Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan Vol. 8 No. 2 Desember 2023 | P-ISSN: 2527-9858
E-ISSN: 2548-1180

Seleksi Fitur Dengan Menggunakan Metode Entropy Pada Algoritma Klasifikasi Naive Bayes Untuk Penyakit Diabetes

Victor Tarigan[1*], Rendy Syahputra[2], Pujo Hari Saputra[3], Ade Yusupa[4]

[1,2,3,4] *Program Studi Teknik Informatika, Jurusan Elektro, Fakultas Teknik, Universitas Sam Ratulangi*

[1]*victortarigan@unsrat.ac.id,* [2]*rendysyahputra@unsrat.ac.id,* [3]*pujoharisaputro@unsrat.ac.id,* [4]*ade@unsrat.ac.id*

Abstrak— Gangguan kesehatan yang kerap terjadi pada masyarakat salah satunya adalah Diabetes yang merupakan penyakit yang disebabkan kadar gula darah yang tinggi. Saat ini konsep data *mining* banyak digunakan di berbagai macam aspek. Salah satu luaran konsep data mining adalah klasifikasi. Data mining adalah proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data. Klasifikasi adalah tipe analisis data yang dapat membantu menentukan kelas label dari sampel yang ingin diklasifikasi. Ada beberapa algoritma klasifikasi yang dapat digunakan untuk menentukan hasil klasifikasi berdasarkan atribut atau fitur yang ada. Algoritma Naïve Bayes adalah salah satu algoritma klasifikasi yang sering digunakan untuk proses klasifikasi dengan data yang banyak dan kompleks dan efektif untuk mengklasifikasikan data medis, termasuk dalam klasifikasi penyakit diabetes. Untuk mendapatkan hasil klasifikasi dalam konsep data mining ada beberpa langkah yang harus dijalankan proses data mining, antara lain : input data, Pre-processing / cleaning, proses data mining, dan post processing Diantara tahapan-tahapan Pre-processing di atas, pada penelitian ini akan difokuskan pada seleksi fitur. Salah satu metode untuk seleksi fitur adalah dengan menggunakan metode Entropy. Diharapkan dengan menghilangkan fitur dari data yang ada dan memiliki nilai informasi rendah, akurasi klasifikasi dapat ditingkatkan dan dapat membantu dalam upaya pencegahan dan pengobatan dini penyakit diabetes**.**

Kata kunci— Klasifikasi, Seleksi Fitur, Entropy, Naive Bayes, Diabetes

*Abstract***—** One of the health problems that often occurs in society is diabetes, which is a disease caused by high blood sugar levels. Currently, the concept of data mining is widely used in various aspects. One of the outcomes of the data mining concept is classification. Data mining is a process of collecting and processing data that aims to extract important information from the data. Classification is a type of data analysis that can help determine the class label of the sample you want to classify. There are several classification algorithms that can be used to determine classification results based on existing attributes or features. The Naïve Bayes algorithm is a classification algorithm that is often used for classification processes with large and complex data and is effective for classifying medical data, including the classification of diabetes. To obtain classification results in the data mining concept, there are several steps that must be carried out in the data mining process, including: data input, pre-processing / cleaning, data mining process, and post processing. Among the pre-processing stages above, in this research will focus on feature selection. One method for feature selection is to use the Entropy method. It is hoped that by eliminating features from existing data that have low information value, classification accuracy can be increased and can help in efforts to prevent and treat diabetes early.

Keywords— Classification, Feature Selection, Entropy, Naive Bayes, Diabetes

I. PENDAHULUAN

Masyarakat salah satunya adalah Diabetes yang merupakan penyakit yang disebabkan kadar gula darah yang tinggi(Rizky Adhi Nugroho, 2019). Hal ini menjadi tantangan yang berat pada sistem pelayanan kesehatan di Indonesia. Salah satu cara untuk mencegah diabetes adalah dengan melakukan diagnosis dini pada pasien yang memiliki risiko terkena diabetes. Oleh karena itu, klasifikasi diabetes mellitus tipe 2 menjadi penting dalam upaya pencegahan dan pengobatan dini. Saat ini konsep data mining banyak digunakan di berbagai macam aspek.

Pemanfaatan jumlah data historis yang melimpah dengan berbagai macam tipe data, Konsep data mining dapat dimanfaatkan untuk menghasilkan luaran yang memiliki nilai yang tinggi untuk dapat dimanfaatkan sebaik mungkin bagi yang memiliki data tersebut. Salah satu luaran konsep data mining adalah klasifikasi. Klasifikasi adalah tipe analisis data yang dapat membantu menentukan kelas label dari sampel yang ingin diklasifikasi. Tujuan dari proses klasifikasi ini sendiri adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data.

Ada beberapa algoritma klasifikasi yang dapat digunakan untuk menentukan hasil klasifikasi berdasarkan atribut atau fitur yang ada. Algoritma Naïve Bayes adalah salah satu algoritma klasifikasi yang sering digunakan untuk proses klasifikasi dengan data yang banyak dan kompleks dan efektif untuk mengklasifikasikan data medis, termasuk dalam klasifikasi penyakit diabetes[1].

Untuk mendapatkan hasil klasifikasi dalam konsep data mining ada beberpa langkah yang harus dijalankan proses data mining, antara lain : *input data, Pre-processing / cleaning,* proses *data mining,* dan *post processing*[2]. Dari tahapan-tahapan tersebut, tahap *Pre-processing / cleaning* memegang peranan dalam proses data mining untuk mendapatkan hasil yang lebih optimal dikarenakan proses ini dimanfaatkan untuk memperbaiki data mentah yang sudah diinput sebelumnya yang nantinya akan membantu mengurangi error hasil proses klasifikasi. Adapun jenis tahapan proses *Pre-processing* ini adalah memilih data yang relevan, pembersihan data, seleksi fitur, dan lain-lain.

Diantara tahapan-tahapan Pre-processing di atas, pada penelitian ini akan difokuskan pada seleksi fitur adalah suatu teknik dalam data mining yang dilakukan untuk memilih fitur atau atribut yang paling relevan dan signifikan dalam dataset. Tujuan seleksi fitur ini dilakukan adalah untuk untuk mengurangi dimensi data, memperbaiki keakuratan model, mempercepat waktu pelatihan model, dan meningkatkan interpretabilitas model. Hal ini sudah dibuktikan pada penelitian sebelumnya, diantaranya penelitian yang dilakukan oleh [3] menunjukkan bahwa menggunakan seleksi fitur dapat meningkatkan akurasi model dan mengurangi waktu pelatihan

model pada tugas klasifikasi email spam. Penelitian ini menggunakan metode seleksi fitur filter untuk memilih fitur-fitur yang paling relevan dalam dataset email spam. Setelah melakukan seleksi fitur, akurasi model meningkat sekitar 4,38%.

Adapun beberapa metode yang digunakan untuk menseleksi fitur, salah satunya adalah dengan menggunakan metode entropy. Metode ini bekerja dengan memilih subset variabel yang paling informatif dalam dataset.Penelitian sebelumnya yang sudah dilakukan [4] dalam penelitiannya bahwa dengan seleksi fitur dengan menggunakn entropy akurasi dapat meningkat serta jumlah fitur yang dikurangi menjadi 50 % dan otomatis mempercepat proses klasifikasi, penelitian yang dilakukan oleh [5] bahwa menggunakan metode seleksi fitur berdasarkan entropy dapat meningkatkan akurasi model hingga 91,5% pada dataset yang sama dan penelitian yang dilakukan oleh [6] menunjukkan bahwa menggunakan metode seleksi fitur berdasarkan entropy dapat meningkatkan akurasi model hingga 71,38% pada dataset yang digunakan.

