pertama adalah pembersihan data (cleaning data), di mana peneliti menghapus data yang tidak relevan, termasuk 4 data mahasiswa yang sudah tidak aktif, agar analisis hanya berfokus pada data yang valid dan terkini. Selanjutnya, peneliti memilih atribut-atribut yang dianggap berpengaruh terhadap hasil penelitian, memastikan bahwa hanya fitur-fitur penting yang digunakan untuk proses klasifikasi. Tahapan ini merupakan dasar yang krusial dalam menjamin bahwa data yang diolah benar-benar mencerminkan kondisi yang akan dianalisis lebih lanjut.

Setelah dilakukan pra-pemrosesan dan evaluasi atribut yang berpengaruh terhadap konsentrasi mahasiswa, data yang digunakan sejumlah 156 data dengan rincian mahasiswa yang mengambil konsentrasi Data Science berjumlah 59 orang, mahasiswa yang mengamnbil konsentrasi RPL berjumlah 55 dan mahasiswa dengan konsentrasi Multimedia berjumlah 42 orang. Kemudian terdiri dari 8 atribut, yaitu Nama, NIM, IPK, Jenis Kelamin, Nilai Mata Kuliah Data Science, Nilai Mata Kuliah RPL, Nilai Mata Kuliah Multimedia, dan Konsentrasi Mahasiswa sebagai label.

Pemilihan konsentrasi seorang mahasiswa tentu sangat bergantung pada sejauh mana mahasiswa tersebut menguasai mata kuliah yang menjadi inti dari konsentrasi tersebut (Wiwit Supriyanti dkk, 2016). Hal inilah yang mendasari peneliti dalam menetapkan atribut yang akan digunakan.

4.4. Splitting Data

Pada tahapan ini, peneliti membagi dataset menjadi dua bagian yaitu data latih dan data uji. Pembagian 75% untuk data training dan 25% untuk data testing. Pembagian 75%-25% adalah praktik yang umum dalam machine learning dan data science. Pembagian ini memberikan keseimbangan yang baik antara jumlah data yang cukup untuk melatih model dan jumlah data yang memadai untuk menguji model. Dengan data latih yang cukup besar (75%), model dapat belajar dengan baik dan menghindari underfitting. Sementara itu, 25% data uji memberikan cukup data untuk mengevaluasi performa model secara akurat.

Mengingat jumlah data yang peneliti miliki, pembagian ini optimal karena memastikan bahwa model mendapatkan informasi yang cukup selama pelatihan, namun tetap memiliki data yang cukup untuk validasi. Hal ini penting untuk meminimalkan risiko overfitting, di mana model mungkin tampil sangat baik pada data latih tetapi gagal untuk menggeneralisasi dengan baik pada data baru.

Salah satu penelitian yang dilakukan oleh Karo dkk pada tahun 2023. Dalam penelitian ini, peneliti melakukan evaluasi model prediksi menggunakan metode Support Vector Machine (SVM) dengan berbagai rasio pembagian data. Hasil penelitian menunjukkan bahwa pembagian data dengan rasio 75:25 menghasilkan tingkat akurasi tertinggi, yaitu 79%. Rasio pembagian data 80:20 memberikan tingkat akurasi sedikit lebih rendah, yaitu 78%, sedangkan rasio 70:30 menghasilkan akurasi terendah, yaitu 77%. Temuan ini mendukung penggunaan rasio pembagian 75:25 sebagai pilihan optimal untuk mencapai akurasi model yang lebih baik.

4.5. Pemilihan dan Penerapan Algoritma

Pada tahapan ini proses dimulai dengan evaluasi beberapa algoritma machine learning, termasuk Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes, berdasarkan kriteria seperti akurasi dan kemampuan generalisasi. Setelah algoritma dipilih, masing-masing diterapkan pada dataset dengan parameter optimal, melibatkan pelatihan model dengan data latih dan pengujian menggunakan data uji yang telah dipisahkan.

4.6. Eksperimen dan Pengujian Model

Pada bagian ini akan dilakukan proses data mining dengan menggunakan Cross Validation serta algoritma Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes sebagai perhitungannya. Proses ini akan didukung dengan tools dan pengolahan data yaitu Google Colab.

Tahapan ini dilakukan untuk memberikan hasil yang ingin dicapai secara optimal yaitu tiga kelas (Data Science, RPL dan Multimedia). Pada penelitian ini menggunakan tiga model yaitu Support Vector Machine (SVM), K-Nearest Neighbor) dan Naïve Bayes yang masing-masing memiliki performance yang memiliki fungsi sebagai validasi untuk dapat dijadikan acuan dalam mencari keakuratan data. Untuk melakukan pemodelan data dipilih beberapa atribut yang mendukung untuk pengambilan keputusan pada output yang diinginkan. Penelitian kali ini menggunakan 6 atribut sebagai pendukung keputusan, dapat dilihat pada tabel 4.2.

Tabel 4. 2. Atribut Data

No	Atribut	Keterangan
1.	Nim	id
2.	Jenis Kelamin	Atribut
3.	IPK	Atribut
4.	Nilai Mata Kuliah Data Science	Atribut
5.	Nilai Mata Kuliah RPL	Atribut
6.	Nilai Mata Kuliah Multimedia	Atribut

Kemudian setelah atribut yang akan digunakan sudah ditentukan sebagai komponen pendukung hasil yang diinginkan maka dilakukan pemodelan data, pemodelan pertama menggunakan Support Vector Machine (SVM), kedua menggunakan K-Nearest Neighbor (KNN) dan terakhir menggunakan Naïve Bayes.

```
from sklearn.svm import SVC
from sklearn.metrics import classification_report
# Inisialisasi model SVM dengan kernel linier
svm_model = SVC(kernel='linear')
# Latih model SVM dengan data training
svm_model.fit(X_train, y_train)
```

Berikut adalah code untuk pemodelan dengan algoritma K-Nearest Neighbor pada pemrograman Python.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report
# Inisialisasi model KNN dengan jumlah tetangga (k)
yang diinginkan
```

