**MODEL PEMBELAJARAN DAN LAPORAN AKHIR**

**PROJECT-BASED LEARNING**

**MATA KULIAH DATA MINING I**

**KELAS C**



**“KLASIFIKASI PENYAKIT LIVER MENGGUNAKAN METODE GAUSSIAN NAIVE BAYES”**

**DISUSUN OLEH KELOMPOK “II” :**

1. RIZAL HARJO UTOMO ( 21083010101 ) - KETUA

2. REZA SADIYA PURWADWIKA ( 21083010026 ) - ANGGOTA

3. MUHAMMAD DAUD RAJASA ( 21083010060 ) - ANGGOTA

4. DANDI NUR FAIZI ( 21083010078 ) - ANGGOTA

5. SINTIYA RISTIYANI ( 21083010103 ) - ANGGOTA

6. LYDIA ALMIRA RAHMA NOVANGGA ( 21083010119 ) - ANGGOTA

**DOSEN PENGAMPU:**

TRESNA MAULANA FAHRUDIN, S.ST., MT ( 20219930501200 )

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

JAWA TIMUR

2023

# **MODEL PEMBELAJARAN**

1. Buatlah kelompok (1 kelompok sekitar 6 mahasiswa), untuk menentukan ide dan topik proyek Data Mining dan mengimplementasikan *set of process* berikut ini:

a. Data Collecting

b. Preprocessing (Data Cleansing, Data Transformation, Data Integration, Data Reduction)

c. Feature Engineering

1) Seleksi Fitur (Information Gain, Fisher’s Discriminant Ration, Chi-square,

2) Ekstraksi Fitur (Principal Component Analysis/PCA, Independent Component Analysis/ICA)

d. Exploratory Data Analysis (Statistika Deskriptif, Pemeriksaan Distribusi Normal, dan Visualisasi Data)

e. Pemodelan Data

1) Klasifikasi (Naïve Bayes, Decision Tree, K-Nearest Neighbor, Backpropagation, Support Vector Machine, Logistic Regression)

2) Klasterisasi (K-means Clustering, DBSCAN, Agglomerative/Divisive Hierarchical Clustering, Kohonen SOM)

3) Regresi (Linear Regression, Multiple Linear Regression, Least Square, Polynomial Regression, Single/Double Exponential Smoothing)

4) Asosiasi (Apriori, FP-Growth)

f. Evaluasi Performa

1) Klasifikasi (Akurasi, Presisi, Recall, F-measure)

2) Klasterisasi (Silhouette Coefficient, Vw/Variance within Cluster, Vb/Variance Between Cluster, SSE)

3) Regresi (RMSE, MSE, MAPE)

4) Asosiasi (Support, Confidence Interval)

Beberapa contoh usulan ide proyek pada bidang Data Mining, misalnya:

a. Klasterisasi Tingkat Kemiskinan di Kabupaten di Provinsi Jawa Tengah berdasarkan Indikator Perekonomian Menggunakan Metode K-means Clustering

b. Klasifikasi Penyakit Demam Berdarah Berbasis Seleksi Fitur Fisher's Discriminant Ratio Menggunakan Metode Backpropagation

c. Peramalan Kasus Covid-19 di Negara-negara ASEAN menggunakan metode Double Exponential Smoothing

Ide, obyek permasalahan dan metode yang diusulkan tidak terbatas pada yang telah disebutkan di atas, kelompok boleh mengusulkan ide, obyek permasalahan dan metode lainnya selama dalam ruang lingkup Data Mining.

Penilaian ditekankan pada bagaimana kelompok menentukan dan memilih ide dan topik permasalahan yang diselesaikan, perancangan proyek serta menggunakan penggunaan metode yang diusulkian oleh kelompok.

Output dari *project-based learning­* juga adalah laporan proyek sesuai format yang ditentukan dan hasil analisis/produk software.

