Klasifikasi Penyakit Liver Menggunakan Metode Elbow Untuk Menentukan K Optimal pada Algoritma K-Nearest Neighbor (K-NN)

Subtitle as needed ***(paper subtitle)***

Ihya’ Nashirudin Abrar[1]\*, Asrul Abdullah[2], Sucipto[3]

Program Studi Teknik Informatika[1], [2], [3]

Universitas Muhammadiyah Pontianak

Pontianak, Indonesia

Emai: 181220082@unmuhpnk.ac.id[1], asrul.abdullah@unmuhpnk.ac.id[2], sucipto@unmuhpnk.ac.id[3]

***Abstract*—** **In the world of health diagnosing the liver disease is not easy, but several medical records record the patient's symptoms and, the diagnosis is very useful for researchers because existing medical records can be used to help diagnose liver disease, by carrying out data mining methods where data mining can be used to examine the additional value of a collection of information as information that cannot be seen physically by using the K-Nearest Neighbor (K-NN) calculation. This method works more effectively than the other methods because this method tries to classify new information from unclear classes by choosing the k amount that is closest to the new information. Most of the k classes that accompany it are selected as the hope class for new information and use the Elbow Method in determining the error rate in determining the optimal k value, the aim is to get the best level of accuracy from the K-Nearest Neighbor (K-NN) method, the test results show the 3rd test is the best with a training accuracy of 80.5% testing 78.9% and accuracy after tuning on the training data 82.2% and testing 77.1% without balancing the data because after balancing the data it becomes overfitting it could be because the resampling data does not match the actual, data so the model is too complex in training the amount of noise which results in overfitting.**

***Keywords—*** ***Classification, Machine Learning, Elbow Method, KNN, Liver.***

***Abstrak*—Dalam dunia kesehatan mendiagnosis penyakit hati tidaklah mudah, namun ada beberapa rekam medis yang mencatat gejala penyakit pasien dan diagnosisnya ini sangat bermanfaat bagi peneliti karena rekam medis yang ada dapat digunakan untuk membantu mendiagnosa penyakit hati, dengan melakukan metode data mining yang mana data mining dapat digunakan untuk memeriksa nilai tambahan dari kumpulan informasi sebagai informasi yang tidak dapat dilihat secara fisik dengan menggunkan perhitungan K-Nearest Neighbor (K-NN). Metode ini bekerja lebih efektif dibandingkan dengan metode lain, karena metode ini mencoba untuk mengelompokkan informasi baru dari kelas yang tidak jelas dengan memilih jumlah k yang paling dekat dengan informasi baru. Sebagian besar dari kelas k yang menyertainya dipilih sebagai kelas harapan untuk informasi baru dan menggunkan Elbow Method dalam menentukan error rate dalam menentukan nilai k optimal, tujuanya adalah untuk mendapatkan tingkat akurasi terbaik dari metode K-Nearest Neighbor (K-NN), hasil pengujian menunjukkan pengujian ke-3 menjadi yang terbaik dengan akurasi training 80,5% testing 78,9% dan akurasi setelah Tuning pada data training 82,2% dan testing 77,1% tanpa melakukan penyeimbangan data karena setelah melakukan penyeimbangan data menjadi overfitting bisa jadi karena data resampling tidak sesuai dengan data sesunguhnya sehingga model terlalu kompleks dalam melatih jumlah noise yang mengakibatkan overfitting.**

***Kata Kunci— Klasifikasi, Pembelajaran Mesin, Metode Siku, KNN, Hati.***

# Pendahuluan

Fungsi organ hati bagi tubuh manusia berperan sangatlah penting, yaitu menghilangkan racun dalam darah hingga berfungsi untuk membantu proses pencernaan, jika organ hati ini tidak dapat berfungsi dengan baik dapat menyebabkan masalah serius pada tubuh, Penyakit liver ini menjadi salah satu penyakit yang mematikan di dunia. Organisasi Kesehatan Dunia (WHO) menyebut bahwa 1,34 juta kematian yang terjadi pada 2015 disebabkan penyakit hati. oleh karena itu prediksi akurat penyakit sangat penting dalam industri kesehatan dan medis. Untuk mengambil keputusan penting yang efektif dalam menganalisis dan memprediksi penyakit yang diderita. [1]

Dalam dunia perawatan kesehatan saat ini, mendiagnosis penyakit hati sangatlah sulit. Namun, ada beberapa rekam medis yang mencatat gejala penyakit pasien dan diagnosisnya sebagai penyakit hati. Hal-hal seperti ini tentunya sangat bermanfaat bagi peneliti. Rekam medis yang ada dapat digunakan untuk membantu mendiagnosa penyakit pasien [2], Hasil tes fungsi hati dapat dijadikan acuan untuk mendiagnosis ada tidaknya penyakit hati. Tes-tes ini termasuk serum transaminase, alkaline phosphatase, bilirubin total, bilirubin terkonjugasi, protein total, albumin, dan rasio albumin terhadap globulin.

