ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 1, Maret 2022

OPTIMASI ALGORITMA DATA MINING MENGGUNAKAN BACKWARD ELIMINATION UNTUK KLASIFIKASI PENYAKIT DIABETES

Muhammad Abid Wiratama[1], Windha Mega Pradnya[2]

[1,2] Amikom Yogyakarta University

email: abidwiratama16@gmail.com[1], windha@amikom.ac.id[2]

Abstrak

Diabetes Mellitus (DM) adalah penyakit metabolik kronis yang ditandai dengan peningkatan kadar gula darah, yang dengan seiring berjalannya waktu akan menyebabkan kerusakan pada organ tubuh lainnya. Menurut situs resmi Word Health Organization (WHO) sekitar 422 juta orang di seluruh dunia menderita diabetes. Di Indonesia sendiri menurut situs databox, Indonesia menempati urutan ke-5 kasus diabetes terbanyak di dunia. Keterlambatan dalam diagnosis penyakit diabetes adalah satu penyebab terjadinya lonjakan jumlah kematian maka dari itu tindakan awal yang harus dilakukan adalah deteksi dini.. Dari banyaknya kasus tersebut maka dihasilkan data pasien diabetes yang dapat diolah. Tujuan penelitian ini adalah untuk mengetahui algoritma klasifikasi terbaik dari akurasi dan nilai AUC tertinggi untuk melakukan deteksi dini penyakit diabetes. Algoritma yang akan teliti adalah algoritma KNN, Naïve Bayes, dan C4.5. Algoritma akan dilakukan optimasi menggunakan metode backward elimination. Metode penelitian penelitian ini akan diselesaikan menggunakan metode CRISP-DM. Hasil penelitian adalah model sebelum dioptimasi adalah algoritma KNN akurasi 92,8% dan auc 0,942, algoritma Naïve Bayes akurasi 88,0% dan auc 0,912, , algoritma C4.5 akurasi 96,7% dan auc 0,956, sedangkan hasil model setelah dioptimasi adalah algoritma KNN akurasi 97,6%dan auc 0,973, algoritma Naïve Bayes akurasi 89,4% dan auc 0,958, algoritma C4.5 akurasi 97,5% dan auc 0,988. Kesimpulan dari penilitian ini adalah algoritma terbaik dari akurasi adalah algoritma KNN yang sudah dioptimasi dengan akurasi 0,976 dan dari auc yang dihasilkan adalah algoritma C4.5 yang sudah dioptimasi dengan nilai auc 0,988.

Kata kunci: Diabetes Mellitus, Klasifikasi, Backward Elimination, KNN, Naïve Bayes, C4.5

Abstract

Diabetes Mellitus (DM) is a chronic metabolic disease characterized by increased blood sugar levels, which over time will cause damage to other body organs. According to the official website of the World Health Organization (WHO), about 422 million people worldwide suffer from diabetes. In Indonesia, according to the databox site, Indonesia ranks the 5th most cases of diabetes in the world. The delay in diagnosing diabetes is one of the causes of the spike in the number of deaths, therefore the initial action that must be taken is early detection. The purpose of this study was to determine the best classification algorithm with the highest accuracy and AUC value for early detection of diabetes. The algorithms that will be studied are the KNN, Naïve Bayes, and C4.5 algorithms. The algorithm will be optimized using the backward elimination method. The research method of this research will be completed using the CRISP-DM method. The results of the study are the model before being optimized is KNN algorithm accuracy 92.8% and auc 0.942, Naïve Bayes algorithm 88.0% accuracy and auc 0.912,, C4. 5 algorithm accuracy 96.7% and auc 0.956, while the results of the model after being optimized is KNN algorithm accuracy 97.6% and auc 0.973, Naïve Bayes algorithm 89.4% accuracy and auc 0.958, C4. 5 algorithm accuracy 97.5% and auc 0.988. The conclusion from this research is the best algorithm of accuracy is the algorithm KNN which is done with accuracy 0,976 and of the auc that is produced is the algorithm C4.5 which is done with auc values 0,988.

Keywords : *Diabetes Mellitus, Classification, Backward Elimination, KNN, Naïve Bayes, C4.5*

Diterima Redaksi: 04-03-2022 | **Selesai Revisi:** 25-03-2022 | **Diterbitkan Online:** 31-03-2022
DOI: http://dx.doi.org/10.23887/janapati.v11i1.45282