Proses menghilangkan fitur yang memiliki nilai informasi rendah, akurasi klasifikasi dapat ditingkatkan. Oleh karena itu, penelitian tentang penggunaan metode Entropy pada seleksi fitur untuk meningkatkan akurasi klasifikasi diabetes mellitus dengan algoritma Naive Bayes menjadi penting untuk dilakukan. Dengan menggunakan metode Entropy pada seleksi fitur, diharapkan dapat meningkatkan akurasi klasifikasi diabetes mellitus dengan algoritma Naive Bayes sehingga dapat membantu dalam upaya pencegahan dan pengobatan dini penyakit diabetes.

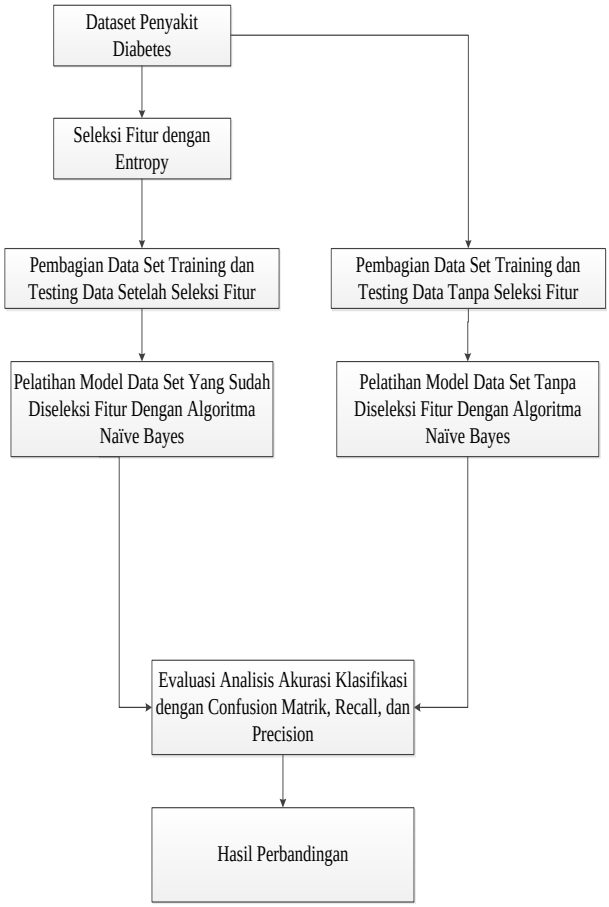
Seleksi fitur adalah suatu proses untuk memilih sejumlah fitur yang merupakan subset dari fitur yang lama atau fitur aslinya sehingga diperoleh fitur-fitur yang paling berpengaruh (signifikan) terhadap akurasi klasifikasi. Proses seleksi fitur dapat mengurangi jumlah noise dan fitur yang kurang relevan, sehingga diharapkan dapat meningkatkan akurasi.

Proses seleksi fitur juga merupakan salah satu strategi yang digunakan untuk melakukan reduksi dimensi terhadap fitur-fitur yang digunakan dalam proses data mining [7] Proses seleksi fitur dapat memberikan beberapa keuntungan dalam proses data mining, antara lain banyak algoritma data mining akan bekerja lebih baik pada fitur yang dimensinya lebih rendah, Hal ini disebabkan karena seleksi fitur dapat mengurangi jumlah noise dan fitur yang kurang relevan. Keuntungan lain adalah dimensi yang lebih rendah akan membentuk model yang lebih mudah dipahami serta lebih mudah divisualisasikan.

Keuntungan yang tidak kalah penting adalah dimensi yang rendah akan menghemat waktu dan memori yang digunakan oleh algoritma data mining [7]

II. METODOLOGI PENELITIAN

Tahap metodologi merupakan tahap dalam menjabarkan tahapan atau alur penelitian yang akan dilakukan. Alur metodologi penelitian terdapat pada Gambar 1.



Gambar 1.Metodologi Penelitian

A. *Dataset Penyakit Diabetes*

Dataset adalah suatu database didalam memori (in-memory). Dataset memiliki semua karateristik, fitur dan fungsi dari database biasa. Dataset dapat memiliki banyak tabel, dan table dapat memiliki hubungan(relationship). Dalam melakukan penelitian pastinya akan membutuhkan dataset yang nantinya akan diolah oleh suatu algoritma, penelitian ini menggunakan menggunakan dataset dari website Kaggle. Dataset yang digunakan merupakan dataset public yang dipublikasi pada tahun 2015 dengan 22 fitur yang dapat dilihat pada tabel 1:

Tabel 1. Atribut yang digunkaan dalam penelitian

Variabel	Keterangan
HighBP	Bernilai 1 jika tekanan darah tinggi atau bernilai 0 jika tekanan darah rendah
HighChol	Bernilai 1 jika memiliki kolesterol atau bernilai 0 jika tidak bernilai kolesterol
CholCheck	Bernilai 0 tidak ada kolesterol dalam 5 tahun terakhir dan bernilai 1 jika ada kolesterol
BMI	Nilai Body Mask Indek Pasien
Smoker	Apakah pernah merokok sebanyak 100 buah batang rokok. Jika ada bernilai 1, jika tidak bernilai 0. Catatan : 100 buah sama

Variabel	Keterangan
	dengan 5 bungkus rokok
Stroke	Bernilai 1 jika memiliki riwayat stroke dan bernilai 0 jika tidak memiliki riwayat stroke
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity	Aktivitias Fisik Berat dalam 30 hari terakhir. Jika ada bernilai 1, jika tidak bernilai 0
Fruits	Apakah mengkonsumsi buah-buahan setiap harinya. Jika Ia maka bernilai 1, jika tidak bernilai 0
Veggies	Apakah mengkonsumsi sayur-sayuran setiap harinya. Jika Ia maka bernilai 1, jika tidak bernilai 0
HvyAlcoholConsump	Pria dewasa mengkonsumsi minuman keras 14 kali setiap minggunya dan Wanita mengkonsumsi minuman keras 7 kali setiap minggunya, jika ia bernilai 1 dan 0 jika tidak
AnyHealthcare	Memiliki asuransi Kesehatan, jika ia bernilai 1 dan 0 jika tidak
NoDocbcCost	Pernah ingin ke dokter tapi tidak memiliki uang. Jika pernah bernilai 1, jika tidak benilai 0
GenHlth	Nilai Kesehatan saat ini. 1 = Sempurna 2= sangat baik 3= baik 4 = wajar 5 = tidak baik
MentHlth	Berapa hari mengalami Kesehatan mental
PhysHlth	Berapa hari mengalami gangguang psikologis
DiffWalk	Apakah kesulitan dalam menaiki tangga? Jika ia bernilai 1 dan bernilai 0 jika tidak
Sex	Jenis kelamin laki-laki atau perempuan. Bernilai 1 apabila jenis kelamin laki-laki dan 0 apabila memiliki jenis kelamin perempuan
Age	Umur pasien. Ada 13 kategori umur. Bernilai 1 jika umur 18-24 2 jika umur 25-29 3 jika umur 30-34 4 jika umur 35-39 5 jika umur 40-44 6 jika umur 45-49 7 jika umur 50-54 8 jika umur 55-59 9 jika umur 60-64 10 jika umur 65-69 11 jika umur 70-74 12 jika umur 75-79 13 jika umur 80 ke atas 14 jika mengolak memberikan informasi umur
Education	Level Pendidikan dari skala 1 sampai 6
Income	Level pendapatan dari skala 1 sampai 8 1 Kurang dari $10,000 2 ($10,000 - $15,000) 3 ($15,000 - $20,000) 4 ($20,000 - $25,000) 5 ($25,000 - $35,000) 6 ($35,000 - $50,000) 7 ($50,000 - $75,000) 8 Lebih dari $75,000
Diabetes	Level diabetes 0 = tidak ada diabetes 1 = prediabetes 2 = diabetes

