Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Implementasi Algoritma Naïve Bayes Menggunakan Pemilihan Atribut Information Gain Pada Penyakit Diabetes

Implementation Of The Naïve Bayes Algorithm Using Information Gain Attribute Selection In Diabetes Disease

Edwin Arizal Mandalika[1], Moh. Dasuki[2], Dr. Reni Umilasari,[3]
[1]Mahasiswa Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
Email: eamandalika@gmail.com
[2]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
Email: moh.dasuki22@unmuhjember.ac.id
[3]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
Email: reni.umilasari@unmuhjember.ac.id

Abstrak

Diabetes adalah penyakit kronis yang menimbulkan gangguan metabolisme dan dikenali melalui kadar gula darah yang tinggi. Dengan memanfaatkan teknologi informasi dan komunikasi maka saat ini proses diagnosis secara dini penyakit diabetes dapat dilakukan dengan cara mengolah *dataset* rekam medis. Tujuan dari klasifikasi adalah untuk mengidentifikasi pola atau hubungan di antara data yang ada sehingga dapat dikategorikan ke dalam kelas yang tepat. Algoritma *Naïve Bayes* merupakan salah satu algoritma yang biasa diterapkan dalam klasifikasi untuk menentukan kelas data. Algoritma ini melakukan klasifikasi probabilistik yang digunakan untuk menentukan kelas suatu item berdasarkan beberapa Atribut. Optimasi algoritma *Naïve Bayes* dengan teknik pemilihan atribut agar tingkat akurasi pemodelan yang didapatkan dapat lebih maksimal. Salah satu teknik pemilihan atribut yang terbukti meningkatkan akurasi adalah *Information Gain*. Data yang digunakan yaitu dataset penyakit diabetes yang berasal dari Kaggle berjumlah 70.692 data meliputi 35.346 record orang yang terkena diabetes dan 35.346 record orang tidak terkena diabetes. Hasil dari penelitian dengan menggunakan metode Naïve Bayes dengan k-fold cross validation menunjukkan hasil yang cukup baik dengan akurasi tertinggi sebesar 73% pada data uji. Dengan pemilihan atribut *Information Gain* menghasilkan akurasi tertinggi sebesar 74% pada uji.
Kata Kunci: Diabetes, *Naïve Bayes, Information Gain*

Diabetes is a chronic disease that causes metabolic disorders and is recognized by high blood sugar levels. By using information and communication technology, currently the process of early diagnosis of diabetes can be carried out by processing medical record datasets. The purpose of classification is to identify patterns or relationships between existing data so that it can be categorized into appropriate classes. The Naïve Bayes algorithm is an algorithm that is usually applied in classification to determine data classes. This algorithm performs probabilistic classification which is used to determine the class of an item based on several attributes. Optimization of the Naïve Bayes algorithm using attribute selection techniques so that the level of modeling accuracy obtained can be maximized. One attribute selection technique that has been proven to increase accuracy is Information Gain. The data used is a diabetes dataset originating from Kaggle totaling 70,692 data including 35,346 records of people who have diabetes and 35,346 records of people who do not have diabetes. The results of research using the Naïve Bayes method with k-fold cross validation show quite good results with the highest accuracy of 73% on test data. Selecting the Information Gain attribute produces the highest accuracy of 74% in the test.
Keywords: *Diabetes, Naïve Bayes, Information Gain*

Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

1. PENDAHULUAN

Diabetes adalah penyakit kronis yang menimbulkan gangguan metabolisme dan dikenali melalui kadar gula darah yang tinggi. Ini menjadi penyebab utama masalah jantung, kebutaan dan kegagalan ginjal. (Kementerian Kesehatan RI., 2020). Adapun teknik diagnosis diabetes yang sering dilakukan adalah dengan pengukuran tingkat glukosa dalam darah. Dimana apabila konsentrasi glukosa dalam darah melampaui batas normal maka orang tersebut akan dikategorikan sebagai penderita diabetes. Namun pada penyakit diabetes terdapat fase asimtomatik yang cukup lama yaitu kondisi dimana penyakit sudah terdeteksi tetapi belum menunjukkan gejala klinis pada pasien (Karyadiputra & Setiawan, 2022)