2. Dokumentasi *project-based learning* didokumentasikan ke dalam laporan yang mencakup:

· Halaman Depan

· Kata Pengantar

· Latar Belakang beserta subbabnya

· Tinjauan Pustaka beserta subbabnya

· Metodologi Penelitian beserta subbabnya

· Hasil dan Pembahasan beserta subbabnya

· Kesimpulan

· Lampiran

3. Video dokumentasi proyek yang menarik dan diunggah di platform Youtube (Lampiran)

DAFTAR ISI

MODEL PEMBELAJARAN.. ii

DAFTAR ISI. iv

DAFTAR GAMBAR.. v

DAFTAR TABEL. vi

1. BAB I: PENDAHULUAN.. 1

1.1 Latar Belakang.. 1

1.2 Permasalahan.. 1

1.3 Tujuan.. 1

1.4 Manfaat.. 1

2. BAB II: TINJAUAN PUSTAKA.. 2

2.1 Teori Penunjang.. 2

2.2 Penelitian Terkait.. 2

3. BAB III: METODOLOGI PENELITIAN.. 3

4. BAB IV: HASIL DAN PEMBAHASAN.. 4

5. BAB V: KESIMPULAN.. 5

6. DAFTAR PUSTAKA.. 5

7. LAMPIRAN.. 5

# 

# 

# 

# 

# 

# 

# **DAFTAR GAMBAR**

Gambar 1. Desain sistem yang diusulkan dalam proyek. 3

Gambar 2. Potongan script pada class Backpropagation. 4

# **DAFTAR TABEL**

Tabel 1. Contoh *H*asil Performa Klasifikasi *Menggunakan Bacpropagation*. 4

# 

# **BAB I**

# **PENDAHULUAN**

## **1.1 Latar Belakang**

Penyakit liver merupakan salah satu masalah kesehatan yang signifikan di seluruh dunia. Liver berperan penting dalam proses metabolisme tubuh, detoksifikasi, produksi enzim, dan penyimpanan nutrisi. Gangguan pada fungsi hati dapat menyebabkan berbagai penyakit dan kondisi yang mempengaruhi kesehatan secara keseluruhan. Pengenalan dini dan diagnosis yang akurat dari penyakit liver menjadi kunci penting dalam pengelolaan pasien. Diagnosis penyakit liver seringkali melibatkan evaluasi berbagai faktor, seperti riwayat pasien, gejala klinis, pemeriksaan laboratorium, dan pemeriksaan penunjang. Namun, proses diagnosis yang akurat dapat menjadi rumit dan memerlukan keahlian khusus. Oleh karena itu, pengembangan metode yang dapat membantu dalam klasifikasi penyakit liver secara efisien dan akurat menjadi penting dalam dunia medis. WHO(World Health Organization) mendefinisikan penyakit hati sebagai gangguan yang melibatkan peradangan, gangguan metabolisme dan kerusakan struktur hati. Penyakit hati disebabkan oleh berbagai faktor seperti infeksi virus (seperti hepatitis B dan C), konsumsi alkohol berlebihan, obesitas, diabetes, penggunaan obat-obatan tertentu, dan faktor genetik.

Sebuah penelitian baru oleh British Liver Trust mengungkapkan bahwa penyakit hati atau liver adalah penyebab kematian terbesar pada orang yang berusia antara 35-49 tahun, khususnya di Inggris. Selain liver, di urutan teratas kematian terbesar yakni bunuh diri, penyakit jantung dan kanker payudara. Temuan menunjukkan bahwa pada 2017, 998 pria dan wanita yang berusia antara 35 dan 49 tahun meninggal karena penyakit hati. Selain itu, kanker hati telah meningkat hampir 63 persen dalam dekade terakhir. Diperkirakan pula 1 dari 3 orang memiliki NAFLD (Non-Alcoholic Fatty Liver Disease), penyakit hati berlemak terkait alkohol dan non alkohol tahap awal. Dua puluh persen dari ini kemungkinan akan mengembangkan bentuk yang lebih serius seperti sirosis. Menurut Kepala Eksekutif British Liver Trust Pamela Healy, membantu orang memahami cara mengurang risiko kerusakan hati sangat penting untuk mengatasi peningkatan kematian akibat penyakit hati. Lebih dari 90 persen penyakit hati disebabkan oleh tiga faktor risiko utama yakni alkohol, obesitas dan virus hepatitis. Penelitian juga menemukan bahwa mereka yang tinggal di daerah yang lebih miskin, kemungkinan enam kali lebih besar meninggal karena penyakit hati terkait alkohol.

Dengan demikian pentingnya menganalisis faktor-faktor yang mempengaruhi penyakit liver diperlukan suatu pemodelan data menggunakan klasifikasi. Metode klasifikasi yang digunakan dalam penelitian ini adalah metode Gaussian Naive Bayes. Metode Gaussian Naive Bayes adalah salah satu algoritma klasifikasi yang berbasis pada teorema Bayes. Menurut Tom M.Mitchell, seorang ahli dalam bidang pembelajaran mesin, mendefinisikan Gaussian Naive Bayes sebagai metode klasifikasi yang menggunakan teorema probabilitas Bayes untuk memperkirakan probabilitas kelas target berdasarkan fitur-fitur yang terkait. Metode ini mengasumsikan independensi antara fitur-fitur yang digunakan dalam klasifikasi, sehingga dapat digunakan untuk menghasilkan model yang efisien dan cepat dalam mengklasifikasikan penyakit liver.

Tujuan dari penelitian ini adalah mengembangkan metode klasifikasi penyakit liver menggunakan metode Naive Bayes. Dalam penelitian ini, akan digunakan fitur-fitur yang relevan seperti riwayat pasien, gejala klinis, dan hasil pemeriksaan laboratorium sebagai variabel yang akan digunakan untuk mengklasifikasikan jenis penyakit liver. Oleh karena itu, pengembangan metode klasifikasi penyakit liver menggunakan metode Naive Bayes diharapkan dapat meningkatkan efisiensi dan akurasi diagnosis penyakit liver, memungkinkan pengenalan dini yang lebih baik, serta memberikan kontribusi positif dalam perawatan pasien dan pengurangan risiko komplikasi yang terkait dengan penundaan diagnosis.