Data set ini memiliki 3 class yaitu tidak ada diabetes yang diberikan nilai 0, prediabetes yaitu 1, dan diabetes dengan nilai 2. Jumlah dataset ini sendiri sebanyak 253.680 data, dengan jumlah data yang tidak memiliki diabetes sebanyak 213.703 data, sedangkan untuk data yang memiliki kelas prediabetes sebanyak 4.631 data, dan yang memiliki diabetes sebanyak 35.346 data.

B. Pembagian Data

Data yang diperoleh dari teknik pengumpulan data akan dibagi menjadi data *training* dan data *testing* menggunakan metode *split validation* dengan rasio perbandingan 90:10, 80:20, dan 70:30. Berikut adalah penjelasan kedua pembagian:

a. Data *training* merupakan data yang diperoleh 90%, 80%, dan 70% dari dataset yang diacak dengan menggunakan Bahasa pemrograman Pyhton. Data *training* nantinya akan digunakan sebagai pengenalan pola dengan menggunakan algoritma *Naïve Bayes*
b. Data *testing* merupakan data yang diperoleh 10% ,20%, dan 30% dari dataset yang diacak. Nantinya data *testing* akan digunakan sebagai data uji dari algoritma *Naïve Bayes* untuk mendapatkan nilai *confusion matrix*.

C. Seleksi Fitur

Proses seleksi fitur dibutuhkan untuk menyeleksi fitur/atribut yang memiliki dampak kecil dalam proses klasifikasi. Tujuan dari proses seleksi fitur yang utama adalah

dengan menghemat waktu proses klasifikasi dengan mengurangi fitur yang tidak berdampak tersebut dan dapat meningkatkan akurasi hasil proses klasifikasi.

Metode yang digunakan untuk proses seleksi fitur ini sendiri adalah dengan menggunakan metode Entropy. Metode Entropy dapat mengurangi fitur dari dataset penyakit diabetes yang memiliki fitur yang cukup banyak yaitu 21 fitur. Dengan menggunakan Bahasa pemrograman Pyhton dapat menghasilkan nilai entropy masing-masing fitur. Gambar 2 akan menampilkan *source code* untuk proses seleksi dengan menggunkan metode Entropy.

```python
import pandas as pd
import numpy as np
from sklearn.feature_selection import mutual_info_classif

# Muat dataset dari file Excel (XLSX)
file_path = "prediabet.xlsx"
data = pd.read_excel(file_path)

# Pisahkan fitur (X) dan target (y)
X = data.iloc[:, :-1]  # Ambil semua kolom kecuali kolom terakhir
y = data.iloc[:, -1]   # Ambil kolom terakhir sebagai target

# Konversi ke array numpy
X = np.array(X)
y = np.array(y)

# Hitung mutual information (entropy) dari setiap fitur
mutual_info = mutual_info_classif(X, y)

# Tampilkan hasil entropy setiap fitur
for i, entropy in enumerate(mutual_info):
    print(f"Entropy dari fitur ke-{i+1}: {entropy}")
```

Gambar 2. *Soruce Code* Metode *Entropy*

Hasil dari source code metode entropy ini sendiri dapat dilihat pada tabel 2.

Tabel 2. Hasil Entropy

No	Variabel	Nilai Entropy
1	*HighBP*	0.05422
2	*HighChol*	0.04075
3	*CholCheck*	0.04883
4	*BMI*	0.03202
5	*Smoker*	0.02483
6	*Stroke*	0.00581
7	*HeartDiseaseorAttack*	0.01322
8	*PhysActivity*	0.04845
9	*Fruits*	0.03969
10	*Veggies*	0.04263
11	*HvyAlcoholConsump*	0.00183
12	*AnyHealthcare*	0.05029
13	*NoDocbcCost*	0.00200
14	*GenHlth*	0.05603
15	*MentHlth*	0.00309
16	*PhysHlth*	0.01504
17	*DiffWalk*	0.02394
18	*Sex*	0.02362
18	*Age*	0.02674
19	*Education*	0.02388
20	*Income*	0.02375

Dari tabel 2 terlihat nilai *entropy* unutk masing-masing fitur. Dari nilai *entropy* ini akan diuji berapa banyak fitur yang akan dihilangkan agar mendapatkan nilai akurasi yang terbaik dengan menghapus nilai *entropy* terkecil satu persatu pada saat proses klasifikasi dengan menggunakan algoritma *Naïve Bayes.*

Dari tabel 2, dapat dilihat 5 fitur yang memiliki nilai entropy terkecil yaitu *HeartDiseaseorAttack*, *Stroke*, *MentHlth*, *NoDocbcCost*, dan *HvyAlcoholConsump*. Tabel 3 akan ditampilkan data fitur yang memiliki nilai entropy dari yang paling besar sampai terkecil.

Tabel 3. Peringkat Hasil Entropy

Peringkat	Variabel	Nilai Entropy
1	*GenHlth*	0.05603
2	*HighBP*	0.05422
3	*AnyHealthcare*	0.05029
4	*CholCheck*	0.04883
5	*PhysActivity*	0.04845
6	*Veggies*	0.04263
7	*HighChol*	0.04075
8	*Fruits*	0.03969
9	*BMI*	0.03202
10	*Age*	0.02674
11	*Smoker*	0.02483
12	*DiffWalk*	0.02394
13	*Education*	0.02388
14	*Income*	0.02375
15	*Sex*	0.02362
16	*PhysHlth*	0.01504
17	*HeartDiseaseorAttack*	0.01322
18	*Stroke*	0.00581
18	*MentHlth*	0.00309
19	*NoDocbcCost*	0.00200
20	*HvyAlcoholConsump*	0.00183

D. Pengujian Komparasi Seleksi Fitur Pada Algoritma Naïve Bayes

Pengujian komparasi yang dilakukan nantinya yaitu pertama akan dianalisis dengan mengukur nilai akurasi yang didapatkan dari klasifikasi dengan mengukur algoritma Naïve Bayes dan kedua, dataset akan melewati seleksi fitur metode Entropy yang nantinya akan dipakai sebagai komparasi. Source Code Bahasa pemrograman Pyhton untuk algoritma Naïve Bayes dapat dilihat pada gambar 3.

```python
file_path = "prediabet.xlsx"
data = pd.read_excel(file_path)

# Pisahkan fitur (X) dan target (y)
X = data.iloc[:, :-1]  # Ambil semua kolom kecuali kolom terakhir
y = data.iloc[:, -1]   # Ambil kolom terakhir sebagai target

# Konversi ke array numpy
X = np.array(X)
y = np.array(y)

# Bagi dataset menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Inisialisasi model Naive Bayes
naive_bayes_model = GaussianNB()

# Latih model dengan data latih
naive_bayes_model.fit(X_train, y_train)

# Lakukan prediksi pada data uji
y_pred = naive_bayes_model.predict(X_test)
```