```
knn_model = KNeighborsClassifier(n_neighbors=5)  #
Contoh: Gunakan 5 tetangga
# Latih model KNN dengan data training
knn_model.fit(X_train, y_train)

# Prediksi menggunakan data testing (disesuaikan
dengan data testing Anda)
X_test = # Masukkan data testing di sini
y_test = # Masukkan label data testing di sini
y_pred = knn_model.predict(X_test)
# Tampilkan laporan klasifikasi
print("Classification Report (KNN):")
print(classification_report(y_test, y_pred))
```

Kemudian berikut adalah kode pemodelan menggunakan algoritma Naïve Bayes.

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
# Inisialisasi model Naive Bayes
nb_model = MultinomialNB()
# Latih model Naive Bayes dengan data training
nb_model.fit(X_train, y_train)
# Prediksi menggunakan data testing (disesuaikan
dengan data testing Anda)
X_test = # Masukkan data testing di sini
```

```
y_test = # Masukkan label data testing di sini
y_pred = nb_model.predict(X_test)
# Tampilkan laporan klasifikasi
print("Classification Report (Naive Bayes):")
print(classification_report(y_test, y_pred))
```

4.7. Evaluasi Kinerja Model

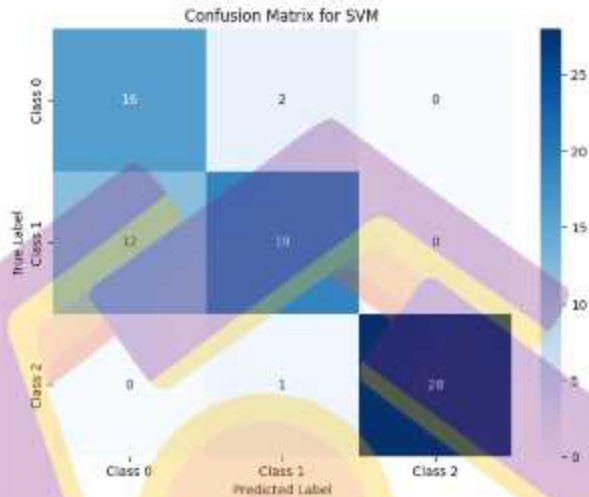
4.7.1. Menentukan Confusion Matrix

Dataset yang digunakan pada penelitian ini sejumlah 156 data setelah melewati tahapan preprocessing yang awalnya berjumlah 160 data dengan 10 atribut. Dataset dibagi menjadi 75% data latih atau sejumlah 117 data dan 25% data uji atau sebanyak 39 data. Selain itu, dataset melibatkan 6 atribut yaitu diantaranya NIM, jenis kelamin, IPK, nilai mata kuliah data science, nilai mata kuliah multimedia dan nilai mata kuliah RPL.

Kelas dalam dataset dibagi menjadi 3 kelas yang berbeda yaitu kelas mahasiswa yang memilih konsentrasi Data Science, kelas mahasiswa yang memilih konsentrasi RPL dan kelas mahasiswa yang memilih konsentrasi multimedia. Perbandingan algoritma SVM, KNN dan Naïve Bayes dilakukan untuk menentukan algoritma mana yang paling efektif untuk kasus pemilihan konsentrasi mahasiswa.

Untuk melakukan evaluasi pada ketiga algoritma tersebut yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes setelah dilakukan split data kemudian langkah selanjutnya adalah dengan menggunakan Confusion Matrix. Gambar 4.1 adalah gambar confusion matrix dengan algroritma

SVM, kemudian gambar 4.2 confusion matrix dengan algoritma KNN dan confusion matrix dengan Naïve Bayes dapat dilihat pada gambar 4.3.

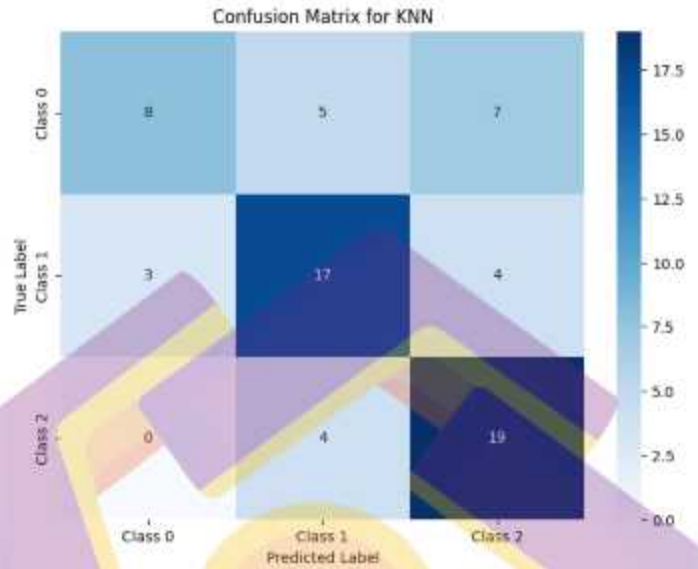


Gambar 4. 1. Confusion Matrix SVM

Dapat kita lihat pada gambar 4.1 tersebut confusion matrix untuk SVM didapatkan nilai untuk kelas 0 (Data Science), TP sejumlah 16 yang berarti itu adalah jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi Data Science dan diprediksi dengan benar. Untuk TN mendapatkan 4 yang berarti itu adalah jumlah mahasiswa yang benar-benar bukan Data Science (baik RPL atau Multimedia) dan diprediksi dengan benar sebagai bukan Data Science. Kemudian FP sejumlah 12 yang berarti jumlah mahasiswa yang sebenarnya bukan Data Science tetapi diprediksi sebagai Data Science. Sedangkan untuk FN sejumlah 0 yang berarti jumlah mahasiswa yang sebenarnya Data Science tetapi diprediksi sebagai bukan Data Science.

Kemudian untuk kelas 1 (RPL) didapatkan nilai TP sejumlah 2 yang berarti itu adalah jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi RPL dan diprediksi dengan benar. Selain itu, untuk nilai TN adalah 19 yang berarti itu jumlah mahasiswa yang benar-benar bukan RPL (baik Data Science atau Multimedia) dan diprediksi dengan benar sebagai bukan RPL. Untuk FP sejumlah 1 yang berarti jumlah mahasiswa yag sebenarnya bukan RPL tetapi diprediksi sebagai RPL dan FN sejumlah 10 yang berarti itu adalah jumlah mahasiswa yang sebenarnya RPL tetapi diprediksi sebagai bukan RPL.