Salah satu strategi yang dapat digunakan untuk mengantisipasi penyakit hati adalah dengan memanfaatkan data mining. Data mining adalah perkembangan siklus yang digunakan untuk memeriksa nilai tambahan dari kumpulan informasi sebagai informasi yang tidak dapat dilihat secara fisik [3], Pada skripsi ini, penulis menggunakan perhitungan K-Nearest Neighbor (K-NN). Strategi urutan informasi bekerja agak efektif dibandingkan dengan teknik karakterisasi informasi yang berbeda. Perhitungan ini mencoba untuk mengelompokkan informasi baru dari kelas yang tidak jelas memilih k yang paling dekat dengan informasi baru.

Sebagian besar dari kelas informasi k yang menyertainya dipilih sebagai kelas harapan untuk informasi baru. k biasanya ditentukan sebagai angka ganjil untuk mencoba tidak menunjukkan jumlah jarak yang sama dalam sistem pengelompokan untuk informasi baru dan menggunkan Elbow Method dalam menentukan error rate dalam menentukan nilai k optimal, tujuanya adalah untuk mendapatkan tingkat akurasi terbaik dari metode K-Nearest Neighbor (K-NN).

Keakuratan metode K-Nearest Neighbor (K-NN) dalam mendiagnosa sebuah penyakit telah teruji dari penelitian terdahulu yang menggunakan metode ini dalam memprediksi atau mengklasifikasi sebuah penyakit. Seperti pada hasil penelitian yang telah dilakukan oleh M. Syukri Mustafa, dan I Wayan Simpen dalam memprediksi pasien yang terjangkit penyakit diabetes dan mendapatkan hasil akurasi terbaik sebesar 93,33% dengan error rate 6,67% [4].

Berdasarkan permasalahan yang dapat ditemukan diatas, penulis tertarik dalam membangun sebuah apliaksi menggunakan metode K-Nearest Neighbor (K-NN) dan metode Elbow untuk menentukan K optimal untuk pemodelan klasifikasi penyakit liver dengan menggunakan data acuan test fungsi hati dari data sekunder yang diambil dari UCI Machine Learning Repository. Data yang digunakan yaitu ILDP (Indian Liver Patient Dataset).

# Tinjauan Pustaka

## Penelitian Terdahulu

Dari penelitian Popon Handayan (2019) dalam jurnalnya Liver Disease Prediction Using Decision Tree and Neural Network Methods dengan memperoleh skor yang dibuat oleh Confusion Matrix dan ROC Curve, didapatkan hasil bahwa Date Test yang dilakukan oleh C4.5 - algoritma memberikan skor yang lebih baik, dimana akurasi model algoritma C4.5 adalah 75,56% dengan nilai AUC 0,898 dan skor akurasi algoritma neural network adalah 74,17% dengan nilai AUC 0,671. Berdasarkan nilai tersebut didapatkan selisih akurasi sebesar 1,39n, selisih AUC sebesar 0,227% [1].

Sedangkan dari penelitian Citra Nurina Prabiantissa (2021) dari jurnalnya yang berjudul “Klasifikasi pada Dataset Penyakit Hati Menggunakan Algoritma Support Vector Machine, K-NN, dan Naïve Bayes” dengan hasil penelitian menunjukkan bahwa dari ketiga algoritma, SVM memiliki rata-rata performa paling baik dengan akurasi sebesar 82,36%.

Dari beberapa penelitian diatas dapat diambil kesamaan pada klasifikasi penyakit liver oleh karena itu saya mencoba mencari metode yang lebih baik dengan menyempurnakan algoritma K-Nearest Neighbor (K-NN) menambahkan metode elbow (siku) agar mendapatkan hasil k yang lebih optimal dan pada pengujian bisa lebih focus dalam pengolahan data bukan pencarian k optimal.

## Klasifikasi

Klasifikasi adalah proses pencarian pola yang menjelaskan atau memisahkan konsep atau kelas informasi untuk memperkirakan kelas objek yang tidak diketahui. Dalam klasifikasi ini, record disebut training set, yang terdiri dari beberapa properti, properti tersebut dapat berupa continuous atau categorical, salah satu properti menunjukkan kelas dari record tersebut. [5].