Gambar 3. Koding Program Pyhton Algoritma Naïve Bayes

Analisis perbandingan dilakukan berdasarkan dari nilai akurasi, *recall*, dan, *precision* yang didapat dari nilai *confusion matrix* dengan menggunakan bahasa pemrograman Pyhton. Koding program untuk pencarian nilai akurasi ini dapat dilihat pada gambar 4.

```
# Tampilkan laporan klasifikasi (precision, recall, f1-score, dll.)
print("Laporan Klasifikasi:")
print(classification_report(y_test, y_pred))
cm = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(cm)
# Tampilkan confusion matrix dalam bentuk heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, cmap="Blues", fmt="d", xticklabels=np.unique(y), yticklabels=np.unique(y))
plt.xlabel("Prediksi")
plt.ylabel("Nilai Sebenarnya")
plt.title("Confusion Matrix")
plt.show()
```

Gambar 4. Koding Program Pyhton Pencarian Nilai Akurasi

III. HASIL DAN PEMBAHASAN

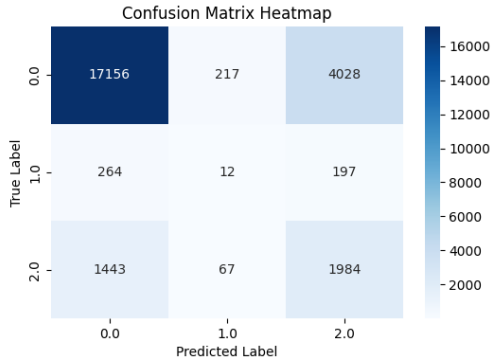
Terdapat beberapa hal yang ditekankan dalam pembahasan penelitian ini, yaitu mendapatkan akurasi proses klasifikasi dengan algoritma *Naïve Bayes* tanpa seleksi fitur dan mendapatkan hasil akurasi setelah proses seleksi fitur. Di bawah ini akan ditampilkan proses pengujian tersebut.sehingga mendapatkan hasil akurasi untuk masing-masing pengujian klasifikasi dengan menggunakna algoritma *Naïve Bayes*

A. Hasil Pengujian Algoritma Klasifikasi Naïve Bayes Tanpa Seleksi Fitur

Hasil pengujian ini dilakukan tanpa menghilangkan fitur dari dataset penyakit diabaetes yang artinya 21 fitur tetap digunakan untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 4 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

Hasil pengujian ini dilakukan tanpa menghilangkan fitur dari dataset penyakit diabaetes yang artinya 21 fitur tetap digunakan untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 4 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

1. Pengujian Tanpa Seleksi Fitur Perbandingan 90:10
   Pengujian ini akan dilakukan proses klasifikasi dengan 90% data training dan 10% data testing. Pada gambar 5 akan ditampilkan *Heat Map* dari Confusion matrix hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 5. *Heat Map Confusion Matrix* Tanpa Seleksi Fitur 90:10.

Dari gambar 5 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{17156+12+1984}{25368}\right) * 100\ \% = 75.5\%$$

$$precision \text{ kelas } 1 = \left(\frac{17156}{17156+4245}\right) = 0.80$$

$$precision \text{ kelas } 2 = \left(\frac{12}{12+461}\right) = 0.025$$

$$precision \text{ kelas } 3 = \left(\frac{1984}{1984+1510}\right) = 0.57$$

Jadi rata-rata Precision: (0.8 + 0.025+ 0.57) /3 = 0.46 * 100 = 46%

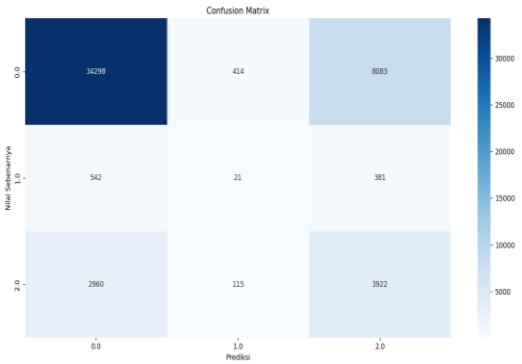
$$Recall \text{ kelas } 1 = \left(\frac{17156}{17156+1707}\right) = 0.91$$

$$Recall \text{ kelas } 2 = \left(\frac{12}{12+284}\right) = 0.04$$

$$Recall \text{ kelas } 3 = \left(\frac{1984}{1984+4225}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.04+ 0.32) /3 = 0.42 * 100 = 42%

2. Pengujian Tanpa Seleksi Fitur Perbandingan 80:20
   Pengujian ini akan dilakukan proses klasifikasi dengan 80% data training dan 20% data testing. Pada gambar 6 akan ditampilkan *Heat Map* dari Confusion matrix hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton.



Gambar 6. *Heat Map Confusion Matrix* Tanpa Seleksi Fitur 80:20.

Dari gambar 6 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$Akurasi = \left(\frac{34298+21+3922}{50736}\right) * 100\ \% = 75.37\%$$

$$precision \text{ kelas } 1 = \left(\frac{34298}{8497+34298}\right) = 0.80$$

$$precision \text{ kelas } 2 = \left(\frac{21}{21+923}\right) = 0.02$$

$$precision \text{ kelas } 3 = \left(\frac{3922}{3922+3075}\right) = 0.56$$

Jadi rata-rata Precision: (0.8 + 0.02 + 0.56) /3 = 0.46

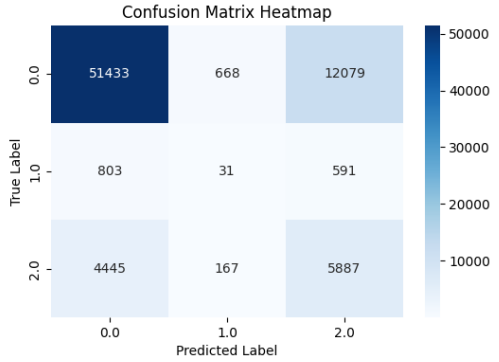
$$Recall \text{ kelas } 1 = \left(\frac{34298}{3502+34298}\right) = 0.91$$

$$Recall\ kelas\ 2 = \left(\frac{21}{21+529}\right) = 0.038$$

$$Recall\ kelas\ 3 = \left(\frac{3922}{3922+8464}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.038+ 0.32) /3 = 0.42

3. Pengujian Tanpa Seleksi Fitur Perbandingan 70:30
Pengujian ini akan dilakukan proses klasifikasi dengan 70% data training dan 30% data testing. Pada gambar 7 akan ditampilkan *Heat Map* dari Confusion matrix hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 7. *Heat Map Confusion Matrix* Tanpa Seleksi Fitur 70:30.