Untuk kelas 2 (Multimedia) mendapatkan nilai TP sejumlah 0 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi multimedia dan diprediksi dengan benar. Nilai TN 28 yang berarti jumlah mahasiswa yang benar-benar bukan Multimedia (baik Data Science atau RPL) dan diprediksi dengan benar sebagai bukan Multimedia. FP sejumlah 0 dan FN sejumlah 4 yang berarti jumlah mahasiswa yang sebenarnya multimedia tetapi diprediksi sebagai bukan multimedia.

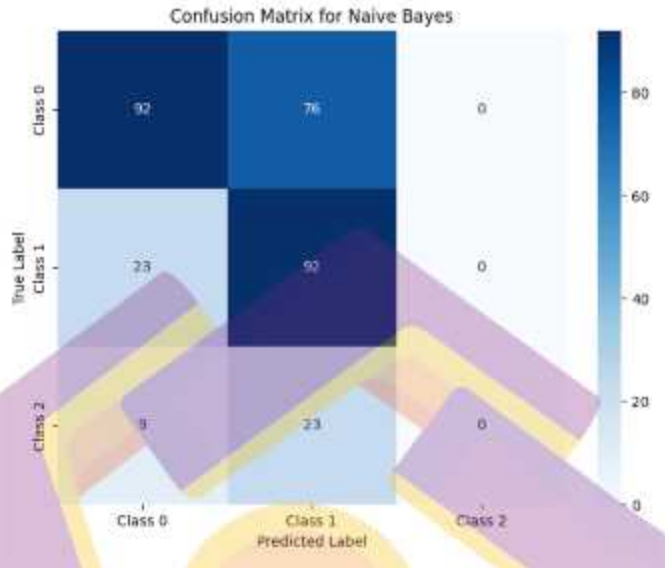


Gambar 4. 2. Confusion Matrix KNN

Untuk algoritma KNN, pada kelas 0 (Data Science) didapatkan nilai TP sejumlah 8 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi Data Science dan diprediksi dengan benar. TN mendapatkan 15 yang berarti jumlah mahasiswa yang benar-benar bukan Data Science (baik RPL atau Multimedia) dan diprediksi dengan benar sebagai bukan Data Science. FP sejumlah 3 yang berarti jumlah mahasiswa yang sebenarnya bukan Data Science tetapi diprediksi sebagai Data Science. Sedangkan FN sejumlah 5 yang berarti jumlah mahasiswa yang sebenarnya Data Science tetapi diprediksi sebagai bukan Data Science.

Pada kelas 1 (RPL), KNN mendapatkan nilai TP sejumlah 5 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi RPL dan diprediksi dengan benar. TN mendapatkan 17 yang berarti jumlah mahasiswa yang benar-benar bukan RPL (baik Data Science atau Multimedia) dan diprediksi dengan benar sebagai bukan RPL. FP sejumlah 4 yang berarti jumlah mahasiswa yang sebenarnya bukan RPL tetapi diprediksi sebagai RPL. Sedangkan FN sejumlah 5 yang berarti jumlah mahasiswa yang sebenarnya RPL tetapi diprediksi sebagai bukan RPL.

Untuk kelas 2 (Multimedia), KNN mendapatkan nilai TP sejumlah 7 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi Multimedia dan diprediksi dengan benar. TN mendapatkan 19 yang berarti jumlah mahasiswa yang benar-benar bukan Multimedia (baik Data Science atau RPL) dan diprediksi dengan benar sebagai bukan Multimedia. FP sejumlah 4 yang berarti jumlah mahasiswa yang sebenarnya bukan Multimedia tetapi diprediksi sebagai Multimedia. Sedangkan FN sejumlah 1 yang berarti jumlah mahasiswa yang sebenarnya Multimedia tetapi diprediksi sebagai bukan Multimedia.



Gambar 4. 3. Confusion Matrix Naive Bayes

Pada gambar 4.3 adalah tampilan confusion matrix untuk model Naïve Bayes. Untuk algoritma Naive Bayes, pada kelas 0 (Data Science) didapatkan nilai TP sejumlah 92 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi Data Science dan diprediksi dengan benar. TN mendapatkan 76 yang berarti jumlah mahasiswa yang benar-benar bukan Data Science (baik RPL atau Multimedia) dan diprediksi dengan benar sebagai bukan Data Science. FP sejumlah 23 yang berarti jumlah mahasiswa yang sebenarnya bukan Data Science tetapi diprediksi sebagai Data Science. Sedangkan FN sejumlah 9 yang berarti jumlah mahasiswa yang sebenarnya Data Science tetapi diprediksi sebagai bukan Data Science.

Pada kelas 1 (RPL), Naive Bayes mendapatkan nilai TP sejumlah 76 yang berarti jumlah mahasiswa yang benar-benar termasuk dalam konsentrasi RPL dan diprediksi dengan benar. TN mendapatkan 92 yang berarti jumlah mahasiswa yang benar-benar bukan RPL (baik Data Science atau Multimedia) dan diprediksi dengan benar sebagai bukan RPL. FP sejumlah 9 yang berarti jumlah mahasiswa yang sebenarnya bukan RPL tetapi diprediksi sebagai RPL. Sedangkan FN sejumlah 23 yang berarti jumlah mahasiswa yang sebenarnya RPL tetapi diprediksi sebagai bukan RPL.

Untuk kelas 2 (Multimedia), Naive Bayes tidak memiliki data yang tersedia, sehingga tidak ada nilai TP, TN, FP, maupun FN yang dapat diinterpretasikan.