## Data Mining

Data mining adalah proses menganalisis dan mengekstrak pengetahuan secara otomatis menggunakan satu atau lebih teknik dalam mempelajari computer [6], Data mining adalah istilah yang sering digunakan untuk menggambarkan informasi kumpulan data. Data mining adalah metode yang terlibat dengan mengekstraksi dan mengenali data bermanfaat dan informasi terkait dari kumpulan data besar yang menggunakan berbagai prosedur terukur, numerik, terkomputerisasi, dan AI (Artificial Intelligence) [7].

## Knowledge Discovery in Database (KDD)

KDD adalah salah satu metode paling populer yang berfokus pada penemuan umum pengetahuan atau informasi dari data, termasuk proses penyimpanan dan penggunaan data, algoritma yang efektif dan efisien untuk pemrosesan data besar, interpretasi dan visualisasi data. [8].

Knowledge Discovery in Database (KDD) adalah proses mengidentifikasi informasi dan pola yang berguna dalam informasi. Informasi ini terkandung dalam database besar yang sebelumnya tidak diketahui dan berpotensi berguna. Penambangan data adalah salah satu langkah dalam rangkaian proses berulang di KDD [9].

## K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) adalah cara untuk mengklasifikasikan objek berdasarkan data pelatihan terdekat (tetangga). Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean. diperlukan suatu sistem klasifikasi sebagai suatu sistem yang dapat mencari informasi [8], adapaun rumus jarak atau Euclidean sebagai berikut.

(1)

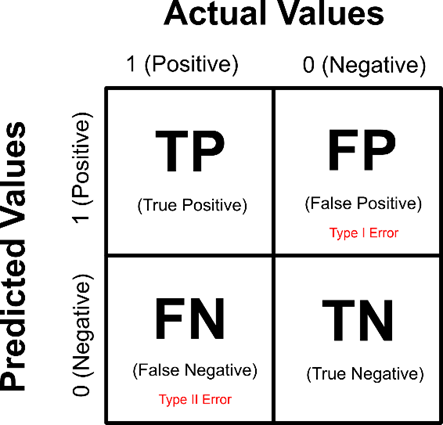
Dimana a = a1,a2, .. , an dan b = b1, b2, .., bn menyatakan nilai n atribut dari dua register [9]. Untuk atribut dengan nilai kelas. Skor diprediksi untuk suatu jenis berdasarkan peringkat tertinggi dari tetangga sekitarnya.

## Elbow Method

Elbow method atau metode siku digunakan untuk memilih jumlah cluster atau kelompok yang optimal. Algoritma siku digunakan untuk menentukan jumlah kelompok yang akan dibentuk [10], Metode elbow diimplementasikan dengan cara menentukan data optimal dan melihat grafik dari nilai k yang disematkan [11].

## Confusion Matrix

Confusion matrix memberikan keputusan yang dibuat selama pelatihan dan pengujian dan juga Confusion Matrix memberikan perkiraan kinerja klasifikasi berdasarkan apakah objek tersebut benar atau salah. [1]. Confusion matrix adalah tabel yang memberikan klasifikasi kumpulan data uji yang benar dan kumpulan data uji yang salah. Contoh matriks kebingungan klasifikasi biner [12].



Gambar 1. Confusion Matrix

1. True Positive (TP): TP adalah data positif yang diprediksi benar.
2. True Negative (TN): TN adalah data negatif yang diprediksi benar.
3. False Positive (FP) — Type I Error: FP adalah data negatif yang diprediksi sebagai data positif.
4. False Negative (FN) — Type II Error: FN adalah data positif yang diprediksi sebagai data negatif.

# Metode Penelitian

## Tinjauan Pustaka

Pada tahap ini akan dilakukan pencarian terkait penelitian ini. Tinjauan pustaka dilakukan untuk mengumpulkan bahan referensi, tinjauan yang digunakan dapat berupa jurnal ilmiah terdahulu, buku dan bahan-bahan yang dapat digunakan untuk mendukung penyelesaian penelitian disertasi.

## Mendefinisikan Masalah

Permasalahan dalam penelitian inia dalah bagaimana cara mengatasi sulitnya mendiaknosa penyakit hati kecuali dilakuakan oleh ahli dalam bidang medis atau dokter, oleh karena dalam penelitian ini penulis membangun sebuah aplikasi yang dapat mengklasifikasi untuk memprediksi penyakit hati pada bidang Machine Learning sebagai process. Masalah ini memiliki output berupa binary class dengan nilai Disease (1) dan No Disease (2).

## Pengumpulan Data

Pada tahapan ini bertujuan untuk mendapatkan data yang baik dengan mencari dataset yang sesuai dengan penelitian karena sulitnya mendapatkan data dari Indonesia karena kode etik medis yang mengharuskan kerahasiaan pasien jadi penelitian ini menggunakan dataset yang memang digunakan untuk penelitian yaitu dataset ILPD (Indian Liver Patient Dataset).

