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

Ihya' Nashirudin Abrar[1], Asrul Abdullah[2], Sucipto[3]
Program Studi Teknik Informatika[1], [2], [3]
Universitas Muhammadiyah Pontianak
Pontianak, Kalimantan Barat, Indonesia
Email: 181220082@unmuhpnk.ac.id[1], asrul.abdullah@unmuhpnk.ac.id[2], sucipto@unmuhpnk.ac.id[3]

***Abstract*— Diagnosing liver disease in the field of healthcare is not an easy task. However, by utilizing medical records as datasets and applying data mining methods such as K-Nearest Neighbor (K-NN), we can analyze and extract knowledge automatically. The K-NN method has proven to be more effective compared to other methods as it clusters new information by selecting the nearest neighbors based on the value of k. In this study, we employed the Elbow method to determine the optimal value of k by measuring the error rate. The test results revealed that the optimal value of k was found to be 4, with the lowest error rate. In the third test, we achieved a training accuracy of 80.5% and a testing accuracy of 78.9%. After fine-tuning the training data, we were able to improve the accuracy to 82.2% for training and 77.1% for testing. However, in the fourth test, we encountered overfitting issues due to data imbalance caused by inappropriate resampling, resulting in a model that was overly complex and prone to excessive noise.**

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

***Abstrak*— Diagnosa penyakit *liver* dalam dunia kesehatan merupakan tugas yang tidak mudah. Namun, dengan menggunakan rekam medis sebagai dataset dan menerapkan metode *data mining* seperti *K-Nearest Neighbor* (K-NN), kita dapat menganalisis dan mengambil pengetahuan secara otomatis. Metode K-NN ini terbukti lebih efektif daripada metode lainnya karena mengelompokkan informasi baru dengan memilih tetangga terdekat berdasarkan nilai *k*. Dalam penelitian ini, kami menggunakan metode *Elbow* untuk menentukan nilai *k* yang optimal dengan mengukur tingkat kesalahan. Hasil pengujian menunjukkan bahwa nilai *k* optimal yang ditemukan adalah 4, dengan tingkat kesalahan yang terendah. Pada pengujian ketiga, kami mencapai akurasi training sebesar 80,5% dan akurasi *testing* sebesar 78,9%. Setelah melakukan penyesuaian pada data *training*, kami berhasil meningkatkan tingkat akurasi menjadi 82,2% untuk training dan 77,1% untuk testing. Namun, pada pengujian keempat setelah melakukan penyeimbangan data, kami mengalami masalah *overfitting* karena proses *resampling* tidak sesuai dengan data asli, sehingga model menjadi terlalu kompleks dan cenderung menghasilkan *noise* yang berlebihan.**

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

I. PENDAHULUAN

Fungsi vital dari organ *liver* dalam tubuh manusia meliputi eliminasi racun dalam darah serta kontribusinya dalam memfasilitasi proses pencernaan. Gangguan fungsi *liver* dapat memiliki konsekuensi serius bagi kesehatan tubuh manusia [1], Penyakit *liver* merupakan salah satu penyakit yang berpotensi mematikan di seluruh dunia. Menurut World Health Organization (WHO), pada tahun 2015, tercatat sebanyak 1,34 juta kematian disebabkan oleh penyakit *liver*. Oleh karena itu, prediksi yang akurat terkait penyakit ini memiliki kepentingan yang besar dalam bidang kesehatan dan kedokteran. Hal ini memungkinkan pengambilan keputusan yang efektif dalam analisis dan prediksi penyakit yang diderita oleh individu [2].

Dalam industri perawatan kesehatan saat ini, diagnosis penyakit *liver* masih menjadi tantangan yang kompleks. Meskipun demikian, terdapat rekam medis yang mencatat gejala penyakit dan diagnosis terkait dengan penyakit *liver*. Informasi yang terdapat dalam rekam medis dapat menjadi sumber yang berharga dalam proses diagnosa penyakit *liver* pada pasien [3], Selain itu, tes fungsi *liver* juga digunakan sebagai acuan dalam mendiagnosis adanya kelainan pada *liver*. Tes-tes ini melibatkan pengukuran *enzim serum transaminase*, *alkaline phosphatase*, total *bilirubin*, *bilirubin* terkonjugasi, total protein, albumin, dan rasio albumin terhadap globulin[4].

Salah satu strategi yang dapat diimplementasikan untuk mengatasi penyakit *liver* adalah melalui penerapan teknik *data mining*. *Data mining* merupakan suatu proses yang digunakan untuk menggali nilai tambahan dari kumpulan informasi yang tidak terlihat secara kasat mata [5], Dalam penelitian ini, penulis menerapkan metode *K-Nearest Neighbor* (K-NN). Pendekatan berbasis urutan informasi ini terbukti lebih efektif dibandingkan dengan teknik karakterisasi informasi lainnya. Metode perhitungan ini bertujuan untuk mengklasifikasikan informasi baru ke dalam kelas yang tepat dengan memilih *k* tetangga terdekat dari informasi baru tersebut [6].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

Banyaknya tetangga (*k*) yang dipilih dari setiap kelas informasi biasanya ditentukan sebagai angka ganjil. Hal ini dilakukan untuk menghindari kemungkinan terjadinya jumlah jarak yang sama dalam sistem pengelompokan terhadap informasi baru. Selain itu, dalam penentuan nilai *k* yang optimal, digunakan metode *Elbow Method* yang membantu dalam menentukan tingkat kesalahan (*error rate*) [7], Tujuan dari menggunakan metode ini adalah untuk mencapai tingkat akurasi yang optimal dalam metode *K-Nearest Neighbor* (K-NN).

Metode *K-Nearest Neighbor* (K-NN) telah terbukti memiliki tingkat keakuratan yang tinggi dalam mendiagnosis penyakit berdasarkan penelitian sebelumnya. Sebagai contoh, dalam sebuah penelitian yang dilakukan oleh M. Syukri Mustafa dan I Wayan Simpen, metode K-NN digunakan untuk memprediksi penyakit diabetes pada pasien. Hasil penelitian tersebut menunjukkan tingkat akurasi terbaik sebesar 93,33% dengan tingkat kesalahan (*error rate*) sebesar 6,67% [8].

