ANALISIS METODE *DECISION TREE* DAN *NAÏVE BAYES* PADA PASIEN PENYAKIT *LIVER*

Damar Adji Sodikin[1], El Thaariq Is'ad[2], Rangga Prayoga[3], Ahmad Nur Ihsan Purwanto[4]

[1]Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ. damar.adji.s@students.esqbs.ac.id,
[2]Sekolah Tinggi Ilmu Manajemen dan Ilmu .Komputer ESQ. el.thaariq.i@students.esqbs.ac.id,
[3]Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ. rangga.surya.p@students.esqbs.ac.id,
[4]Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ. ahmadnur.ihsan@students.esqbs.ac.id

Abstrak*— Penyakit liver disebabkan oleh konsumsi alkohol, penyakit hati berlemak, genetika dari orang tua, penyakit diabetes, obesitas, serta bahan kimia dalam obat. Hati adalah organ dalam manusia yang terbesar dan terpenting. Pemeriksaan bagian organ dalam, termasuk organ paru-paru, organ jantung, kulit, otak, sistem saraf, serta lambung dapat memberikan tanda mengenai penyebab penyakit liver. Fungsi hati adalah untuk mendetoksifikasi racun dari dalam tubuh. Data yang diteliti merupakan hasil telaah dari 583 data yang diperoleh dengan 416 orang dinyatakan "positif" penyakit liver dan sisanya 167 orang "negatif" penyakit hati (liver). Oleh karenanya dibutuhkan sebuah analisis data mining yang menggunakan algoritma pohon keputusan dengan nilai akurasi optimal sebesar 70.29%, sedangkan metode Naive Bayes memiliki nilai akurasi optimal sebesar 70.29%, 67.05%. Dapat disimpulkan bahwa metode pohon keputusan merupakan salah satu metode yang dapat memecahkan masalah penentuan penyakit liver.

Keywords* — *Decision Tree*, Naïve Bayes, *liver

Abstract— Liver infection is caused by liquor utilization, greasy liver illness, hereditary qualities from guardians, diabetes, corpulence, and chemicals in drugs. The liver is the biggest and most critical human inner organ. Examination of parts of the body, counting the lungs, heart, skin, brain, apprehensive framework, and stomach, can give clues around the cause of liver infection. The work of the liver is to detoxify poisons from the body. The information examined was the result of a audit of 583 information gotten with 416 individuals pronounced "positive" for liver illness and the remaining 167 individuals "negative" for liver malady. Subsequently, we require a information mining examination framework that employments the choice tree calculation strategy with an ideal exactness esteem of 70.29%, whereas the Credulous Bayes strategy has an ideal exactness esteem of 70.29%, 67.05%. It can be concluded that the choice tree strategy is one strategy that can fathom the issue of deciding liver infection.

Keywords* — *Decision Tree*, *Naïve Bayes*, *liver

I. PENDAHULUAN

Gangguan *liver*/hati adalah penyakit yang timbul di organ hati manusia, dimana kesehatan organ hati sangat vital bagi tubuh. *Liver* atau hati mengubah zat yang beracun menjadi sebuah nutrisi, dimana selanjutnya digunakan tubuh untuk mengontrol hormon dalam tubuh [1]. Selain itu, juga berfungsi untuk memproduksi protein yang membantu pembekuan darah, dan memecah sel darah merah. Penyebab penyakit liver adalah