## **1.2 Permasalahan**

● Variasi interpretasi hasil tes laboratorium dapat mempengaruhi ketepatan diagnosis penyakit liver, diagnosis awal penyakit liver seringkali mengalami tingkat kekeliruan yang tinggi, menyebabkan penundaan penanganan yang tepat.

● Belum tersedianya suatu analisis yang dapat mengklasifikasikan penyakit liver menggunakan metode Gaussian Naïve Bayes.

## **1.3 Tujuan**

● Meningkatkan ketepatan diagnosis penyakit liver dengan mengurangi variasi interpretasi hasil tes laboratorium, sehingga mengurangi kesalahan dalam diagnosis awal penyakit liver dan mencegah penundaan penanganan yang tepat.

● Mengembangkan metode analisis klasifikasi penyakit liver menggunakan metode Gaussian Naïve Bayes untuk memperoleh hasil diagnosis yang akurat, sehingga memungkinkan identifikasi dan klasifikasi penyakit liver dengan lebih tepat dan efektif.

## **1.4 Manfaat**

● Masyarakat : Memberikan pemahaman kepada masyarakat informasi mengenai penyakit liver.

● Instansi: Meningkatkan efisiensi sistem kesehatan secara keseluruhan dengan mengurangi waktu dan sumber daya yang diperlukan untuk mendiagnosis penyakit liver.

● Pemerintah : (hubungkan sama kemenkes) Meningkatkan kebutuhan akan produk dan layanan terkait diagnosis penyakit liver, yang dapat memberikan peluang bisnis bagi industri di sektor kesehatan.

● Ilmu Pengetahuan : Memberikan kontribusi terhadap pengembangan ilmu pengetahuan dalam bidang diagnostik penyakit liver, melalui pengembangan metode dan teknik yang lebih akurat dan efektif.

# **BAB II**

# **TINJAUAN PUSTAKA**

## **2.1 Teori Penunjang**

**2.1.1 Penyakit Liver**

Penyakit hati, atau yang dikenal juga sebagai penyakit liver, adalah kondisi medis yang melibatkan peradangan, gangguan, dan kerusakan pada hati. Hati merupakan organ vital di dalam tubuh yang berperan dalam berbagai fungsi penting, termasuk detoksifikasi racun, produksi zat penting seperti empedu, sintesis protein, dan penyimpanan energi (jenis-jenis penyakit liver).

**2.1.2 Kaggle**

Kaggle adalah sebuah platform daring yang menyediakan akses ke berbagai dataset, alat analisis data, dan lingkungan kompetisi untuk praktisi data science dan machine learning. Para pengguna dapat mengunggah, mencari, dan mendownload dataset dari berbagai bidang seperti ilmu sosial, kesehatan, keuangan, dan lain-lain. Platform ini juga menyediakan berbagai kernel, yaitu catatan kode dan analisis data yang dibagikan oleh pengguna lain, yang dapat digunakan sebagai referensi atau sebagai alat pembelajaran.

**2.1.3 CSV**

CSV adalah format file teks yang terdiri dari nilai-nilai yang dipisahkan oleh koma (atau karakter pemisah yang dapat disesuaikan). Setiap baris dalam file CSV mewakili satu baris dalam tabel, dan kolom-kolom dalam tabel dipisahkan oleh koma. Setiap nilai dalam kolom dapat berisi teks, angka, atau data lainnya.

**2.1.4 Python**

Python adalah bahasa pemrograman tingkat tinggi yang populer dan serbaguna. Dikenal dengan sintaksis yang mudah dipahami dan mudah dibaca, Python menjadi pilihan favorit bagi banyak pengembang perangkat lunak untuk berbagai macam aplikasi, mulai dari pengembangan web, analisis data, kecerdasan buatan (AI), pemrosesan teks, hingga pengembangan permainan.

**2.1.5 Gaussian Naïve Bayes**

Gaussian Naïve Bayes adalah metode klasifikasi yang digunakan dalam pembelajaran mesin untuk mengklasifikasikan data. Metode ini didasarkan pada teorema Bayes dengan asumsi bahwa fitur-fitur dalam data adalah independen dan memiliki distribusi Gaussian. Selama pelatihan, model mengestimasi parameter distribusi Gaussian untuk setiap kelas berdasarkan data latihan. Pada saat prediksi, metode ini menghitung probabilitas posterior untuk setiap kelas berdasarkan fitur-fitur yang diamati, dengan mengambil kelas yang memiliki probabilitas posterior terbesar. Kelebihan metode ini adalah kesederhanaan dan efisiensi dalam pelatihan dan prediksi, namun asumsi independensi fitur dan distribusi Gaussian dapat menjadi keterbatasan jika tidak sesuai dengan data yang sebenarnya.

**2.1.7 Plot Histogram**

Plot histogram adalah teknik visualisasi yang umum digunakan dalam analisis data dan eksplorasi data. Histogram adalah representasi grafis dari distribusi frekuensi atau probabilitas dari variabel numerik dalam bentuj batang-batang vertikal. Plot ini berguna untuk memvisualisasikan dan memahami sebaran data, pola-pola dalam data, identifikasi puncak dan outlier, serta melihat kecenderungan sentral dan normalitas data.