Dari gambar 7 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{51433+31+5887}{76104}\right) * 100\ \% = 75.36\%$$

$$precision\ \text{kelas}\ 1 = \left(\frac{51433}{51433+12747}\right) = 0.80$$

$$precision\ \text{kelas}\ 2 = \left(\frac{31}{31+1394}\right) = 0.02$$

$$precision\ \text{kelas}\ 3 = \left(\frac{5887}{5887+4612}\right) = 0.56$$

Jadi rata-rata Precision: (0.8 + 0.02+ 0.56) /3 = 0.46

$$Recall\ \text{kelas}\ 1 = \left(\frac{51433}{51433+5248}\right) = 0.91$$

$$Recall\ \text{kelas}\ 2 = \left(\frac{31}{31+835}\right) = 0.04$$

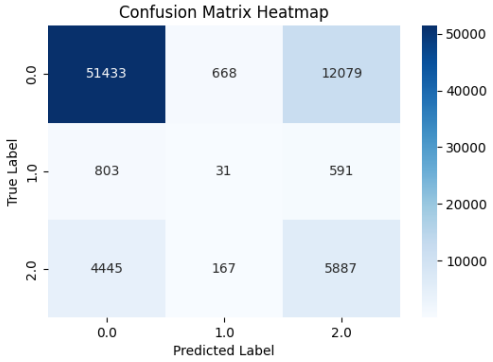
$$Recall\ \text{kelas}\ 3 = \left(\frac{5887}{5887+4225}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.04+ 0.32) /3 = 0.42

B. Hasil Pengujian Algoritma Klasifikasi Naïve Bayes Pengurangan Satu Fitur

Hasil pengujian ini dilakukan dengan menghilangkan satu fitur berdasarkan tabel 4 yang berada di peringkat terakhir, dimana peringkat terakhir yaitu fitur *HvyAlcoholConsump*, sehingga akan digunakan 20 fitur untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 4 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

1. Pengujian Pengurangan 1 Fitur Perbandingan 90:10
Pengujian akan dilakukan dengan mengurangi 1 fitur dengan menggunakan perbandingan 90:10. Pada gambar 8 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 8. *Heat Map Confusion Matrix* Pengurangan Satu Fitur 70:30.

Dari gambar 8 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{17324+11+1916}{25368}\right) * 100\ \% = 75.89\%$$

$$precision\ \text{kelas}\ 1 = \left(\frac{17324}{17324+4077}\right) = 0.81$$

$$precision\ \text{kelas}\ 2 = \left(\frac{11}{11+462}\right) = 0.02$$

$$precision\ \text{kelas}\ 3 = \left(\frac{1916}{1916+1578}\right) = 0.55$$

Jadi rata-rata Precision: (0.81 + 0.02+ 0.55) /3 = 0.46

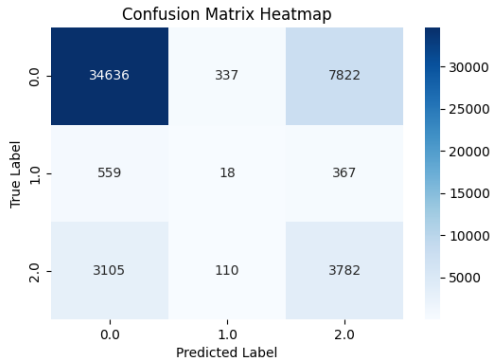
$$Recall\ \text{kelas}\ 1 = \left(\frac{17324}{17324+1791}\right) = 0.91$$

$$Recall\ \text{kelas}\ 2 = \left(\frac{11}{11+230}\right) = 0.05$$

$$Recall\ \text{kelas}\ 3 = \left(\frac{1916}{1916+4096}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.05+ 0.32) /3 = 0.43

2. Pengujian Pengurangan 1 Fitur Perbandingan 80:20
Pengujian akan dilakukan dengan mengurangi 1 fitur dengan menggunakan perbandingan 80:20. Pada gambar 9 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 9. *Heat Map Confusion Matrix* Pengurangan Satu Fitur 80:20.

Dari gambar 9 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{34636+18+3782}{50736}\right) * 100\ \% = 75.76\%$$

$$precision \text{ kelas } 1 = \left(\frac{34636}{34636+8159}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{18}{18+926}\right) = 0.02$$

$$precision \text{ kelas } 3 = \left(\frac{3782}{3782+3215}\right) = 0.54$$

Jadi rata-rata Precision: (0.81 + 0.02+ 0.54) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{34636}{34636+3664}\right) = 0.90$$

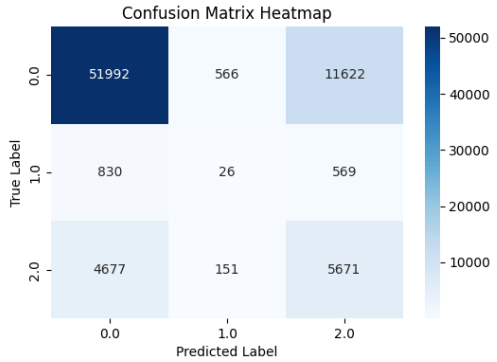
$$Recall \text{ kelas } 2 = \left(\frac{18}{18+447}\right) = 0.04$$

$$Recall \text{ kelas } 3 = \left(\frac{3782}{3782+8189}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.05+ 0.32) /3 = 0.42

3. Pengujian Pengurangan 1 Fitur Perbandingan 70:30

Pengujian akan dilakukan dengan mengurangi 1 fitur dengan menggunakan perbandingan 70:30. Pada gambar 10 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 10. *Heat Map Confusion Matrix* Pengurangan Satu Fitur 70:30.

Dari gambar 10 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{51992+26+5671}{76104}\right) * 100\ \% = 75.8\%$$

$$precision \text{ kelas } 1 = \left(\frac{51992}{51992+12188}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{26}{26+1399}\right) = 0.02$$

$$precision \text{ kelas } 3 = \left(\frac{5671}{5671+4828}\right) = 0.54$$

Jadi rata-rata Precision: (0.81 + 0.02+ 0.54) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{51992}{51992+5507}\right) = 0.90$$

$$Recall \text{ kelas } 2 = \left(\frac{26}{26+717}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{5671}{5671+12191}\right) = 0.32$$

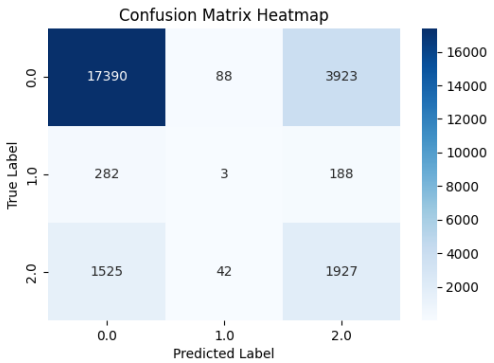
Jadi rata-rata Recall : (0.91 + 0.05+ 0.32) /3 = 0.42

C. Hasil Pengujian Algoritma Klasifikasi Naïve Bayes Pengurangan Dua Fitur

Hasil pengujian ini dilakukan dengan menghilangkan satu fitur berdasarkan tabel 4 yang berada di peringkat terakhir, dimana peringkat terakhir yaitu fitur *HvyAlcoholConsump* dan *NoDocbcCost*, sehingga akan digunakan 19 fitur untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 4 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

1. Pengujian Pengurangan 2 Fitur Perbandingan 90:10

Pengujian akan dilakukan dengan mengurangi 2 fitur dengan menggunakan perbandingan 90:10. Pada gambar 11 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 11. *Heat Map Confusion Matrix* Pengurangan Dua Fitur 90:10.

