# BAB II
# TINJAUAN PUSTAKA

## A. Hasil Penelitian Terdahulu

Penelitian terdahulu yang dilakukan oleh (Abdurrahman & Wijaya, 2019) dengan judul "Analisis Klasifikasi Kelahiran *Caesar* Menggunakan Algoritma Naïve Bayes" penelitian ini menggunakan data yang ada di UCI *Machine Learning Repository,* dataset ini terdiri dari 80 data ibu hamil dengan lima atribut, yaitu umur, jumlah tenaga medis, waktu melahirkan, tekanan darah, dan masalah jantung. Tujuan penelitian ini adalah untuk mendukung dunia kesehatan, khususnya dalam menentukan apakah proses kelahiran memerlukan tindakan operasi *Caesar* atau tidak, dengan menyediakan model klasifikasi untuk prediksi medis. Harapannya, hasil penelitian ini dapat memberikan dukungan yang signifikan bagi dunia kesehatan, terutama dalam kasus ibu melahirkan, dengan menyediakan model klasifikasi yang dapat digunakan untuk prediksi dan pengambilan keputusan medis. Kesimpulan dari penelitian ini menunjukkan bahwa klasifikasi kelahiran menggunakan algoritma Naïve Bayes masih menghasilkan nilai yang kurang memuaskan. Rata-rata data terklasifikasi dengan benar adalah sebesar 67,16%, sementara rata-rata data terklasifikasi salah adalah sebesar 32,84%. Persentase tertinggi dan terendah dari data terklasifikasi dengan benar masing-masing adalah 100% dan 55%, sedangkan persentase tertinggi dan terendah dari data terklasifikasi salah masing-masing adalah 45% dan 0%. Nilai precision dan recall tertinggi tercapai pada hasil uji dataset dengan percentage split sebesar 95%, yaitu masing-masing bernilai 1. Saran untuk penelitian selanjutnya, disarankan untuk mengimplementasikan algoritma klasifikasi lain sebagai pembanding guna meningkatkan performa klasifikasi.

Penelitian terdahulu yang dilakukan oleh (Hidayat et al., 2019) dengan judul "Implementasi Algoritma K-Nearest Neighbor dan *Probabilistic Neural Network*

untuk Analisis Opini Masyarakat Terhadap Toko Online di Indonesia" Penelitian ini memiliki tujuan untuk menganalisis masalah-masalah yang sering menjadi keluhan masyarakat di toko online, yang menyebabkan perilaku komplain dan kurangnya kepercayaan masyarakat dalam bertransaksi. Kepercayaan dalam bertransaksi di toko online dianggap sebagai faktor utama dalam kepuasan pelanggan. Data penelitian diambil dari survei *iPrice E-commerce Merchants Award* (iEMA) 2018, dengan fokus pada dua toko online yang populer di Twitter, yaitu Blibli dengan 474.700 pengikut dan Lazada dengan 363.600 pengikut. Akun resmi toko online tersebut menyajikan ulasan melalui komentar yang berisi opini masyarakat terhadap isu-isu yang sedang berlangsung. Teknik text mining digunakan untuk menganalisis masalah-masalah tersebut dan mengidentifikasi pandangan masyarakat terhadap toko online. Hasil penelitian, menggunakan metode pembagian data dengan 10 K pada K-Fold Cross Validation, menunjukkan perbandingan akurasi antara KNN (K-Nearest Neighbors) dan PNN (Probabilistic Neural Network) terhadap data Lazada dan Blibli. Untuk data Lazada, akurasi KNN lebih tinggi daripada PNN, dengan akurasi KNN mencapai 71.57%, sedangkan PNN sebesar 66.71%. Demikian pula, pada data Blibli, akurasi KNN juga lebih tinggi daripada PNN, dengan akurasi KNN sebesar 68.29%, sedangkan PNN sebesar 65.29%. Oleh karena itu, dapat disimpulkan bahwa hasil akurasi menggunakan algoritma KNN pada data Lazada dan Blibli menunjukkan performa yang lebih baik dibandingkan dengan PNN.