1. Fitur dan Keterangan

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | ***Fitur*** | ***Keterangan*** | ***Jenis Data*** |
| 1 | Age | Umur Pasien | Numerikal |
| 2 | Gender | Jenis Kelamin Pasien | Kategorikal |
| 3 | TB (Total Bilirubin) | Pigmen berwarna kuning kecoklatan yang ada didalam empedu, darah dan tinja. | Numerikal |
| 4 | DB (Direct Bilibruin) | Pigmen berwarna jingga kuning sisa dari perombakan sel darah merah (langsung) | Numerikal |
| 5 | Alkaline (Alkaline Phosphatase) | Enzim hidrolase yang terutama ditemukan pada sebagian besar organ tubuh, terutama di tulang, tulang dan plasenta. | Numerikal |
| 6 | Alamine (Alnine Aminotransferase) | Enzim yang sering dijumpai di serum darah dan berbagai jaringan tubuh, tetapi sering dikaitkan dengan kerusakan hati | Numerikal |
| 7 | Aspartate (Aspartate Aminotransferase) | Enzim yang berkaitan dengan kinerja organ hati | Numerikal |
| 8 | TP (Total Proteins) | Berisi Albumin dan Globulin | Numerikal |
| 9 | ALB (Albumin) | Protein Utama pada darah yang diproduksi hati | Numerikal |
| 10 | A/G (Ratio Albumin Globulin Ratio) | Perbandingan albumin dan globulin yang merupakan konstituen utama protein yang ditemukan dalam darah | Numerikal |
| 11 | Class | Menderita liver/ tidak menderita liver | Kategorikal |

## Proses Data Mining

Diagram

Description automatically generated

Gambar 2. Diagram Alir Data Mining

1. Exploratory Data Analysis (EDA): Langkah pertama yang dialakukan setelah mendapatkan dataset yaitu pengolahan data mining sesuai dengan metode yang digunakan pada penelitian ini yaitu Knowledge Discovery in Database (KDD) dengan diawali dengan EDA (Exploratory Data Analysis) yaitu tahap memahami data untuk menganalisis setiap fitur yang ada dalam dataset EDA memiliki beberapa tipe yaitu Univariate Analysis, Bivariate Analysis, dan Multivariate Analysis.
2. Preprocessing: Preprocessing yaitu menyiapkan dataset sebelum dilatih oleh model untuk mendapatkan hasil atau akurasi yang maksimal, dengam menangani Missing Value merubahnya menjadi nilai median, lalu handling duplicate data dengan menghapus data tersebut, setelah itu melakukan penanganan outlier dengan menggunkan perhitungan IQR, penaganan data Imbalance dengan library SMOTE dan Tomek, dan yang terakhir yaitu Label Encoding dan Resampling data.
3. Permodelan KNN: Pada penelitian ini pemodelan diawali dengan mencari nilai k terbaik menggunakan elbow method setelah didapatkan k optimal kemudian dilakukan permodelan KNN dan dilanjutkan dengan peningkatan atau Tuning menggunakan GridSearchCV.
4. Evaluation Model: Setelah mendapatkan hasil akurasi dari pemodelan KNN peneliti mengevaluasi hasil tersebut apakah hasil permodelan sesuai dengan keinginan peneliti.

## Pengujian

Pengujian pada penelitian ini menggunkaan *Confusion Matrix* dan Skor AUC, dilakukan pada beberapa percobaan:

1. Percobaan 1: Pada pengujian pertama ini hanya dilakukan preprocessing, splitting data, elbow method knn, dan pemodelan knn lalu pengecekan akurasi.
2. Percobaan 2: Pada pengujian kedua ini dilakukan penanganan outlier, preprocessing, splitting data, elbow method knn, dan pemodelan knn lalu pengecekan akurasi.
3. Percobaan 3: Pada pengujian ketiga ini dilakukan penanganan outlier, preprocessing, splitting data, normalisasi data, elbow method knn, dan pemodelan knn lalu pengecekan akurasi.
4. Percobaan 4: Pada pengujian keempat ini dilakukan penanganan outlier, preprocessing, splitting data, normalisasi data, menangani data tidak seimbang, elbow method knn, dan pemodelan knn lalu pengecekan akurasi

# Hasil dan Pembahasan

## Hasil Explore Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah proses memeriksa dan menganalisis data secara keseluruhan. Tujuannya adalah untuk menemukan pola, anomali, menguji hipotesis, dan memverifikasi asumsi. Dengan EDA, peneliti dapat menemukan kesalahan lebih awal, mengidentifikasi outlier, menentukan hubungan antar data, dan menemukan faktor penting dari data tersebut. Proses ini sangat berguna untuk analisis statistik.