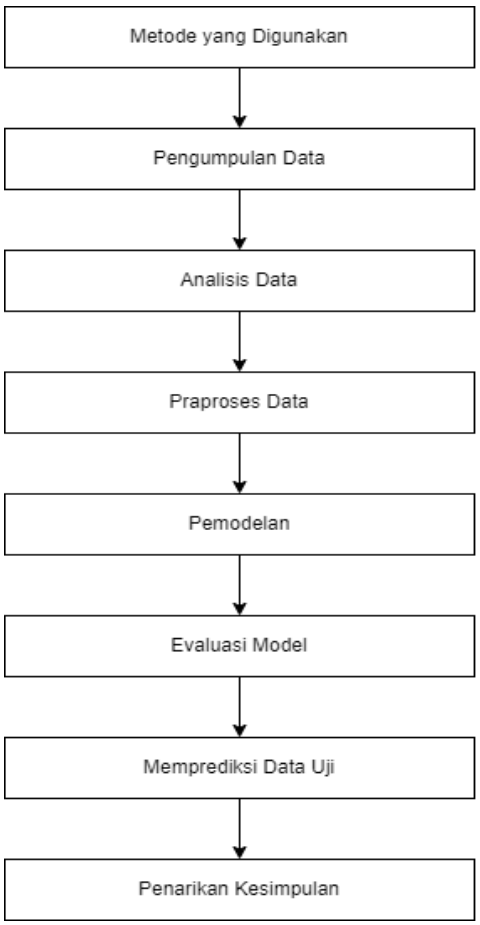
konsumsi alkohol berlebihan, penumpukan lemak pada hati, faktor genetik, diabetes, dan obesitas tubuh. Dampak kerusakan hati antara lain peradangan, penggumpalan darah, serta gagal hati. Fungsi hati adalah menyaring seluruh darah dari usus melalui vena portal, kemudian menyimpan dan mengatur komponen makanan yang diterima melalui vena portal. Komponen makanan ini kemudian dilepaskan ke aliran darah sesuai kebutuhan. Hati juga akan menjaga kebutuhan organ tubuh terutama otak terhadap zat-zat beracun yang mau tidak mau akan diserap (didetoksifikasi) oleh usus, misalnya amonia dari usus merupakan zat yang sangat beracun. [2]. Penyakit hati, yang menyebabkan kematian karena dianggap sebagai *silent killer* tanpa gejala. Terdapat 28 juta penderita penyakit liver di Indonesia, menjadikan penyakit liver termasuk dalam 10 penyakit dengan angka kematian tertinggi, sehingga angka kematian meningkat setiap tahunnya. [3]. Diagnosis adalah identifikasi ciri-ciri suatu penyakit atau kondisi atau membedakan suatu penyakit atau kondisi dengan penyakit atau kondisi lainnya. Evaluasi dapat dilakukan melalui pemeriksaan fisik, pengujian laboratorium, atau cara lain dan dapat juga dibantu dengan program komputer yang dirancang untuk meningkatkan proses pengambilan keputusan. [4]. Dalam industri perawatan kesehatan, salah mendiagnosis penyakit pasien adalah tanggung jawab terberat bagi seorang profesional perawatan kesehatan. Kesalahan diagnosis dapat membahayakan kesehatan pasien, bahkan berujung pada kematian [5]. Penambangan data, melibatkan pencarian informasi bisnis yang berharga dari database yang sangat besar [6]. Data mining adalah proses mengekstraksi informasi atau menarik data dalam database yang besar. Dan dalam jurnal data mining dikenal juga dengan *Knowledge Discovery in Databases* (KDD) digunakan untuk mengembangkan masalah identifikasi penyakit liver menggunakan metode pohon keputusan dan membandingkan Naïve Bayes untuk menentukan metode apa yang paling akurat [7]. Sumber data penelitian ini diperoleh dari https://www.kaggle.com/datasets/uciml/indian-liver-patient-records. Data tersebut adalah hasil pemeriksaan terhadap 583 orang yang berasal dari wilayah Andhra Pradesh, India. [8].

II. METODE PENELITIAN

Metodologi penelitian dilaksanakan sebagai pedoman bagi peneliti agar mencapai hasil dan tujuan yang sesuai.

1. Desain Kajian

Kajian ini merupakan penelitian eksperimental yang bertujuan untuk membuat perbandingan antara algoritma *Decision Tree* dengan *Naive bayes*. Studi eksperimental ini didasarkan pada alur pemecahan masalah yang ditunjukkan pada Gambar 1 sebagai berikut:



Gambar 1. Metode penelitian

2. Klasifikasi Metode

Dalam metode penelitian ini digunakan klasifikasi algoritma *Decision Tree* dan *Naive bayes*. Untuk membandingkan metode mana yang

dapat menyelesaikan masalah dalam menentukan penyakit *liver* dari *data-data* yang ada.

3. Pengumpulan *Data*

Data yang diambil secara langsung oleh peneliti disebut sebagai sumber primer, sedangkan sumber *data* yang digunakan disebut sumber sekunder [9]. *Data* yang dipakai adalah *data* sekunder karena didapat dari *data* yang digunakan pada penelitian yang bersumber dari: *https://www.kaggle.com/datasets/uciml/indian-liver-patient-records*.