Penelitian sebelumnya yang dilakukan oleh (Rizaldi & Mustakim, 2020) yang berjudul "Perbandingan Teknik Pembagian Data untuk Klasifikasi Sarana Akses Air pada Algoritma K-Nearest Neighbor dan *Naïve Bayes Classifier*" Penelitian ini bertujuan untuk mengklasifikasikan kelayakan akses air berdasarkan profil sekolah dengan menggunakan indikator tertentu. Data yang digunakan berasal dari Sekolah Menengah Atas (SMA) dan Sekolah Menengah Kejuruan (SMK) di Provinsi Kepulauan Riau, dengan total 215 data sesuai dengan cakupan kewenangan yang

relevan. Hasil dari penelitian ini menunjukkan bahwa teknik pembagian data terbaik adalah K-Medoid. Secara berurutan, pada algoritma K-Nearest Neighbor (KNN) dengan parameter K=10, menghasilkan akurasi sebesar 89,39%, recall sebesar 79,44%, dan presisi sebesar 81,77%. Di sisi lain, Naïve Bayes Classifier menghasilkan akurasi sebesar 40,91%, recall sebesar 55,19%, dan presisi sebesar 45,69%. Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa percobaan yang menggunakan data Sarana Akses Air dapat memanfaatkan K-Medoid sebagai teknik pembagian data yang efektif untuk klasifikasi Sarana Akses Air pada algoritma klasifikasi. Teknik K-Nearest Neighbor dengan parameter K=10 menunjukkan performa yang lebih baik dibandingkan dengan Naïve Bayes Classifier dalam konteks klasifikasi kelayakan akses air berdasarkan profil sekolah.

Penelitian sebelumnya yang dilakukan oleh (Sulastri et al., 2020) yang berjudul "Analisis Perbandingan Klasifikasi Prediksi Penyakit Hepatitis Dengan Menggunakan Algoritma K-Nearest Neighbor*, Naïve Bayes* Dan *Neural Network*" penelitian ini bertujuan untuk mendiagnosis penyakit pasien berdasakan rekam medis pasien, penelitian ini memiliki manfaat yang signifikan Untuk para ahli kesehatan, memungkinkan mereka untuk menggunakan catatan rekam medis yang ada sebagai panduan dalam membuat keputusan mengenai diagnosis penyakit pasien adalah langkah yang sangat bernilai. Data yang digunakan dalam penelitian ini berasal dari UCI Machine Learning, khususnya data pasien hepatitis dengan total 155 rekam medis, terdiri dari 19 variabel penjelas dan 1 variabel respon. Hasil penelitian menunjukkan bahwa menggunakan algoritma Naïve Bayes, model klasifikasi mencapai tingkat akurasi terbaik pada percobaan 2, yaitu sebesar 76,92%, dengan tingkat error sebesar 23,01%. Penggunaan Algoritma Neural Network menghasilkan model klasifikasi dengan tingkat akurasi tertinggi pada percobaan 1, mencapai 82,97%, dengan tingkat error sebesar 17,03%. Sementara itu, dengan algoritma K-Nearest Neighbor, model klasifikasi mencapai tingkat akurasi terbaik pada percobaan 3, yaitu sebesar 93%, dengan tingkat error sebesar

7%. Hasil ini menunjukkan potensi penggunaan algoritma klasifikasi dalam menganalisis data rekam medis untuk mendukung proses diagnosis penyakit. Algoritma K-Nearest Neighbor menonjol dengan tingkat akurasi yang tinggi, memberikan harapan bahwa pendekatan ini dapat membantu para ahli kesehatan dalam membuat keputusan yang lebih tepat dan efektif terkait dengan penyakit hepatitis pada pasien.