Berdasarkan permasalahan yang dapat ditemukan diatas, perlu dilakukan penelitian untuk 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 *liver* dari data sekunder yang diambil dari UCI Machine Learning Repository. Data yang digunakan yaitu ILDP (Indian Liver Patient Dataset).

II. TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Berdasarkan penelitian yang dilakukan oleh Popon Handayan pada tahun 2019 dalam jurnalnya "Liver Disease Prediction Using Decision Tree and Neural Network Methods", menggunakan skor yang dihasilkan oleh *Confusion Matrix* dan ROC *Curve*, ditemukan bahwa pengujian data menggunakan algoritma C4.5 memberikan hasil yang lebih baik. Akurasi model algoritma C4.5 mencapai 75,56% dengan nilai AUC (*Area Under the Curve*) sebesar 0,898, sedangkan akurasi algoritma *neural network* mencapai 74,17% dengan nilai AUC sebesar 0,671. Selisih akurasi antara kedua algoritma tersebut adalah 1,39%, sedangkan selisih AUC sebesar 0,227%[2].

Berdasarkan Penelitian Muhammad Rizki Fahdia (2020) dalam jurnalnya yang berjudul "Perbandingan Algoritma Klasifikasi Untuk Prediksi Penyakit *Liver*" mendapatkan hasil *Decision Tree* (C4.5) diperoleh dengan skor terbaik tingkat akurasi (72,56%) dan AUC (0,594) setelah meningkatkan kinerja dengan ekstraksi fitur dan pemilihan fitur. Peningkatannya adalah 73,24% (akurasi) dan 0,602 (AUC) [9].

Menurut penelitian yang dilakukan oleh Intan Setiawati pada tahun 2019 dalam jurnal berjudul "Implementasi Decision Tree Untuk Mendiagnosis Penyakit Liver", ditemukan bahwa Algoritma *Decision Tree C4.5* menghasilkan tingkat akurasi sebesar 72,67%. Selain itu, penelitian juga menunjukkan bahwa dari 11 variabel yang terkait dengan penyakit *liver* pada dataset ILPD, hanya 2 variabel (*alanine aminotransferase*) yang memiliki peran kritis dalam menentukan adanya penyakit [10].

Dari Penelitian Endah Patimah (2021) dalam jurnalnya yang berjudul "Klasifikasi Penyakit *Liver* dengan Menggunakan Metode *Decision Tree*", Berdasarkan hasil tersebut dapat disimpulkan bahwa akurasi yang paling tinggi adalah dengan *k-fold cross-validation* dan *standard scaler* dengan akurasi sebesar 0,733. [11].

Sedangkan dari penelitian Citra Nurina Prabiantissa (2021) dari jurnalnya yang berjudul "Klasifikasi pada Dataset Penyakit *liver* Menggunakan Algoritma *Support Vector Machine*, K-NN, dan Naïve Bayes" Hasil penelitian menunjukkan bahwa SVM memiliki kinerja rata-rata terbaik diantara ketiga algoritma dengan akurasi sebesar 82,36%. [12].

Berdasarkan beberapa penelitian sebelumnya yang telah dijelaskan, terdapat kesamaan dalam klasifikasi penyakit *liver*. Dalam upaya untuk meningkatkan metode *K-Nearest Neighbor* (K-NN), penelitian ini bertujuan untuk menyempurnakan algoritma tersebut dengan memperkenalkan metode *elbow* (siku) guna mendapatkan hasil yang lebih optimal dalam penentuan nilai *k*. Selain itu, pada tahap pengujian, penelitian ini akan fokus pada pengolahan data yang lebih terfokus, bukan hanya pada pencarian *k* yang optimal.

B. Klasifikasi

Klasifikasi adalah proses pencarian pola yang menjelaskan atau memisahkan konsep atau kelas informasi untuk memperkirakan kelas objek yang tidak diketahui [13]. 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 [14].

C. Data Mining

Data mining adalah proses menganalisis dan mengekstrak pengetahuan secara otomatis menggunakan satu atau lebih teknik dalam mempelajari komputer [15][16], *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*) [17].

D. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database 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].

E. 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

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

klasifikasi sebagai suatu sistem yang dapat mencari informasi [18], adapaun rumus jarak atau *Euclidean* seperti yang ditunjukkan Persamaan (1).

$$euc = \sqrt{((a_1 - b_1)^2 + \cdots + (a_n - b_n)^2)} \quad (1)$$

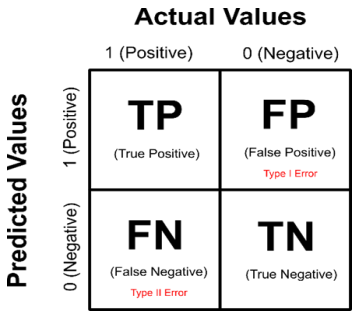
Dimana $a = a_1, a_2, \ldots, a_n$ dan $b = b_1, b_2, \ldots, b_n$ menyatakan nilai *n* atribut dari dua *register* [19]. Untuk atribut dengan nilai kelas. Skor diprediksi untuk suatu jenis berdasarkan peringkat tertinggi dari tetangga sekitarnya [20].

F. *Elbow Method*

Elbow method atau metode siku digunakan untuk memilih jumlah kluster atau kelompok yang optimal. Algoritma siku digunakan untuk menentukan jumlah kelompok yang akan dibentuk [21]. Metode *elbow* diimplementasikan dengan cara menentukan data optimal dan melihat grafik dari nilai *k* yang disematkan [22].

G. *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. [2]. *Confusion matrix* adalah tabel yang memberikan klasifikasi kumpulan data uji yang benar dan kumpulan data uji yang salah. Contoh *confusion matrix* klasifikasi biner [23][24].



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.

