P-ISSN : 2622-1594
E-ISSN : 2685-449X

Pendekatan Data Science untuk Deteksi Dini Diabetes Menggunakan Naive Bayes Classifier

Norma Ningsih[a*], Aprianto[b], Angeline[c]
[a]*Teknologi Rekayasa Internet, Politeknik Elektronika Negeri Surabaya*
[b]*Sistem Informasi, Universitas Dinamika*
[c]*Sistem Informasi, Universitas Dinamika*
E-mail: norma@pens.ac.id, 18410100002@dinamika.ac.id, 18410100228@dinamika.ac.id

***Abstrak*—Diabetes merupakan penyakit yang memiliki gejala dimana kadar gula darah berada diatas normal yang disebabkan karena kurangnya insulin dalam darah seseorang. Umumnya diabetes disebabkan karena adanya gangguan metabolisme dalam tubuh selama periode yang cukup lama. Diabetes merupakan penyakit yang berbahaya dengan jumlah penderita yang terus meningkat setiap tahun. Hal ini disebabkan karena kurangnya kesadaran pola hidup sehat dan deteksi dini penyakit yang sering tertunda. Penelitian ini membuat sistem klasifikasi yang dapat melakukan pendeteksian dini terhadap penyakit diabetes. Metode yang digunakan adalah *naïve bayes classifier* dengan *Laplacian smoothing*. Penelitian ini menggunakan 100 data dari data random, data tersebut dibagi menjadi 80% data latih dan 20% data uji. Berdasarkan hasil pengujian menggunakan *confusion matrix* diperoleh nilai ukuran testing set yang digunakan adalah 40% testing set dan sisanya 60% sebagai training set merupakan hasil yang paling ideal. Dari pembagian dataset tersebut diperoleh nilai akurasi sebesar 70%.**

Kata Kunci*—Diabetes, *Naïve bayes classifier*, klasifikasi, *Laplacian smoothing

I. PENDAHULUAN

Diabetes atau Diabetes Melitus merupakan salah satu penyakit yang ditandai dengan gejala lebihnya kadar gula darah diatas ambang normal yaitu sama atau lebih dari 200 mg/dl dan melebihi kadar gula darah puasa yang diatas atau sama dengan 126 mg/dl. Penyakit diabetes ini merupakan salah satu penyakit berbahaya karena sering tidak disadari oleh penderitanya dan saat diketahui sudah terjadi komplikasi dalam dirinya, dari situ diabetes ini juga sering dikenal sebagai *silent killer* [1].

Jumlah penyandang diabetes dunia pada 2019 adalah sebanyak 463 juta penderita dan akan di prediksi melonjak naik di tahun 2045 dengan jumlah 700 juta penderita (atau meningkat sebanyak 51%) [2]. Masih dalam data yang sama, Indonesia menduduki peringkat ketujuh negara dengan penyandang diabetes terbesar di dunia dari 211 negara yang terdata oleh IDF dengan jumlah penyandang sebesar 10,7 juta orang. Sedangkan untuk kawasan Pasifik Barat, Indonesia menduduki peringkat kedua dibawah negara China dari 36 negara yang terdata dengan tingkat prevalensi sebesar 6,2% dari jumlah populasi orang dewasa sebesar 172,2 juta jiwa pada 2019.

Faktor-faktor yang memiliki andil dalam penyakit diabetes antara lain adalah faktor genetis atau keturunan dari keluarga, faktor usia, faktor gaya hidup (terkait pada pola makan, aktivitas sehari-hari) dan faktor riwayat terkena penyakit diabetes gestasional pada wanita saat hamil [3]. Sebagian besar penderita diabetes tidak menyadari bahwa dirinya berisiko atau sudah terdiagnosa mengalami diabetes, hal ini disebabkan oleh minimnya pengetahuan mengenai gejala-gejala yang terjadi membuat seakan tidak terjadi apa-apa. 75% dari total penyandang diabetes di Indonesia belum menyadari bahwa dirinya menyandang diabetes. 25% sisanya sudah menyadari mereka menyandang diabetes dengan 17% pasien menjalani terapi diabetes dan 8% sisanya tidak menjalankan terapi [4].