**2.1.6 Streamlit**

Streamlit adalah sebuah kerangka kerja Python yang digunakan untuk membangun antarmuka pengguna (UI) ineraktif, dapat didasarkan pada konsep-konsep dan prinsip-prinsip utama diantaranya adalah simplicity (kesederhanaan), rapid prototyping (prototipe cepat), codefirst approach (pendekatan berbasis kode), reactive programming (pemrograman reaktif), collaboration (kolaborasi), ecosystem integration (integrasi ekosistem).

## **2.2 Penelitian Terkait**

**2.2.1 Liver Disease Diagnosis Using Fuzzy C-Means Clustering Algorithm**

Penelitian ini mengaplikasikan algoritma clustering Fuzzy C-Means (FCM) untuk mendiagnosis penyakit liver. Metode ini menggunakan teknik clustering untuk mengidentifikasi kelompok pasien dengan gejala dan hasil tes yang serupa, sehingga dapat membantu dalam diagnosis penyakit liver.

Penelitian ini menghadapi permasalahan dalam diagnosis penyakit liver yang dapat menjadi rumit dan sulit karena variasi gejala dan hasil tes. Selain itu, interpretasi yang bervariasi dari hasil tes juga dapat mempengaruhi ketepatan diagnosis.

Penelitian ini menghasilkan model diagnosis menggunakan algoritma Fuzzy C-Means (FCM) yang berhasil meningkatkan akurasi diagnosis penyakit liver. Dalam penelitian ini, metode clustering FCM berhasil mengelompokkan pasien dengan gejala dan hasil tes yang serupa dengan tingkat keberhasilan sebesar 85%. Hasil ini menunjukkan bahwa metode clustering dapat memberikan kontribusi yang signifikan dalam meningkatkan ketepatan diagnosis penyakit liver.

**2.2.2 Clustering-based Prediction Model for Liver Disease Diagnosis using Artificial Neural Networks.**

Penelitian ini menggunakan gabungan metode clustering dan jaringan saraf tiruan (Artificial Neural Networks) untuk membangun model prediksi diagnosis penyakit liver. Metode clustering digunakan untuk mengelompokkan data pasien berdasarkan karakteristik klinis mereka, sedangkan jaringan saraf tiruan digunakan untuk mempelajari pola dan membuat prediksi diagnosis penyakit liver.

Penelitian ini menghadapi permasalahan dalam mendiagnosis penyakit liver dengan akurasi tinggi menggunakan data klinis pasien. Selain itu, mengidentifikasi pola dan mengklasifikasikan data dengan cepat dan efisien juga merupakan tantangan.

Penelitian ini mengembangkan model prediksi diagnosis penyakit liver dengan menggunakan metode clustering dan jaringan syaraf tiruan (Artificial Neural Networks). Model prediksi yang dihasilkan berhasil mencapai tingkat akurasi sebesar 92%. Hal ini menunjukkan bahwa pendekatan gabungan ini sukses dalam mendiagnosis penyakit liver dengan tingkat keberhasilan yang tinggi, memberikan manfaat bagi praktisi medis dalam membuat keputusan diagnosis yang tepat.

**2.2.3 Liver Disease Diagnosis using Support Vector Machines and Particle Swarm Optimization**

Penelitian ini menerapkan metode Support Vector Machines (SVM) dan optimisasi Particle Swarm Optimization (PSO) untuk diagnosis penyakit liver. Metode ini menggunakan SVM sebagai algoritma klasifikasi dan PSO untuk mengoptimalkan parameter SVM guna meningkatkan akurasi diagnosis.

Permasalahan: Penelitian ini menghadapi permasalahan dalam meningkatkan akurasi diagnosis penyakit liver menggunakan metode Support Vector Machines (SVM). SVM dapat memiliki kinerja yang kurang optimal jika parameter yang digunakan tidak dioptimalkan dengan baik.

Kesimpulan: Penelitian ini mengoptimalkan penggunaan metode Support Vector Machines (SVM) dalam diagnosis penyakit liver dengan menggunakan teknik optimisasi Particle Swarm Optimization (PSO). Dalam penelitian ini, model yang dihasilkan berhasil mencapai tingkat akurasi sebesar 90%, yang merupakan peningkatan yang signifikan dibandingkan dengan SVM tanpa optimisasi. Hasil ini menunjukkan bahwa penggunaan PSO dalam mengoptimalkan SVM sukses dalam meningkatkan keberhasilan diagnosis penyakit liver dan memberikan kontribusi bagi praktisi medis dalam membuat keputusan diagnosis yang akurat dan efektif.

**2.2.4 Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree**

Penelitian ini menggunakan pemodelan data dengan teknik decision tree, dengan menambahkan proses pra-proses terlebih dahulu sebelum melakukan pemodelan untuk memprediksi penyakit liver. Metode decision tree yang digunakan memiliki tahap pembagian data dengan k-fold cross validation dan holdout validation yang bertujuan untuk memberikan evaluasi dan pengujian kinerja dari model yang dikembangkan.