PENDAHULUAN

Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah yang melebihi batas normal [1]. Menurut situs resmi *Word Health Organization (WHO)* sekitar 422 juta orang di seluruh dunia menderita diabetes, mayoritas yang tinggal di negara berpenghasilan rendah dan menengah dan 1,5 juta kematian secara langsung dikaitkan dengan dengan diabetes setiap tahun. Menurut situs resmi *Word Health Organization (WHO)* sekitar 422 juta orang di seluruh dunia menderita diabetes, mayoritas yang tinggal di negara berpenghasilan rendah dan menengah dan 1,5 juta kematian secara langsung dikaitkan dengan dengan diabetes setiap tahun [2]. Di Indonesia sendiri kasus diabetes tidak kalah banyaknya. Dari situs data databox kita dapat lihat Indonesia berada di posisi ke-5 negara dengan kasus diabetes tertinggi setelah Amerika Serikat. Dengan banyaknya kasus diabetes di Indonesia dihasilkan data dari pasien diabetes dan dari data pasien diabetes tersebut bisa diolah menggunakan teknik data mining untuk melakukan deteksi dini penyakit diabetes.

Dari latar belakang tersebut, dapat dirumuskan masalah seperti berikut :

1. Berapa tingkat akurasi dan nilai AUC yang didapatkan algoritma C4.5, Naïve Bayes, dan K-Nearest Neighbor dengan dikombinasikan atau tidak dengan backward elimination ?
2. Apakah ada perbedaan yang signifikan akurasi antara sebelum dan setelah algoritma dilakukan optimasi menggunakan backward elimination ?
3. Apa algoritma terbaik untuk klasifikasi penyakit diabetes berdasarkan akurasi dan nilai auc yang dihasilkan?

Tujuan penelitian ini adalah untuk mengetahui algoritma terbaik dari segi nilai akurasi dan nilai AUC tertinggi untuk melakukan klasifikasi atau deteksi dini penyakit diabetes dan algoritma yang dibandingkan adalah C4.5, Naïve Bayes, dan K-Nearest Neighbor. Pengambilan data yang dilakukan adalah dengan menggunakan *dataset statistic* yaitu pengambilan data dari sebuah bank data. Dalam penelitian ini digunakan dataset "*Early stage diabetes risk prediction dataset"* yang diambil dari situs *UCI Machine Learning Repository* dengan 520 record dan juga 17 atribut didalamnya.

Penelitian ini dibuat berdasarkan penelitian yang sudah dilakukan atau penelitian terdahulu yaitu Penelitian [3] berjudul "Optimasi *Backward Elimination* untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritma k-Nearest Neighbor (k-NN) dan Naïve Bayes". Penelitian ini bertujuan untuk optimasi fitur Backward Elimination pada klasifikasi kepuasan pelanggan dengan algoritma k-NN dan Naïve Bayes. Hasilnya didapati akurasi sebesar 97,28% dan AUC 0.996.

Penelitian lainnya [4] adalah penelitian yang berjudul "Klasifikasi Nasabah Asuransi Jiwa Menggunakan Algoritma Naïve Bayes Berbasis *Backward Elimination".* Penelitian ini bertujuan untuk mengklasifikasi nasabah asuransi yang lancer dan tidak lancer membayar premi penelitian ini menggunakan algoritma *Naïve Bayes* berbasis *backward elimination.* Hasilnya akurasi sebesar 83,32% menjadi 85,89% setelah dioptimasi.

Penelitian yang dilakukan oleh [5] yang berjudul "Optimasi *Linear Sampling* dan *Information Gain* pada Algoritma *Decision Tree* untuk Diagnosis Penyakit Diabetes". Pada penelitian ini dilakukan optimasi Algoritma *Decision Tree* menggunakan *Information Gain* dan *Linear Sampling* untuk meningkatkan akurasi yang dihasilkan, hasilnya tanpa dilakukan optimasi akurasi algoritma *Decision Tree* yang dihasilkan adalah 90,38% dan setelah dilakukan optimasi menggunakan *Information Gain* dan *Linear Sampling* tingkat akurasi meningkat menjadi 99,04%.

Pada penelitian lainnya yang dilakukan oleh [6] yang berjudul "Penerapan *Particle Swarm Optimization* Untuk Meningkatkan Kinerja Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit Diabetes. Pada penelitian ini bertujuan untuk mengetahui tingkat akurasi dari algoritma *k*-NN setelah dilakukan optimasi menggunakan *Particle Swarm Optimization* (PSO) sebagai seleksi fitur. Hasilnya tingkat akurasi sebelum dilakukan optimasi adalah 75% dan setelah dilakukan optimasi akurasi meningkat menjadi 77,214%.

Pada penelitian selanjutnya yang dilakukan oleh [7] yang berjudul "Klasifikasi Penyakit Diabetes Menggunakan Metode *CFS* dan *ROS* dengan Algoritma J48 Berbasis Adaboost". Penelitian ini bertujuan untuk meningkatkan hasil akurasi untuk klasifikasi penyakit diabetes agar lebih baik dan optimal. Hasilnya setelah dilakukan optimasi menggunakan metode tersebut tingkat akurasi yang dihasilkan adalah 92,3%.