Kondisi tersebut dapat ditangani ketika penyandang diabetes lebih dini mengetahui dirinya terkena diabetes atau setidaknya memahami dirinya berisiko atau tidak dibandingkan dengan pengobatan pasca diagnosis terkena diabetes yang sudah terlambat dan umumnya sudah disertai beragam komplikasi. Kecerdasan buatan atau *Artificial Intelligence* (AI) disini dapat berperan dalam membantu deteksi dini adanya penyakit diabetes dengan menggunakan *Machine Learning* (ML) atau pembelajaran mesin. ML merupakan cabang dari AI dan dapat didefinisikan sebagai proses pemecahan masalah praktis dengan mengumpulkan dataset dan membangun model statistik dari data set tersebut, *Machine* learning mampu beradaptasi dengan data baru secara mandiri untuk menghasilkan keputusan yang andal dari perhitungan sebelumnya [5].

Machine learning dalam hal ini dapat digunakan untuk membantu seseorang mengenali lebih dini mengenai risiko penyakit diabetes. Klasifikasi merupakan salah satu jenis subset dalam machine learning yang dapat digunakan untuk mengklasifikasikan suatu hal kedalam salah satu jenis kategori. Dalam hal ini, machine learning dapat digunakan untuk melakukan klasifikasi terhadap gejala yang diberikan apakah gejalagejala tersebut masuk kedalam penyakit diabetes atau tidak. *Naïve Bayes* merupakan salah satu metode klasifikasi dalam *Machine Learning* yang dapat memprediksi keanggotaan kelas kepada kelas tertentu dengan menggunakan probabilitas. Dasar dari metode *Naïve*

Naskah Masuk : 07 Maret 2023
Naskah Direvisi : 11 April 2023
Naskah Diterima : 12 April 2023
*Corresponding Author : norma@pens.ac.id

DOI : 10.37823/insight.v5i1.300

Bayes adalah Teorama Bayes yang dikembangkan Thomas Bayes yang awalnya digunakan dalam teori probabilitas dan keputusan selama abad ke-18 [6]

Dengan adanya klasifikasi atau deteksi ini, seseorang dapat melakukan pengecekan diabetes lebih dini menggunakan bantuan machine learning sebelum melakukan pengecekan medis lebih lanjut jika hasil klasifikasi menunjukan bahwa orang tersebut terdeteksi diabetes. Pengecekan ini lebih terjangkau bagi masyarakat karena dapat diakses dimana saja dan tidak membutuhkan biaya.

II. TINJAUAN PUSTAKA

A. *Diabetes*

Diabetes melitus atau juga dikenal sebagai penyakit kencing manis yaitu penyakit yang disebabkan karena terganggunya metabolisme karbohidrat, protein dan lemak yang umumnya ditandai dengan peningkatan kadar glukosa dalam darah akibat kelainan sekresi insulin atau menurunya kerja insulin yang menyebabkan gangguan pada kinerja metabolisme, kegagalan pada berbagai organ terutama pada organ mata, ginjal, saraf, jantung dan pembuluh darah [7]. Diabetes sendiri merupakan salah satu penyakit yang berpotensi paling banyak menimbulkan komplikasi (memicu timbulnya penyakit lain) karena berkaitan dengan tingginya kadar gula darah sehingga berdampak pada rusaknya pembuluh darah, saraf dan struktur internal lainnya. Komplikasi ini dapat timbul jika penderita diabetes tidak ditangani dengan baik [8]

B. *Machine Learning*

Machine learning (ML) atau pembelajaran mesin merupakan bagian atau subset dari kecerdasan buatan atau *Artificial Intelligence* (AI) yang merupakan simulasi dari kecerdasan manusia oleh sistem komputer. ML sendiri adalah kombinasi dari ilmu komputer dan statistik, ilmu komputer berfokus pada pemecahan masalah dan identifikasi apakah masalah dapat diselesaikan pada semua tahapan. Sedangkan statistik pada sisi pemodelan data, hipotesis dan mengukur keandalan [9]

C. *Naïve Bayes*

Naïve Bayes merupakan salah satu metode klasifikasi dalam Machine *Learning* yang dapat memprediksi keanggotaan kelas kepada kelas tertentu dengan menggunakan probabilitas. Dasar dari metode *Naïve Bayes* adalah Teorama *Bayes* yang dikembangkan Thomas Bayes yang awalnya digunakan dalam teori probabilitas dan keputusan selama abad ke-18 [6].