Penelitian sebelumnya yang dilakukan oleh (Sugawara & Nikaido, 2014) yang berjudul "Analisis Perbandingan Algoritma SVM Dan K-NN Untuk Klasifikasi Anime Bergenre Drama" Penelitian ini memiliki tujuan untuk menganalisis perbandingan antara dua algoritma, yaitu algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN), yang akan dilatih dan diuji menggunakan dataset anime. Dataset ini terdiri dari 12.294 data dengan dua kelas genre, yaitu drama dan non-drama. Data anime berasal dari sumber publik di Kaggle. Dataset anime terdiri dari 7 atribut yang dibagi menjadi dua bagian, yaitu 80% untuk data pelatihan dan 20% untuk data uji. Hasil dari algoritma K-Nearest Neighbor (KNN) menunjukkan nilai akurasi pelatihan sebesar 100% dan nilai akurasi uji sebesar 84%, disisi lain, hasil dari algoritma Support Vector Machine (SVM) menunjukkan nilai akurasi pelatihan sebesar 83% dan nilai akurasi uji sebesar 82%. Dalam konteks klasifikasi anime ini, algoritma K-Nearest Neighbor (KNN) menunjukkan akurasi uji yang lebih baik dibandingkan dengan Support Vector Machine (SVM), meskipun selisih keduanya cukup tipis. Hasil ini menunjukkan bahwa, dalam kasus ini, model yang dikembangkan dengan menggunakan algoritma KNN mampu memberikan performa yang lebih baik dalam mengklasifikasikan data uji anime dibandingkan dengan model yang dikembangkan menggunakan algoritma SVM. Meskipun nilai akurasi pelatihan KNN mencapai 100%, penting untuk memperjantungkan keseimbangan antara akurasi pelatihan dan uji untuk memastikan bahwa model dapat secara efektif menggeneralisasi pola yang telah dipelajari ke data baru..

Penelitian sebelumnya yang dilakukan oleh (Setiyorini & Asmono, 2018) yang berjudul "Komparasi Metode *Decision Tree, Naïve Bayes* dan K-Nearest Neighbor pada Klasifikasi Kinerja Siswa" penelitian ini bertujuan untuk menganalisis atau mengevaluasi faktor-faktor yang mempengaruhi kinerja siswa, dengan fokus pada metode Decision Tree, Naive Bayes, dan K-Nearest Neighbor. Dataset yang digunakan berasal dari UCI Machine Learning Repository dan terdiri dari 30 atribut dan 1 kelas. Hasil dari penelitian ini menunjukkan bahwa penggunaan metode Decision Tree menghasilkan akurasi sebesar 78,85%. Metode Naive Bayes memberikan akurasi sebesar 77,69%, sedangkan metode K-Nearest Neighbor (KNN) mencapai akurasi tertinggi, yaitu sebesar 79,31%. Setelah dilakukan perbandingan, dapat disimpulkan bahwa metode K-Nearest Neighbor (KNN) memberikan akurasi tertinggi dibandingkan dengan metode Decision Tree dan Naive Bayes dalam konteks penelitian ini. Hasil ini menunjukkan bahwa, dalam analisis faktor-faktor yang mempengaruhi kinerja siswa, metode K-Nearest Neighbor (KNN) lebih efektif dalam menghasilkan prediksi yang akurat. Dengan demikian, dapat dianggap bahwa K-Nearest Neighbor (KNN) adalah pilihan yang lebih baik dalam memodelkan hubungan antara faktor-faktor yang mempengaruhi kinerja siswa dalam dataset yang digunakan.

Penelitian sebelumnya yang dilakukan oleh (andiani et al., 2019) yang berjudul "Analisis Penyakit Jantung Menggunakan Metode K-NN dan *Random Forest*" Penelitian ini memiliki tujuan untuk membantu ahli jantung dalam menyusun dan menggolongkan data Penelitian ini bertujuan untuk merinci pola atau tema dalam data, memberikan makna pada analisis yang dilakukan, menjelaskan kategori atau pola dalam data, serta mencari hubungan antar data. Data penelitian diambil dari Kaggle dalam bentuk file CSV dengan total data sebanyak 1025, di mana 526 pasien memiliki penyakit jantung dan 499 pasien tidak memiliki penyakit jantung. Data ini berasal dari tahun 1988 dan berasal dari empat database: Cleveland, Hongaria, Swiss, dan Long ReversibleDefect Beach V. Hasil uji coba menunjukkan bahwa metode K-Nearest Neighbor (KNN) dapat digunakan untuk mengklasifikasi

data penyakit jantung dengan tingkat akurasi sebesar 93%. Sebagai perbandingan, metode Random Forest menghasilkan tingkat akurasi sebesar 72%. Dengan demikian, berdasarkan hasil penelitian ini, metode KNN terbukti lebih efektif dalam mengklasifikasi data penyakit jantung dibandingkan dengan metode Random Forest. Penting untuk dipahami bahwa temuan ini dapat memberikan wawasan yang berharga bagi para peneliti, praktisi kesehatan, dan pemangku kepentingan lainnya dalam mengidentifikasi metode terbaik untuk klasifikasi penyakit jantung. Dengan tingkat akurasi yang tinggi, K-Nearest Neighbor (KNN) dapat dianggap sebagai pilihan yang lebih optimal dalam konteks analisis penyakit jantung pada dataset yang digunakan.