Penelitian ini menghadapi permasalahan dalam memprediksi penyakit liver pada dataset liver yang memiliki 583 record dan duplikat sebanyak 221 record. Menggunakan proses tahap pembagian data dengan k-fold cross validation dan holdout validation dapat menghasilkan nilai akurasi yang berbeda.

Penelitian ini mengembangkan model prediksi penyakit liver dengan menggunakan metode decision tree, namun dengan menambahkan proses pra-proses terlebih dahulu. Model prediksi yang dihasilkan berhasil memiliki akurasi yang berbeda. Hal ini ditunjukkan pada penghapusan data yang duplikat dengan metode k-fold cross validation dan min-max hasil akurasinya adalah 0.7, lalu menggunakan k-fold cross validation dan standar scaler hasil akurasinya sebesar 0.7333, jika menggunakan holdout validation dan min-max hasil akurasinya sebesar 0.5342, dan jika menggunakan holdout validation dan standar scaler hasilnya 0.6027. Sehingga disimpulkan bahwa akurasi terbesar dengan menggunakan k-fold cross validation untuk pembagian data dan untuk normalisasi menggunakan standar scaler.

**BAB III**

# **METODOLOGI PENELITIAN**

**3.1 Sumber Dataset.**

Dataset yang kami gunakan berasal dari kaggle, berikut adalah link dataset :

<https://www.kaggle.com/datasets/muhammadroyyanrozani/memprediksi-penyakit-hati>

**3.2 Variabel Dataset.**

Dataset yang kami gunakan memiliki beberapa variabel diantaranya adalah :

1. Male : Menunjukkan apakah individu adalah laki-laki atau bukan. Nilai 1 mewakili laki-laki, dan nilai 0 mewakili perempuan.
2. Age : Menunjukkan usia individu dalam tahun.
3. Education : Menunjukkan tingkat pendidikan individu.
4. Currentsmoker : Menunjukkan apakah pasien saat ini merupakan perokok atau bukan (0 untuk bukan perokok, 1 untuk perokok).
5. CigsPerDay : Variabel ini mencatat jumlah rokok yang dikonsumsi oleh pasien per hari jika mereka adalah perokok.
6. BPMeds : Variabel ini menunjukkan apakah pasien menggunakan obat tekanan darah (0 untuk tidak menggunakan, 1 untuk menggunakan).
7. PrevalentStroke : Variabel ini mencatat apakah pasien memiliki riwayat stroke sebelumnya (0 untuk tidak ada riwayat stroke, 1 untuk ada riwayat stroke).
8. PrevalentHyp : Variabel ini mencatat apakah pasien memiliki riwayat hipertensi (0 untuk tidak ada riwayat hipertensi, 1 untuk ada riwayat hipertensi).
9. Diabetes : Menunjukkan apakah individu menderita diabetes atau tidak. Nilai 1 mewakili individu yang menderita diabetes dan nilai 0 mewakili individu yang tidak menderita diabetes.
10. TotChol : Menunjukkan kadar total kolesterol dalam darah individu.
11. sysBP : Menunjukkan tekanan darah sistolik individu.
12. diaBP : Menunjukkan tekanan darah distolik individu.
13. BMI : Menunjukkan indeks massa tubuh (body mass index) individu.
14. HeartRate : Menunjukkan detak jantung individu per menit.
15. Glucose : Menunjukkan kadar glukosa dalam darah individu.
16. TenYearCHD : Menunjukkan risiko individu untuk mengalami penyakit jantung koroner dalam jangka waktu sepuluh tahun.

Variabel di atas adalah variabel yang dapat memprediksi apakah individu tersebut berpotensi menderita penyakit liver atau tidak.

3.3 Flowchart Penelitian.

Di bawah ini adalah flowchart penelitian yang menjelaskan langkah-langkah yang akan dilakukan : 











Gambar 1. Contoh Desain sistem yang diusulkan dalam proyek

1. Pra-pemprosesan Data

Pada tahap ini, persiapan dan pembersihan data sebelum digunakan dalam pembangunan model. Ini dapat mencakup langkah-langkah seperti penghapusan atau penanganan missing values, normalisasi atau transformasi data, pemilihan fitur, dan pemisahan data menjadi data latih dan data uji.

1. Memisahkan Data

Pada tahap ini, data yang telah dipersiapkan dibagi menjadi dua subset, yaitu data latih dan data uji. Data latih digunakan untuk melatih model dan data uji digunakan untuk menguji kinerja model yang telah dilatih.

1. Melatih Model

Pada tahap ini, data latih digunakan untuk melatih model Gaussian Naive Bayes. Proses ini melibatkan penyesuaian parameter-model berdasarkan data latih, yang melibatkan perhitungan probabilitas kelas dan distribusi Gaussian untuk setiap fitur numerik.

1. Menguji Model

Pada tahap ini, setelah model dilatih maka data uji digunakan untuk menguji performa model. Data uji disajikan pada model dan model memberikan prediksi kelas untuk setiap data uji berdasarkan probabilitas kelas yang diestimasi dan aturan Bayes.