D. *Laplacian Smoothing*

Laplacian smoothing atau laplace smoothing merupakan salah satu metode atau algoritma pemulusan (smoothing) tertua yang digunakan. Menurut [10] laplacian smoothing merupakan metode yang berpengaruh untuk mencegah masalah probabilitas nol. Metode *laplacian smoothing* juga dikenal sebagai add-one smoothing yaitu menambahkan angka 1 pada setiap frekuensi *token* yang didapat [11].

E. *Confusion Matrix*

Confusion matrix merupakan tabel pencatat hasil kerja klasifikasi. *Confusion matrix* melakukan pengujian untuk memperkirakan objek yang benar dan salah. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya. *Confusion matrix* berisi informasi aktual *(actual*) dan prediksi (*predicted*) pada sistem klasifikasi [12].

$$Akurasi = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad \dots (1)$$

Rumus dari akurasi adalah hasil penjumlahan antara *True Positive* dan *True Negative* dibagi dengan hasil penjumlahan antara *True Positive, False Positive, False Negative,* dan *True Negative.*

III. METODE DAN INTI PENELITIAN

A. *Dataset*

Pada penelitian ini dataset yang digunakan merupakan hasil *survey* pada 100 orang sehingga diperoleh 100 *record dataset*. Pada studi kasus deteksi dini menggunakan machine learning dengan metode *Naïve Bayes classifier* ini, kolom yang digunakan pada dataset mengacu pada beberapa gejala dari penderita diabetes [13]. Kolom dataset yang digunakan sebagai parameter inputan ada 9, dimana 8 kolom sebagai gejala diabetes dan 1 kolom sebagai kolom hasil klasifikasi. Adapun kolom yang digunakan dalam dataset dijelaskan dalam tabel 1.

TABEL I
PENJELASAN TABEL DATASET

No	Nama Kolom	Penjelasan
1	Usia	Usia dari penderita diabetes, dibagi menjadi 3 kategori yaitu 20-40, 40-50 dan 50-60
2	Jkel	Jenis kelamin penderita diabetes yaitu Pria dan Wanita
3	Banyak_kencing	Gejala penderita diabetes yaitu apakah sering buang air kecil ? (ya / tidak)
4	Turun bb	Gejala penderita diabetes yaitu apakah mengalami penurunan berat badan yang cukup derastis akhir-akhir ini? (ya/tidak)
5	Luka_sukar	Gejala penderita diabetes yaitu apakah mengalami luka yang sulit untuk sembuh / kering? (ya/tidak)
6	Kesemutan	Gejala penderita diabetes yaitu apakah sering mengalami kesemutan? (ya/tidak)
7	Lemas	Gejala penderita diabetes yaitu apakah sering mengalami lemas atau letih akhir-akhir ini? (ya/tidak)
8	Kulit_gatal	Gejala penderita diabetes yaitu apakah mengalami gatalgatal pada kulit atau kulit kering? (ya/tidak)
9	Keturunan	Gejala penderita diabetes yaitu apakah mempunyai Riwayat diabetes dalam keluarga? (ya/tidak)
10	Hasil	Hasil klasifikasi diabetes yaitu terdeteksi sebagai penderita diabetes (ya/tidak)

B. *Metode yang diusulkan*

Naïve Bayes Classifier (NBC) merupakan salah satu metode yang digunakan untuk klasifikasi dalam ruang lingkup Machine Learning. Cara kerja NBC adalah melakukan perhitungan probabilitas untuk memprediksi peluang dimasa depan berdasarkan pengalaman sebelumnya

 DOI : 10.37823/insight.v5i1.300

[14]. Perhitungan probabilitas yang dimaksud adalah menghitung nilai probabilitas setiap kelas dari perbandingan data yang akan diklasifikasi dengan data yang digunakan sebagai dataset testing sesuai jumlah kelas yang ada. Kelas yang dimaksudkan adalah variabel yang digunakan untuk melakukan klasifikasi, dalam studi kasus ini yang dimaksud kelas adalah gejala-gejala diabetes yang digunakan. Dataset testing yang dimaksud adalah kumpulan data yang akan digunakan sebagai acuan klasifikasi terhadap input data gejala baru.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \ldots\ldots\ldots\ldots \quad (2)$$

dimana C adalah kelas target, X adalah data, P(C) adalah probabilitas kelas (probabilitas sebelumnya), P(X) adalah probabilitas prediktor (probabilitas sebelumnya), dan P(XC) adalah probabilitas berdasarkan kondisi hipotesis, dan P(CX) adalah probabilitas hipotesis berdasarkan kondisi (probabilitas posterior).