Dari gambar 11 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{17390+3+1927}{25368}\right) * 100\ \% = 76.16\%$$

$$precision \text{ kelas } 1 = \left(\frac{17390}{17390+4011}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{3}{3+470}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{1927}{1927+1567}\right) = 0.55$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.55) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{17390}{17390+1807}\right) = 0.91$$

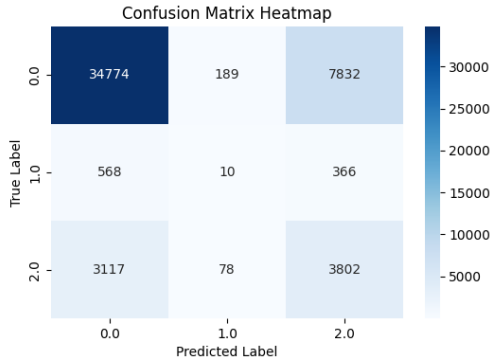
$$Recall \text{ kelas } 2 = \left(\frac{3}{3+130}\right) = 0.02$$

$$Recall \text{ kelas } 3 = \left(\frac{1927}{1927+4111}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.02+ 0.32) /3 = 0.42

2. Pengujian Pengurangan 2 Fitur Perbandingan 80:20

Pengujian akan dilakukan dengan mengurangi 2 fitur dengan menggunakan perbandingan 80:20. Pada gambar 12 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 12. *Heat Map Confusion Matrix* Pengurangan Dua Fitur 80:20.

Dari gambar 12 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{34774+10+3802}{50736}\right) * 100\ \% = 76.05\%$$

$$precision \text{ kelas } 1 = \left(\frac{34774}{34774+8021}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{10}{10+934}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{3802}{3802+3195}\right) = 0.54$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.54) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{34774}{34774+3685}\right) = 0.90$$

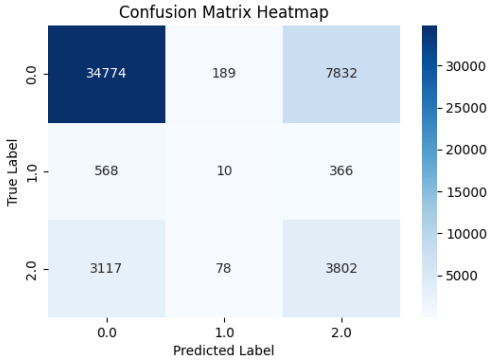
$$Recall \text{ kelas } 2 = \left(\frac{10}{10+267}\right) = 0.04$$

$$Recall \text{ kelas } 3 = \left(\frac{3802}{3802+8198}\right) = 0.32$$

Jadi rata-rata Recall : (0.9 + 0.04+ 0.32) /3 = 0.42

3. Pengujian Pengurangan 2 Fitur Perbandingan 70:30

Pengujian akan dilakukan dengan mengurangi 2 fitur dengan menggunakan perbandingan 70:30. Pada gambar 13 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 13. *Heat Map Confusion Matrix* Pengurangan Dua Fitur 80:20.

Dari gambar 13 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{52175+14+5707}{76104}\right) * 100\ \% = 76.07\%$$

$$precision \text{ kelas } 1 = \left(\frac{52175}{52175+12005}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{14}{14+1411}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{5707}{5707+4792}\right) = 0.54$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.54) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{52175}{52175+5525}\right) = 0.90$$

$$Recall \text{ kelas } 2 = \left(\frac{14}{14+429}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{5707}{5707+12254}\right) = 0.32$$

Jadi rata-rata Recall : (0.9 + 0.02+ 0.32) /3 = 0.42

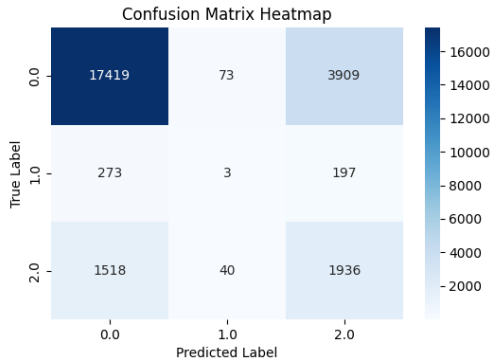
D. Hasil Pengujian Algoritma Klasifikasi Naïve Bayes Pengurangan Tiga Fitur

Hasil pengujian ini dilakukan dengan menghilangkan tiga fitur berdasarkan tabel 4 yang berada di 3 peringkat terakhir, dimana 3 peringkat terakhir yaitu fitur *HvyAlcoholConsump, NoDocbcCost, dan MentHlth* sehingga akan digunakan 18 fitur untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 3 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

1. Pengujian Pengurangan 3 Fitur Perbandingan 90:10

Pengujian akan dilakukan dengan mengurangi 3 fitur dengan menggunakan perbandingan 90:10. Pada gambar 14 akan

ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 14. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 90:10.

Dari gambar 14 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{17419+3+1936}{25368}\right) * 100\ \% = 76.31\%$$

$$precision \text{ kelas } 1 = \left(\frac{17419}{17419+3982}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{3}{3+470}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{1936}{1936+1558}\right) = 0.55$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.55) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{17419}{17419+1791}\right) = 0.91$$

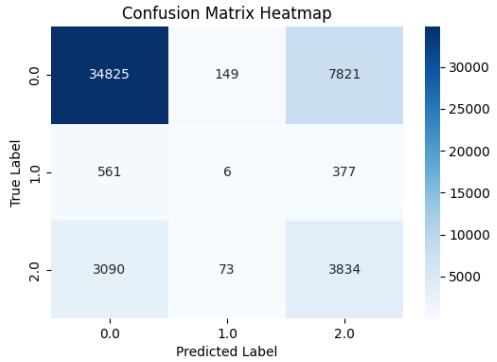
$$Recall \text{ kelas } 2 = \left(\frac{3}{3+113}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{1936}{1936+4106}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.02+ 0.32) /3 = 0.42

2. Pengujian Pengurangan 3 Fitur Perbandingan 80:20

Pengujian akan dilakukan dengan mengurangi 3 fitur dengan menggunakan perbandingan 80:20. Pada gambar 15 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 15. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 80:20.

Dari gambar 15 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{34825+6+3834}{50736}\right) * 100\ \% = 76.21\%$$

$$precision \text{ kelas } 1 = \left(\frac{34825}{34825+7970}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{6}{6+938}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{3834}{3834+3163}\right) = 0.55$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.55) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{34825}{34825+3651}\right) = 0.91$$

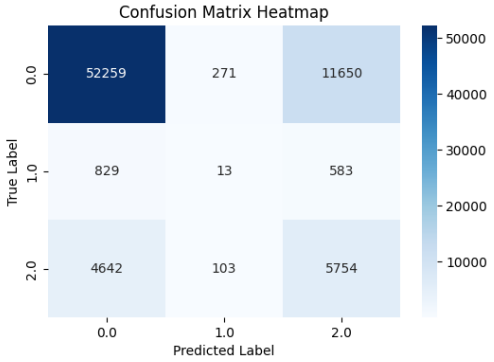
$$Recall \text{ kelas } 2 = \left(\frac{6}{6+222}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{3834}{3834+8198}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.02+ 0.32) /3 = 0.42

3. Pengujian Pengurangan 3 Fitur Perbandingan 70:30

Pengujian akan dilakukan dengan mengurangi 3 fitur dengan menggunakan perbandingan 70:30. Pada gambar 16 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 16. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 70:30.