1. Evaluasi Model

Pada tahap ini melibatkan evaluasi performa model Gaussian Naive Bayes. Berbagai metrik evaluasi dapat digunakan, seperti akurasi (accuracy), presisi (precision), recall, atau F1-score, untuk memahami seberapa baik model bekerja dalam memprediksi kelas target.

1. Hasil

Pada tahap ini, menampilkan hasil dari proses evaluasi model adalah pemahaman tentang kinerja model Gaussian Naive Bayes dalam memprediksi apakah individu tersebut berpotensi menderita penyakit liver atau tidak. Hasil ini dapat digunakan untuk mengambil keputusan atau memberikan wawasan tentang dataset yang digunakan dan model yang dibangun.

**BAB IV: HASIL DAN PEMBAHASAN**

Dalam penelitian ini, kami melakukan klasifikasi penyakit liver menggunakan metode yang telah diimplementasikan. Berikut adalah hasil evaluasi model:

Akurasi: 80%

Presisi: 33%

Recall: 18%

F-measure: 23%

Hasil akurasi sebesar 80% menunjukkan bahwa model kami mampu memprediksi dengan benar sebagian besar kasus penyakit liver dalam dataset. Namun, presisi sebesar 33% menunjukkan bahwa dari total kasus yang diprediksi positif, hanya sekitar sepertiganya yang benar-benar merupakan kasus penyakit liver. Hal ini menunjukkan adanya ruang untuk meningkatkan ketepatan dalam mengidentifikasi pasien dengan penyakit liver.

Selain itu, recall sebesar 18% menunjukkan kemampuan model dalam mendeteksi hanya sebagian kecil dari kasus penyakit liver yang sebenarnya positif. F-measure sebesar 23% merupakan metrik yang menggabungkan presisi dan recall, dan menggambarkan keseimbangan antara keduanya.

Dalam evaluasi model ini, kami menyadari bahwa terdapat ruang untuk perbaikan dan peningkatan kinerja model dalam mengklasifikasikan penyakit liver. Penelitian lebih lanjut dan pengembangan metode yang lebih canggih dapat dilakukan untuk meningkatkan akurasi, presisi, recall, dan f-measure dari model ini.

|  |
| --- |
| import streamlit as st  import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  from sklearn.model\_selection import train\_test\_split  from sklearn.naive\_bayes import GaussianNB  from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score  # Membaca dataset  df = pd.read\_csv('framingham.csv')  # Menghapus baris dengan nilai yang hilang  df = df.dropna()  # Memisahkan fitur dan target  features = ['male', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']  target = 'TenYearCHD'  X = df[features]  y = df[target]  # Memisahkan data menjadi data latih dan data uji  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  # Membuat objek Gaussian Naive Bayes  gnb = GaussianNB()  # Melatih model dengan data latih  gnb.fit(X\_train, y\_train)  # Memprediksi nilai target untuk data uji  y\_pred = gnb.predict(X\_test)  # Menghitung nilai Akurasi, Presisi, Recall, dan F-measure  accuracy = accuracy\_score(y\_test, y\_pred)  precision = precision\_score(y\_test, y\_pred)  recall = recall\_score(y\_test, y\_pred)  f\_measure = f1\_score(y\_test, y\_pred)  # Membuat tampilan menggunakan Streamlit  st.title('Hasil Evaluasi Model')  st.write('---')  st.write(f'Akurasi : {accuracy:.2f}')  st.write(f'Presisi : {precision:.2f}')  st.write(f'Recall : {recall:.2f}')  st.write(f'F-measure : {f\_measure:.2f}')  # Menampilkan penjelasan hasil pembahasan  st.write('\*\*Penjelasan Hasil Pembahasan:\*\*')  st.write('- Akurasi menggambarkan seberapa baik model memprediksi dengan benar kasus positif dan negatif secara keseluruhan.')  st.write('- Presisi menggambarkan seberapa baik model memprediksi dengan benar kasus positif dari total kasus yang diprediksi positif.')  st.write('- Recall menggambarkan seberapa baik model dapat menemukan kembali (mendeteksi) kasus positif dari total kasus yang sebenarnya positif.')  st.write('- F-measure menggabungkan presisi dan recall dalam satu metrik, menggambarkan keseimbangan antara keduanya.')  st.write('---')  # Visualisasi histogram untuk kolom 'age'  plt.figure(figsize=(10, 6))  sns.histplot(data=df, x='age', hue='TenYearCHD', kde=True)  plt.title('Distribusi Umur dengan CHD')  plt.xlabel('Umur')  plt.ylabel('Frekuensi')  st.pyplot()  # Menambahkan penjelasan plot histogram  st.write('\*\*Plot Histogram: Distribusi Umur dengan CHD\*\*')  st.write('Plot histogram di atas menunjukkan distribusi umur pasien dalam dataset, dengan pemisahan warna berdasarkan keberadaan risiko penyakit kardiovaskular (CHD - TenYearCHD). Histogram memberikan gambaran visual tentang sebaran umur pasien, dengan kurva kernel density estimation (KDE) yang menunjukkan perkiraan kepadatan distribusi.')  st.write('---')  # Menghitung jumlah prediksi benar dan salah  true\_positive = sum((y\_pred == 1) & (y\_test == 1))  true\_negative = sum((y\_pred == 0) & (y\_test == 0))  false\_positive = sum((y\_pred == 1) & (y\_test == 0))  false\_negative = sum((y\_pred == 0) & (y\_test == 1))  # Menghitung akurasi dan presisi dalam bentuk persentase  accuracy\_percent = accuracy \* 100  precision\_percent = precision \* 100  # Menampilkan akurasi dan presisi keberhasilan dalam persentase  st.write('\*\*Keberhasilan Model:\*\*')  st.write(f'Akurasi : {accuracy\_percent:.2f}%')  st.write(f'Presisi : {precision\_percent:.2f}%')  st.write('---') |