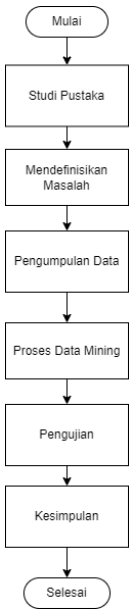
4) *False Negative (FN) — Type II Error: FN* adalah data positif yang diprediksi sebagai data negatif.

H. *Overfitting*

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 [25][26].

III. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini, dapat dilihat pada diagram alir (*flowchart*) yang ditampilkan pada Gambar 2.



Gambar 2 Diagram Alir Penelitian

A. *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 lain yang dapat digunakan untuk mendukung penyelesaian penelitian.

B. *Mendefinisikan Masalah*

Permasalahan dalam penelitian ini adalah bagaimana cara menggunakan metode *Elbow* dalam algoritma *K-Nearest Neighbors* (KNN) untuk melakukan klasifikasi dan memprediksi penyakit hati (*liver disease*) menggunakan teknik *Machine Learning*. Masalah ini melibatkan *output binary class* dengan nilai *Disease* (0) dan *No Disease* (1).

C. *Pengumpulan Data*

Dalam tahap ini, tujuannya adalah untuk mendapatkan dataset yang sesuai dengan penelitian. Mengingat sulitnya mendapatkan data dari Indonesia karena aturan etika medis yang mengharuskan kerahasiaan pasien, penelitian ini akan menggunakan dataset ILPD (*Indian Liver Patient Dataset*) yang memang telah digunakan dalam penelitian sebelumnya.fitur dan keterangan yang terdapat dalam dataset ILPD adalah seperti Tabel I.

TABLE I. FITUR DAN KETERANGAN

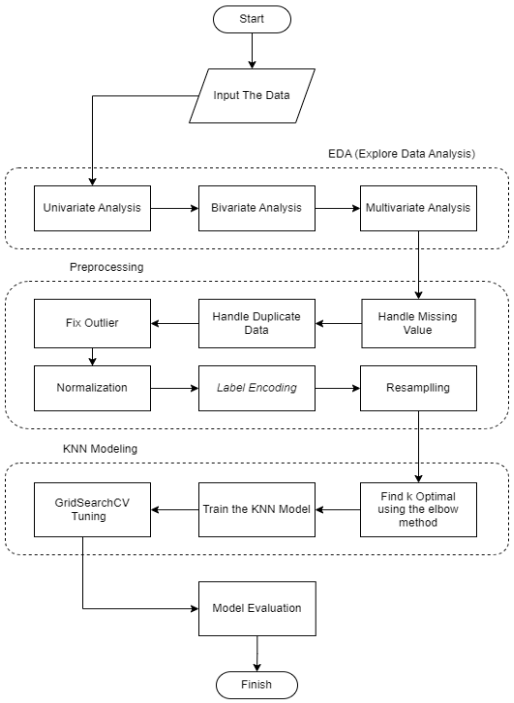
No	Fitur	Keterangan	Jenis Data
1	Age	Umur Pasien	Numerikal
2	Gender	Jenis Kelamin Pasien	Kategorikal

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

No	Fitur	Keterangan	Jenis Data
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 *liver*	Numerikal
7	Aspartate (Aspartate Aminotransferase)	Enzim yang berkaitan dengan kinerja organ *liver*	Numerikal
8	TP (Total Proteins)	Berisi Albumin dan Globulin	Numerikal
9	ALB (Albumin)	Protein Utama pada darah yang diproduksi *liver*	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 atau tidak menderita liver	Kategorikal

D. Proses Data Mining

Proses *data mining* adalah cara untuk mengeksplorasi data dan menemukan pola atau informasi yang berguna dari data tersebut. Prosesnya dapat disederhanakan menjadi beberapa langkah seperti pada Gambar 3.



Gambar 3. 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*[27].

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 [28], penaganan *data Imbalance* dengan *library* SMOTE dan Tomek [29], 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 [30].

4) *Evaluation Model:* Setelah mendapatkan hasil akurasi dari pemodelan KNN peneliti mengevaluasi hasil tersebut apakah hasil permodelan sesuai dengan keinginan peneliti.

E. Pengujian

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

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

1) Percobaan 1: Pada percobaan pertama ini hanya dilakukan *preprocessing, splitting data, elbow method knn*, dan pemodelan knn lalu pengecekan akurasi.

2) Percobaan 2: Pada percobaan kedua ini dilakukan penanganan *outlier, preprocessing, splitting data, elbow method knn*, dan pemodelan knn lalu pengecekan akurasi.

3) Percobaan 3: Pada percobaan ketiga ini dilakukan penanganan *outlier, preprocessing, splitting data, normalisasi data, elbow method knn*, dan pemodelan *knn* lalu pengecekan akurasi.

4) Percobaan 4: Pada percobaan keempat ini dilakukan penanganan *outlier, preprocessing, splitting data, normalisasi data*, menangani data tidak seimbang, *elbow method knn*, dan pemodelan *knn* lalu pengecekan akurasi

F. Kesimpulan

Kesimpulan didapat berdasarkan dari analisis hasil pengujian yang telah dilakukan dan juga dari hasil analisis selama pembangunan aplikasi, Kesimpulan berupa hasil dari penelitian sesuai yang tergambar pada diagram alir metode penelitian.

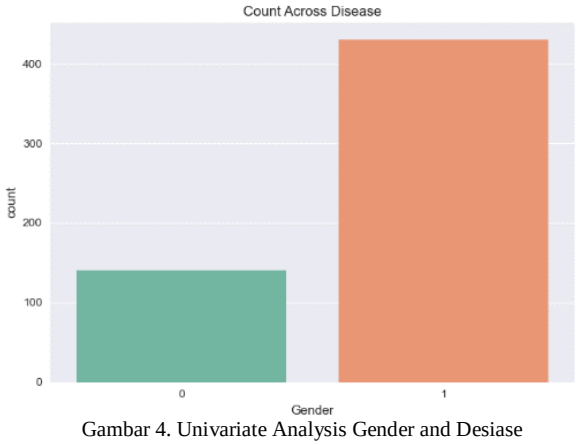
IV. HASIL DAN PEMBAHASAN

Hasil dan Pembahasan diambil sesuai dengan Diagram alir *Data Mining* pada Gambar 3, konsep dasarnya yaitu EDA, *Preprocessing*, KNN *Modeling*, dan Evaluasi Model.