Dalam studi kasus ini juga digunakan teknik pemulusan *Laplacian Smoothing* untuk menghindari nilai nol (0) pada hasil probabilitas yang akan mempengaruhi hasil klasifikasi nantinya. Metode *Naïve Bayes* dengan laplacian smoothing dapat digambarkan dengan menambahkan nilai 1 pada perhitungan probabilitas. Bagaimana cara kerja metode *Naïve Bayes* dalam melakukan klasifikasi apakah seseorang mengidap diabetes atau tidak dapat dilihat pada tabel 2.

TABEL II
TABEL GEJALA

Usia (20-40)	Turun Berat Badan (Ya/Tidak)	Keturunan (Ya/ Tidak)	Hasil (Ya/ Tidak)
50-60	Ya	Ya	???

Konsep dari klasifikasi *Naïve Bayes* didasarkan atas probabilitas dari semua kelas gejala terhadap masing-masing kelas hasil. Dari tabel gejala diatas misalnya (gambaran diberikan menggunakan 3 parameter gejala untuk menyederhanakan kasus), diberikan 3 buah parameter untuk memprediksi apakah seseorang mengidap penyakit diabetes atu tidak. *Naïve Bayes classifier* menghitung probabilitas dari gejala tersebut pada kelas ya dan tidak (sesuai atribut pada kelas hasil yaitu ya dan tidak). Probabilitas untuk gejala dengan usia 50-60, turun berat badan dan memiliki keturunan diabetes dicari dari dataset dimana probabilitas itu akan dikalikan dengan probabilitas untuk hasil "ya" dan "tidak".

Untuk hasil prediksi ya, dari dataset akan dihitung berapa buah data yang memenuhi kondisi usia=50-60; turun_bb=ya;keturunan=ya;hasil=ya; terhadap jumlah dataset. Hasil dari prediksi "ya" terhadap gejala tersebut dapat misalkan sebagai a/n dengan a sebagai jumlah data yang sesuai dan n sebagai jumlah dataset keseluruhan. Hal yang sama dilakukan untuk hasil prediksi "tidak" yaitu dengan mencari probabilitas terhadap data yang memenuhi kondisi usia=50-60; turun_bb=ya; keturunan=ya; hasil= tidak;. Hasil prediksi "tidak" terhadap kondisi tersebut dapat dimisalkan sebagai b/n dengan b sebagai jumlah data yang sesuai dengan kondisi dan n sebagai jumlah dataset keseluruhan. Dari perhitungan probabilitas terhadap kelas hasil "ya" dan "tidak", kemudian akan diambil nilai tertinggi dari keduanya sebagai hasil klasifikasi. Hasil probabilitas tertinggi menandakan bahwa sebuah kondisi pada hasil tertentu akan condong kearah hasil tersebut.

Dalam studi kasus ini menggunakan pemulusan dengan metode *Laplacian smoothing* untuk menghindari nilai 0 pada hasil probabilitas. Nilai 0 dihindari karena akan mempengaruhi hasil klasifikasi. Jika nilai 0 pada salah satu kelas gejala muncul dan nilai itu dikalikan dengan sejumlah kelas gejala yang lain, maka hasil akhir perhitungan akan menghasilkan nilai 0 (nilai apapun jika dikali dengan nilai 0 akan menghasilkan nilai 0). Penerapan *laplacian smoothing* dilakukan dengan menambahkan nilai 1 pada pembilang dan nilai "x" pada penyebutnya. Contohnya adalah probabilitas terhadap a/n. Hasil dari probabilitas tersebut dapat berpotensi menghasilkan nilai 0 jika pembilangnya (a) bernilai 0. Maksud dari nilai 0 adalah tidak ada data dengan kondisi tertentu yang cocok pada dataset testing, sehingga untuk menghindari dari jumlah data yang sesuai kondisi tersebut tidak ditemukan akan ditambahkan 1 pada pembilangnya dan ditambahkan nilai "x" pada penyebutnya.