Dari gambar 16 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{52259+13+5754}{76104}\right) * 100\ \% = 76.25\%$$

$$precision \text{ kelas } 1 = \left(\frac{52259}{52259+11921}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{13}{13+1412}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{5754}{5754+4745}\right) = 0.55$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.55) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{52259}{52259+5471}\right) = 0.91$$

$$Recall \text{ kelas } 2 = \left(\frac{13}{13+374}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{5754}{5754+12233}\right) = 0.32$$

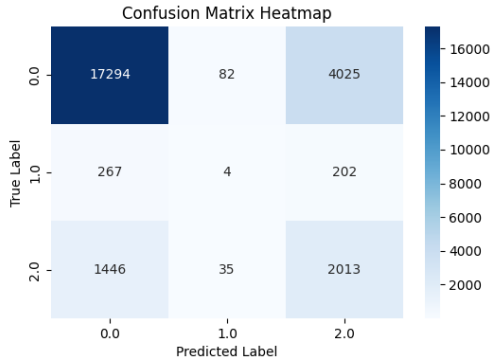
Jadi rata-rata Recall : (0.91 + 0.01+ 0.32) /3 = 0.42

E. Hasil Pengujian Algoritma Klasifikasi Naïve Bayes Pengurangan Tiga Fitur

Hasil pengujian ini dilakukan dengan menghilangkan empat fitur berdasarkan tabel 4 yang berada di 4 peringkat terakhir, dimana 4 peringkat terakhir yaitu fitur *HvyAlcoholConsump*, *NoDocbcCost*, *MentHlth*, dan *Stroke*. Sehingga akan digunakan 17 fitur untuk proses klasifikasi dengan menggunakan Algoritma Naïve Bayes.Proses pengujian ini akan terbagi menjadi 3 kali uji, yaitu pengujian dengan perbandingan 90:10, 80:20, dan 70:30.

1. Pengujian Pengurangan 4 Fitur Perbandingan 90:10

Pengujian akan dilakukan dengan mengurangi 4 fitur dengan menggunakan perbandingan 90:10. Pada gambar 17 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 17. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 90:10.

Dari gambar 17 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{17294+4+2013}{25368}\right) * 100\ \% = 76.12\%$$

$$precision \text{ kelas } 1 = \left(\frac{17294}{17294+4107}\right) = 0.81\ * 100$$

$$precision \text{ kelas } 2 = \left(\frac{4}{4+469}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{2013}{2013+1481}\right) = 0.58$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.58) /3 = 0.46 * 100 = 46%

$$Recall \text{ kelas } 1 = \left(\frac{17294}{17294+1713}\right) = 0.91$$

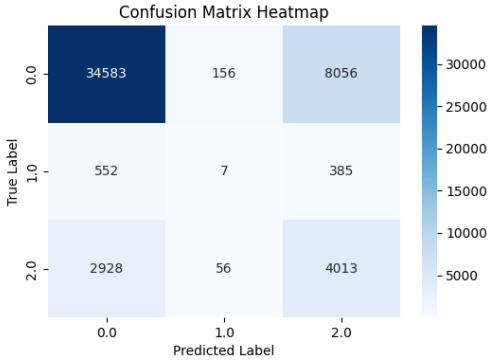
$$Recall \text{ kelas } 2 = \left(\frac{4}{4+117}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{2013}{2013+4227}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.03+ 0.32) /3 = 0.42 * 100 = 42%

2. Pengujian Pengurangan 4 Fitur Perbandingan 80:20

Pengujian akan dilakukan dengan mengurangi 4 fitur dengan menggunakan perbandingan 80:20. Pada gambar 18 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 18. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 80:20.

Dari gambar 18 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{34583+7+4013}{50736}\right) * 100\ \% = 76.09\%$$

$$precision \text{ kelas } 1 = \left(\frac{34583}{34583+8212}\right) = 0.81$$

$$precision \text{ kelas } 2 = \left(\frac{7}{7+937}\right) = 0.01$$

$$precision \text{ kelas } 3 = \left(\frac{4013}{4013+2984}\right) = 0.57$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.57) /3 = 0.46

$$Recall \text{ kelas } 1 = \left(\frac{34583}{34583+3480}\right) = 0.91$$

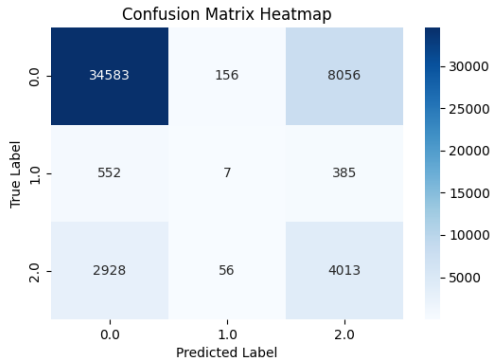
$$Recall \text{ kelas } 2 = \left(\frac{7}{7+212}\right) = 0.03$$

$$Recall \text{ kelas } 3 = \left(\frac{4013}{4013+8441}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.03+ 0.32) /3 = 0.42

3. Pengujian Pengurangan 4 Fitur Perbandingan 70:30

Pengujian akan dilakukan dengan mengurangi 4 fitur dengan menggunakan perbandingan 70:30. Pada gambar 19 akan ditampilkan *Heat Map* dari *Confusion matrix* hasil klasifikasi yang didapatkan dari bahasa pemrograman Pyhton



Gambar 19. *Heat Map Confusion Matrix* Pengurangan Tiga Fitur 70:30.

Dari gambar 19 dapat dicari nilai akurasi, *precision*, dan *recall* di bawah ini

$$\text{Akurasi} = \left(\frac{51881+12+5998}{76104}\right) * 100\ \% = 76.07\%$$

$$\textit{precision}\ \text{kelas}\ 1 = \left(\frac{51881}{51881+12299}\right) = 0.81$$

$$\textit{precision}\ \text{kelas}\ 2 = \left(\frac{12}{12+1413}\right) = 0.01$$

$$\textit{precision}\ \text{kelas}\ 3 = \left(\frac{5998}{5998+4501}\right) = 0.57$$

Jadi rata-rata Precision: (0.81 + 0.01+ 0.57) /3 = 0.46

$$\textit{Recall}\ \text{kelas}\ 1 = \left(\frac{51881}{51881+5218}\right) = 0.91$$

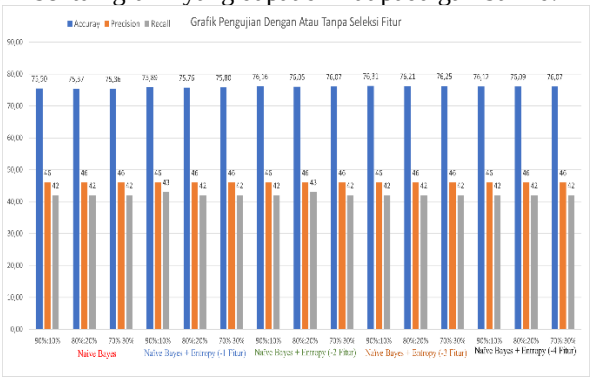
$$\textit{Recall}\ \text{kelas}\ 2 = \left(\frac{12}{12+385}\right) = 0.03$$

$$\textit{Recall}\ \text{kelas}\ 3 = \left(\frac{5998}{5998+12610}\right) = 0.32$$

Jadi rata-rata Recall : (0.91 + 0.03+ 0.32) /3 = 0.42

F. Hasil Rekap Pengujian Klasifikasi Dengan Seleksi Fitur Atau Tanpa Seleksi Fitur

Dari beberapa pengujian sebelumnya akan ditampilkan dalam bentuk grafik yang dapat dilihat pada gambar 20.