Penelitian sebelumnya yang dilakukan oleh (LOUIS MADAERDO SOTARJUA & DIAN BUDHI SANTOSO, 2022) yang berjudul "Perbandingan Algortima KNN, *Decision Tree, dan Random Forest* pada data *Imbalanced Class* untuk Klasifikasi Promosi Karyawan" Tujuan dari penelitian ini adalah untuk menganalisis performa model machine learning pada data klasifikasi karyawan. Data yang digunakan dalam penelitian ini merupakan data dengan kelas yang tidak seimbang (imbalanced class), sehingga diterapkan teknik Synthetic Minority Over-Sampling Technique (SMOTE) untuk menangani ketidakseimbangan tersebut. Berdasarkan hasil performa model klasifikasi dari berbagai algoritma yang digunakan pada penelitian ini, model K-Nearest Neighbor (KNN) menunjukkan hasil performa terbaik dengan akurasi sebesar 86,57% berdasarkan metrik evaluasi yang digunakan. Oleh karena itu, pada penelitian ini, model KNN dianggap sebagai model klasifikasi yang lebih baik dibandingkan dengan algoritma Decision Tree yang memiliki akurasi sebesar 85,29% dan Random Forest yang memiliki akurasi sebesar 86,37%. Temuan ini memberikan wawasan penting bahwa dalam kasus data imbalanced class pada klasifikasi karyawan, model KNN dapat menjadi pilihan yang efektif untuk meningkatkan performa dan akurasi prediksi. Hasil ini dapat bermanfaat bagi pemangku kepentingan yang berkepentingan dalam

pengelolaan dan pengambilan keputusan terkait dengan data karyawan pada konteks penelitian ini.

Penelitian sebelumnya yang dilakukan oleh (Syukri Mustafa & Wayan Simpen, 2019) yang berjudul "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba" Penelitian ini bertujuan untuk menguji Penelitian ini bertujuan untuk memprediksi kemungkinan seorang pasien baru di Puskesmas Manyampa, Kabupaten Bulukumba, terkena penyakit diabetes melitus atau tidak dengan menggunakan analisis data mining menggunakan algoritma K-Nearest Neighbor (KNN). Aplikasi yang dihasilkan memanfaatkan beberapa atribut dalam klasifikasi data mining, seperti usia, anak ke, jumlah saudara, jenis kelamin, status perkawinan, pekerjaan, dan riwayat keluarga yang terkena diabetes melitus.

Data pelatihan terdiri dari 200 data pasien yang telah menjalani pemeriksaan dalam 2 tahun terakhir. Pengujian dilakukan dengan menggunakan data uji atau sampel sebanyak 104 data pasien yang diambil dari luar data pelatihan, dengan tujuan membandingkan hasil prediksi Status Diabetes Melitus Pasien dari sistem yang dirancang dengan hasil Status Diabetes Melitus Pasien yang sesuai dengan data. Hasil akurasi yang diperoleh dari pengujian tersebut adalah sebesar 68,30%. Meskipun akurasi tersebut memberikan suatu prediksi, penting untuk diingat bahwa hasil ini mungkin memerlukan evaluasi lebih lanjut dan validasi menggunakan data yang lebih besar dan representatif. Selain itu, penelitian ini memberikan dasar untuk pengembangan lebih lanjut dalam meningkatkan performa dan ketepatan prediksi sistem untuk membantu dalam pencegahan dan manajemen penyakit diabetes melitus.