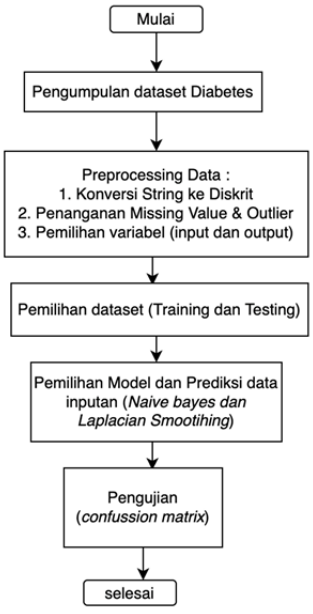
Kembali kepada contoh a/n untuk kelas gejala usia=50-60 dan hasil=ya. Penerapan lapacian smoothing adalah menambah nilai 1 pada a dan nilai x pada n. Nilai x adalah jumlah dari atribut unik pada kelas a yaitu usia. Pada kelas usia memiliki 3 atribut nilai yaitu usia 20-40, 40-50 dan 50-60. Sehingga pada penyebut n ditambahkan 3 dan bentuk dari *laplacian smoothingnya*

$$\frac{a+1}{n+x} = \frac{a+1}{n+3} \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Bentuk perhitungan dengan *laplacian smoothing* dapat dilihat pada rumus berikut ini.

$$P(W_i|\text{class}) = \frac{\text{freq}(W_i,\text{class})+1}{N_{class}+V_{class}} \ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Laplacian Smoothing menambahkan nilai 1 pada setiap frekuensi kelas dan menambahkan V class yaitu jumlah atribut pada kelas tertentu untuk menghasilkan probabilitas tanpa hasil nol.



Gambar. 1. Blok Diagram Sistem

 DOI : 10.37823/insight.v5i1.300

Pada gambar 1 menunjukkan blok diagram system yang menjelasakan tahapan dalam melakukan penelitian ini. Dimulai dari pengumpulan dataset, dilanjutkan pada *preprocessing data* yang terdiri dari 3 proses. Kemudian dilakukan pemilihan dataset yang digunakan untuk proses *training* dan *testing.* Pemilihan model dengan menggunakan metode *naïve bayes* dan *Laplacian smoothing*. Tahapan yang terakhir yaitu melakukan evaluasi atau pengujian dengan *confusion matrix* untuk menguji tingkat keberhasilan dari mode;l yang digunakan.

Tahapan Penelitian dapat dijelaskan sebagai berikut :

A. Import Data

Pertama dilakukan import data yang akan digunakan untuk training model klVasifikasi beserta modul yang akan digunakan untuk membantu proses klasifikasi.

B. Konversi nilai String

Konversi nilai kedalam bentuk nilai diskrit karena model *Naïve Bayes* tidak dapat menerima *input* string. Konversi dilakukan untuk mengubah atribut dalam sebuah kelas untuk diwakili dengan angka, contohnya adalah pada kelas usia terdapat 3 atribut yaitu usia 20-40, 40-50 dan 50-60 tahun. Atribut tersebut diubah menjadi angka 0 mewakili usia 20-40, 1 mewakili 40-50 dan angka 2 mewakili 50-60 tahun.

C. Missing Value

Setelah itu dilakukan penanganan terhadap *missing value* dan *outlier* yang akan mengganggu jalannya proses klasifikasi. Sebelumnya akan dilakukan pengecekan penyimpangan atribut (*outliers*) pada setiap kelas gejala sesuai pada kolom dataset. Setelah semua atribut kelas sesuai dengan yang diharapkan, kemudian dilakukan pengecekan terhadap *missing value* untuk mengetahui apakah terdapat nilai yang kosong dari data

D. Penetapan Model

Setelah data siap untuk dilakukan klasifikasi, model untuk klasifikasi akan diuji menggunakan *dataset* yang terbagi menjadi *training set* dan *testing set*. Pengujian dilakukan dengan melakukan perulangan pengujian model dengan memilih ukuran *testing set* yang berbeda untuk mendapatkan tingkat akurasi terbaik yang kemudian nantinya akan digunakan sebagai model prediksi terhadap data inputan.

E. Pengujian Model

Confusion matrix melakukan pengujian untuk memperkirakan objek yang benar dan salah.