Klasifikasi data adalah proses pengelompokkan data ke dalam kategori atau kelas yang telah ditentukan sebelumnya. Tujuan utama dari klasifikasi adalah untuk mengidentifikasi pola atau hubungan di antara data yang ada sehingga dapat dikategorikan ke dalam kelas yang tepat. Algoritma *Naïve Bayes* merupakan salah satu algoritma yang biasa diterapkan dalam klasifikasi untuk menentukan kelas data. Algoritma ini melakukan klasifikasi probabilistik yang digunakan untuk menentukan kelas suatu item berdasarkan beberapa Atribut yang melekat pada item tersebut.

Maka dari hal tersebut perlu adanya optimasi algoritma *Naïve Bayes* dengan teknik pemilihan atribut agar tingkat akurasi pemodelan yang didapatkan dapat lebih maksimal. Salah satu teknik pemilihan atribut yang terbukti meningkatkan akurasi adalah *Information Gain*. *Information Gain* adalah sebuah teknik seleksi atribut yang mengeliminasi atribut yang tidak berpengaruh dan sering digunakan untuk mendapatkan informasi mengenai atribut mana yang paling berpengaruh terhadap suatu kelas dengan cara merangking (Setiyorini & Asmono, 2019). Kelebihan teknik pemilihan atribut dengan *Information Gain* yaitu baik digunakan dalam memilih atribut khususnya dalam menangani data dengan dimensi tinggi (Pratikaningtyas et al., 2019). Namun kelemahannya tidak efektif dalam mengatasi data dengan atribut yang memiliki nilai yang hilang.

2. TINJAUAN PUSTAKA

A. Penyakit Diabetes

Menurut (Mustafa & Simpen, 2019) bahwa penyakit diabetes yang juga dikenal sebagai penyakit gula adalah kondisi kronis yang berkembang sebagai akibat dari anomali sekresi insulin dalam peningkatan kadar glukosa yang tidak menentu. Diabetes menyebabkan kenaikan persentase gula pada darah yang meningkatkan risiko konsekuensi seperti penyakit stroke, jantung, kebutaan, gagal ginjal dan kematian. Adapun ia menyebutkan pengertian lain bahwa diabetes adalah sejenis penyakit yang mematikan dan memperburuk kondisi kadar gula darah.

B. Klasifikasi

Menurut (Wijaya & Dwiasnati, 2020) proses identifikasi suatu fungsi atau model yang dapat menggambarkan dan membedakan data menjadi berbagai kelas atau konsep. Proses ini melibatkan evaluasi terhadap karakteristik dari objek dan menentukan objek pada kelas yang sudah ditetapkan sebelumnya. Contohnya seorang calon penerima beasiswa dikategorikan sebagai layak menerima atau tidak layak menerima. Lalu klasifikasi adalah suatu metode pembelajaran untuk memprediksi nilai dari sekelompok attribut dalam menggambarkan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Rahayuningsih, 2019). Kemudian menurut (Febriani & Sulistiani, 2021) bahwa klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

C. *Naïve Bayes*

Naïve Bayes adalah algoritma yang memakai metode probabilitas atau kemungkinan dan statistik. Algoritma ini memiliki cara yang sangat sederhana untuk mengklasifikasikan data dengan asumsi

Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

sederhana tentang atribut klasifikasi. *Naïve Bayes* sering digunakan dalam memecahkan masalah dalam belajar mesin dan dikenal mempunyai akurasi yang lebih baik meskipun dengan menggunakan perhitungan yang sederhana (Hayuningtyas, 2019). *Naïve Bayes* memiliki kelebihan di mana hanya memerlukan jumlah data pelatihan yang kecil untuk menetapkan perkiraan parameter dalam proses penggolongan kemudian mudah diimplementasikan dan dapat memberikan hasil yang baik pada banyak kasus.