A. 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.

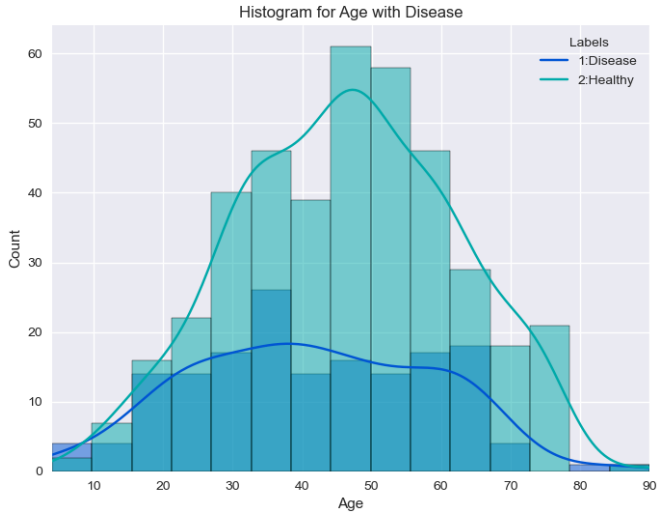
1) Analisis *Univariate: Exploratory Data Analysis* yang pertama adalah *Univariate Analysis* yaitu membaca satu set variabel, Tujuan dari analisis univariat adalah untuk mendapatkan data, mendefinisikan, meringkas, dan menganalisis pola yang ada di dalamnya.



Gambar 4. Univariate Analysis Gender and Desiase

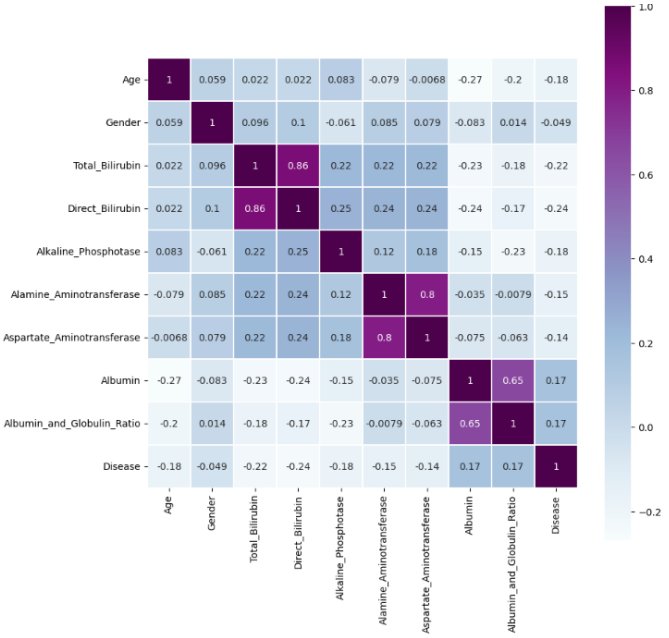
Pada Gambar 4 menjelaskan bahwa lebih banyak data pria daripada wanita di dalam dataset yaitu pria berjumlah 430 sedangkan wanita 140.

2) Analisis *Bivariate:* Gambar 5 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 5. Bivariate Analysis Age and Disease

3) Analisis *Multivariates:* Analisis *multivariates* adalah bentuk teknik analisis statistik yang lebih kompleks dan digunakan ketika ada lebih dari dua variabel dalam kumpulan data, Dari Gambar 6 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.



Gambar 6. *Multivariate Analysis* Setiap Fitur

Pada Gambar 6 menjelaskan *correlation* dari setiap fitur dan dari gambar tersebut dapat diambil kesimpulan bahwa korelasi

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

fitur dengan label kurang bagus, korelasi terbaik ada pada *Albumin* dan *Albumin and Globulin Ratio* dan korelasi terburuk pada fitur *Direct Bilirubin*.

B. 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.*

1) *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.

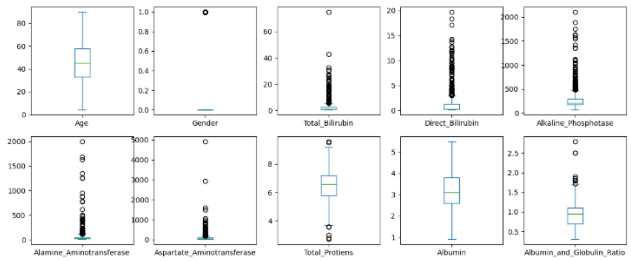
```
Age                          0
Gender                       0
Total_Bilirubin              0
Direct_Bilirubin             0
Alkaline_Phosphotase         0
Alamine_Aminotransferase     0
Aspartate_Aminotransferase   0
Total_Protiens               0
Albumin                      0
Albumin_and_Globulin_Ratio   4
Dataset                      0
```

Gambar 7. Hasil Pengecekan *Missing Value*

Gambar 6 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 menghapus akan tetapi pada penelitian ini penulis merubah data yang hilang menjadi *median*.

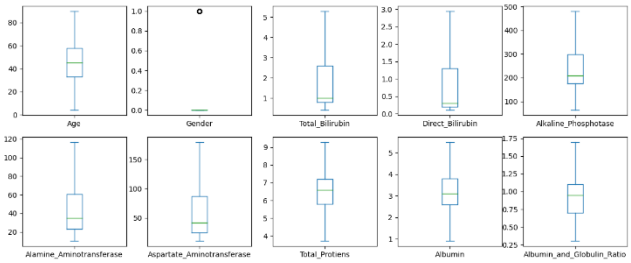
2) *Duplicate Data:* Seperti namanya, *Duplicate data* adalah data yang serupa atau kumpulan data dapat terdiri dari beberapa objek data (duplikat). Selama pemrosesan, hampir selalu ada tumpang tindih antara data. Pada penelitian ini pengolahan dilakukan dengan menghapus data duplikat.