Tabel 2. 1 Penelitian Terdahulu

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| 1 | (Abdurrahman & Wijaya, 2019) | *Naïve Bayes* | Algoritma Naïve Bayes masih menghasilkan nilai yang kurang memuaskan. Rata-rata data terklasifikasi dengan benar adalah sebesar 67,16%, sementara rata-rata data terklasifikasi salah adalah sebesar 32,84%. Persentase tertinggi dan terendah dari data terklasifikasi dengan benar masing-masing adalah 100% dan 55%, sedangkan persentase tertinggi dan terendah dari data terklasifikasi salah masing-masing adalah 45% dan 0%. Nilai precision dan recall tertinggi tercapai pada hasil uji dataset dengan percentage split sebesar 95%, yaitu masing-masing bernilai 1. Saran untuk penelitian selanjutnya, disarankan untuk mengimplementasikan algoritma klasifikasi lain sebagai pembanding guna meningkatkan performa klasifikasi |
| 2 | (Hidayat et al., 2019) | K-Nearest Neighbor *dan Neural Network* | Hasil penelitian, menggunakan metode pembagian data dengan 10 K pada K-Fold Cross Validation, |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | | | menunjukkan perbandingan akurasi antara KNN (K-Nearest Neighbors) dan PNN (Probabilistic Neural Network) terhadap data Lazada dan Blibli. Untuk data Lazada, akurasi KNN lebih tinggi daripada PNN, dengan akurasi KNN mencapai 71.57%, sedangkan PNN sebesar 66.71%. Demikian pula, pada data Blibli, akurasi KNN juga lebih tinggi daripada PNN, dengan akurasi KNN sebesar 68.29%, sedangkan PNN sebesar 65.29%. Oleh karena itu, dapat disimpulkan bahwa hasil akurasi menggunakan algoritma KNN pada data Lazada dan Blibli menunjukkan performa yang lebih baik dibandingkan dengan PNN. |
| 3 | (Rizaldi & Mustakim, 2020) | K-Nearest Neighbor *dan Naïve Bayes* | Hasil dari penelitian ini menunjukkan bahwa teknik pembagian data terbaik adalah K-Medoid. Secara berurutan, pada algoritma K-Nearest Neighbor (KNN) dengan parameter K=10, menghasilkan akurasi sebesar 89,39%, recall sebesar 79,44%, dan presisi sebesar 81,77%. Di sisi lain, |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | | | Naïve Bayes Classifier menghasilkan akurasi sebesar 40,91%, recall sebesar 55,19%, dan presisi sebesar 45,69%. Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa percobaan yang menggunakan data Sarana Akses Air dapat memanfaatkan K-Medoid sebagai teknik pembagian data yang efektif untuk klasifikasi Sarana Akses Air pada algoritma klasifikasi. Teknik K-Nearest Neighbor dengan parameter K=10 menunjukkan performa yang lebih baik dibandingkan dengan Naïve Bayes Classifier dalam konteks klasifikasi kelayakan akses air berdasarkan profil sekolah. |
| 4 | (Sulastri et al., 2020) | K-Nearest Neighbor, *Naïve Bayes dan Neural Network* | Hasil penelitian menunjukkan bahwa menggunakan algoritma Naïve Bayes, model klasifikasi mencapai tingkat akurasi terbaik pada percobaan 2, yaitu sebesar 76,92%, dengan tingkat error sebesar 23,01%. Penggunaan Algoritma Neural Network menghasilkan model klasifikasi dengan tingkat akurasi tertinggi pada percobaan 1, mencapai |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | | | 82,97%, dengan tingkat error sebesar 17,03%. Sedangkan, dengan algoritma K-Nearest Neighbor, model klasifikasi mencapai tingkat akurasi terbaik pada percobaan 3, yaitu sebesar 93%, dengan tingkat error sebesar 7%. |
| 5 | (Sugawara & Nikaido, 2014) | K-Nearest Neighbor *dan Support Vector Machine* | Hasil dari algoritma K-Nearest Neighbor (KNN) menunjukkan nilai akurasi pelatihan sebesar 100% dan nilai akurasi uji sebesar 84%. Sementara itu, hasil dari algoritma Support Vector Machine (SVM) menunjukkan nilai akurasi pelatihan sebesar 83% dan nilai akurasi uji sebesar 82%. Dalam konteks klasifikasi anime ini, algoritma K-Nearest Neighbor (KNN) menunjukkan akurasi uji yang lebih baik dibandingkan dengan Support Vector Machine (SVM), meskipun selisih keduanya cukup tipis |
| 6 | (Setiyorini & Asmono, 2018) | K-Nearest Neighbor, *Naïve Bayes dan Decision Tree* | Hasil dari penelitian ini menunjukkan bahwa penggunaan metode Decision Tree menghasilkan akurasi sebesar 78,85%, dengan menggunakan metode Naive Bayes, |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | | | diperoleh akurasi sebesar 77,69%, sedangkan, dengan metode K-Nearest Neighbor (KNN), akurasi yang dicapai sebesar 79,31%. Setelah dilakukan perbandingan, dapat disimpulkan bahwa metode K-Nearest Neighbor (KNN) memberikan akurasi tertinggi dibandingkan dengan metode Decision Tree dan Naive Bayes dalam konteks penelitian ini. |
| 7 | (andiani et al., 2019) | K-Nearest Neighbor *dan Random Forest* | Hasil uji coba menunjukkan bahwa metode K-Nearest Neighbor (KNN) dapat digunakan untuk mengklasifikasi data penyakit jantung dengan tingkat akurasi sebesar 93%. Sebagai perbandingan, metode Random Forest menghasilkan tingkat akurasi sebesar 72%. Dengan demikian, berdasarkan hasil penelitian ini, metode KNN terbukti lebih efektif dalam mengklasifikasi data penyakit jantung dibandingkan dengan metode Random Forest. |
| 8 | (LOUIS MADAERDO | K-Nearest Neighbor, | Berdasarkan hasil performa model klasifikasi dari model algoritma yang |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | SOTARJUA & DIAN BUDHI SANTOSO, 2022) | *Decision Tree dan Random Forest* | Berdasarkan hasil performa model klasifikasi dari model algoritma yang digunakan pada penelitian ini, maka model KNN memiliki hasil performa yang terbaik nilai yaitu nilai akurasinya 86,57% metriks evaluasinya. Sehingga pada penelitian ini, model KNN adalah model klasifikasi yang lebih baik digunakan pada penelitian ini, dibandingkan dengan algoritma *Decision Tree* akurasi 85,29% dan *Random Forest* akurasi 86.37%. |
| 9 | (Syukri Mustafa & Wayan Simpen, 2019) | K-Nearest Neighbor | sistem dapat melakukan prediksi berdasarkan kedekatan histori data yang ada dengan data baru, menentukan apakah pasien diprediksi terkena diabetes atau tidak. Data pelatihan terdiri dari 200 data pasien yang telah menjalani pemeriksaan dalam 2 tahun terakhir. Pengujian dilakukan dengan menggunakan data uji atau sampel sebanyak 104 data pasien yang diambil dari luar data pelatihan, dengan tujuan membandingkan hasil prediksi |