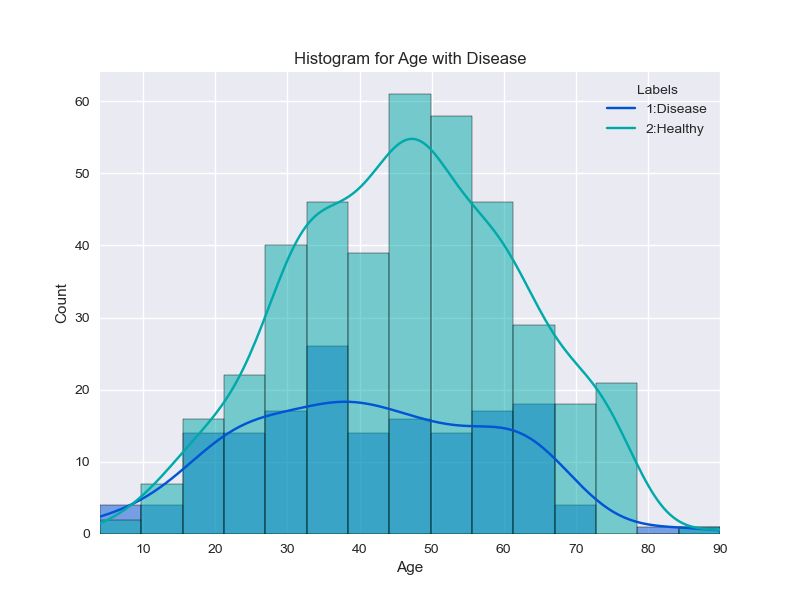
### Univariate Analysis: Exploratory Data Analysis yang pertama adalah univariate analysis yaitu membaca satu set variable, Tujuan dari analisis univariat adalah untuk mendapatkan data, mendefinisikan, meringkas, dan menganalisis pola yang ada di dalamnya. Data dibawah menjelaskan bahwa lebih banyak data pria daripada wanita di dalam dataset yaitu pria berjumlah 430 sedangkan wanita 140.

**Chart, bar chart

Description automatically generated**

Gambar 3. Univariate Analysis Gender and Desiase

### Bivariate Analysis: Dibawah ini adalah contoh dari bivariate analysis yaitu gambar salah satu plot fitur age dengan dataset, dari data tersebut dapat diambil kesimpulan bahwa orang mulai terkena penyakit liver pada usia 25-50 tahun.

****

Gambar 4. Bivariate Analysis Age and Disease

* + 1. Multivariates Analysis: Analysis multivariat adalah bentuk teknik analisis statistik yang lebih kompleks dan digunakan ketika ada lebih dari dua variabel dalam kumpulan data, Dari gambar dibawah kita bisa melihat korelasi antara setiap fitur rata-rata kurang bagus bisa terlihat dari nilai yang mendekati 1 dan -1 dan dari warna yang mana semakin pucat berarti korelasi tidak baik.

**Calendar

Description automatically generated with low confidence**

Gambar 5. Multivariate Analysis Setiap Fiture

## Hasil Preprocessing

Pada tahapan ini akan dilakukan langkah setelah tahapan EDA yaitu preprocessing, preprocessing adalah proses untuk mempersiapkan data sebelum dilatih oleh model biasa disebut juga Data Cleaning.

### Missing value: Menurut definisinya, Missing Value adalah ketiadaan data pada suatu entri atau observasi. Dalam dunia data science, Missing Value sangat penting dalam proses perselisihan data (data wrangling) sebelum dilakukan analisis dan prediksi data. Data wrangling merupakan proses untuk menyederhanakan data atau membersihkan data dari data yang tidak berguna sehingga data siap digunakan dalam analisis.

|  |
| --- |
| Age 0  Gender 0  Total\_Bilirubin 0  Direct\_Bilirubin 0  Alkaline\_Phosphotase 0  Alamine\_Aminotransferase 0  Aspartate\_Aminotransferase 0  Total\_Protiens 0  Albumin 0  Albumin\_and\_Globulin\_Ratio 4  Dataset 0 |

Gambar 6. Hasil Pengecekan Missing Value

Gambar diatas adalah daftar kategorik mana saja yang terdapat Missing value, setelah melakukan analisis terdapat missing value hanya ada pada data Albumin and Globulin Ratio, kemudian dilakukan penanganan bisa dengan meng drop atau menghapus akan tetapi pada penelitian ini penulis merubah data yang hilang menjadi median.

### Duplicate Data: Seperti namanya, Duplicate data adalah data yang serupa atau kumpulan data dapat terdiri dari beberapa objek data (duplikat). Selama pemrosesan, hampir selalu ada tumpang tindih antara data. Pada penelitian ini pengolahan dilakukan dengan membuang data duplikat.