Adapun Rumus persamaan *Naïve Bayes* yaitu sebagai berikut:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Keterangan:

X = Data Uji
H = Hipotesis bahwa data *X* termasuk ke dalam kelas tertentu
$P(H|X)$ = Probabilitas hipotesis *H* dengan acuan *X*
$P(H)$ = Probabilitas anggapan dasar *H*
$P(X|H)$ = Probabilitas Data *X* dengan acuan Kondisi Hipotesis *H*
$P(X)$ = Probabilitas Data *X*

D. Information Gain

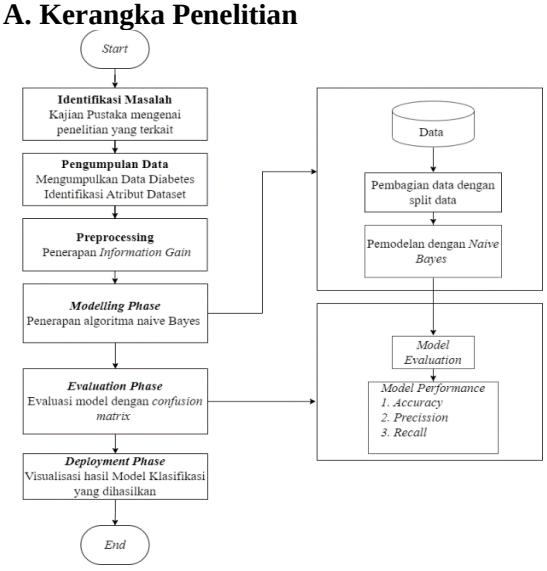
Information Gain biasa dipakai untuk menentukan urutan atribut yang paling mempengaruhi kelas. Untuk menghitung nilai *gain* atribut pertama-tama dihitung *entropi* total dari target variabel (Setiyorini & Asmono, 2019). Kemudian *entropi* dari setiap nilai atribut dikalkulasikan dan dikurangkan dari *entropi* total. Hasil dari operasi ini adalah *gain entropi* dari setiap atribut. Atribut dengan nilai *gain entropi* tertinggi menunjukkan bahwa atribut tersebut memiliki informasi terbanyak untuk memprediksi target variabel. Kelebihan teknik pemilihan atribut dengan *Information Gain* yaitu adalah baik digunakan dalam memilih atribut khususnya dalam menangani data dengan dimensi tinggi (Pratikaningtyas et al., 2019). Adapun kelemahannya tidak efektif dalam mengatasi data dengan atribut yang memiliki nilai yang hilang dan memilih atribut dengan banyak kategori dapat membuat proses klasifikasi lambat.

E. Confusion Matrix

Menurut (Hasanah et al., 2019) *Confussion matrix* adalah metode penghitungan yang membandingkan hasil klasifikasi dengan data sebenarnya. Matrik ini menunjukkan tingkat akurasi dalam persentase dan berguna sebagai acuan untuk menilai performa algoritma klasifikasi.

3. METODE PENELITIAN

A. Kerangka Penelitian



Gambar 1. Kerangka Penelitian
Sumber: (pemikiran sendiri)

Adapun penjelasan dari prosedur penelitian yang terdapat pada Kerangka Penelitian pada Gambar 1. yaitu:

1. Identifikasi Masalah: Pada tahapan pertama penelitian ini dilakukan identifikasi masalah dari Kajian *library* yaitu aktivitas berupa pengumpulan data penelitian sebelumnya yang bersangkutan dengan topik yang akan di angkat baik berupa jurnal buku dan lain lain.
2. Tahapan Pengumpulan Data: Tahapan selanjutnya adalah tahapan pemahaman data dengan langkah meliputi :
   a. Melakukan pencarian dan pengumpulan dataset penyakit diabetes dari *Kaggle*
   b. Identifikasi *dataset* yaitu kegiatan mengidentifikasi jenis atribut yang akan digunakan.