| No | Peneliti | Metode | Hasil |
|---|---|---|---|
| | | | Status Diabetes Militus Pasien dari sistem yang dirancang dengan hasil Status Diabetes Militus Pasien yang sesuai dengan data. Hasil akurasi yang diperoleh dari pengujian tersebut sebesar 68,30%. |

B. Landasan Teori

1. Data Mining

Data mining merupakan pemisahan pola yang menarik dari suatu data yang besar menggunakan matematika, statistik, machine learning maupun kecerdasan buatan (Sugawara & Nikaido, 2014). Proses data mining melibatkan Penerapan analisis data yang cermat dan eksploitasi algoritma machine learning bertujuan untuk mengungkapkan pola atau relasi yang ada dalam data. Hal ini bermanfaat untuk melakukan prediksi, meningkatkan efisiensi bisnis, dan memperoleh landasan keputusan yang lebih matang. Tujuan utama dari data mining adalah untuk menemukan hubungan atau pola yang tersembunyi dalam data yang mungkin sulit atau tidak mungkin ditemukan dengan metode analisis statistik konvensional (Sugawara & Nikaido, 2014). Data mining melibatkan beberapa tahap, termasuk pemahaman masalah, pengumpulan data, pemrosesan data, transformasi data, dan analisis data menggunakan teknik data mining seperti regresi, klasifikasi, clustering, dan association rule. Hasil analisis data mining digunakan untuk mengembangkan model prediktif atau deskriptif yang dapat membantu dalam mengambil keputusan di masa depan.