### Outlier: Outlier adalah titik data yang nilainya signifikan berbeda dari nilai-nilai pada populasi tertentu. Walaupun definisi ini tampak sederhana, menentukan titik data yang merupakan outlier sebenarnya cukup subjektif dan tergantung pada studi dan jumlah data yang tersediacara melihat data outlier bisa dengan box plot.

**Box and whisker chart

Description automatically generated**

Gambar 7. Box Plot Pengecekan Outlier

Diatas adalah box plot untuk melihat ada tidaknya data outlier, bisa dilihat bahwa banyak data outlier di setiap fitur sehingga penulis melakukan penanganan dengan merubah data outlier menjadi nilai mean, dari beberapa kali peneliti melakukan uji coba pembatasan data mungkin bukan solusi yang baik dalam masalah ini untuk menangani outlier, karena TB, DB dan parameter tersebut sangat signifikan untuk deteksi atau prediksi Penyakit Hati, karena batas waktu dan kurangnya pengetahuan domain di lapangan, peneliti ingin melanjutkan lebih jauh dengan membatasi outlier dengan nilai mean.

**Box and whisker chart

Description automatically generated**

Gambar 8. Box Plot Pengecekan Outlier Setelah Dilakukan Penanganan

Diatas adalah *plot box* pengecekan *outlier* setelah dilakukan penanganan menggunakan nilai mean meskipun tidak banyak yang berubah karena takut merusak hasil prediksi akan tetapi diharapkan dapat sedikit meningkatkan akurasi sebelum data di latih.

### Imbalance Data: Imbalanced Data adalah masalah umum dalam klasifikasi pembelajaran mesin, di mana ada hubungan yang tidak proporsional antar kelas. Itu ditemukan di banyak bidang, termasuk diagnosis medis.

**Chart, bar chart

Description automatically generated**

Gambar 9. Data Imbalance

Diatas adalah plot persebaran data y label atau dataset yang berisi data liver dan not liver dari data diatas bisa dilihat bahwa ada ketidak seimbangan data antara 1 dan 2 yang mana lebih banyak data yang terjangkit daripada yang tidak.

**Chart, bar chart

Description automatically generated**

Gambar 10. Data Balance

Ditampilkan plot distribusi data y-label setelah pemrosesan data yang tidak seimbang dengan SMOTE Tomek. SMOTE adalah metode oversampling yang digunakan untuk menyeimbangkan data dengan membuat representasi komposit dari kelas minoritas. Sedangkan Tomek Links adalah metode subsampling yang digunakan untuk menghapus data dari kelas mayoritas dengan karakteristik yang mirip. Namun, Tautan Tomek hanya menghapus instans yang didefinisikan sebagai "Tautan Tomek", sehingga data yang dianalisis tidak seimbang. Oleh karena itu, metode ini dapat dikombinasikan dengan metode lain untuk meningkatkan kinerja. Dalam pekerjaan ini, metode gabungan SMOTE dan Tomek Links digunakan untuk menyeimbangkan data yang tidak seimbang dan diterapkan pada tiga material menggunakan metode klasifikasi JST. Hasil analisis menunjukkan bahwa penerapan kombinasi metode SMOTE dan Tomek-Links memberikan kinerja yang lebih baik dibandingkan dengan metode SMOTE dan metode Tomek-Links saja untuk analisis klasifikasi KNN.

### Label Encoding: Machine learning dan deep learning membutuhkan semua variabel input dan output menjadi numerik. Oleh karena itu, jika data yang digunakan berupa kategori, hal pertama yang harus dilakukan adalah merubah tipe data tersebut menjadi numerik. Proses ini biasa disebut dengan encoding atau encode. Setelah tahap encoding selesai dilakukan, tahap selanjutnya bisa dilanjutkan ke proses modeling (model fitting dan evaluating).

**Chart, waterfall chart, treemap chart

Description automatically generated**

Gambar 11. Data Distribution

Diatas adalah plot persebaran data distribusi dan dapat dilihat ada data object yaitu pada data Gender atau jenis kelamin oleh karena itu peneliti menggunakan label encoding library Python dengan scikit-learn data yang ada akan otomatis diurutkan berdasarkan abjad alfabet agar data dapat diproses saat melakukan modeling data.