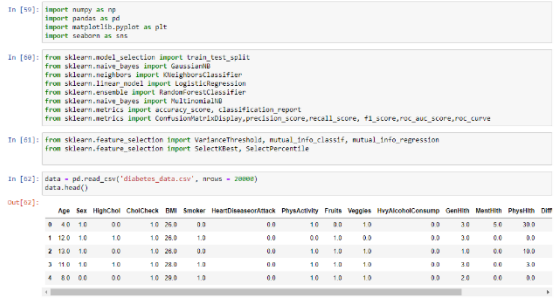
Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

3. Tahapan *Preprocessing:* Tahapan *Preprocessing* dalam penelitian ini menggunkan teknik *Information Gain* untuk pemilihan Atribut yang optimal dalam klasifikasi penyakit diabetes. Pada tahap ini atribut yang memiliki pengaruh signifikan terhadap klasifikasi akan dipilih dari dataset sementara atribut yang dianggap tidak relevan atau memiliki dampak yang minim akan dihilangkan.
4. Tahapan Pemodelan: Tahapan ini merupakan tahapan penerapan algoritma yang dipilih untuk melakukan pemodelan. Adapun Teknik pemodelan pada penelitian ini menerapkan algoritma *Naïve Bayes* dan seleksi Atribut *Information Gain*.
5. Tahapan Evaluasi (*Evaluation Phase)*: Pada tahapan ini melakukan evaluasi model apakah hasil pemodelan sudah sesuai dengan tujuan pada fase awal.
6. Tahapan Penyebaran *(Deployment Phase)*: Tahap terakhir adalah tahap penyebaran dimana pengetahuan berupa pemodelan klasifikasi yang dibuat pada proses *Data mining* akan divisualisasikan agar informasi yang didapatkan dapat lebih mudah dipahami.

B. Tahapan Pengumpulan

Pada tahapan ini dilakukan proses pengumpulan data yang akan digunakan

pada penelitian. Adapun data yang digunakan diperoleh dari Situs *Kaggle* yaitu *dataset* penyakit diabetes yang berjumlah 70.692 data meliputi 35.346 *record* orang yang terkena diabetes dan 35.346 *record* orang tidak terkena diabetes.



Gambar 2. Contoh Sebagaian Data
Sumber: (*python)*

C. Metode Yang Digunakan

Pemodelan dengan algoritma *Naïve Bayes* mempunyai kelemahan yaitu probabilitasnya tidak dapat memberikan informasi yang cukup mengenai tingkat akurasi klasifikasi. Selain itu keakuratan hasil klasifikasi dapat dipengaruhi oleh pemilihan atribut yang digunakan dalam data. Oleh karena itu diperlukan optimasi pada *Naïve Bayes Classifier* untuk mengatasi kelemahan tersebut. Dalam penelitian ini metode yang akan diusulkan untuk mengatasi kelemahan tersebut adalah dengan menerapkan teknik *Information Gain* untuk memilih atribut mana yang akan digunakan pada proses pemodelan. Adapun proses pertama dari *Information Gain* adalah dengan memilih Atribut-Atribut yang paling berpengaruh dan relevan dari data. Hasil dari *Information Gain* adalah Ranking atribut yang paling berpengaruh terhadap kelas data. Setelah proses *Information Gain* selesai maka selanjutnya adalah proses pemodelan menggunakan *Naïve Bayes* dengan hasil atribut yang dipilih menggunakan *Information Gain*.

D. Eksperimen dan Pengujian Metode

Pada tahapan eksperimen dan pengujian metode dilakukan proses percobaan pemodelan dengan pembagian data training dan data testing menggunakan *Split Data* dan evaluasi hasil pemodelan menggunakan *Confusion Matrix.*

E. Evaluasi Model

Pada evaluasi model pada penelitian ini akan menggunakan confusion matrix meliputi uji *Accuracy*, *Recall* dan *precision*. Adapun Hasil dari penelitian ini menghasilkan data keluaran hasil pemodelan klasifikasi. *dataset* data dengan algoritma *Naïve Bayes* dan optimasi *Information Gain* berupa hasil akurasi, *precission* dan *Recall*.