Pada penelitian yang dilakukan oleh [8] yang berjudul "Optimasi Algoritma Naïve Bayes Berbasis *Particle Swarm Optimization* (PSO) dan *Stratified* Untuk Meningkatkan Akurasi Prediksi Penyakit Diabetes". Hasilnya sebelum dilakukan optimasi akurasi yang didapatkan

adalah 75,40% dan nilai AUC sebesar 0,829 dan setelah dilakukan optimasi nilai akurasi meningkat menjadi 90% dan nilai AUC 0,924.

Dari penelitian yang sudah dilakukan belum terdapat kasus yang membahas mengenai penyakit diabetes dan juga dikarenakan kasus diabetes sudah marak baik di Indonesia maupun seluruh dunia maka dari itu dengan banyaknya data pasien diabetes dapat dimanfaatkan untuk mendeteksi dini penyakit tersebut, maka dari itu penelitian ini dibuat.

STUDI LITERATUR

A. Diabetes

Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah yang melebihi batas normal [1]. Kondisi tubuh yang normal akan menghasilkan insulin dengan sendirinya yang berfungsi untuk mencukupi kadar gula secara normal. Ketika tubuh menghasilkan insulin yang tidak mencukupi kebutuhan, maka kadar gula darah akan naik dan terjadilah penyakit diabetes. Adapun gejala yang sering dialami oleh penderita penyakit diabetes adalah rasa haus dan lapar yang tak kunjung hilang, dan terlalu sering buang air kecil [9]. Diabetes diklasifikasikan menjadi beberapa tipe yaitu diabetes tipe 1, diabetes tipe 2, dan diabetes gestasional [10]. Diabetes tipe 1 dapat terjadi dikarenakan terdapat kerusakan sel autoimun pada pankreas yang menyebabkan produksi insulin hilang. Dan diabetes tipe 2 dapat terjadi dikarenakan hilangnya kemampuan tubuh untuk merespon insulin dan menyebabkan tubuh tidak bisa menyerap glukosa dan menyebabkan tumpukan gula dalam darah. Penyakit diabetes yang berkelanjutan dapat menyebabkan komplikasi, yang sering terjadi adalah stroke, amputasi, kebutaan, kecacatan, dan lain-lain [10].

B. CRISP-DM

Cross-Industry Standart Process for Data Mining (CRISP-DM) adalah salah satu model atau framework dalam data mining yang awalnya (1996) dibagun oleh 5 perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA. Framework ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan methodology standard non-proprietary bagi data mining [11]. Menurut CRISP-DM, data mining memiliki siklus hidup yang terdiri dari enam fase, dan fase tersebut bersifat adaptif, yaitu fase berikutnya sangat bergantung pada hasil yang terkait dengan fase sebelumnya. Ketergantungan paling signifikan antar fase ditunjukan oleh panah.Berikut adalah fase CRISP-DM :

Dalam siklus CRISP-DM terdapat 6 (enam) fase yaitu sebagai berikut [11]:

1. **Fase pemahaman bisnis** (*Business Understanding Phase)*
   a. Menentukan tujuan dan kebutuhan proyek secara detaul dalam bisnis.
   b. Mengubah tujuan dan batasan menjadi rumusan masalah data mining.
   c. Menyiapkan strategi untuk mencapai tujuan.
2. **Fase pemahaman data** (*Data Understanding Phase)*
   a. Mengumpulkan data
   b. Menggunakan analisis penyelidikan data
   c. Mengevaluasi kualitas data
3. **Fase pengolahan data** (*Data Preparation Phase)*
   a. Menyiapkan data awal.
   b. Pilih kasus dan variabel yang akan dianalisis.
   c. Jika dibutuhkan, lakukan perubahan beberapa variabel.
   d. Siapkan data awal untuk pemodelan.
4. **Fase pemodelan** (*Modeling Phase)*
   a. Memilih dan implementasi model.
   b. Jika dibutuhkan, Kembali ke proses pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan kebutuhan teknik data mining.
5. **Fase evaluasi** (*Evaluation Phase)*
   a. Evaluasi satu atau lebih model yang dilakukan pada fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum dipublish.
   b. Menentukan apakah model memenuhi tujuan atau tidak
   c. Menentukan apakah masalah atau penelitian yang tidak tertangani dengan baik.
   d. Mengambil keputusan yang berkaitan dengan penggunaan hasil dari data mining.
6. **Fase penyebaran** (*Deployment Phase)*
   a. Menggunakan model yang dihasilkan.
   b. Contoh sederhana penyebaran : pembuatan laporan.