### Normalisasi: Normalisasi data adalah proses penting dalam penambangan data untuk memastikan konsistensi catatan dalam kumpulan data. Proses ini melibatkan transformasi data, atau mengubah data asli menjadi bentuk yang memungkinkan pemrosesan data yang efisien. Tujuan utama normalisasi data adalah untuk menghilangkan redundansi data (pengulangan) dan standarisasi data untuk memastikan aliran data yang lebih baik. Normalisasi data digunakan untuk menskalakan karakteristik data ke rentang yang lebih kecil, mis. misalnya -1 ke 1 atau 0 ke 1. Ini umumnya berguna untuk algoritma klasifikasi. Pada penelitian ini, beberapa percobaan dengan metode normalisasi yang berbeda dilakukan untuk menemukan metode normalisasi yang paling akurat. Namun, peneliti menggunakan MinMaxScaler. Metode normalisasi min-max mengubah kumpulan data dalam skala dari 0 (min) menjadi 1 (max) seperti yang ditunjukkan pada contoh gambar di atas.

## Hasil KNN Modeling

Setelah data melewati proses preprocessing data kemudian dilatih untuk bisa digunakan untuk memprediksi menggunakan metode K-Nearest Neighbor (K-NN).

1. *Elbow method:* Langkah terpenting dalam *machine learning* yang menggunakan metode *K-Nearest Neighbor* (K-NN) adalah menentukan nilai K yang optimal. Dengan kata lain, berapa banyak *cluster* yang harus dibagi menjadi data nilai optimal untuk k mengurangi efek *noise* pada klasifikasi, tetapi membuat batas antar kelas kurang jelas. *Metode Elbow* membantu data scientist untuk memilih jumlah *cluster* yang optimal untuk clustering KNN. Ini adalah salah satu metode paling populer untuk menentukan nilai optimal K.

**Chart, line chart

Description automatically generated**

Gambar 12. Elbow Method

Pada penelitian ini pemodelan diawali dengan mencari nilai k terbaik menggunakan *elbow method*, pada gambar diatas dapat diambil kesimpulan k terbaik adalah 4 karena setelah k 4 *error rate* meningkat, peneliti juga mencoba mengetes k optimal dengan me-*looping* setiap k dari 1 sampai 45 menggunakan nilai akurasi data *training* dan *testing*, seperti contoh gambar di bawah.

**Chart

Description automatically generated**

Gambar 13. Hasil mencari K optimal

Dari hasil me-*looping* setiap k 1 sampai 45 didapat hasil k optimal di K = 4 dengan *Best accuracy* 0.7894736842105263, dengan ini peneliti memutuskan bahwa k optimal untuk penelitian ini adalah k 4.

1. Pemodelan *K-Nearest Neighbor* (K-NN): Setelah berhasil menentukan k optimal langkah selanjutnya yaitu pemodelan metode KNN menggunakan K = 4 dan mendapatkan tingkat accuracy seperti tampilan dibawah.

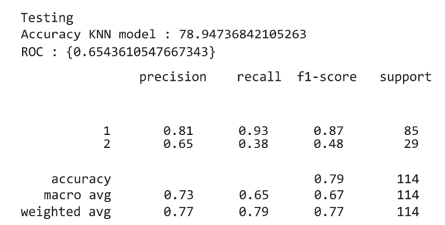
A screenshot of a computer

Description automatically generated with low confidence

Chart, treemap chart

Description automatically generated

Gambar 14. Confusion matrix Training



Chart, treemap chart

Description automatically generated

Gambar 15. Confusion matrix Testing

1. *Hyperparameter Tuning* dan *GridSearchCV*: Setelah mendapatkan hasil akurasi dari pemodelan KNN peneliti juga mencoba menggunakan *Hyperparameter Tuning* untuk melihat apakah akurasi akan meningkat atau tidak, Hasil *grid search best param*s menggunakan jarak *n\_neighboars* 1 sampai 21 dan kembali lagi ke 1, kemudian menggunakan *K-Fold* 10 untuk mencari baris terbaik untuk K lalu melakukan pengulangan sebanyak 3 kali dan mendapatkan hasil seperti digambar yaitu *best* *params* untuk menghitung jarak menggunakan perhitungan *metric manhattan* dan k atau *n\_neighbors* terbaik adalah 2 dan *weights* terbaik adalah *uniform.*

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

Gambar 16. Confusion matrix Training Tuning

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

Gambar 17. Confusion matrix Testing Tuning

Diatas adalah hasil akurasi setelah melakukan *hyperparameter tuning* menggunakan *GridSearchCV* dari hasil evaluasi mendapatkan hasil peningkatan di *training* akan tetapi mengalami penurunan saat memprediksi data *testing*.