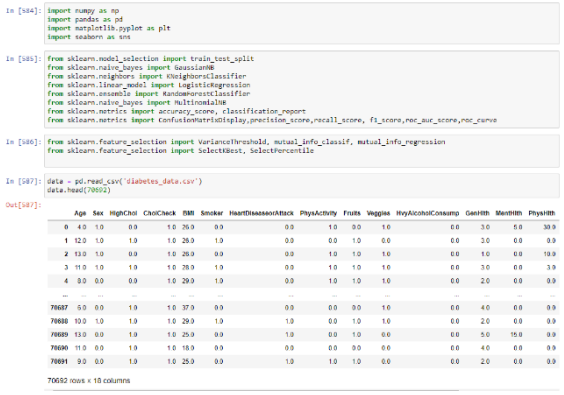
4. HASIL DAN PEMBAHASAN

A. Tahap *Import* Data

Tahap pertama yang dilakukan adalah memasukkan atau meng-*import* data ke *Jupyter Notebook*. Data yang diperoleh dari Situs *Kaggle* yaitu mengenai *dataset* penyakit

Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

diabetes yang berjumlah 70.692. Data meliputi 35.346 *record* orang yang terkena diabetes dan 35.346 *record* orang tidak terkena diabetes.



Gambar 3. *Import* Data dan Dataset
Sumber: (*python)*

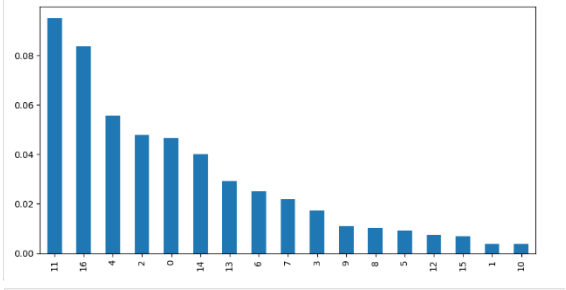
B. Menghitung *Information Gain* Pada *Python*

Proses perhitungan *Information Gain* dimulai dengan mengukur ketidakpastian (atau entropi) dari data pada label sebelum dan sesudah mengambil fitur tertentu. Semakin rendah entropi, semakin pasti dan informatif data tersebut. Proses ini dilakukan menggunakan *library sklearn.feature_selection* dengan modul *mutual_info_classif.*

```
In [108]: mi = mutual_info_classif(X_train_unique, y_train)

In [120]: len(mi)
Out[120]: 17

In [121]: mi
Out[121]: 11    0.095123
          16    0.083720
          4     0.055641
          2     0.047948
          0     0.046608
          14    0.040024
          13    0.029237
          6     0.025057
          7     0.022024
          3     0.017176
          9     0.011017
          8     0.010202
          5     0.009123
          12    0.007378
          15    0.006953
          1     0.003845
          10    0.003732
          dtype: float64

In [122]: mi = pd.Series(mi)
          mi.index = X_train_unique.columns

In [123]: mi.sort_values(ascending=False, inplace=True)
```

Gambar 4. *Information Gain* di *Python*
Sumber: *(python)*

Langkah terakhir adalah memvisualisasikan daftar atribut dengan nilai *gain*-nya menggunakan *library matplotlib.*



Gambar 5. *Visualisasi Information Gain*
Sumber: *(python)*

Library scikit-learn digunakan untuk melakukan seleksi atribut dengan memanfaatkan metode *mutual information*.

C. Klasifikasi *Naïve Bayes*

Tahap pelatihan dilakukan terhadap dataset normal dan dataset yang sudah diseleksi fiturnya menggunakan metode *Information Gain*. Pada dataset normal, seluruh fitur akan digunakan untuk diklasifikasi terhadap label. Sedangkan dataset yang sudah diseleksi fitur menggunakan 30% fitur dengan tingkat entropi terendah atau fitur yang memiliki tingkat korelasi dengan label tertinggi.

i. Split Data

Split data dilakukan dengan melibatkan pemecahan dataset menjadi subset atau sebagian data berdasarkan nilai fitur tertentu. *Split data* yang dimaksud pada tahap ini adalah pembagian atribut dan label. Atribut dan label masing masing disimpan pada variabel *X* dan *y*.

```
X = data.drop('Diabetes', axis = 1)
y = data['Diabetes']

X.shape, y.shape

((70692, 17), (70692,))
```

Gambar 6. Tahap *Split* Data
Sumber: *(python)*

ii. 2-*fold Cross Validation*

Berikut adalah penerapan algoritma *naïve bayes* menggunakan seleksi fitur *information gain* dengan k-*fold cross validation.*