C. Algoritma K-Nearest Neighbor

Pengklasifikasi *k-nearest-neighbor* mencari ruang pola untuk k tupel pelatihan yang paling dekat dengan tupel yang tidak diketahui. Tupel pelatihan k ini adalah k "tetangga terdekat" dari tupel yang tidak diketahui .Untuk menghitung jarak digunakan rumus *eucledian distance* untuk menentukan tetangga terdekat

[12]. Rumus *eucledian distance* adalah pada rumus 1 sebagai berikut :

$$d = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

Dimana :
d : jarak *Euclidean,*
x_{2i} : nilai pada data *testing* ke *-i,*
x_{1i} : nilai pada data *training* ke *-i,*
p : banyaknya atribut

D. Algoritma Naïve Bayes

Klasifikasi Bayesian didasarkan pada teorema Bayes, yang dijelaskan selanjutnya. Studi yang membandingkan algoritma klasifikasi telah menemukan pengklasifikasi Bayesian sederhana yang dikenal sebagai naïve. Pengklasifikasi Bayesian agar sebanding dalam kinerja dengan pohon keputusan dan pengklasifikasi jaringan saraf terpilih. Pengklasifikasi Bayesian juga menunjukkan akurasi dan kecepatan tinggi ketika diterapkan ke database besar [12]. Untuk membangun *Naïve Bayes* harus menghitung probabilitas setiap atribut yang diberi *class.* Persamaan teorema *Bayes* pada rumus 2 adalah sebagai berikut [12] :

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2)$$

Dimana :
X : Data dengan *class* yang belum diketahui,
H : Hipotesis data X merupakan suatu *class* spesifik,
P(H|X) : Probabilitas hipotesis berdasarkan kondisi,
P(H) : Probabilitas hipotesis H,
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H,
P(X) : Probabilitas X

E. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang merupakan pengembangan dari algoritma pendahulunya yaitu algoritma ID3, yang dimana perkembangan yang terlihat adalah bisa mengatasi *missing* data, data kontinu, dan *pruning* [12]. Terdapat perhitungan yang terjadi untuk memilih atribut akar. Pemilihan tersebut didasari oleh nilai gain tertinggi dari atribut yang ada. Untuk menghitung nilai gain digunakan rumus 3 berikut:

$$Gain = Entrophy\ (S) - Zi; \quad = i \quad Entrophy(Si) \quad (3)$$

Dimana :
S : himpunan kasus,
A : atribut,
N : jumlah partisi atribut A,
\Si\ : jumlah kasus pada partisi ke-i,
$|S^r|$: jumlah kasus dalam S

F. Backward Elimination

Backward Elimination adalah metode feature selection yang termasuk kedalam tipe wrapper yang bertujuan untuk mengoptimalkan kinerja model dari suatu algoritma [13]. Untuk lebih jelasnya dapat kita lihat pada contoh berikut [12]:
Initial attribute set:
{*A*1, *A*2, *A*3, *A*4, *A*5, *A*6}
=> {*A*1, *A*3, *A*4, *A*5, *A*6}
=> {*A*1, *A*4, *A*5, *A*6}
=> Reduced attribute set:
{*A*1, *A*4, *A*6}

G. K-Fold Cross Validation

Validasi adalah proses pengujian performa algoritma. Umumnya validasi dilakukan dengan cara mengulang proses perhitungan beberapa kali [14]. Pada penelitian ini proses validasi dilakukan dengan menggunakan metode *K-Fold Cross Validation. K-Fold Cross Validation* merupakan metode validasi yang membagi data ke dalam *k* bagian dan dari bagian yang terbagi dilakukan klasifikasi. Dengan menggunakan metode ini akan dilakukan percobaan sebanyak *k* kali. Setiap percobaan akan menggunakan satu data *testing* dan *k-1* bagian akan menjadi data *training* [15]. Contohnya jika *k* yang digunakan 10 maka untuk data *testing* sebanyak 10% dan data *training* 90% dari total data.

H. Uji Paired T-Test

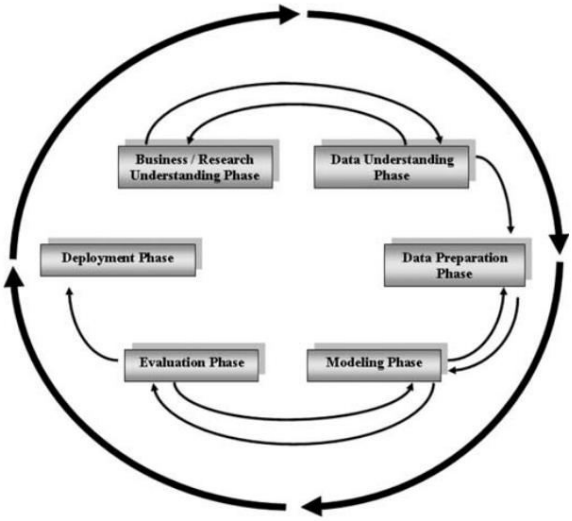
Uji paired t-test atau yang biasa disebut uji t berpasangan adalah sebuah metode pengujian hipotesis yang dimana data yang digunakan tidak bebas dalam artian berpasangan [15]. Ciri khas dari paired t-test adalah objek penelitian yang digunakan diberikan perlakuan yang berbeda. Walaupun menggunakan individu yang sama, peneliti tetap menggunakan 2 macam data sample, yaitu data dengan perlakuan a dan data dengan perlakuan b [15].
Hipotesis dari sebuah kasus dapat ditulis seperti berikut [15]:
H0 = μ1 - μ2 = 0
H1 = μ1 - μ2 ≠ 0
Ha yaitu adalah bahwa selisih sebenarnya dari kedua rata-rata tidak sama dengan nol.