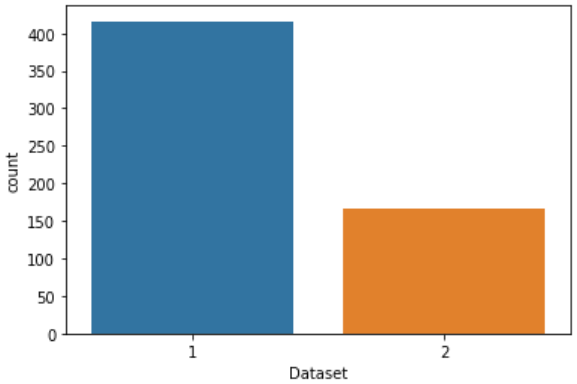
Kumpulan *data* tersebut berisi sebanyak 416 catatan pasien *liver*/hati dan 167 catatan pasien tanpa penyakit *liver*/hati yang dikumpulkan dari Timur Laut Andhra Pradesh, India. Kolom "*Dataset*" adalah label yang digunakan untuk membagi kelompok menjadi pasien *liver* (penyakit *liver*) dan yang tidak terinfeksi. Kumpulan *data* ini berisi 441 catatan pria dan 142 catatan pasien wanita. Di mana atributnya meliputi *age of the patient*, *Gender of the patient*, *total billirubin*, *direct billirubin, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, total Protiens, Albumin, Albumin and Globulin Ratio.*

Tabel 1. Parameter atribut *Data*set

Atribut	**Nilai Rujukan**
Age	<20 Anak, <=20 - <50 Muda, dan >=500 Dewasa
Gender	Male (laki-laki) dan Female (perempuan)
Total Bilirubin	<=1 Normal dan >1 Abnormal
Direct Bilirubin	<=0.2 Normal dan >0.2 Abnormal
Alkaline Phosphotase	<=30 - <=120 Normal dan >120 Tinggi
Alamine Aminotransferase	<47 Normal dan >=47 Abnormal
Total Protiens	<6 Rendah, <=6 <=8 Normal, >8 Tinggi
Albumin	<3.4 Rendah, <=3.4 - <=4.8 Normal, >4.8 Tinggi
Ration Albumin and Globulin Ratio	>1 Normal dan <= 1
Dataset	1 (Positif) dan 2 (Negatif)

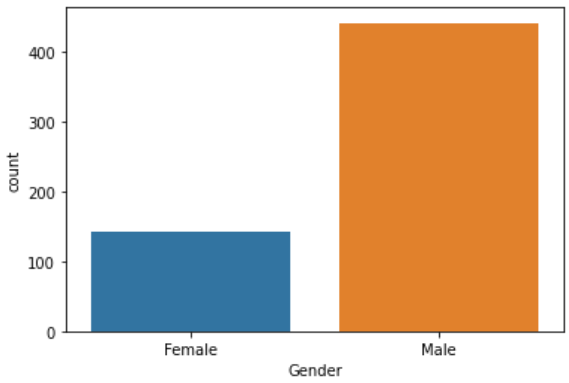
4. Analisis *Data*

Pengklasteran data dilakukan berdasarkan hasil prediksi pasien dari atribut dataset yang terdapat pada kelas atribut. Berikut grafik dari pengelompokan *data*nya :



Gambar 3. Pengelompokan *Data* Positif dan Negatif

Kemudian diperoleh 583 *data* dengan rincian sebanyak 416 pasien terinfeksi penyakit hati/*liver* dan sejumlah 167 orang *negative* dari penyakit tersebut. Selanjutnya pengelompokan *data* berdasarkan gender. Berikut hasil pengelompokannya :