Jurnal Smart Teknologi
Vol. 5, No. 4, Mei 2024, Halaman 430 – 438
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

```
# Implementasi Information Gain feature selection
percentile = 30  # persentil fitur yang digunakan
sel = SelectPercentile(score_func=mutual_info_classif, percentile=percentile)

# K-fold cv
stratified_kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

# model Naive Bayes
model = GaussianNB()

# Perform k-fold cross-validation
y_pred = cross_val_predict(model, sel.fit_transform(X, y), y, cv=stratified_kfold)

# Print hasil per fold
for fold, (train_idx, test_idx) in enumerate(stratified_kfold.split(X, y)):
    fold_y_true = y.iloc[test_idx]
    fold_y_pred = y_pred[test_idx]

    print(f"Fold {fold + 1}")
    print('Model performance for Cross-Validation set (Diabetes)')
    print("- Accuracy: {:.2f}".format(accuracy_score(fold_y_true, fold_y_pred)))
    print('- Precision: {:.2f}'.format(precision_score(fold_y_true, fold_y_pred)))
    print('- Recall: {:.2f}'.format(recall_score(fold_y_true, fold_y_pred)))

    # Confusion Matrix
    cm = confusion_matrix(fold_y_true, fold_y_pred)
    print("Confusion Matrix:")
    print(cm)
    print('----------------------------------')

# hitung dan print hasil rata-rata
print("Average Results with Information Gain:")
print(f"- Accuracy: {accuracy_score(y, y_pred):.2f}")
print(f"- Precision: {precision_score(y, y_pred):.2f}")
print(f"- Recall: {recall_score(y, y_pred):.2f}")
```

Gambar 7. Penerapan *Naive Bayes* Dengan *K-Fold Cross Validation*
Sumber: *(python)*

Berikut adalah tabel *confusion matrix* dengan rincian akurasi, presisi, dan *recall* pada model *naïve bayes* menggunakan seleksi fitur *information gain* pada *fold 2*.

Tabel 1. *Confusion Matrix fold* 2 *Naïve Bayes*

	Kelas Aktual	
Kelas Prediksi	**Positif**	**Negatif**
Positif	TP = 13350	FP = 5104
Negatif	FN = 4323	TN = 12569

Sumber: *(python)*

TP = 13350 data yang mempunyai nilai positif. benar dan diprediksi sebagai positif.
FP = 5104 data yang mempunyai nilai negatif benar tetapi salah diprediksi sebagai positif.
FN = 4323 data yang mempunyai nilai positif benar tetapi salah diprediksi sebagai negatif.
TN =12569 data yang mempunyai nilai negatif benar dan diprediksi dengan benar sebagai negatif

Berdasarkan hasil percobaan di atas, nilai akurasi paling tinggi terletak pada *fold*-1 dengan *information gain* sebesar 74%, nilai presisi paling tinggi terletak pada *fold*-2 tanpa *information gain* sebesar 73% dan *recall* tertinggi terletak pada *fold*-2 dengan *information gain* sebesar 76%.

iii. 5-*Fold Cross Validation*

Berikut adalah tabel *confusion matrix* dengan rincian akurasi, presisi, dan *recall* pada model *naïve bayes* menggunakan seleksi fitur *information gain* pada *fold 5*.

Tabel 2. *Confusion Matrix fold* 5 *Naïve Bayes*

	Kelas Aktual	
Kelas Prediksi	**Positif**	**Negatif**
Positif	TP = 5364	FP = 2036
Negatif	FN = 1705	TN = 5033

Sumber: *(python)*

TP = 5364 data yang mempunyai nilai positif. benar dan diprediksi sebagai positif.
FP = 2036 data yang mempunyai nilai negatif benar tetapi salah diprediksi sebagai positif.
FN = 1705 data yang mempunyai nilai positif benar tetapi salah diprediksi sebagai negatif.
TN = 5033data yang mempunyai nilai negatif benar dan diprediksi dengan benar sebagai negatif

Berdasarkan hasil percobaan di atas, nilai akurasi paling tinggi terletak pada *fold*-1 dengan *information gain* sebesar 74%, nilai presisi paling tinggi terletak pada *fold*-1 dengan *information gain* sebesar 73% dan *recall* tertinggi terletak pada *fold*-1 dengan *information gain* sebesar 77%.

iv. 10-*Fold Cross Validation*

Berikut adalah tabel *confusion matrix* dengan rincian akurasi, presisi, dan *recall* pada model *naïve bayes* menggunakan seleksi fitur *information gain* pada *fold 10*.