4.7.2. Algoritma Support Vector Machine

Dalam penelitian ini, digunakan metode klasifikasi untuk mengelompokkan konsentrasi mahasiswa ke dalam tiga kelas, yaitu :

- Kelas 0 : konsentrasi mahasiswa Data Science
- Kelas 1 : konsentrasi mahasiswa RPL (Rekayasa Perangkat Lunak)
- Kelas 2 : konsentrasi mahasiswa Multimedia

Berikut alur kerja dari algoritma Support Vector Machine dapat dilihat pada gambar 4.4.

Gambar 4. 4. Flowchart Alur Kerja SVM

Alur kerja algoritma Support Vector Machine :

- Input Data Training

  Data training dimasukkan ke dalam sistem. Data training terdiri dari beberapa atribut yaitu NIM, Jenis Kelamin, IPK, Nilai Mata Kuliah Data Science, Nilai Mata Kuliah RPL, Nilai Mata Kuliah Multimedia dan

Konsentrasi Mahasiswa sebagai label atau kelasnya yang akan digunakan untuk melatih model Support Vector Machine.

- Proses Penerapan Fungsi Kernel SVM

  Pada tahapan ini, fungsi kernel ditetapkan pada data training. Fungsi kernel berguna untuk memetakan data ke dalam ruang atribut yang lebih tinggi agar data menjadi lebih mudah untuk dipisahkan oleh hyperlane.

- Proses Perhitungan Squential Training SVM

  Setelah fungsi kernel diterapkan, proses training dilakukan menggunakan algoritma Sequential Minimal Optimization (SMO) atau algoritma lain yang relevan. Algoritma ini bertujuan untuk menemukan hyperplane optimal yang memaksimalkan margin antara kelas-kelas data.

- Proses Testing SVM

  Model yang sudah dilatih kemudian diuji menggunakan data uji untuk mengevaluasi performanya. Data uji ini tidak digunakan selama tahap training.

- Hasil Klasifikasi SVM

  Hasil dari proses testing ditampilkan, yang menunjukkan bagaimana model SVM mengklasifikasikan data uji ke dalam kelas-kelas yang sesuai.

Evaluasi performa klasifikasi dilakukan dengan menghitung akurasi, presisi, dan recall. Metrik-metrik ini memberikan informasi tentang seberapa baik model dapat mengklasifikasikan data. Setelah dilakukan pengolahan data, berikut adalah classification report dari algoritma Support Vector Machine (SVM).

```
Classification Report:
              precision    recall  f1-score   support
```

0	0.55	1.00	0.71	16
1	0.67	0.17	0.27	12
2	0.00	0.00	0.00	4
accuracy			0.56	32
macro avg	0.41	0.39	0.33	32
weighted avg	0.53	0.56	0.46	32

Algoritma SVM menghasilkan akurasi sebesar 56% yang menunjukkan bahwa lebih dari setengah prediksi yang dibuat oleh model ini benar. Namun, masih ada ruang untuk peningkatan karna bisa jadi kurang optimal disebabkan oleh data yang digunakan memiliki atribut yang tidak linear dan SVM membutuhkan turning parameter yang lebih baik.

Dari hasil analisis, dapat disimpulkan bahwa klasifikasi mahasiswa ke dalam berbagai kelas program studi menunjukkan variasi dalam performa model yang dikembangkan. Kelas 0, yang merupakan konsentrasi Mahasiswa Data Science, menunjukkan presisi yang relatif baik sebesar 55%, yang diiringi dengan nilai recall dan f1-score yang tinggi. Namun, ketika melihat kelas 1, yang mewakili konsentrasi Mahasiswa Rekayasa Perangkat Lunak (RPL), meskipun memiliki presisi yang lebih tinggi sebesar 67%, namun nilai recall dan f1-score yang rendah mengindikasikan adanya ketidakakuratan dalam klasifikasi.

Sedangkan, kelas 2 yang mencakup Mahasiswa Multimedia menunjukkan tantangan yang lebih besar, dengan presisi dan recall yang rendah, bahkan mencapai 0%, menunjukkan bahwa model kurang mampu mengidentifikasi dan mengklasifikasikan data ke dalam kelas ini. Hal ini menggambarkan perlunya peninjauan lebih lanjut terhadap faktor-faktor yang mempengaruhi kinerja model, serta kemungkinan penyempurnaan dalam teknik klasifikasi yang digunakan.

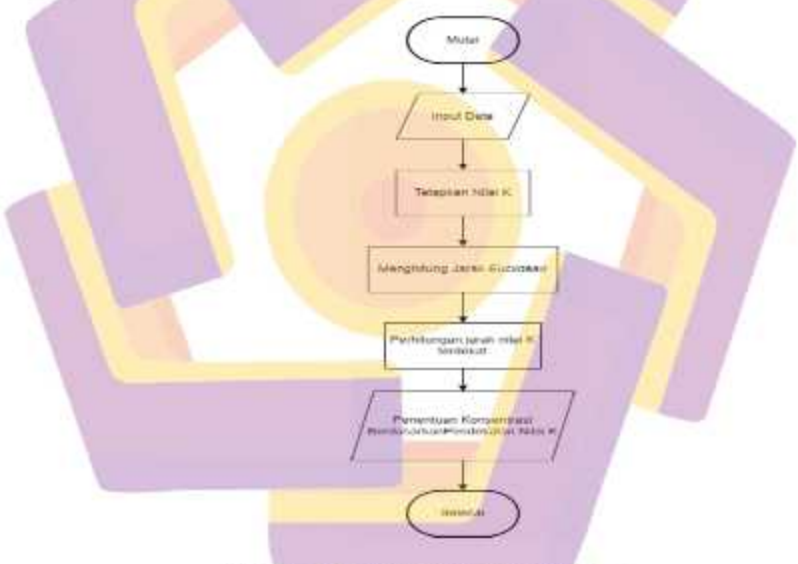
Beberapa hasil analisis yang didapat :

- Adanya ketidakseimbangan kelas, misalnya performa untuk konsentrasi Data Science sangat baik yang dapat dilihat dari recall 100% tetapi sangat buruk untuk Multimedia (recall : 0%). Hal ini bisa terjadi karena distribusi data yang tidak merata atau kurangnya representasi yang baik dari beberapa kelas.
- Kesulitan klasifikasi kelas minoritas, model SVM memiliki kesulitan besar dalam mengklasifikasikan kelas dengan jumlah sample yang sedikit (seperti kelas Multimedia). Hal ini sering terjadi pada algoritma SVM ketika ada kelas yang sangat kurang terwakili dalam data.
- Beberapa atribut yang digunakan mungkin tidak cukup informative untuk membedakan antara kelas. Menambahkan lebih banyak atribut yang lebih relevan bisa membantu.
- Jumlah data yang relatif kecil dapat membatasi kemampuan model untuk belajar dan generalisasi dengan baik. Penambahan data dapat membantu meningkatkan performa.