Gambar 4. Pengelompokan *Data* Pria dan Wanita`

Dan didapatkan *data* pria sebanyak 441 orang dan wanita 142 orang pada *data* ini. Terakhir mengecek *missing value* yang ada pada *data*. Berikut adalah gambar hasil pengecekannya:

```
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 5. Hasil pemeriksaan pada *missing value*

Ternyata terdapat 4 *data record missing value* pada *albumin* dan *globulin ratio*. Untuk mengatasi *missing value* dari variabel berikut dan untuk memiliki hasil yang maksimal dalam memprediksi maka *data* tersebut akan didrop nantinya.

5. Praproses *Data*

Praproses *data* ini tujuannya untuk menormalisasi *data* yang berbeda bentuknya dan juga mengisi kekosongan *data* yang ada pada kolom, serta *drop* kolom yang tidak perlu. Pada tahapan ini juga mengubah tipe *data* menjadi satu tipe *data* yang sama, mengubah tipe *data* (mentransformasi) kolom kategori menjadi numerik agar *data* dapat diproses ke tahapan selanjutnya. Setelah mentransformasi *data* maka *data* akan dibagi yaitu *data training* dan validasi atau *testing*. *Data training* sebesar 80% sedangkan *data* validasi atau *testing* sebesar 20%.

6. Pemodelan

Pemodelan terbagi menjadi dua proses yaitu memisahkan *data training* dan *data testing* kemudian membuat modelnya. Pada praproses *data* sudah dibagi sebesar 80% *data training* dan 20% *data testing*. Pembagian *data* menggunakan *model splitting sklearn model train_test_split*. Setelah *data* terbagi selanjutnya adalah pemodelan kita menggunakan model *Decision Tree* dan *naive bayes*. Algoritma Pohon Keputusan dapat dipakai untuk meramalkan atau mengelompokkan sebuah peristiwa dengan pembuatan *decision tree* dimana algoritma ini yang dikembangkan oleh J. Ross [10]. Naïve Bayes atau multinomial naïve bayes merupakan metodeyang digunakan untuk mengklasifikasikan sekumpulan dokumen [11].

7. Evaluasi Model

Tahap evaluasi model bertujuan untuk meningkatkan model dengan penskalaan fitur untuk mengetahui keberhasilan dari sistem yang dibuat

8. Memprediksi *data* uji

Memprediksi *data* uji bertujuan untuk melihat hasil, serta memprediksi nilai selanjutnya yang akan menjadi nilai *data liver* selanjutnya dan hasil nilainya akan diekspor menjadi *data frame*. Dari 10 kali iterasi didapatkan beberapa kali *data* yang sama. Oleh karena itu model yang telah dibuat sudah cukup untuk dapat melakukan *data* uji sebenarnya.

9. Penarikan kesimpulan

Analisis hasil pengujian sistem yang dilakukan pada tahap ini, setelah melakukan seluruh rangkaian langkah-langkah yang telah dilakukan. Kelebihan dan kekurangan dari metode yang dibuat dapat diturunkan dari ini. Tahap akhir ini menentukan tindakan yang akan dilakukan oleh peneliti berikutnya untuk lebih mengembangkan penelitian sebelumnya. Hasil

yang didapatkan dari prediksi kasus terbaru penyakit *liver*, terdapat *data* yang tidak seimbang dari keseluruhan *data*, artinya *data* yang ada saat ini harus dievaluasi kembali untuk mendapatkan nilai yang merata.

III. HASIL DAN PEMBAHASAN

1. Praproses *Data*

Data yang digunakan adalah kasus 5 tahun lalu di India. Berikut hasil dari proses EDA:

- *Drop Column*

Menghapus *record* dengan *missing value* dengan perintah *df.dropna(inplace=True)*

Sebelum dihapus :

```
# cek missing values
df.isna().sum()
```

```
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*

Dan setelah dihapus hasilnya seperti berikut :

```
# cek missing values
df.isna().sum()
```

```
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  0
Dataset                     0
```

Gambar 7. Hasil Pengecekan *Missing value*

- Transformasi *data* menjadi numerik

Dalam *data* terdapat 1 kolom *data* bertipe *data* kategori. Berikut gambar *data*nya :

```
#   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
0   Age                         579 non-null    int64
1   Gender                      579 non-null    object
2   Total_Bilirubin             579 non-null    float64
3   Direct_Bilirubin            579 non-null    float64
4   Alkaline_Phosphotase        579 non-null    int64
5   Alamine_Aminotransferase    579 non-null    int64
6   Aspartate_Aminotransferase  579 non-null    int64
7   Total_Protiens              579 non-null    float64
8   Albumin                     579 non-null    float64
9   Albumin_and_Globulin_Ratio  579 non-null    float64
10  Dataset                     579 non-null    int64
```