Pada library *scikit*-learn modul klasifikasi dengan *Naïve Bayes* memiliki beberapa jenis seperti *Gaussian Classifier*, *Bernoulli Classifier* dan *Multinomial Classifier*. *Gaussian Classifier* cocok diterapkan terhadap data berjenis kontinu yang memiliki distribusi normal. *Bernoulli Classifier* cocok digunakan untuk data yang bersifat binary (memiliki nilai benar/salah atau 0/1). Multinomial classifier cocok digunakan terhadap data yang memiliki beberapa atribut seperti usia memiliki 3 atribut [15]. Berdasarkan tiga jenis model klasifikasi naïve bayes, model yang paling cocok digunakan dalam kasus ini adalah model klasifikasi Multinomial NB dimana dapat menampung beberapa atribut dan tidak terkait dengan statistik persebaran data (distribusi data).

TABEL III
DATA REAL

No	usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan	hasil
1	20-40	wanita	ya	ya	ya	ya	ya	tidak	ya	ya
2	40-50	wanita	tidak	ya	tidak	tidak	ya	ya	tidak	tidak
3	20-40	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
4	50-60	wanita	ya	tidak	ya	ya	tidak	ya	ya	ya
5	40-50	pria	ya	ya	tidak	ya	ya	ya	tidak	ya
6	20-40	pria	ya	tidak	tidak	tidak	tidak	ya	ya	tidak
7	50-60	wanita	tidak	ya	ya	tidak	ya	ya	tidak	ya
8	50-60	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
9	20-40	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
10	50-60	pria	ya	ya	tidak	ya	tidak	ya	ya	ya
11	40-50	pria	tidak	tidak	ya	tidak	ya	ya	ya	ya
12	50-60	wanita	ya	tidak	tidak	ya	tidak	ya	ya	ya
13	40-50	wanita	tidak	ya	ya	ya	tidak	tidak	tidak	tidak
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
97	50-60	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
98	20-40	pria	tidak	ya	tidak	ya	ya	ya	ya	ya
99	50-60	pria	ya	tidak	ya	tidak	ya	tidak	ya	ya
100	40-50	wanita	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
101	50-60	wanita	ya	tidak	ya	tidak	ya	tidak	ya	ya

IV. HASIL EKSPERIMEN DAN PENELITAN

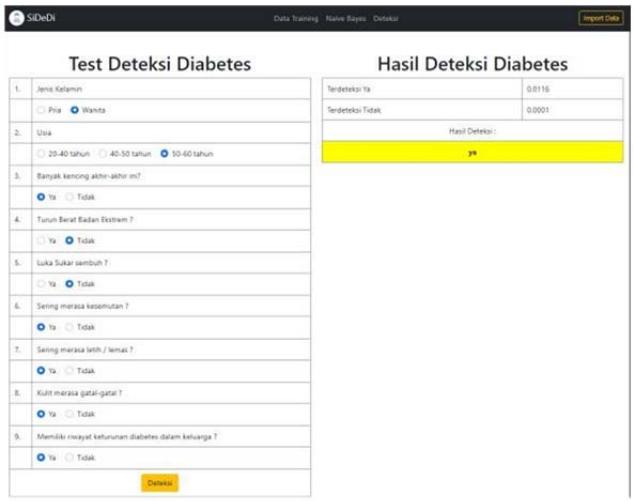
Implementasi dari deteksi dini diabetes menggunakan metode *Naïve Bayes* classifier diterapkan menggunakan bahasa Python 3 dengan tools Jupyter Notebook serta bantuan dari beberapa *open source library* yang membantu dalam pembentukan *source code*.

Pada gambar 2 menunjukkan aplikasi yang digunakan oleh pengguna untuk menginputkan data sesuai dengan atribut yang dibutuhkan dalam proses prediksi mengunakan *naïve bayes* seperti usia, jenis kelamin hingga apakah ada keturunan diabetes atau tidak, banyak kencing, turun berat badan, luka sulit untuk sembuh, kesemutan, lemas, kulit gatal dan apakah ada keturunan penyakit diabetes atau tidak.