4.7.3 Algoritma K-Nearest Neighbor

Pemodelan menggunakan algoritma K-Nearest Neighbor (KNN) ini tujuannya adalah untuk dapat mengklasifikasikan objek baru berdasarkan atribut dan training sample yang Dimana nantinya hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada K-Nearest Neighbor (KNN).

Berikut alur kerja dari algoritma K-Nearest Neighbor dapat dilihat pada gambar 4.5.



Gambar 4. 5. Flowchart Alur Kerja KNN

Alur Kerja Algoritma K-Nearest Neighbor

- Input data

  Data yang akan digunakan untuk proses klasifikasi dimasukkan ke dalam sistem. Data ini terdiri dari data latih (training data) dan data uji (testing data).

- Tetapkan Nilai K

  Nilai K ditetapkan. Nilai K adalah jumlah tetangga terdekat yang akan dipertimbangkan dalam proses klasifikasi.

- Menghitung Jarak Euclidean

  Untuk setiap data uji, hitung jarak Euclidean antara data uji tersebut dengan setiap data latih.

- Perhitungan Jarak Nilai K Terdekat

  Setelah jarak Euclidean dihitung, urutkan semua data latih berdasarkan jaraknya dari data uji. Pilih K data terdekat (nilai K terdekat) dari hasil pengurutan ini.

- Penentuan Konsentrasi Berdasarkan Pendekatan Nilai K

  Tentukan kelas atau konsentrasi dari data uji berdasarkan mayoritas kelas dari K tetangga terdekat. Misalnya, jika mayoritas dari K tetangga terdekat adalah dari kelas A, maka data uji diklasifikasikan sebagai kelas A.

Dalam pengolahan data mahasiswa untuk klasifikasi konsentrasi, dilakukan analisis menggunakan beberapa metrik evaluasi, termasuk presisi (precision),

recall, dan f1-score. Hasil dari pengolahan data tersebut direkap dalam classification report berikut:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.62      0.67        13
           1       0.56      0.50      0.53        10
           2       0.64      0.88      0.74         8