3) *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*.



Gambar 8. Box Plot Pengecekan Outlier

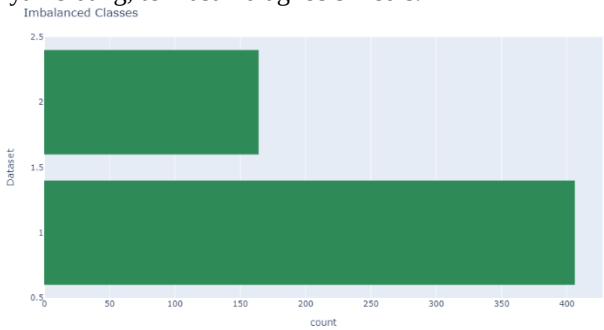
Gambar 8 adalah *box plot* untuk melihat ada tidaknya data *outlier*, bisa dilihat bahwa banyak data outlier di setiap fitur sehingga penulis melakukan penanganan dengan merubah data *outlier* menjadi nilai *mean*, dari beberapa kali peneliti melakukan uji coba pembatasan data mungkin bukan solusi yang baik dalam masalah ini untuk menangani *outlier*, karena TB, DB dan parameter tersebut sangat signifikan untuk deteksi atau prediksi Penyakit *liver*, karena batas waktu dan kurangnya pengetahuan domain di lapangan, peneliti ingin melanjutkan lebih jauh dengan membatasi *outlier* dengan nilai *mean*.



Gambar 9. *Box Plot* Pengecekan *Outlier* Setelah Dilakukan Penanganan

Gamabr 9 adalah *plot box* pengecekan *outlier* setelah dilakukan penanganan menggunakan perhitungan IQR membatasi dari diatas Q3 dan dibawah Q1 dan merubah nilai menjadi *mean* meskipun tidak banyak yang berubah karena dikhawatirkan merusak hasil prediksi akan tetapi diharapkan dapat sedikit meningkatkan akurasi sebelum data di latih.

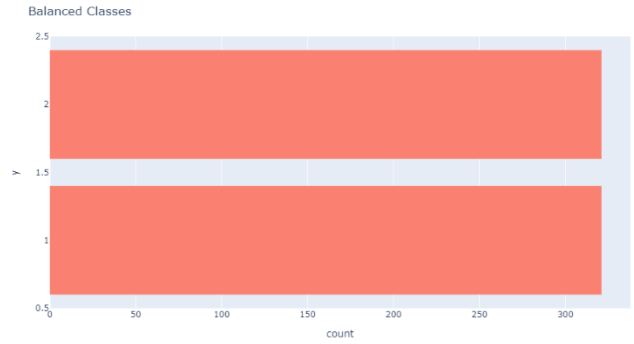
4) *Imbalance Data: Imbalanced Data* adalah masalah umum dalam klasifikasi pembelajaran mesin, di mana ada hubungan yang tidak proporsional antar kelas. Itu ditemukan di banyak bidang, termasuk diagnosis medis.



Gambar 10. *Data Imbalance*

Gambar 10 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.

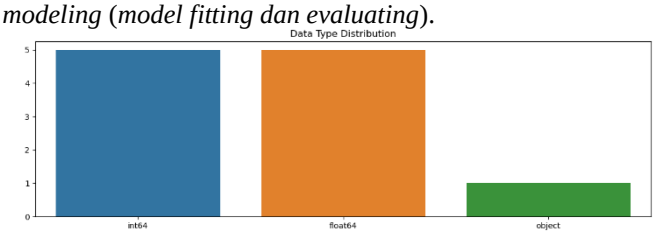
p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023



Gambar 11. *Data Balance*

Gambar 11 menampilkan *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.

5) *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*).



Gambar 12. *Data Distribution*

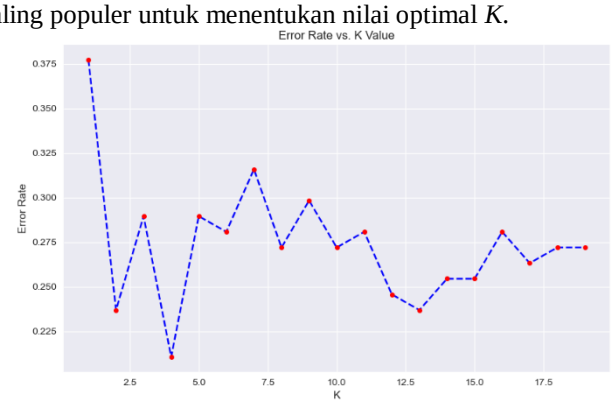
Gambar 12 adalah *plot* persebaran data distribusi dan dapat dilihat ada data *object* yaitu pada data *Gender* atau jenis kelamin oleh karena itu peneliti menggunakan *label encoding library Python* dengan *scikit-learn* data yang ada akan *otomatis* diurutkan berdasarkan abjad alfabet agar data dapat diproses saat melakukan *modeling data*.

6) *Normalization:* 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, 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*).

C. Hasil KNN Modeling

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

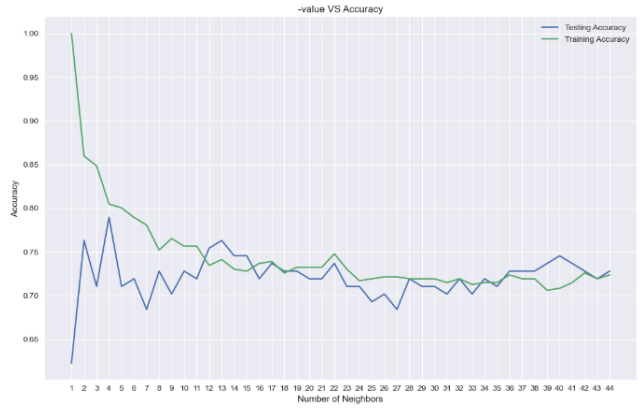
1) *Elbow method:* Langkah terpenting dalam *machine learning* yang menggunakan metode *K-Nearest Neighbor* (K-NN) adalah menentukan nilai *K* yang optimal. Dengan kata lain, berapa banyak *cluster* yang harus dibagi menjadi data nilai optimal untuk k mengurangi efek *noise* pada klasifikasi, tetapi 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*.