Interpretasi

a) Langkah pertama adalah menentukan nilai α atau alpha atau taraf signifikansi.
b) Bandingkan t hitung dengan t tabel
c) Apabila :
t hitung > t tabel → H_0 ditolak atau ada beda yang signifikan
t hitung < tabel → H_0 diterima atau tidak ada beda yang signifikan

METODE

Penelitian ini menggunakan *framework data mining* yaitu CRISP-DM atau *cross industry standart process for data mining.* Menurut CRISP-DM, data mining memiliki siklus hidup yang terdiri dari enam fase, dan fase tersebut bersifat adaptif, yaitu fase berikutnya sangat bergantung pada hasil yang terkait dengan fase sebelumnya. Ketergantungan paling signifikan antar fase ditunjukan oleh panah. Keenam fase tersebut dapat dilihat pada gambar 1 berikut :



Gambar 1. Tahapan CRISP-DM

1. *Business Understanding*

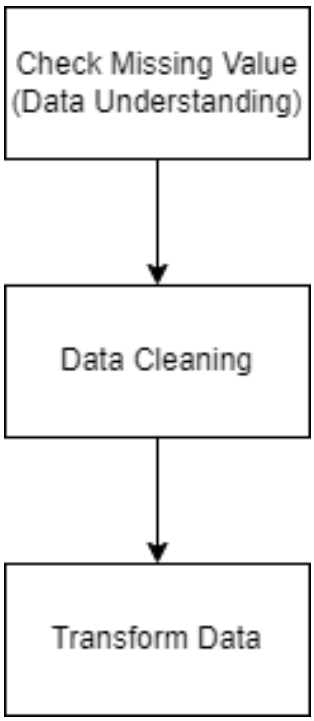
Tahap ini dilakukan untuk mendapatkan pemahaman tujuan berdasarkan perspektif bisnis untuk kemudian diubah menjadi definisi masalah data mining. Setelah itu ditentukan bagaimana solusi yang tepat dalam menangani permasalahan tersebut. Adapun masalah penelitian ini adalah penyakit diabetes menjadi 10 besar penyakit dengan kematian tertinggi dan angka tersebut meningkat setiap tahunnya. Keterlambatan diagnosis menjadi penyebab utama lonjakan kematian. Solusi yang ditawarkan yaitu penerapan data mining menggunakan algoritme k-NN, C4.5, dan Naïve Bayes untuk mencari algoritma terbaik dengan Backward Elimination dalam pengurangan atribut dan peningkatan akurasi untuk deteksi dini penyakit diabetes.

2. *Data Understanding*

Setelah melalui tahapan pemahaman bisnis, selanjutnya kita akan tahapan data understanding atau pemahaman data. Pada tahap ini kita akan mengumpulkan semua data yang akan digunakan pada penelitian yaitu dataset Early Stage Diabetes Risk Prediction Dataset. Setelah itu akan ditampilkan beberapa sample data dari dataset tersebut untuk mengetahui isi dari dataset tersebut. Setelah ditampilkan langkah selanjutnya adalah penjabaran data dan juga check missing value. Dari penjabaran atribut tersebut pemahaman terhadap data yang akan digunakan lebih baik. Setelah penjabaran atribut dilakukan, langkah selanjutnya adalah memeriksa kualitas dari *dataset* diabetes yang akan digunakan meliputi data *missing*. *Missing value* adalah terdapat data yang hilang.

3. *Data Preparation*

Tahapan ketiga adalah data preparation atau persiapan data. Dari pemahaman data kita akan lebih mengetahui struktur maupun pola dari data tersebut, maka dari itu untuk persiapan data akan dilakukan beberapa langkah untuk mempersiapkan data untuk tahap selanjutnya yaitu modelling. Pada tahap ini yang akan dilakukan adalah data cleaning dan transform data.