    accuracy                           0.65        31
   macro avg       0.64      0.66      0.64        31
weighted avg       0.65      0.65      0.64        31
```

Dari Classification tersebut dapat dilihat bahwa konsentrasi Data Science (kelas 0) memiliki presisi sebesar 0.73, recall sebesar 0.62, dan f1-score sebesar 0.67. Dari 13 sampel, sebagian besar dapat diklasifikasikan dengan benar, meskipun recallnya agak rendah. Konsentrasi RPL (kelas 1) memiliki presisi sebesar 0.56, recall sebesar 0.50, dan f1-score sebesar 0.53. Meskipun presisi dan f1-score cukup rendah, jumlah sampel yang diklasifikasikan dengan benar cukup signifikan. Konsentrasi Multimedia (kelas 2) memiliki presisi sebesar 0.64, recall sebesar 0.88, dan f1-score sebesar 0.74. Dari 8 sampel, mayoritas dapat diklasifikasikan dengan baik, terutama dalam hal recall.

Dari nilai akurasi sebesar 0.65, dapat disimpulkan bahwa model memiliki tingkat keberhasilan yang cukup baik dalam mengklasifikasikan konsentrasi mahasiswa, namun masih terdapat ruang untuk peningkatan khususnya dalam hal presisi dan recall untuk beberapa kelas.

Dari hasil evaluasi model klasifikasi konsentrasi mahasiswa menggunakan algoritma KNN, ditemukan bahwa akurasi keseluruhan model sekitar 67.74%,

menunjukkan kinerja yang moderat. Ketepatan (presisi) model bervariasi tergantung pada kelasnya. Kelas Data Science (kelas 0) memiliki presisi sekitar 69.23%, menunjukkan bahwa sebagian besar prediksi yang diklasifikasikan sebagai Data Science adalah benar. Namun, presisi untuk kelas RPL (kelas 1) hanya sekitar 50%, menunjukkan bahwa model memiliki kesulitan dalam mengidentifikasi kelas RPL dengan akurasi yang tinggi. Di sisi lain, presisi untuk kelas Multimedia (kelas 2) sekitar 70%, menunjukkan performa yang lebih baik dalam mengklasifikasikan kelas Multimedia. Pengukuran recall, atau kemampuan model dalam mengidentifikasi kelas sebenarnya, juga bervariasi. Kelas Data Science memiliki recall sekitar 64.29%, sedangkan kelas RPL hanya sekitar 50%. Namun, kelas Multimedia memiliki recall yang signifikan, sekitar 87.5%, menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali kelas Multimedia. Dari hasil ini, dapat disimpulkan bahwa model memiliki kinerja yang cukup baik dalam mengidentifikasi kelas Multimedia, namun masih perlu peningkatan dalam mengklasifikasikan kelas RPL dengan lebih akurat.

Hasil analisis menggunakan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi mahasiswa ke dalam tiga kelas (Data Science, RPL, dan Multimedia) menunjukkan performa yang bervariasi. Model KNN mencapai akurasi keseluruhan sebesar 65%, dengan rincian precision, recall, dan f1-score yang berbeda untuk setiap kelas. Kelas Data Science memiliki precision 0.73, recall 0.62, dan f1-score 0.67, menunjukkan bahwa meskipun model cukup baik dalam memprediksi mahasiswa Data Science, masih ada beberapa yang teridentifikasi secara salah. Kelas RPL menunjukkan performa yang lebih rendah dengan precision 0.56, recall

0.50, dan f1-score 0.53, mengindikasikan bahwa model sering salah dalam mengidentifikasi mahasiswa RPL. Sebaliknya, kelas Multimedia memiliki performa terbaik dengan recall 0.88 dan f1-score 0.74, menunjukkan model sangat efektif dalam mengenali mahasiswa Multimedia.

Kelebihan utama dari model ini adalah kemampuannya dalam mengidentifikasi mahasiswa kelas Multimedia dengan sangat baik. Namun, kelemahannya terletak pada rendahnya akurasi untuk kelas RPL, yang disebabkan oleh precision dan recall yang rendah. Atribut yang digunakan dalam model ini mencakup nama, NIM, IPK, serta nilai mata kuliah Data Science, RPL, dan Multimedia. IPK dan nilai spesifik mata kuliah memberikan informasi penting untuk prediksi, namun atribut seperti nama dan NIM tidak berpengaruh pada hasil klasifikasi. Untuk meningkatkan performa model, disarankan untuk menambahkan atribut lain seperti minat mahasiswa, aktivitas ekstrakurikuler, dan rekomendasi dosen. Selain itu, teknik penyeimbangan data dan optimisasi nilai K dalam KNN juga dapat diterapkan untuk mengurangi bias dan meningkatkan akurasi prediksi.

4.7.4 Algoritma Naïve Bayes

Untuk pemodelan dengan Naïve Bayes atribut-atribut yang digunakan terlebih dahulu dicari nilai probabilitas kemunculan setiap nilai atribut tersebut. Alur kerja dari algoritma Naïve Bayes dapat dilihat pada gambar 4.6.

Gambar 4. 6. Flowchart Alur Kerja Naive Bayes

Alur Kerja Naïve Bayes ;

- Input Data Uji

  Data uji dimasukkan ke dalam sistem. Data ini akan digunakan untuk menguji model Naive Bayes yang sudah dilatih.

- Baca Data Latih

  Model membaca data latih yang sudah digunakan sebelumnya untuk melatih algoritma Naive Bayes. Data ini berisi informasi tentang atribut dan kelas.

- Hitung Probabilitas Kemunculan Setiap Nilai Atribut

  Model menghitung probabilitas kemunculan setiap nilai atribut dalam data latih. Ini termasuk menghitung probabilitas setiap atribut untuk setiap kelas.

- Hitung Probabilitas Bersyarat untuk Setiap Kelas

  Model kemudian menghitung probabilitas bersyarat untuk setiap kelas berdasarkan data latih. Probabilitas bersyarat ini adalah peluang sebuah fitur muncul dalam suatu kelas tertentu.

- Kalikan Semua Variabel Per Kelas

  Probabilitas bersyarat untuk setiap atribut dikalikan bersama untuk mendapatkan probabilitas gabungan untuk setiap kelas.

- Bandingkan Hasil Per Kelas

  Hasil probabilitas gabungan untuk setiap kelas dibandingkan untuk menentukan kelas mana yang memiliki probabilitas tertinggi.

- Tentukan Kelas Mana Data Uji Berdasarkan Hasil Kali Kelas yang Lebih Besar

  Data uji diklasifikasikan ke dalam kelas dengan probabilitas tertinggi berdasarkan perhitungan sebelumnya.

- Output Hasil Klasifikasi

  Hasil klasifikasi ditampilkan, menunjukkan kelas mana data uji termasuk.

Berikut Classification Report nya :

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.91      0.85       101
           1       0.89      0.77      0.83        99

    accuracy                           0.84       200
   macro avg       0.85      0.84      0.84       200
weighted avg       0.85      0.84      0.84       200
```

Dari Classification Report tersebut dapat dilihat presisi dari konsentrasi Data Science adalah 0.80 yang berarti semua prediksi yang menyatakan mahasiswa memilih konsentrasi Data Science 80% benar. Kemudian untuk presisi konsentrasi RPL adalah 0.89 yang berarti semua prediksi yang menyatakan mahasiswa memilih RPL 89% benar.

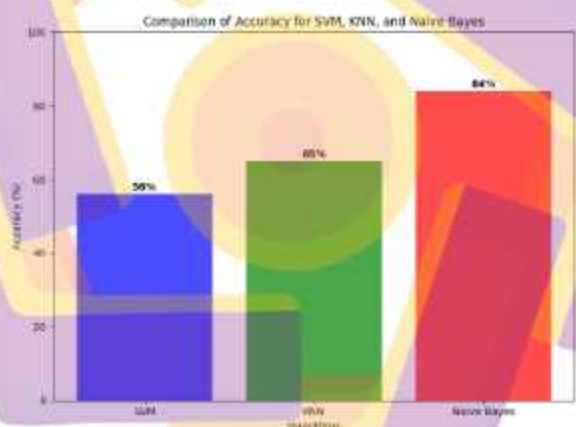
Algoritma Naïve Bayes menunjukkan performa yang sangat baik untuk konsentrasi Data Science dan RPL dengan akurasi keseluruhan 84%. Namun, model tidak berhasil mengklasifikasikan kelas multimedia yang mungkin disebabkan oleh kurangnya data/atribut yang kurang informatif.

Hasil analisis dari algoritma Naïve Bayes ini adalah tidak adanya kelas Multimedia, hal ini bisa disebabkan oleh ketidakseimbangan dalam distribusi data pelatihan. Tidak cukup data untuk kelas ini sehingga model tidak belajar untuk bisa mengenali kelas ini. Model ini bekerja berdasarkan asumsi

independensi antar atribut. Jika atribut yang digunakan tidak memberikan informasi yang cukup untuk membedakan kelas Multimedia dengan yang lain. Model mungkin tidak bisa mengklasifikasikannya dengan benar.