## Hasil Pengujian

1. Table Pengujian

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **Pengujian** | **Akurasi KNN** | | **Tuning KNN** | |
| **Latih** | **Uji** | **Latih** | **Uji** |
| 1 | Pengujian 1 | 72,6% | 74,6% | 75,2% | 73,7% |
| 2 | Pengujian 2 | 73,5% | 73,7% | 75,0% | 72,8% |
| 3 | Pengujian 3 | **80,5%** | **78,9%** | **82,2%** | **77,1%** |
| 4 | Pengujian 4 | 98,8% | 69,2% | 100% | 62,3% |

Dari data tabel pengujian akurasi dapat dilihat perbandingan akurasi dari setiap pengujian dalam pengujian ketiga memiliki tingkat akurasi yang lebih baik dari keempat percobaan karena akurasi yang dihasilkan seimbang yang mana berperan penting dalam memprediksi sedangkan untuk percobaan keempat saat melakukan penyeimbangan data mengalami peningkatan akurasi di data *training* akan tetapi pada data *testing* mengalami penurunan signifikan bisa diambil kesimpulan bahwa data setelah melakukan penyeimbangan data mengalami data *Overfitting* oleh karena itu peneliti mengambil percobaan ke 3 sebagai *best* akurasi dari penelitian ini.

Overfitting terjadi ketika model terlalu kompleks untuk jumlah noise dari data pelatihan. Solusi yang mungkin bisa dilakukan adalah, dengan menyederhanakan model seperti membatasi fitur yang tidak berkorelasi, mengumpulkan lebih banyak data pelatihan, dan mengurangi noise dalam data pelatihan [13].

# Kesimpulan

Kesimpulan dari penelitian ini adalah hasil optimal dari k menggunakan *elbow method* pada pengujian ketiga didapatkan hasil k = 4 hal ini sangat evektif dalam membantu penulis dalam mencari nilai k optimal.

Dari hasil ujicoba akurasi dapat diambil kesimpulan bahwa data percobaan ketiga yang menjadi akurasi terbaik dengan akurasi *training* 80,5% *testing* 78,9% dan akurasi setelah *Tuning* pada data *training* 82,2% *testing* 77,1%, walaupun pada data percobaan keempat dengan menyeimbangkan data mengalami banyak peningkatan pada data *training* 98,8% dan setelah *tuning* menjadi 100% akan tetapi pada data *testing* mengalami penurunan yang sangat signifikan hingga 69,2% dan setelah *tuning* menjadi 62% bisa diambil kesimpulan bahwa data setelah penyeimbangan data menjadi *Overfitting,* dimana model mencoba mempelajari semua detail, termasuk *noise* pada data, dan mencoba memasukkan semua informasi, sehingga prediksi menjadi tidak baik.

##### References

[1] A. Noviriandini, P. Handayani, and Syahriani, “Prediksi Penyakit Liver Dengan Menggunakan Metode,” *Pros. TAU SNAR-TEK Semin. Nas. Rekayasa dan Teknol.*, no. November, pp. 75–80, 2019.

[2] Rudianto, “Penentuan Penyakit Peradangan Hati Dengan Menggunakan Neural Network Backpropagation,” *Indones. J. Comput. Inf. Technol. Vol 1 No 1*, vol. 1, no. 1, pp. 27–33, 2016.

[3] A. Muzakir and R. A. Wulandari, “Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree,” *Sci. J. Informatics*, vol. 3, no. 1, pp. 19–26, 2016, doi: 10.15294/sji.v3i1.4610.

[4] M. S. Mustafa and I. W. Simpen, “Implementasi Algoritma K-Nearest Neighbor ( KNN ) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba,” *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. VIII, no. 1, pp. 1–10, 2019, [Online]. Available: https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1 -10

[5] G. A. Marcoulides, *Discovering Knowledge in Data: an Introduction to Data Mining*, vol. 100, no. 472. 2005. doi: 10.1198/jasa.2005.s61.

[6] M. M. Hidayat, “Data Mining Data mining,” *Min. Massive Datasets*, vol. 2, no. January 2013, pp. 5–20, 2015, [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book\_part

[7] D. Turban, Efraim ; Aronson, Jay E ; Liang, Ting peng ; Prabantini, “Decision Support Systems And Intelligent Systems : ( Sistem Pendukung Keputusan Dan Sistem Cerdas ) / Efraim Turban,” 2005.

[8] A. Fitria and H. Azis, “Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[9] M. Lestari, “Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung,” *Fakt. Exacta*, vol. 7, no. September 2010, pp. 366–371, 2014.

[10] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.

[11] R. Yuliana Sari, H. Oktavianto, and H. Wahyu Sulistyo, “Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia,” *J. Smart Teknol.*, vol. 3, no. 2, pp. 2774–1702, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[12] D. Putra and A. Wibowo, “Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes,” *Pros. Semin. Nas. Ris. Dan Inf. Sci.*, vol. 2, pp. 84–92, 2020.

[13] A. Géron, *Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow*. 2019.