Gambar 8. Info *data*

Data gender kemudian diubah menjadi tipe *data* numerik dengan perintah :

le = LabelEncoder() df.Gender = le.fit_transform(df.Gender)

Berikut hasil dari *data* yang telah di ubah :

```
#   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
0   Age                         579 non-null    int64
1   Gender                      579 non-null    int64
2   Total_Bilirubin             579 non-null    float64
3   Direct_Bilirubin            579 non-null    float64
4   Alkaline_Phosphotase        579 non-null    int64
5   Alamine_Aminotransferase    579 non-null    int64
6   Aspartate_Aminotransferase  579 non-null    int64
7   Total_Protiens              579 non-null    float64
8   Albumin                     579 non-null    float64
9   Albumin_and_Globulin_Ratio  579 non-null    float64
10  Dataset                     579 non-null    int64
```

Gambar 9. Info *data*

Data gender berubah menjadi *integer*.

2. Modelling

Membuat sebuah model *machine learning* yang akan memprediksi angka '*New Case*' dengan metode *Decision Tree* dan *naive bayes*, dengan langkah-langkah berikut:

- Splitting

Membagi *data*set menjadi dua bagian yaitu *data* tes dan *data* uji dengan ketentuan yang sudah dijelaskan pada tahap praproses *data*.

- Evaluation

Proses pengukuran hingga mana tujuan dari program ini telah tercapai. Disini kita akan melakukan *feature scaling* atau juga bisa disebut normalisasi menggunakan fungsi

MinMaxScaler() dari *library Scikit-learn*. Setelah itu kita mengubah *data* yang telah di normalisasi ke dalam sebuah bentuk *DataFrame* dan melakukan pemodelan kembali dengan *data* yang sudah di normalisasi tadi. Dan mendapatkan hasil 57.034% dari akurasi *Decision Tree* yang sudah di normalisasi.

- Predict Test

Digunakan untuk *testing* sebuah model, sebagai simulasi *testing* untuk dunia nyata. Disini kami melakukan prediksi dengan *data* uji, lalu mengubah hasilnya menjadi kedalam bentuk sebuah *DataFrame*.

IV. HASIL

Hasil dari *Decision Tree* seperti berikut :

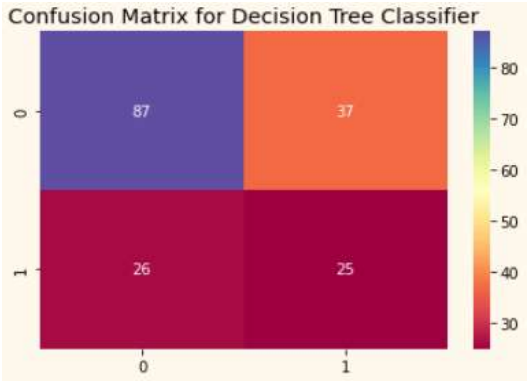
Classification metrics :

	precision	recall	f1-score	support
1	0.74	0.78	0.76	82
2	0.40	0.35	0.38	34
accuracy			0.66	116
macro avg	0.57	0.57	0.57	116
weighted avg	0.64	0.66	0.65	116

Gambar 10. Klasifikasi *Matrix Decision Tree*

	Actual	Prediction
31	1	1
485	1	1
202	2	2
454	2	1
500	1	1
397	1	1
561	1	1
581	1	1
404	1	2
14	1	1
231	1	2
309	1	1

Gambar 11. Prediksi Aktual *Decision Tree*



Gambar 12. *Matrix Confusion* dari *Decision Tree*

Hasil dari *Naive bayes* seperti berikut :

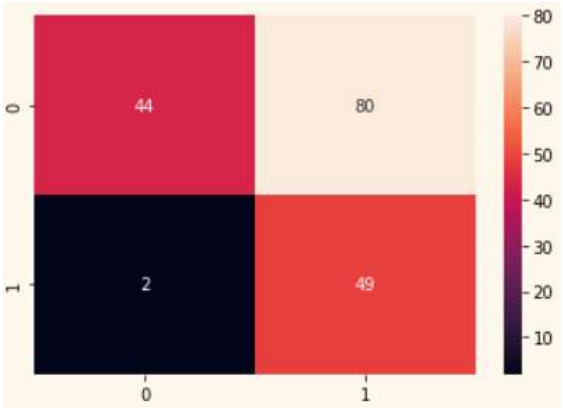
Classification metrics :

	precision	recall	f1-score	support
1	0.76	1.00	0.87	45
2	0.00	0.00	0.00	14
accuracy			0.76	59
macro avg	0.38	0.50	0.43	59
weighted avg	0.58	0.76	0.66	59