4.8. Analisis Hasil

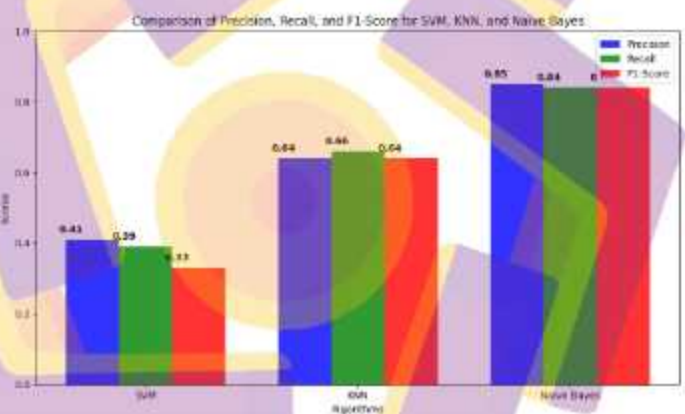
Pada penelitian ini, tiga algoritma machine learning, yaitu Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes, telah diterapkan untuk memprediksi konsentrasi mahasiswa. Hasil dari setiap algoritma dianalisis berdasarkan metrik akurasi, presisi, recall, dan F1-score.



Gambar 4. 7. Perbandingan Akurasi SVM, KNN dan NB

Pada gambar 4.7 dapat dilihat perbandingan hasil akurasi dari ketiga algoritma yang digunakan yaitu SVM, KNN dan Naïve Bayes. Terlihat pada gambar 4.7 Dari hasil yang diperoleh, algoritma Naive Bayes menunjukkan performa terbaik dengan akurasi mencapai 84%. Hal ini menunjukkan bahwa Naive Bayes mampu menangkap pola dalam data yang relevan untuk memprediksi

konsentrasi mahasiswa secara efektif. Algoritma KNN menempati posisi kedua dengan akurasi 65%, yang menunjukkan bahwa meskipun KNN memiliki kemampuan dalam mengklasifikasikan data dengan mempertimbangkan kedekatan antara data, hasilnya kurang optimal dibandingkan dengan Naive Bayes. SVM, dengan akurasi 56%, menunjukkan performa yang paling rendah dalam penelitian ini. Kinerja SVM yang lebih rendah ini dapat disebabkan oleh kurangnya pemilihan parameter yang tepat atau kompleksitas data yang tidak dapat ditangani dengan baik oleh SVM.



Gambar 4. 8. Perbandingan Presisi, Recall dan F-1 Score

Dapat dilihat pada gambar 4.8 yang menampikan metric evaluasi tambahan seperti presisi, recall dan F-1 Score. Untuk algoritma SVM menghasilkan F-1 Score sebesar 0.33, kemudian untuk algoritma KNN F-1 Score nya sebesar 0.64 dan untuk F-1 Score dari algoritma Naïve Bayes mendapati posisi paling tinggi yaitu sebesar 0.84. Dalam hal presisi dan recall, Naive Bayes juga menunjukkan hasil yang lebih

baik dibandingkan dengan KNN dan SVM, menegaskan keunggulan algoritma ini dalam penelitian ini. Hasil ini konsisten dengan asumsi bahwa Naive Bayes, yang bekerja berdasarkan asumsi independensi antar fitur, dapat menangani dataset dengan baik meskipun asumsi tersebut tidak sepenuhnya terpenuhi.

Pengaturan parameter pada masing-masing algoritma juga memainkan peran penting dalam kinerja model. Pada KNN, jumlah tetangga yang dipilih mempengaruhi hasil klasifikasi, sementara pada SVM, pilihan kernel dan nilai parameter C sangat berpengaruh. Dalam penelitian ini, mungkin perlu dilakukan optimasi parameter lebih lanjut untuk meningkatkan performa SVM.

Perbandingan hasil antara algoritma SVM, KNN dan Naïve Bayes lebih jelasnya dapat dilihat pada Tabel 4.3.

Tabel 4. 3. Perbandingan Hasil Algoritma SVM, KNN dan Naïve Bayes

Algoritma	Akurasi (%)	Presisi	Recall	F-1 Score
SVM	56%	0.41	0.39	0.33
KNN	65%	0.64	0.66	0.64
NB	84%	0.85	0.84	0.84

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dari ketiga algoritma yang digunakan yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN) dan Naïve Bayes, dapat ditarik kesimpulan :

1. Dari atribut yang digunakan *Naïve Bayes* terbukti sebagai Algoritma yang lebih baik dari *Support Vector Machine* dan *K-Nearest Neighbor*.
2. Dari hasil analisis menggunakan algoritma Support Vector Machine (SVM), terlihat bahwa performa model dalam mengklasifikasikan konsentrasi mahasiswa cenderung bervariasi. Meskipun kelas Data Science (kelas 0) memiliki presisi yang relatif baik sebesar 55%, namun kelas Multimedia (kelas 2) menunjukkan tantangan yang lebih besar dengan presisi dan recall yang rendah bahkan mencapai 0%. Model juga mengalami kesulitan dalam mengidentifikasi kelas RPL (kelas 1) dengan baik, ditandai dengan presisi yang rendah meskipun recall-nya sedikit lebih tinggi daripada kelas Multimedia. Dalam hal akurasi, model SVM mencapai 56%, menunjukkan kinerja yang moderat.
3. Sementara itu, algoritma K-Nearest Neighbor (KNN) menunjukkan akurasi sekitar 67.74%, yang cukup baik dalam

mengklasifikasikan konsentrasi mahasiswa. Namun, presisi dan recall bervariasi tergantung pada kelasnya. Kelas Data Science memiliki presisi dan recall yang relatif tinggi, sedangkan kelas Multimedia menunjukkan kinerja yang lebih rendah, terutama dalam hal recall. Kelas RPL juga mengalami tantangan dalam model KNN, terutama dalam hal presisi.

4. Sementara itu, algoritma Naïve Bayes menunjukkan hasil yang cukup baik dalam mengklasifikasikan konsentrasi mahasiswa, dengan akurasi sekitar 84%. Namun, perlu diperhatikan bahwa dalam kasus ini, model Naïve Bayes menggunakan probabilitas untuk memprediksi kelas, yang mungkin memiliki asumsi yang cukup kuat tentang independensi antaratribut. Meskipun demikian, dengan menggunakan pendekatan ini, model berhasil memprediksi kelas dengan baik, meskipun hasilnya mungkin perlu divalidasi lebih lanjut.