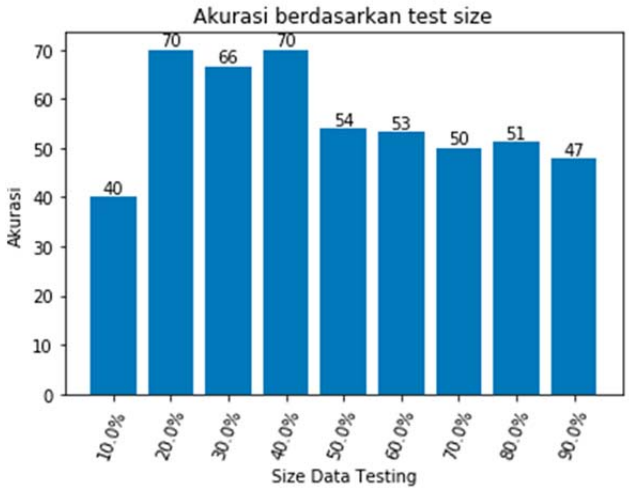
A. Hasil Eksperimen Naïve Bayes

Data yang sudah siap untuk dilakukan klasifikasi, model untuk klasifikasi akan diuji menggunakan dataset yang terbagi menjadi training set dan testing set. Pengujian dilakukan dengan melakukan perulangan pengujian model dengan memilih ukuran testing set yang berbeda untuk mendapatkan tingkat akurasi terbaik yang kemudian nantinya akan digunakan sebagai model prediksi terhadap data inputan. Dari pengulangan pengujian dengan menggunakan ukuran sampel test sebesar 10% hingga 90% didapatkan tingkat akurasi seperti pada gambar dibawah ini.

 DOI : 10.37823/insight.v5i1.300



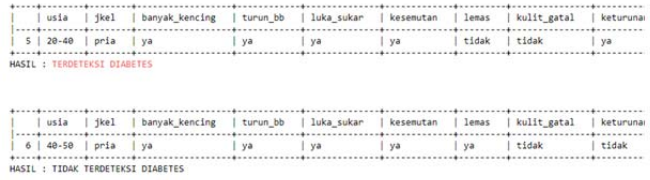
Gambar. 2. Sistem aplikasi website untuk menginputkan data



Gambar. 3. Hasil Pengujian

Dari grafik hubungan jumlah *testing set* dan tingkat akurasi diatas dapat terlihat bahwa tingkat akurasi tertinggi diperoleh menggunakan ukuran *testing set* sebesar 20% dan 40%. Dalam studi kasus ini ukuran testing set yang digunakan adalah 40% *testing set* dan sisanya 60% sebagai *training set.*

Setelah model ditetapkan, kemudian dilakukan pembacaan data inputan yang akan diprediksi dari sebuah file *excel.* Data inputan berisikan gejala-gejala yang sudah ditetapkan. Berikutnya dilakukan konfigurasi output untuk menampilkan hasil prediksi sesuai dengan hasil yang diharapkan. Langkah yang terakhir dilakukan adalah melakukan prediksi yaitu menjalankan konfigurasi output terhadap setiap baris data inputan yang sudah dimuat sebelumnya. Gambaran output dari kumpulan data inputan yang sebelumnya dapat dilihat pada gambar 4.



Gambar. 4. Hasil Output Produksi

```
<---------------------- Sample test size 40.0% ----------------------->

Akurasi
======================
70.0%

Confusion matrix
======================
[[ 7 12]
 [ 0 21]]

Classification Report
======================
              precision    recall  f1-score   support

           0       1.00      0.37      0.54        19
           1       0.64      1.00      0.78        21

    accuracy                           0.70        40
   macro avg       0.82      0.68      0.66        40
weighted avg       0.81      0.70      0.66        40
```

Gambar. 5. Hasil Kinerja

Pada gambar 5 menunjukkan hasil pengujian menggunakan *confusion matrix* untuk memperkirakan objek dan benar dan salah. Nilai akurasi yaitu 70% dimana akurasi menunjukkan seberapa banyak data aktual yang benar diklasifikasikan oleh sistem dengan ketentuan jumlah data yang benar diklasifikasikan sistem dibagi jumlah data keseluruhan. Presisi menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model melalui perhitungan *Confusion matrix*. *Recall* atau *sensitivity* menggambarkan tingkat keberhasilan sistem dalam menemukan/menghasilkan prediksi yang sesuai dengan *class* sebenarnya. Sedangkan F-1 Score menggambarkan perbandingan rata-rata presisi dan *recall* yang dibobotkan.