Gambar 13. Klasifikasi *Matrix Decision Tree*

	Actual	Prediction
85	1	1
12	2	1
403	2	1
450	1	1
504	1	1
500	1	1
499	1	1
118	1	1
37	1	1
34	2	1
422	1	1
407	1	1

Gambar 14. Prediksi Aktual *Naive bayes*



Gambar 15. *Matrix Confusion* dari *Naive bayes*

V. SIMPULAN DAN SARAN

Secara singkat, kajian ini memanfaatkan data pasien dengan penyakit hati/liver dengan memanfaatkan Pohon Keputusan / Decision Tree dan Naïve Bayes untuk mencari algoritma terbaik dalam mengidentifikasi penyakit liver. Gunanya mengevaluasi efisiensi pada kedua teknik tersebut, yaitu teknik Split Validation dan Cross Validation digunakan untuk mengukur kinerja dari kedua metode tersebut. Dari pengukuran tersebut disimpulkan bahwa pada metode pohon keputusan dalam klasifikasinya menghasilkan akurasi sebesar 70,29%. termasuk dalam klasifikasi ekuitas. *Naïve Bayes* menghasilkan akurasi sebesar 67,05%. Dengan demikian dapat disimpulkan bahwa metode untuk menyelesaikan masalah penentuan penyakit *liver* adalah dengan menggunakan pohon keputusan.

DAFTAR PUSTAKA

[1] Rahman NT. Analisa Algoritma Decision Tree dan Naive Bayes pada Pasien Penyakit Liver. Jurnal Fasilkom. 2020 Aug 13;10(2):144-51.

[2] Pujiyanta A, Pujiantoro A. Sistem Pakar Penentuan Jenis Penyakit Hati dengan Metode Inferensi Fuzzy Tsukamoto. Jurnal Informatika. 2012;6(1):617-29.

[3] Cahyanti FL, Sarasati F, Astuti W, Firasari E. KLASIFIKASI DATA MINING DENGAN ALGORITMA MACHINE LARNING UNTUK PREDIKSI PENYAKIT LIVER. Technologia: Jurnal Ilmiah. 2023 Apr 1;14(2):134-9.

[4] Adler J. Diagnosa penyakit dengan gejala demam pada manusia berbasis mobile: Knowledge based system. Komputika: Jurnal Sistem Komputer. 2017 Oct 31;6(2):51-8.

[5] Rahman NT. Analisa Algoritma Decision Tree dan Naive Bayes pada Pasien Penyakit Liver. Jurnal Fasilkom. 2020 Aug 13;10(2):144-51.

[6] Khormarudin AN. Teknik Data Mining: Algoritma K-Means Clustering. J. Ilmu Komput. 2016:1-2.

[7] Siregar AM, Kom S, Puspabhuana MK, Kom S, Kom M. Data Mining: Pengolahan Data Menjadi Informasi dengan RapidMiner. CV Kekata Group; 2017.

[8] Learning, U. M. (2017, September 20). Indian liver patient records. Kaggle. https://www.kaggle.com/datasets/uciml/indian-liver-patient-records.

[9] Adriansa M, Yulianti L, Elfianty L. Analisis Kepuasan Pelanggan Menggunakan Algoritma C4. 5. Jurnal Teknik Informatika UNIKA Santo Thomas. 2022 Jun 22:115-21.

[10] Yuliza R. Sistem Pakar Akurasi dalam Mengidentifikasi Penyakit Gingivitis pada Gigi Manusia dengan Metode Naive Bayes. Jurnal Sistim Informasi dan Teknologi. 2023:27-32.

[11] Wibisono A. Filtering Spam Email Menggunakan Metode Naive Bayes. Jurnal Teknologi Pintar. 2023 Jun 6;3(4).