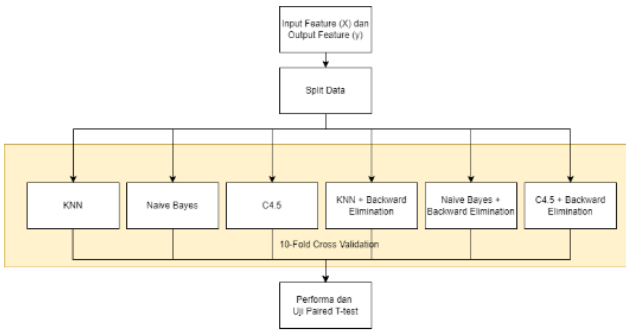
Gambar 2. Langkah Data Preparation

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 1, Maret 2022

Dari tahap data understanding kita dapat mengetahui kualitas dari data dengan check missing value, setelah itu kita dapat mengetahui jumlah dari data missing yang ada pada data, lalu untuk mengatasi masalah tersebut kita dapat melakukan langkah data cleaning. Data cleaning dilakukan untuk mengatasi missing value, hal tersebut dapat diatasi dengan menggunakan library Pandas pada bahasa pemograman python dengan cara mengapus baris data yang memiliki missing value sehingga missing value akan dapat diatasi

4. *Modelling*

Setelah memalui tahapan data preparation maka data siap untuk diolah pada tahapan modelling, pada tahap ini ada beberapa Langkah yang akan dilakukan yaitu input x dan y, split data, dan selanjutnya adalah modelling dan uji t. Dan untuk penggambarannya adalah seperti gambar 2 berikut :



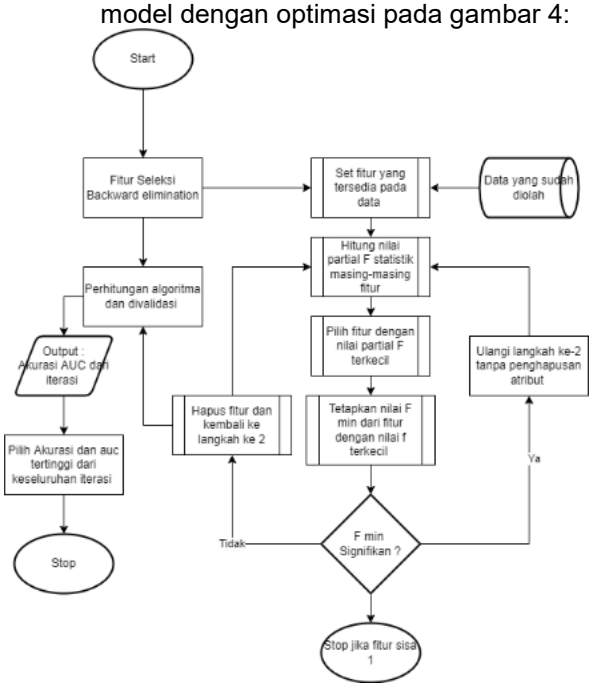
Gambar 3. Alur Modelling

Setelah memalui tahapan *data preparation* maka data siap untuk diolah pada tahapan *modelling* seperti pada gambar 3. Pada tahapan *modelling* ada beberapa langkah yang dilakukan adalah sebagai berikut :

1. Input Feature (X) dan Output Feature (y)
   Pada langkah ini yang dilakukan adalah menentukan atribut dan class. Untuk melakukannya akan digunakan library Pandas.
2. Split Data
   Langkah selanjutnya adalah split data. Split data dilakukan untuk membagi data training dan data testing. Untuk melakukannya, akan digunakan module train_test_split pada package sklearn.model_selection.
3. Modelling
   Langkah ketiga adalah tahap modelling. Pada langkah modelling ini terdapat dua model. Model tersebut akan divalidasi menggunakan 10-fold validation untuk mendapatkan akurasi yang optimal. Model tersebut adalah sebagai berikut :
   - Model Tanpa Optimasi

Gambar 4. Flowchart Model Tanpa Optimasi

Selanjutnya berikut adalah flowchat model dengan optimasi pada gambar 4:



Gambar 5. Flowchart Model Dengan Optimasi

3. Performa

Dari langkah modelling didapati performa berupa akurasi dan nilai AUC yang akan dibahas lebih lanjut pada tahap selanjutnya yaitu Evaluation.

4. Uji Paired T-Test
   Sebelum masuk ke tahap evaluation akan dilakukan perumusan hipotesa dan penetapan taraf signifikansi yaitu sebagai berikut :
   α = 0,05 atau 5 %
   H0 = Tidak ada perbedaan signifikan sebelum dan sesudah dilakukan optimasi menggunakan backward elimination pada akurasi yang dihasilkan.
   H1 = Ada perbedaan signifikan sebelum dan sesudah dilakukan optimasi menggunakan backward elimination pada akurasi yang dihasilkan.
   Keterangan :
   α = Alpha atau taraf signifikansi
   H0 = Hipotesis null
   H1 = Hipotesis alternatif

5. *Evaluation*
   Tahap kelima adalah evaluation atau evaluasi. Evaluasi dilakukan untuk menentukan apakah modelling sudah mencapai tujuan yang ditetapkan pada tahap business understanding. Evaluasi yang dilakukan dilihat dari akurasi dan nilai AUC yang dihasilkan oleh masing-masing model. Setelah dilakukan evaluasi performa yang dihasilkan maka akan dilakukan uji paired t-test untuk mengetahui apakah terdapat perbedaan akurasi dari sebelum dan sesudah dilakukan optimasi.