Gambar 20. Grafik Pengujian Klasifikasi Dengan Atau Tanpa Seleksi Fitur

Dari gambar 20 terlihat hasil klasifikasi untuk dataset penyakit diabetes apabila dilakukan pengujian tanpa proses seleksi fitur dengan menghasilkan akurasi sebesar 75.5% untuk perbandingan 90% data *training* dan 10 % data *testing*. Untuk Pengujian dengan perbandingan 80% data *training* dan data *testing* sebesar 20% mendapatkan nilai akurasi 75.37% dan pengujian terakhir dengan 70% data *training* dan 30% data *testing* menghasilkan nilai akurasi 75.36% sehingga menghasilkan rata-rata nilai akurasi sebesar 75.41%.

Pengujian selanjutnya adalah dengan menghilangkan satu fitur setelah dilakukan proses seleksi fitur dengan menggunakan metode Entropy. Fitur yang dihapus adalah fitur *HvyAlcoholConsump* sehingga menghasilkan akurasi sebesar 75.89% untuk perbandingan 90% data *training* dan 10 % data *testing*. Untuk Pengujian dengan perbandingan 80% data *training* dan data *testing* sebesar 20% mendapatkan nilai akurasi 75.76% dan pengujian terakhir dengan 70% data *training* dan 30% data *testing* menghasilkan nilai akurasi 75.8% sehingga menghasilkan rata-rata nilai akurasi sebesar 75.71%. Hasil rata-rata dengan menghilangkan 1 fitur memiliki nilai akurasi lebih tinggi dengan selisih 0.29% daripada pengujian klasifikasi tanpa proses seleksi.

Pengujian selanjutnya adalah dengan menghilangkan dua fitur setelah dilakukan proses seleksi fitur dengan menggunakan metode Entropy. Fitur yang dihapus adalah fitur *HvyAlcoholConsump* dan *NoDocbcCost* sehingga menghasilkan akurasi sebesar 76.16% untuk perbandingan 90% data *training* dan 10 % data *testing*. Untuk Pengujian dengan perbandingan 80% data *training* dan data *testing* sebesar 20% mendapatkan nilai akurasi 76.05% dan pengujian terakhir dengan 70% data *training* dan 30% data *testing* menghasilkan nilai akurasi 76.07% sehingga menghasilkan rata-rata nilai akurasi sebesar 76.09%. Hasil rata-rata dengan menghilangkan 2 fitur memiliki nilai akurasi lebih tinggi dengan selisih 0.68% daripada pengujian klasifikasi tanpa proses seleksi.

Pengujian selanjutnya adalah dengan menghilangkan tiga fitur setelah dilakukan proses seleksi fitur dengan menggunakan metode Entropy. Fitur yang dihapus adalah fitur *HvyAlcoholConsump*, *NoDocbcCost*, dan *MentHlth* sehingga menghasilkan akurasi sebesar 76.31% untuk perbandingan 90% data *training* dan 10 % data *testing*. Untuk Pengujian dengan perbandingan 80% data *training* dan data *testing* sebesar 20% mendapatkan nilai akurasi 76.21% dan pengujian terakhir dengan 70% data *training* dan 30% data *testing* menghasilkan nilai akurasi 76.07% sehingga menghasilkan rata-rata nilai akurasi sebesar 76.26%. Hasil rata-rata dengan menghilangkan 3 fitur memiliki nilai akurasi lebih tinggi dengan selisih 0.85% daripada pengujian klasifikasi tanpa proses seleksi.

Pengujian selanjutnya adalah dengan menghilangkan empat fitur setelah dilakukan proses seleksi fitur dengan menggunakan metode Entropy. Fitur yang dihapus adalah fitur *HvyAlcoholConsump* ,*NoDocbcCost*, *MentHlth*, dan *Stroke* sehingga menghasilkan akurasi sebesar 76.12% untuk perbandingan 90% data *training* dan 10 % data *testing*. Untuk Pengujian dengan perbandingan 80% data *training* dan data *testing* sebesar 20% mendapatkan nilai akurasi 76.09% dan pengujian terakhir dengan 70% data *training* dan 30% data *testing* menghasilkan nilai akurasi 76.07% sehingga menghasilkan rata-rata nilai akurasi sebesar 76.09%. Hasil rata-rata dengan menghilangkan 4 fitur memiliki nilai akurasi lebih tinggi dengan selisih 0.68% daripada pengujian klasifikasi tanpa proses seleksi.

IV. KESIMPULAN

Berdasarkan hasil yang diperoleh dari hasil pengujian yang telah dilakukan pada penelitian ini dapat disimpulkan bahwa Ketika menggunakan seleksi fitur dengan menggunakan metode Entropy terbukti dapat meningkatkan nilai akurasi disbanding dengan menggunakan proses klasifikasi dengan menggunakan Algoritma Naïve Bayes konvensional.

Dari proses seleksi fitur sendiri, 4 nilai fitur yang memiliki nilai entropy paling terkecil adalah sebagai berikut HvyAlcoholConsump, NoDocbcCost, MentHlth, dan Stroke sehingga 4 fitur ini tidak digunakan untuk proses klasifikasi dan terbukti dari pengujian apabila setiap fitur dihilangkan nilai akurasi meningkat dan proses peningkatan tertinggi apabila 3 fitur dihilangkan nilai akurasi memiliki peningkatan dengan selisih nilai akurasi sebesar 0.85.

Berbeda halnya dengan nilai akurasi, untuk nilai precision dan recall, perbedaan nilai apabila dengan menggunakan seleksi dan tanpa seleksi fitur tidak memiliki perbedaan nilai yang signifikan seperti halnya dengan nilai akurasi.

REFERENSI

[1] F. D. T. A. Latifah Uswatun Khasanah, Yuki Novia Nasution, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier," *BASIS*, vol. 1, no. 1, pp. 41–50, 2022, doi: 10.32528/justindo.v7i1.4949.

[2] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.

[3] S. H. A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2546–2554, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[4] V. S. M. P. Paryoko, "Seleksi Fitur Pada Klasifikasi Multi-label Menggunakan Proportional Feature Rough Selector," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 4, pp. 2084–2094, 2021, doi: 10.35957/jatisi.v8i4.1259.

[5] A.- Meiriza, E. Lestari, and R. Zulfahmi, "Implementasi Metode Entropy Dan Technique For Order Preference By Similarity To Ideal Solution (Topsis) Dalam Pemilihan Biro Perjalanan Umroh," *JSI J. Sist. Inf.*, vol. 11, no. 1, pp. 77–90, 2019, doi: 10.36706/jsi.v11i1.7686.

[6] C. E. Prawiro, M. Y. H. Setyawan, and S. F. Pane, "Studi Komparasi Metode Entropy dan ROC dalam Menentukan Bobot Kriteria," *J. Tekno Insentif*, vol. 15, no. 1, pp. 1–14, 2021, doi: 10.36787/jti.v15i1.353.

[7] F. A. Andi Akram, Nur Risal, Nur Inayah Yusuf, Andi Baso Kaswar, "Penerapan Data Mining dalam Mengklasifikasikan Tingkat Kasus Covid-19 di Sulawesi Selatan Menggunakan Algoritma Naive Bayes," *Indones. J. Fundam. Scienes*, vol. 7, no. 1, pp. 18–28, 2021.