Tabel 3. *Confusion Matrix fold* 10 *Naïve Bayes*

	Kelas Aktual	
Kelas Prediksi	**Positif**	**Negatif**
Positif	TP = 2673	FP = 986
Negatif	FN = 862	TN = 2548

Sumber: *(python)*

TP = 2673 data yang mempunyai nilai positif. benar dan diprediksi sebagai positif.
FP = 986 data yang mempunyai nilai negatif benar tetapi salah diprediksi sebagai positif.
FN = 862 data yang mempunyai nilai positif benar tetapi salah diprediksi sebagai negatif.
TN = 2548 data yang mempunyai nilai negatif benar dan diprediksi dengan benar sebagai negatif

Berdasarkan hasil percobaan di atas, nilai akurasi paling tinggi terletak pada *fold*-1 dan 2 dengan *information gain* sebesar 74%, nilai presisi paling tinggi terletak pada *fold*-1 dan 2 dengan *information gain* sebesar 73% dan *recall* tertinggi terletak pada *fold*-1 dan 2 dengan *information gain* sebesar 77%.

5. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang dilakukan, kesimpulan yang dapat diambil adalah:

1. Model *Naïve Bayes* dengan *k-fold cross validation*, saat diterapkan pada dataset Diabetes, menunjukkan hasil yang cukup baik dengan akurasi tertinggi sebesar 73% pada data uji.
2. Metode pemilihan atribut *Information Gain* yang diterapkan pada model *Naive Bayes* terhadap dataset Diabetes dengan *k-fold cross validation* menghasilkan akurasi tertinggi sebesar 74% pada uji.
3. Walaupun hasil ini menunjukkan kemajuan, masih ada ruang untuk peningkatan performa model. Penyesuaian lebih lanjut terhadap fitur-fitur yang digunakan serta eksplorasi berbagai teknik peningkatan model dapat meningkatkan hasil prediksi.

B. Saran

Adapun saran yang bisa diterapkan untuk penelitian selanjutnya yaitu:

1. Disarankan menggunakan dataset yang memiliki *noice* agar penggunaan seleksi fitur *information gain* lebih optimal.
2. Penyetelan parameter model secara cermat, serta eksperimen dengan algoritma *machine learning* lainnya yang lebih kompleks, dapat dieksplorasi untuk melihat apakah ada model yang lebih baik cocok dengan pola data tersebut.

6. DAFTAR PUSTAKA

Aini, S. H. A., Sari, Y. A., & Arwan, A. (2018). Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(9), 2546–2554. http://j-ptiik.ub.ac.id

Ainurrohmah. (2021). Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka. PRISMA, Prosiding Seminar Nasional Matematika, 4, 493–499.

Al Karomi, M. A., & Ivandari. (2019). Optimasi algoritma Naïve Bayes dengan Information Gain ratio untuk menangani dataset berdimensi tinggi. IC-Tech Journal of Informatic and Computer Technology, 14(2), 18–24.

Arifin, T., & Ariesta, D. (2019). Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization. Jurnal Tekno Insentif, 13(1), 26–30. https://doi.org/10.36787/jti.v13i1.97

Fauzi, F. A., Furqon, M. T., & Yudistira, N. (2021). Klasifikasi Jenis Tanaman Tembakau di Indonesia menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 5(2), 698–703. http://j-ptiik.ub.ac.id

Febriani, S., & Sulistiani, H. (2021). Analisis Data Hasil Diagnosa Untuk Klasifikasi Gangguan Kepribadian Menggunakan Algoritma C4.5. 89Jurnal Teknologi Dan Sistem Informasi (JTSI), 2(4), 89–95.