Gambar 13. *Elbow Method*

Pada penelitian ini pemodelan diawali dengan mencari nilai *k* terbaik menggunakan *elbow method*, pada Gambar 13 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 Gambar 14.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023



Gambar 14. Hasil Mencari K optimal

Dari hasil perulangan 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.

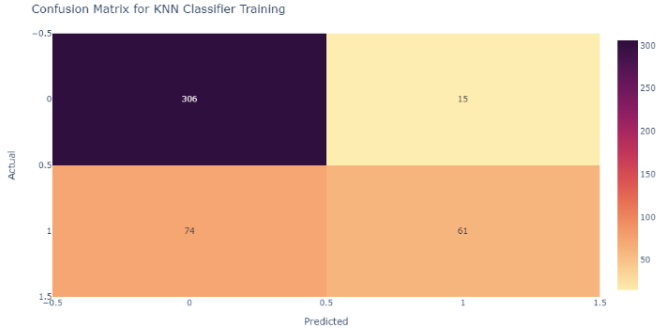
2) Pemodelan *K-Nearest Neighbor* (K-NN): Setelah berhasil menentukan *k* optimal langkah selanjutnya yaitu pemodelan metode KNN menggunakan *K* = 4.

```
Training:
Accuracy KNN model: 80.48245614035088
ROC : {0.7025614399446174}
```

	precision	recall	f1-score	support
1	0.81	0.95	0.87	321
2	0.80	0.45	0.58	135
accuracy			0.80	456
macro avg	0.80	0.70	0.73	456
weighted avg	0.80	0.80	0.79	456

Gambar 15 Matrik Evaluasi *Training*

Gambar 15 adalah Matrik evaluasi untuk data *Training* yang mana mendapatkan hasil Akurasi sebesar 80% dan ROC 0,7 pada label 2 kurang bagus dalam *recall* dan *f1-score* sedangkan untuk label 1 sudah mendapatkan hasil yang bagus.



Gambar 16. *Confusion matrix Training*

Gambar 16 adalah *Confusion Matrix* data *Training*, *training* dapat memprediksi dengan baik untuk label 0 sedangkan untuk label 1 lebih banyak kesalahan, Adapun penjelasan Gambar 16 adalah berikut:

a) *True Positive (TP):* 306 data dengan label 0 terklasifikasi benar sebagai 0

b) *True Negative (TN):* 61 data dengan label 1 terklasifikasi benar sebagai 1

c) *False Positive (FP) — Type I Error:* 15 data dengan label 0 terklasifikasi salah sebagai 1

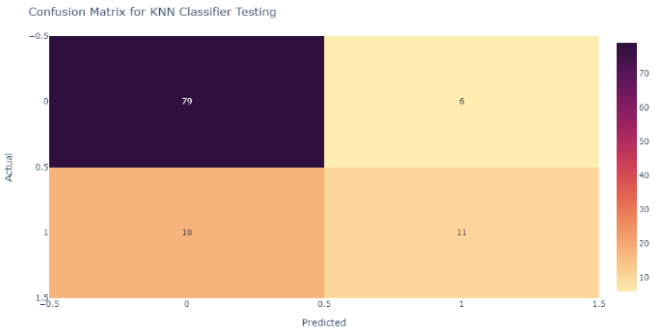
d) *False Negative (FN) — Type II Error:* 74 data dengan label 1 terklasifikasi salah sebagai 0

```
Testing
Accuracy KNN model : 78.94736842105263
ROC : {0.6543610547667343}
```

	precision	recall	f1-score	support
1	0.81	0.93	0.87	85
2	0.65	0.38	0.48	29
accuracy			0.79	114
macro avg	0.73	0.65	0.67	114
weighted avg	0.77	0.79	0.77	114

Gambar 17 Matrik Evaluasi *Testing*

Gambar 17 adalah Matrik evaluasi untuk data *Training* yang mana mendapatkan hasil Akurasi sebesar 78% dan ROC 0,6 pada label 2 sama halnya dengan data *training* kurang bagus dalam *recall* dan *f1-score* sedangkan untuk label 1 sudah mendapatkan hasil yang bagus.



Gambar *18. Confusion matrix Testing*

a) True Positive (TP): 79 data dengan label 0 terklasifikasi benar sebagai 0

b) True Negative (TN): 11 data dengan label 1 terklasifikasi benar sebagai 1

c) False Positive (FP) — Type I Error: 6 data dengan label 0 terklasifikasi salah sebagai 1

d) False Negative (FN) — Type II Error: 18 data dengan label 1 terklasifikasi salah sebagai 0

3) *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 params* menggunakan jarak *n_neighbors* 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 Gambar 17 untuk data *training*, dan Gambar 18 untuk data *testing*. yaitu

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

best params untuk menghitung jarak menggunakan perhitungan *metric manhattan* dan *k* atau *n_neighbors* terbaik adalah 2 dan *weights* terbaik adalah *uniform.*

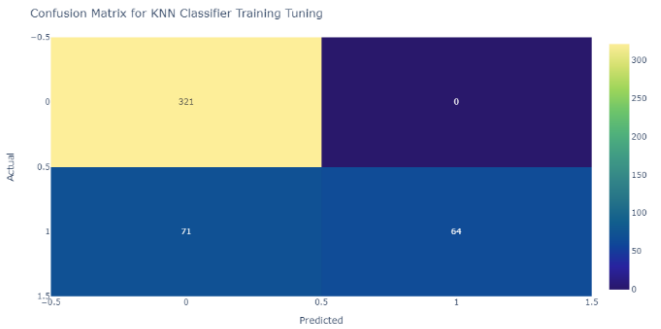
```
Training Tuning:
Accuracy KNN model: 84.4298245614035
ROC : {0.737037037037037}
              precision    recall  f1-score   support

           1       0.82      1.00      0.90       321
           2       1.00      0.47      0.64       135

    accuracy                           0.84       456
   macro avg       0.91      0.74      0.77       456
weighted avg       0.87      0.84      0.82       456
```

Gambar 19 Matrik Evaluasi *Training Tuning*

Gambar 19 adalah Matrik evaluasi untuk data *Training* setelah di *Tuning* yang mana mendapatkan hasil Akurasi sebesar 84% dan ROC 0,7 pada label 2 kurang bagus dalam *recall* dan *f1-score* sedangkan untuk label 1 sudah mendapatkan hasil yang bagus.