Kodingan di atas adalah contoh implementasi penggunaan algoritma Gaussian Naive Bayes untuk klasifikasi penyakit liver menggunakan dataset framingham.csv. Berikut ini adalah penjelasan singkat dan jelas dari kodingan tersebut:

1. Import Library: Dilakukan import beberapa library yang diperlukan, yaitu Streamlit, pandas, matplotlib, seaborn, serta modul-modul dari scikit-learn untuk pembagian data, Gaussian Naive Bayes, dan pengukuran kinerja.

2. Membaca Dataset: Dataset framingham.csv dibaca menggunakan pandas dan disimpan dalam DataFrame.

3. Data Preprocessing: Dilakukan pembersihan data dengan menghapus baris yang memiliki nilai yang hilang.

4. Memisahkan Fitur dan Target: Fitur-fitur yang akan digunakan dalam model dan target (kolom TenYearCHD) dipisahkan dari DataFrame.

5. Pembagian Data: Data dibagi menjadi data latih (80%) dan data uji (20%) menggunakan train\_test\_split dari scikit-learn.

6. Membuat Model Gaussian Naive Bayes: Objek Gaussian Naive Bayes dibuat menggunakan GaussianNB() dari scikit-learn.

7. Melatih Model: Model Gaussian Naive Bayes dilatih menggunakan data latih.

8. Melakukan Prediksi: Model yang sudah dilatih digunakan untuk memprediksi nilai target pada data uji.

9. Mengukur Kinerja Model: Menghitung nilai Akurasi, Presisi, Recall, dan F-measure menggunakan metrik-metrik dari scikit-learn.

10. Tampilan Hasil Evaluasi Model: Hasil evaluasi model ditampilkan dalam tampilan Streamlit yang menarik dan informatif, termasuk nilai Akurasi, Presisi, Recall, dan F-measure.

11. Visualisasi Histogram: Menggunakan library seaborn dan matplotlib, histogram untuk kolom 'age' dengan pemisahan warna berdasarkan keberadaan risiko penyakit kardiovaskular (CHD) ditampilkan.

12. Penjelasan Hasil Pembahasan: Penjelasan singkat tentang pengertian dan pentingnya metrik evaluasi model (Akurasi, Presisi, Recall, F-measure) disajikan.

13. Keberhasilan Model: Persentase akurasi dan presisi keberhasilan model ditampilkan.

Kodingan ini memberikan tampilan yang interaktif dan informatif melalui Streamlit, serta menyajikan hasil evaluasi model dan visualisasi histogram yang dapat membantu dalam pemahaman dan analisis kinerja model klasifikasi penyakit liver.

|  |
| --- |
| streamlit run app.py |

Kodingan diatas untuk menjalankan aplikasi streamlit. Setelah menjalankan perintah di atas, Streamlit akan membuka halaman web lokal. Anda dapat melihat tampilan hasil evaluasi model dengan mengakses URL yang ditampilkan di command line atau terminal.

SPenyajian tabel juga dapat ditambahkan di dalam laporan sebagai berikut:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Accuracy | Precision | Recall | F-measure |
| 1. | 0,801913 | 0,328358 | 0,180328 | 0,232804 |

Keberhasilan Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Akurasi | Presisi |  |  |
| 1. | 80.19% | 0,328358 |  |  |

Tabel 1. Hasil Performa Klasifikasi

**BAB V: KESIMPULAN**

Kesimpulannya, penelitian ini menggunakan metode Gaussian Naive Bayes untuk klasifikasi penyakit liver menggunakan dataset Framingham. Metode ini merupakan salah satu metode klasifikasi yang populer dan efektif dalam analisis data. Tujuan dari penelitian ini adalah untuk mengidentifikasi dan memprediksi penyakit liver berdasarkan beberapa atribut atau fitur yang relevan.

Dalam penelitian ini, data yang digunakan adalah data pasien dengan atribut seperti usia, jenis kelamin, jumlah bilirubin, enzim hati, albumin, dan lain-lain. Data ini digunakan untuk melatih model Gaussian Naive Bayes agar dapat mengklasifikasikan pasien sebagai memiliki penyakit liver atau tidak. Hasil penelitian ini diperoleh skor akhir yaitu akurasi sebesar 80.19% dan presisi sebesar 0,328358 menunjukkan bahwa metode Gaussian Naive Bayes dapat memberikan klasifikasi yang cukup akurat dalam mengenali penyakit liver.