Hartati, S., Ramdhan, N. A., & SAN3, H. A. (2022). Prediksi Kelulusan Mahasiswa Dengan Naïve Bayes dan Feature Selection Information Gain Student Graduation Prediction With Naïve Bayes and Feature. 4(02), 223–235.

Hasanah, R. L., Hasan, M., Pangesti, W. E., Wati, F. F., & Gata, W. (2019). Klasifikasi Penerima Dana Bantuan Desa Menggunakan Metode Knn (K-Nearest Neighbor). Jurnal Techno Nusa Mandiri, 16(1), 1–6. https://doi.org/10.33480/techno.v16i1.25

Hayuningtyas, R. Y. (2019). Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita. 6(1), 18–22.

Marcoulides, G. A. (2005). Discovering Knowledge in Data: an Introduction to Data Mining. In Journal of the American Statistical Association (Vol. 100, Issue 472). https://doi.org/10.1198/jasa.2005.s61

Muqorobin, M., Kusrini, K., & Luthfi, E. T. (2019). Optimasi Metode Naïve Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah. Jurnal Ilmiah SINUS, 17(1), 1. https://doi.org/10.30646/sinus.v17i1.378

Mustafa, M. S., & Simpen, I. W. (2019). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi, VIII(1), 1–10. https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1 -10

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. Jurnal SIMETRIS, 10(2), 421–432.

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. 1(12), 1750–1757.

Pratikaningtyas, R., Soeleman, M. A., Sarjana, P., Informatika, T., & Dian, U. (2019). Klasifikasi Penerbitan Surat Keputusan Tunjangan Profesi Guru Menggunakan Naïve Bayes Berbasis Information Gain. Jurnal Teknologi Informasi, 15(2), 93–102. http://research.pps.dinus.ac.id/index.php/Cyberku/article/view/88

Rahayuningsih, P. A. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining. Jurnal Teknik Informatika Kaputama (JTIK), 3(1).

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. PRISMA, Prosiding Seminar Nasional Matematika, 3, 64–71.

Setiyorini, T., & Asmono, R. T. (2019). Penerapan Metode K-Nearest Neighbor Dan Information Gain Pada Klasifikasi Kinerja Siswa. JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer), 5(1), 7–14. https://doi.org/10.33480/jitk.v5i1.613

Siallagan, R. A., & Fitriyani. (2021). Prediksi Penyakit Diabetes Mellitus. Jurnal Responsif, 3(1), 45–46.

Wijaya, H. D., & Dwiasnati, S. (2020). Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat. Jurnal Informatika, 7(1), 1–7. https://doi.org/10.31311/ji.v7i1.6203

Saputra, V. W. (2019). Klasifikasi Jenis Makanan menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain.

Rizaldy, A., & Santoso, H. A. (2017). Performance Improvement Of Support Vector Machine (SVM) With Information Gain On Categorization Of Indonesian News Documents.

Dewi, Rulita Kumala, dan Moh Dasuki. “Pengadaan Papan Nama Jalan Dusun dalam Meningkatkan Tata Desa Jatisari.” JIWAKERTA: Jurnal Ilmiah Wawasan Kuliah Kerja Nyata, no. 2, 2023,

http://jurnal.unmuhjember.ac.id/index.php/jiwakerta.

Saifudin, Ilham, dan Reni Umilasari. Automatic Aircraft Navigation Using Star Metric Dimension Theory in Fire Protected Forest Areas. no. 2, 2021, hal. 294–304, https://doi.org/10.31764/jtam.v5i2.4331.

Umilasari, R., et al. "Local irregularity chromatic number of vertex shackle product of graphs." IOP Conference Series: Materials Science and Engineering, vol. 821, no. 1, Institute of Physics Publishing, 2020, https://doi.org/10.1088/1757-899X/821/1/012038.

Umilasari, Reni. "Dominating number of distance two of corona product of graphs." Indonesian Journal of Combinatorics, vol. 1, no. 1, 2016, www.ijc.or.id.

---. "Star metric dimension of complete, bipartite, complete bipartite and fan graphs." International Journal of Trends in Mathematics Education Research, vol. 5, no. 2, Juni 2022, hal. 199–205, https://doi.org/10.33122/ijtmer.v5i2.137.