6. *Deployment*
   Tahapan terakhir adalah Deployment atau penyebaran. Dari hasil akurasi dan nilai AUC yang dihasilkan dari model dan sudah dievaluasi pada tahap sebelumnya akan dibuat laporan sederhana dalam bentuk informasi atau pengetahuan dan dipresentasikan sehingga dapat digunakan oleh pengguna.

HASIL DAN PEMBAHASAN

1. *Business Understanding*
   Pada tahap pemahaman bisnis berfokus pada latar belakang masalah dan tujuan penelitian. Diabetes menjadi 10 besar penyakit dengan kematian tertinggi, bahkan angkanya terus meningkat selama beberapa dekade terakhir. Hal tersebut perlu kita teliti penyebabnya, setelah dilakukan penelitian mendalam mengenai penyakit diabetes para peneliti mendapati salah satu penyebab penyakit diabetes menjadi penyakit yang mematikan yaitu dikarenakan keterlambatan dalam diagnosis sehingga hal tersebut memicu komplikasi yang membuat pasien semakin parah. Dari permasalahan tersebut tujuan penelitian ini adalah untuk mencari algoritma terbaik dari segi tingkat akurasi dan nilai AUC untuk deteksi dini penyakit diabetes jika dikombinasikan dan tidak dengan backward elimination, sehingga permasalahan yang sudah diuraikan akan teratasi.

2. *Data Understanding*
   Penelitian ini menggunakan dataset Early Stage Diabetes Risk Prediction Dataset yang didapatkan dari situs UCI Machine Learning Repository. Dataset mempunyai 16 atribut dan 1 class dengan 520 record data. Berikut sampel dataset Early Stage Diabetes Risk Prediction yang terdapat pada tabel 1:

Tabel 1. Sampel Dataset

Age (Tahun)	…	Class
40	…	Positive
58	…	Positive
41	…	Positive
45	…	Positive
60	…	Positive
55	…	Positive
57	…	Positive

Setelah dilakukan tabulasi sampel, langkah selanjutnya adalah penjabaran atribut yaitu digunakan untuk mengetahui definisi dari setiap atribut, tipe data, dan juga value yaitu seperti pada tabel 2 berikut :

Tabel 2. Penjabaran Atribut

Atribut	Tipe Data	Value	Definisi
Partial Paresis	Int64	Yes,No	Kelemahan gerakan volunter
Muscle Stiffness	Int64	Yes,No	Otot terasa kencang
Alopecia	Int64	Yes,No	Rambut rontok
Obesity	Int64	Yes,No	Lemak berlebih
Age	Int64	Yes,No	Umur
Gender	Int64	Yes,No	Jenis Kelamin
Polydipsia	Int64	Yes,No	Haus berlebih
Polyuria	Int64	Yes,No	Sering BAK
Sudden weight	Int64	Yes,No	Turun berat badan tiba-

loss			tiba
Weakness	Int64	Yes,No	Tubuh lemah
Polyphagia	Int64	Yes,No	Nafsu makan berlebih
Genital thrush	Int64	Yes,No	Infeksi area genital
Visual blurring	Int64	Yes,No	Penglihatan kabur
Itching	Int64	Yes,No	Tubuh terasa gatal
Irritability	Int64	Yes,No	Mudah marah
Delayed healing	Int64	Yes,No	Penyembuhan luka yang lambat
Class	Int64	Positive , Negative	Pasien menghidap diabetes atau tidak

Setelah dilakukan penjabaran data selanjutnya adalah check missing value yaitu cek apakah terdapat data *missing* pada dataset. Hasilnya dari pengecekan menggunakan Google Colab hasilnya tidak terdapat missing value pada dataset.

3. *Data Preparation*

Setelah dilakukan check missing value pada tahap data understanding pada tahap ini akan dilakukan data cleaning dan transform data.