5.2. Saran

Untuk pengembangan penelitian selanjutnya maka penulis menyarankan :

1. Penambahan atribut pada penghitungan agar mendapatkan hasil untuk dilakukan perbandingan dengan penelitian sebelumnya.
2. Melakukan analisis yang lebih mendalam terhadap atribut-atribut yang relevan dengan konsentrasi mahasiswa dapat membantu meningkatkan kinerja model.

3. Melibatkan atribut tambahan yang mungkin memiliki hubungan yang kuat dengan konsentrasi mahasiswa, seperti aktivitas ekstrakulikuler, minat karir atau latar belakang pendidikan dapat membantu meningkatkan akurasi prediksi.
4. Dapat dilakukan komparasi menggunakan beberapa algoritma lain yang biasa digunakan pada *Machine Learning*.

DAFTAR PUSTAKA

Sathe, M. T., & Adamuthe, A. C. (2021). *Comparative Study of Supervised Algorithms for Prediction of Students' Performance. International Journal of Modern Education and Computer Science,* 13(1), 1-21. DOI: 10.5815/ijmecs.2021.01.01.

Rabbani, S., Safitri, D., Siregar, F. T. P., Rahmaddeni, & Efrizoni, L. (2024). *Evaluation of Support Vector Machine, Naive Bayes, Decision Tree, and Gradient Boosting Algorithms for Sentiment Analysis on ChatGPT Twitter Dataset. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM),* 7(1), 11-21.

Ramadanty, M. C., Siregar, A. M., & Kusumaningrum, D. S. (2021). *Penerapan Algoritma C4.5 dan K-Nearest Neighbor untuk Klasifikasi Peminatan Program Studi di Perguruan Tinggi Berdasarkan Nilai Rapor. Scientific Student Journal for Information, Technology and Science,* II(1), 76. ISSN: 2715-2766.

Fakhrurrifqi, M., & Wardoyo, R. (2013). *Perbandingan Algoritma Nearest Neighbour, C4.5 dan LVQ untuk Klasifikasi Kemampuan Mahasiswa. International Journal on Cloud Computing: Services and Architecture (IJCCS),* 7(2), 145-154. ISSN: 1978-1520.

Arifin, O., & Sasongko, T. B. (2018). *Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine dan Naïve Bayes Classifier untuk Klasifikasi Jalur Minat SMA. Proceedings of the Seminar Nasional Teknologi Informasi dan Multimedia 2018, Universitas AMIKOM Yogyakarta,* 10 Februari 2018, ISSN: 2302-3805, 1.2-67.

Howay, S., & Suhirman. (2022). *Comparison of SVM, NBC, and KNN Classification Methods in Determining Students' Majors at SMK N02 Manokwari. Journal of Computer Science and Technology Studies (JCSTS),* ISSN: 2709-104X, DOI: 10.32996/jcsts.

Tejawati, A., Septiarini, A., Rismawati, R., & Puspitasari, N. (2023). Comparison of K-Nearest Neighbor and Naive Bayes Methods for Classification of News Content. Jurnal Teknik Informatika (JUTIF), Vol. 4, No. 2, April 2023, hlm. 401-412. DOI: https://doi.org/10.52436/1.jutif.2023.4.2.676. p-ISSN: 2723-3863, e-ISSN: 2723-3871.

Putri, A., Hardiana, C. S., Novfuja, E., Siregar, F. T. P., Rahmaddeni, Y., Fatma, Y., & Wahyuni, R. (2023). Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction. MALCOM: Indonesian Journal of Machine Learning and Computer Science, Vol. 3, Iss. 1, April 2023, pp: 20-26. ISSN(P): 2797-2313, ISSN(E): 2775-8575.

Rabbani, S., Safitri, D., Siregar, F. T. P., Rahmaddeni, & Efrizoni, L. (2024). Evaluation of Support Vector Machine, Naive Bayes, Decision Tree, and Gradient Boosting Algorithms for Sentiment Analysis on ChatGPT Twitter Dataset. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), Vol. 7, No. 1, March 2024, pp. 11–21. p-ISSN: 2614-3372, e-ISSN: 2614-6150.

Mustafa Çakir, Mesut Yilmaz, Mükerrem Atalay Oral, Hüseyin Özgür Kazanci, Okan Oral. "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture." Journal of King Saud University – Science, Vol. 36, No. 1, 2023, pp. 1-10. Available online 12 June 2023.

Salleh, S. A., Khalid, N., Danny, N., Mohd. Zaki, N. A., Ustuner, M., Abd Latif, Z., & Foronda, V. (2024). Support vector machine (SVM) and object based classification in earth linear features extraction: A comparison. *Revue Internationale de Geomatique, 33*(0), 183-199. https://doi.org/10.1016/j.rig.2024.04.001

Tsalera, E., Papadakis, A., & Samarakou, M. (2020). Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm. *Energy Reports, 6*(Supplement 6), 223-230. https://doi.org/10.1016/j.egyr.2020.08.045

Libnao, M., Misula, M., Andres, C., Mariñas, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve bayes algorithm. *Procedia Computer Science, 227*, 316-325. https://doi.org/10.1016/j.procs.2023.10.530.

Hidayanti, I., Kurniawan, T. B., & Afriyudi. (2020). Perbandingan dan Analisis Metode Klasifikasi untuk Menentukan Konsentrasi Jurusan. *Jurnal Ilmiah Informatika Global*, 11(01), 16-25. ISSN Print: 2302-500X, ISSN Online: 2477-3786.

Wibowo, M. F. S., Puspitasari, N. F., & Satya, B. (2022). Penerapan Data Mining dan Algoritma Naïve Bayes untuk Pemilihan Konsentrasi Mahasiswa Menggunakan Metode Klasifikasi. *Journal of Information System Management (JOISM)*, 3(2), 39-47. e-ISSN: 2715-3088.

Supriyanti, W., Kusrini, & Amborowati, A. (2016). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Ketepatan Pemilihan Konsentrasi Mahasiswa. *Jurnal INFORMA Politeknik Indonusa Surakarta*, 1(3), 61-68. ISSN: 2442-7942.