2. Klasifikasi

Klasifikasi merupakan penggolongan atau pengelompokan fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya belum diketahui atau pembagian sesuatu menurut kelas-kelas nya (Dinata et al., 2020). Tujuan utama klasifikasi adalah untuk mengenali pola atau hubungan antara fitur atau atribut dan kelas atau kategori untuk dapat memprediksi kelas atau kategori dari objek atau pengamatan yang belum dikenal (Sulastri et al., 2020). Klasifikasi merupakan suatu bentuk analisis data yang membantu individu untuk memprediksi label atau kategori yang seharusnya diberikan kepada sampel tertentu. Dalam klasifikasi, data historis digunakan sebagai data pelatihan untuk membangun model klasifikasi, dan model ini kemudian diuji menggunakan data pengujian untuk menguji kinerjanya. Beberapa teknik klasifikasi yang umum digunakan dalam data mining adalah pohon keputusan, *Naive Bayes,* K-Nearest Neighbor *(k-NN), Decision Tree, dan Support Vector Machines (SVM).*

3. Algoritma *K-Nearest Neigbor*

Algoritma K-Nearest Neighbor (K-NN) adalah algoritma klasifikasi yang digunakan untuk memprediksi label atau kelas suatu sampel data berdasarkan data yang terdekat dengan sampel tersebut. Algortima K-Nearest Neighbor (KNN) adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya (Syukri Mustafa & Wayan Simpen, 2019). Algoritma K-Nearest Neighbor memiliki tujuan untuk mengklasifikasikan atau mengelompokkan objek data baru berdasarkan jarak data baru tersebut ke beberapa data terdekat dengan berdasarkan atribut dan sampel dari data pelatihan (Yuliarina & Hendry, 2022). Keuntungan dari klasifikasi non-parametrik adalah fleksibilitasnya dalam menangani berbagai bentuk distribusi data dan tidak membutuhkan asumsi

tertentu tentang data. Menurut (Dinata et al., 2020) ada beberapa metode pendekatan (KNN) dengan K tetangga (*neighbor*) terdekat dalam data *uji,* pada penelitian ini menggunakan rumus jarak KNN Euclidean karena jarak Euclidean sangat sederhana, mudah dimengerti, fleksibel dapat digunakan pada data numeriik dan dapat diimplementasikan dengan mudah. Adapun rumus jarak euclidean dapat dilihat dibawah ini (Dinata et al., 2020) pada persamaan (1) :

a. Jarak Euclidean

$$d_i\,(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \ldots\ldots\ldots (1)$$

Keterangan:

a. xi = sampel data
b. yi = data Uji (*uji)*
c. i = variabel data
d. d(x,y) = jarak antara 2 titik x dan y
e. m = dimensi data

*4. Confusion matrix*

*Confusion matrix* adalah metode evaluasi yang umum digunakan dalam bidang kecerdasan buatan dan pembelajaran mesin untuk mengukur kinerja model klasifikas. *Confusion matrix* adalah metode dalam evaluasi yang dapat menentukan kinerja yang didasarkan pada benar dan salah pada sebuah klasifikasi (Rizaldi & Mustakim, 2020). *Confusion matrix* menyajikan metrik evaluasi, termasuk akurasi, precision, dan recall. Akurasi menggambarkan efektivitas keseluruhan algoritma. *Precision* adalah proporsi data yang benar-benar positif dari keseluruhan hasil positif yang diprediksi oleh model terhadap

data sebenarnya. Sedangkan *recall* mengukur seberapa baik model dapat memprediksi data positif dengan benar dari keseluruhan data positif yang sebenarnya. *Confusion matrix* memberikan informasi tentang jumlah prediksi yang benar dan salah untuk setiap kelas target yang ada. *Confusion matrix* pada klasifikasi biner dengan dua kelas umumnya memiliki empat elemen utama:

a. *True Positive* (TP): Jumlah contoh yang diklasifikasikan dengan benar sebagai positif (kelas yang benar adalah positif dan model juga memprediksi positif).
b. *True Negative* (TN): Jumlah contoh yang diklasifikasikan dengan benar sebagai negatif (kelas yang benar adalah negatif dan model juga memprediksi negatif).
c. *False Positive* (FP): Jumlah contoh yang salah diklasifikasikan sebagai positif (kelas yang benar adalah negatif tetapi model memprediksi positif).
d. *False Negative* (FN): Jumlah contoh yang salah diklasifikasikan sebagai negatif (kelas yang benar adalah positif tetapi model memprediksi negatif).