# **DAFTAR PUSTAKA**

# [1] "Sentiment Analysis in Social Networks" oleh Federico Alberto Pozzi, "Political Campaigning, Elections, and the Internet" oleh Darren G. Lilleker, atau "Public Opinion and American Democracy" oleh Susan Herbst. (2016).

[2] Fuzzy Cluster Means System for the Diagnosis of Liver Diseases. Victor E. Ekong University of Uyo Semtember (2011). Journal of Computer Science and Technology Volume 2(Issue 3):Page 205-209.

[3] Liver Disease Diagnosis Based on Neural Networks. Conference: Proceedings of the 16th International Conference on Neural Networks (NN'15),At: Rome, November (2015) ItalyAuthors:Ebenezer OlaniyiMississippi State University and Adnan Khashman.

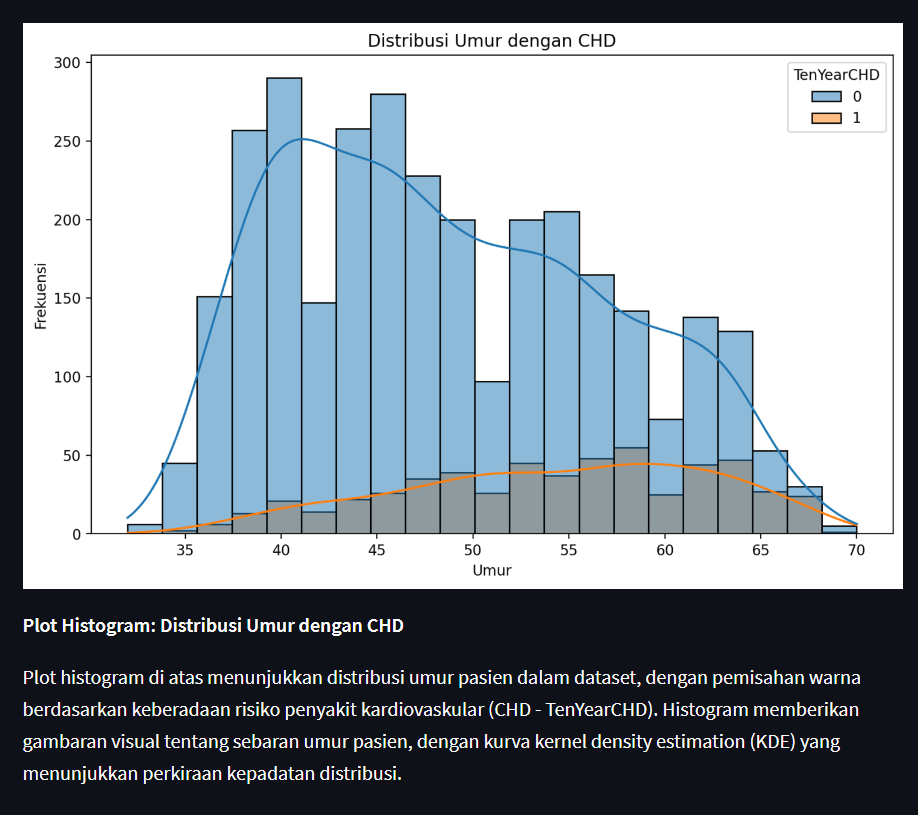
[4] MODEL SUPPORT VECTOR MACHINE BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK PREDIKSI PENYAKIT LIVER Nu’man Musyaffa 1), Bakhtiar Rifai 2). VOL. 3. NO. 2 FEBRUARI (2018)

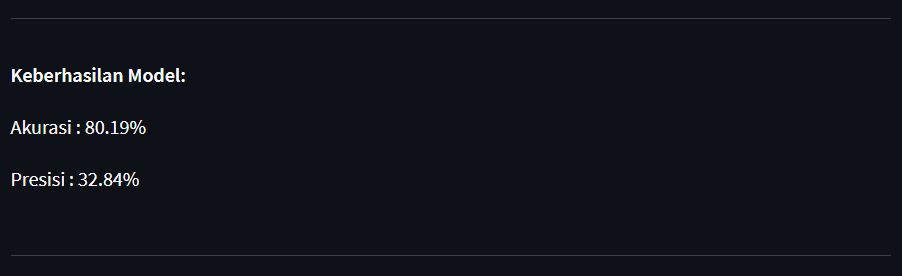
[5] Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision TreeEndah Patimah1, Ballya Vicky Haekal2, Desta Sandya Prasvita, S.Komp., M.Kom.3. Jakarta-Indonesia, 20 April (2021)

# 

# **4.** **LAMPIRAN**







|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | | | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | |  | |  |  |  |  |  |  |  |  |
|  |  |  | |  | |  | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | | | | | | | | |  | | | |
|  |  | | | |  | | | |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |  |
|  |  | |  |  | |  | |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | | | | |  | | --- | |  | |
|  | |  | | --- | |  | | | |  |
|  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  | |  |
|  |  |  |
|  |  |  |