V. KESIMPULAN

Penelitian yang telah dilakukan dengan menggunakan algoritma *Naïve bayes* dan *Laplacian smoothing* mampu untuk digunakan dalam memprediksi deteksi awal dari penyakit diabetes. Berdasarkan hasil pengujian menggunakan *confusion matrix* diperoleh nilai ukuran testing set yang digunakan adalah 40% testing set dan sisanya 60% sebagai training set merupakan hasil yang paling ideal dan dari pembagian dataset tersebut diperoleh nilai akurasi sebesar 70%.

DAFTAR PUSTAKA

[1] *Hestiana, D. W.Journal of Health Education. Journal of Health Education, 2(2), 138–145. https://doi.org/10.1080/10556699.1994. 10603001. 2017*

[2] *International Diabetes Federation. Global Diabetes Data Report 2010-2045.Journal IDF. https://diabetesatlas.org/data/en/world/.2019*

[3] *Sunur, I. C. Mengenal Perbedaan Diabetes Tipe 1 dan Tipe 2. Https://Www.Alodokter.Com/Mengenal-Perbedaan-Diabetes-Tipe-1-Dan-Tipe-2. https://www.alodokter.com/mengenal-perbedaan-diabetes-tipe-1-dan-tipe-2. 2020*

[4] *Rossa, V., & Halidi, R. Lebih dari 70 Persen Orang Indonesia Tak Sadar Terkena Diabetes. https://www.suara.com/health/2019/11/14/052301/lebih-dari-70-persen- orang-indonesia-tak-sadar-terkena-diabetes. 2019*

[5] *Zohuri, B., & Rahmani, F. M. Artificial Intelligence Driven Resiliency with Machine Learning and Deep Learning Components. Journal of Communication and Computer, 15(1), 1–13. https://doi.org/10.17265/1548-7709/2019.01.001. 2019*

[6] *Han, J., Kamber, M., & Pei, J. Data mining: Data mining concepts and techniques. In The Morgan Kaufmann Series in Data Management Systems (3rd ed.). Elesivier. https://doi.org/10.1109/ICMIRA.2013.45. 2012*

[7] *Wahyuni, R., Ma'ruf, A., & Mulyono, E. Hubungan Pola Makan Terhadap Kadar Gula Darah Penderita Diabetes Mellitus.*

 DOI : 10.37823/insight.v5i1.300

Jurnal Medika Karya Ilmiah Kesehatan, 4(2). http://jurnal.stikeswhs.ac.id/index.php/medika. 2019

[8] *Hayat, C. Identifikasi Dini Penyakit Diabetes Melitus Menggunakan Expert System Builder Early Identification of Diabetes Mellitus Disease Using Expert System Builder. Jurnal Teknik Dan Ilmu Komputer, 5(20), 431–445. 2016*

[9] *Pavithra Devi, & Jayanthi, A. A STUDY ON MACHINE LEARNING ALGORITHM IN MEDICAL DIAGNOSIS. International Journal of Advanced Research in Computer Science, 9(4), 42–46. 2018*

[10] *Kilimci, Z. H., & Ganiz, M. C. Evaluation of Classification Models for Language Processing. 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA). 2015*

[11] *Listiowarni, I., & Setyaningsih, E. R. Analisis Kinerja Smoothing pada Naive Bayes untuk Pengkategorian Soal Ujian. Jurnal Teknologi Dan Manajemen Informatika, 4(2). 2018*

[12] *Aprilia, R., Muludi, K., & Aristoteles. Pemetaan Sebaran Asal Siswa Dan Klasifikasi Jarak Asal Siswa Sma Negeri Di Kabupaten Pringsewu Menggunakan Metode Naïve Bayes. Jurnal Komputasi Ilmu Komputer Unila, 4(2), 52–66. 2016*

[13] *Tandra, H. Segala Sesuatu yang Harus Anda Ketahui Tentang Diabetes.Gramedia. 2017*

[14] *Informatikalogi. Algoritma Naive Bayes. https://informatikalogi.com/algoritma- naive-bayes/. 2021*

[15] *Packt.Implementing 3 NaiveBayes classifiers in scikit-learn. https://hub.packtpub.com/implementing-3-naive-bayes-classifiers-in-scikit-learn/. 2018*

 DOI : 10.37823/insight.v5i1.300