Gambar 20. *Confusion matrix Training Tuning*

Gambar 20 adalah *Confusion Matrix* data *Training* setelah *Tuning*, *training* dapat memprediksi dengan baik untuk label 0 sedangkan untuk label 1 lebih banyak kesalahan, Adapun penjelasan Gambar 20 adalah berikut:

a) True Positive (TP): 321 data dengan label 0 terklasifikasi benar sebagai 0

b) True Negative (TN): 64 data dengan label 1 terklasifikasi benar sebagai 1

c) False Positive (FP) — Type I Error: 0 data dengan label 0 terklasifikasi salah sebagai 1

d) False Negative (FN) — Type II Error: 71 data dengan label 1 terklasifikasi salah sebagai 0

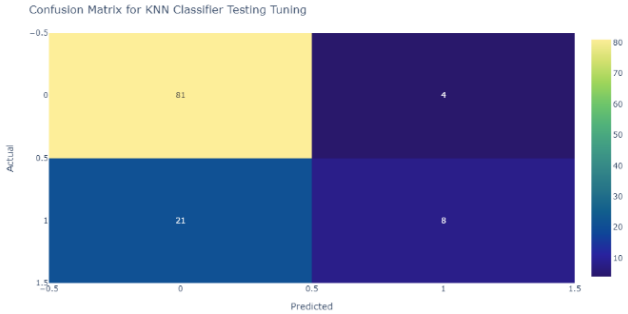
```
Testing Tuning
Accuracy KNN model : 78.0701754385965
ROC : {0.6144016227180528}
              precision    recall  f1-score   support

           1       0.79      0.95      0.87        85
           2       0.67      0.28      0.39        29

    accuracy                           0.78       114
   macro avg       0.73      0.61      0.63       114
weighted avg       0.76      0.78      0.75       114
```

Gambar 21 Matrik Evaluasi *Testing Tuning*

Gambar 21 adalah Matrik evaluasi untuk data *Testing* setelah di *Tuning* yang mana mendapatkan hasil Akurasi sebesar 78% dan ROC 0,6 pada label 2 kurang bagus dalam *recall* dan *f1-score* sedangkan untuk label 1 sudah mendapatkan hasil yang bagus



Gambar *22. Confusion matrix Testing Tuning*

Gambar 22 adalah *Confusion Matrix* data *Testing* setelah *Tuning*, *Testing* dapat memprediksi dengan baik untuk label 0 sedangkan untuk label 1 lebih banyak kesalahan, Adapun penjelasan Gambar 20 adalah berikut:

a) True Positive (TP): 81 data dengan label 0 terklasifikasi benar sebagai 0

b) True Negative (TN): 8 data dengan label 1 terklasifikasi benar sebagai 1

c) False Positive (FP) — Type I Error: 4 data dengan label 0 terklasifikasi salah sebagai 1

d) False Negative (FN) — Type II Error: 21 data dengan label 1 terklasifikasi salah sebagai 0

D. Evaluasi dan Hasil Pengujian

Pengujian dilakukan dalam 4 kali seperti yang tertera pada perancangan pengujian dan didapatkan hasil seperti Table II.

TABLE II. 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 Table II pengujian akurasi dapat dilihat perbandingan akurasi dari setiap pengujian:

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

1) Percobaan 1: Pada pengujian pertama ini hanya dilakukan *preprocessing, splitting data, elbow method knn*, dan pemodelan knn, hasil *elbow method* terbaik pada *k* = 19 dan mendapatkan hasil akurasi 72,6% *training* dan 74,6% *testing*, setelah *Tuning* mendapatkan 75,2% *training* dan 73,7% *testing*.

2) Percobaan 2*:* Pada pengujian kedua ini dilakukan penanganan *outlier*, *preprocessing, splitting data, elbow method knn*, dan permodelan knn hasil *elbow method* terbaik pada k= 20 dan mendapatkan hasil akurasi 73,5% *training* dan 73,7% *testing* setelah *Tuning* mendapatkan 75,0% *training* dan 72,8% *testing*.

3) Percobaan 3*:* Pada pengujian ketiga ini dilakukan penanganan *outlier, preprocessing, splitting data, normalisasi data, elbow method knn*, dan pemodelan knn hasil *elbow method* terbaik pada k= 4 dan mendapatkan hasil akurasi 80,5% *training* dan 78,8% *testing* setelah *Tuning* mendapatkan 82,2% *training* dan 77,1% *testing*.

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 hasil *elbow method* terbaik pada k= 2 dan mendapatkan hasil akurasi 98,8% *training* dan 78,9% *testing* setelah *Tuning* mendapatkan 100% *training* dan 62,3% *testing*.

Dari hasil beberapa percobaan diatas bisa dilihat pada percobaan pertama dan kedua hasil akursai yang didapat tidak jauh berbeda, pada percobaan ketiga setelah data dinormalisasikan mengalami peningkatan yang signifikan didalam akurasi dan sedangkan pada percobaan ke empat saat penyeimbangan data akurasi *training* menjadi sangat baik tetapi *testing* menjadi sangat jelek bisa diambil kesimpulan percobaan ketiga mendapatkan hasil akurasi terbaik dan percobaan ke empat data menjadi *Overfitting*.