Melakukan pengujian untuk memperkirakan objek yang benar dan salah dapat dilihat tabel 2. 2 *confusion matrix*

Tabel 2. 2 *Confusion matrix*

| | NILAI AKTUAL | |
|---|---|---|
| NILAI PREDIKSI | TP | FN |
| | FP | TN |

Adapun rumus perhitungan *confusion matrix* bila dituliskan dapat dilihat dibawah ini :

a. Akurasi mengukur sejauh mana model mampu membuat prediksi yang benar secara keseluruhan (Dinata et al., 2020), dihitung sebagai rasio

antara jumlah prediksi benar (*true positives* dan *true negatives*) dengan jumlah total prediksi. Akurasi memberikan gambaran umum tentang seberapa baik model berkinerja, tetapi dapat memberikan hasil yang bias jika kelas target tidak seimbang, rumus akurasi dapat dilihat pada persamaan (2)

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots \quad (2)$$

b. Presisi mengukur sejauh mana prediksi positif yang dibuat oleh model benar (Dinata et al., 2020), dihitung sebagai rasio antara *true positives* dengan jumlah semua prediksi positif (*true positives* dan *false positives*). Presisi berguna ketika penting untuk menghindari kesalahan positif palsu, dan fokus pada seberapa tepat model dalam memprediksi suatu kelas, rumus presisi dapat dilihat pada persamaan (3)

$$precision = \frac{TP}{TP+FP} \dots\dots\dots\dots\dots \quad (3)$$

c. Recall mengukur sejauh mana model mampu mendeteksi atau "mengingat" instance positif yang sebenarnya (Dinata et al., 2020), dihitung sebagai rasio antara *true positives* (*true positives* dan *false negatives*). Recall berguna ketika penting untuk menghindari kesalahan negatif palsu, dan fokus pada seberapa baik model dapat mengidentifikasi semua instance positif, rumus recall dapat dilihat pada persamaan (4)

$$recall = \frac{TP}{TP+FN} \dots\dots\dots\dots\dots\dots \quad (4)$$

d. F1 score adalah nilai rata-rata harmonis antara presisi dan recall, berguna ketika kita ingin menyeimbangkan keduanya (Dinata et al., 2020). F1 score baik digunakan ketika terdapat ketidakseimbangan antara kelas positif dan negatif, dan kita ingin mencari nilai yang

seimbang antara presisi dan recall, rumus recall dapat dilihat pada persamaan (5)

$$F1 - score = \frac{2*(Precision*Recall)}{(Precision+Recall)}.. \quad (5)$$

*5. Python*

*Python* adalah bahasa pemrograman yang sangat kuat dan populer untuk data mining. *Python* memiliki keunggungulan seperti *readability*, efisien, multifungsi, interoperabilitas, dan memiliki dukungan komunitas yang memadai (Retnoningsih & Pramudita, 2020). *Python* memungkinkan para peneliti dan analis data untuk mengakses, mengolah, dan menganalisis data dengan efisien. *Python* juga mendukung berbagai teknik data mining seperti clustering, klasifikasi, regresi, dan pengelompokan yang dapat diterapkan untuk mengungkap pola dan wawasan yang berharga dari data besar. Membangun model klasifikasi, perlu memuat dataset dan membaginya menjadi data pelatihan dan data *uji*. Setelah model dilatih dengan menggunakan data pelatihan, dapat menggunakan data pengujian untuk mengukur kinerja model, misalnya dengan menggunakan metrik akurasi, presisi, recall, atau F1-score. Proses ini memungkinkan untuk mengembangkan model yang dapat memprediksi kelas atau label yang benar berdasarkan fitur-fitur tertentu, membuka pintu untuk berbagai aplikasi klasifikasi dalam analisis data dan kecerdasan buatan menggunakan *Python*. Kombinasi antara kemudahan penggunaan, dukungan komunitas yang luas, dan ekosistem yang kaya menjadikan *Python* sebagai bahasa yang ideal untuk tujuan data mining dan analisis data yang kompleks.