V. KESIMPULAN

Kesimpulan dari penelitian ini adalah bahwa penggunaan metode *elbow* pada pengujian ketiga menghasilkan nilai optimal k = 4, yang terbukti sangat efektif dalam membantu penulis menemukan nilai k yang optimal. Dalam penelitian ini, klasifikasi penyakit *Liver* menghasilkan *output* dengan nilai *Disease* (1) dan *No Disease* (2).

Dari hasil uji coba akurasi, dapat disimpulkan bahwa percobaan ketiga menghasilkan akurasi terbaik, dengan akurasi *training* sebesar 80,5% dan *testing* sebesar 78,9%. Setelah dilakukan *tuning* pada data *training*, akurasi meningkat menjadi 82,2%, sedangkan akurasi pada data *testing* menjadi 77,1%. Meskipun pada percobaan keempat, data yang telah diseimbangkan menghasilkan peningkatan yang signifikan pada akurasi *training* menjadi 98,8%, dan setelah dilakukan *tuning*, mencapai 100%. Namun, akurasi pada data *testing* mengalami penurunan yang signifikan hingga mencapai 69,2%, dan setelah dilakukan *tuning*, menjadi 62%. Dari hasil ini, dapat disimpulkan bahwa penyeimbangan data pada percobaan keempat mengakibatkan kondisi *overfitting*, di mana model mencoba untuk mempelajari semua detail, termasuk *noise* pada data, dan mencoba untuk memasukkan semua informasi, yang pada akhirnya mengurangi kemampuan prediksi.

REFERENCES

[1] J. Tandi, "Pola Penggunaan Obat Pada Pasien Penyakit Hati Yang Menjalani Rawat Inap Di Rumah Sakit Umum Daerah Undata Palu," *Perspekt. J. Pengemb. Sumber Daya Insa.*, vol. 2, no. 2, pp. 218–223, 2017.

[2] 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.

[3] 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.

[4] I. R. Hikmah and R. N. Yasa, "Perbandingan Hasil Prediksi Diagnosis pada Indian Liver Patient Dataset (ILPD) dengan Teknik Supervised Learning Menggunakan Software Orange," *J. Telemat.*, vol. 16, no. 2, pp. 69–76, 2021.

[5] 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.

[6] J. Suntoro, "22-DATA MINING Algoritma dan Implementasi Menggunakan Bahasa Pemrograman PHP," *DATA Min. Algoritm. dan Implementasi Menggunakan Bhs. Pemrograman PHP*, vol. 9, no. 9, pp. 259–278, 2019.

[7] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.

[8] 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

[9] M. R. F. Rizki, "Perbandingan Algoritma Klasifikasi Untuk Prediksi Penyakit Liver," *Reputasi J. Rekayasa Perangkat Lunak*, vol. 1, no. 2, pp. 82–88, 2020, doi: 10.31294/reputasi.v1i2.109.

[10] I. Setiawati, A. P. Wibowo, and A. Hermawan, "Pendahuluan Tinjauan Pustaka Penelitian Sebelumnya Klasifikasi," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.

[11] E. Patimah, V. B. Haekal, and D. Sandya Prasvita, "Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, vol. 2, no. 1, pp. 655–659, 2021.

[12] Prabiantissa Citra Nurina, "Klasifikasi pada Dataset Penyakit HatiMenggunakan Algoritma Support Vector Machine, K-NN, dan Naïve Bayes," *Semin. Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.*, vol. 1, no. 1, pp. 263–268, 2021.

[13] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," *Fakt. Exacta*, vol. 11, no. 3, pp. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.

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

[15] 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

[16] M. A. Muslim *et al.*, "Data Mining Algoritma C4.5 Disertai contoh kasus dan penerapannya dengan program computer," *Nucl. Phys.*, vol. 13, no. 1, pp. 104–116, 2019.

[17] D. Turban, Efraim ; Aronson, Jay E ; Liang, Ting peng ; Prabantini, "Decision Support Systems And Intelligent Systems : (Sistem

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023

Pendukung Keputusan Dan Sistem Cerdas) / Efraim Turban," 2005.

[18] 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.

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

[20] A. P. W. Anjar Wanto, Muhammad Noor Hasan Siregar, N. L. W. S. R. G. Dedy Hartama, M. R. L. Darmawan Napitupulu, Edi Surya Negara, and C. P. Sarini Vita Dewi, *Data Mining Algoritma & Implementasi*. 2020.

[21] 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.

[22] 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

[23] 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.

[24] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

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

[26] W. A. Firmansyach, U. Hayati, and ..., "Analisa Terjadinya Overfitting Dan Underfitting Pada Algoritma Naive Bayes Dan Decision Tree Dengan Teknik Cross Validation," *JATI (Jurnal Mhs. …*, vol. 7, no. 1, 2023, [Online]. Available: https://ejournal.itn.ac.id/index.php/jati/article/view/6329%0Ahttps://ejournal.itn.ac.id/index.php/jati/article/download/6329/3678

[27] G. Szabo, G. Polatkan, O. Boykin, and A. Chalkiopoulos, *Social Media Data Mining and Analytics [Minería y análisis de datos de medios sociales]*. 2019.

[28] S. M. Faradisa, T. D. Nugrahadi, Muliadi, I. Budiman, and D. Kartini, "Implementasi IQR-SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Diabetes menggunakan K-Nearest Neighbors," vol. 15, pp. 48–60, 2021.

[29] R. Agustika, "Penerapan Kombinasi SMOTE dan Tomek Links untuk Klasifikasi Data Tidak Seimbang dengan Metode Random Forest," 2021, [Online]. Available: http://etd.repository.ugm.ac.id/penelitian/detail/199065

[30] Z. Maisat, E. Darmawan, and A. Fauzan, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM," vol. 13, no. 1, pp. 8–15, 2023.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i2.1643, Copyright ©2023
Submitted : March 6, 2023, Revised : May 17, 2023, Accepted : May 22, 2023, Published : July 1, 2023