Data Cleaning dilakukan untuk membersihkan data missing pada dataset, namun dikarenakan dari check missing value yang sudah dilakukan pada tahap sebelumnya maka langkah ini bisa kita lewati. Langkah selanjutnya setelah data cleaning adalah transform data. Transform data dilakukan untuk mengubah data ke dalam bentuk data yang diinginkan,. Untuk lanjut ketahap berikutnya diperlukan transform data dari data categorical ke data numerical. Berikut hasil dari transform data pada gambar 5 :

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching
0	40	1	0	1	0	1	0	0	0	1
1	58	1	0	0	0	1	0	0	1	0
2	41	1	1	0	0	1	1	0	0	1
3	45	1	0	0	1	1	1	1	0	1
4	60	1	1	1	1	1	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...
515	39	0	1	1	1	0	1	0	0	1
516	48	0	1	1	1	1	1	0	0	1
517	58	0	1	1	1	1	1	0	1	0
518	32	0	0	0	0	1	0	0	1	1
519	42	1	0	0	0	0	0	0	0	0

520 rows × 17 columns

Gambar 6. Hasil Transform Data

4. *Modelling*

Model Tanpa Optimasi

1. Model KNN
   Menggunakan modul KNeighborClassifier dengan k = 1 pada package sklearn.neighbors dipadukan dengan 10-fold cross validation dengan menggunakan modul cross_val_score pada package sklearn.model_selection.
2. Model C4.5
   Menggunakan modul DecisionTreeClassifier pada package sklearn.tree dan dipadukan dengan 10-fold cross validation menggunakan modul cross_val_score pada package sklearn.model_selection.
3. Model Naïve Bayes
   Menggunakan modul GaussianNB pada package sklearn.naive_bayes dan dipadukan dengan 10-fold cross validation menggunakan modul cross_val_score pada package sklearn.model_selection.

Model Dengan Optimasi

1. Model KNN + Backward Elimination
   Modul KNeighborClassifier juga dimasukkan kedalam parameter modul sfs. Nilai AUC juga dapat dicari menggunakan modul sfs dengan cara mengganti parameter scoring menjadi ‘roc_auc’. Plotting akurasi dari modul sfs juga akan ditampilkan untuk dapat dianalisa dengan modul plot_sequential_feature_selection dari package mlxtend.plotting.
2. Model C4.5 + Backward Elimination
   Modul DecisionTreeClassifier juga dimasukkan kedalam parameter modul sfs. Nilai AUC juga dapat dicari menggunakan modul sfs dengan cara mengganti parameter scoring menjadi ‘roc_auc’. Plotting akurasi dari modul sfs juga akan ditampilkan untuk dapat dianalisa dengan modul plot_sequential_feature_selection dari package mlxtend.plotting.
3. Model Naïve Bayes + Backward Elimination
   Modul GaussianNB juga dimasukkan kedalam parameter modul sfs. Nilai AUC juga dapat dicari menggunakan modul sfs dengan cara mengganti parameter scoring menjadi ‘roc_auc’. Plotting akurasi dari modul sfs juga akan ditampilkan untuk dapat dianalisa dengan modul plot_sequential_feature_selection dari package mlxtend.plotting.

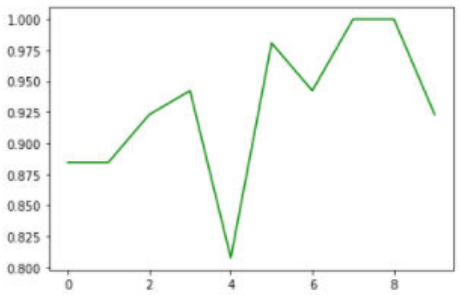
Uji T

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 1, Maret 2022

Sebelum masuk ke tahap evaluation akan dilakukan perumusan hipotesa dan penetapan taraf signifikansi yaitu sebagai berikut :
α = 0,05 atau 5 %
H0 = Tidak ada perbedaan signifikan sebelum dan sesudah dilakukan optimasi menggunakan backward elimination pada akurasi yang dihasilkan.
H1 = Ada perbedaan signifikan sebelum dan sesudah dilakukan optimasi menggunakan backward elimination pada akurasi yang dihasilkan.

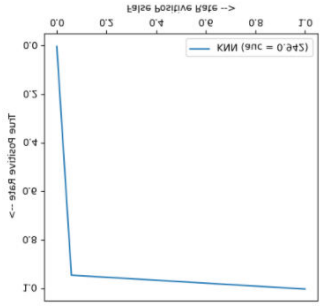
5. *Evaluation*

Model KNN



Gambar 7. *Validasi Akurasi Model KNN*

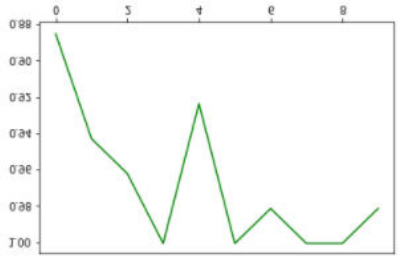
Dari 10 percobaan yang sudah dilakukan dihasilkan performa seperti gambar 6 dan untuk akurasi didapatkan dari rata-rata 10 percobaan yang dilakukan dan untuk akurasi model knn adalah **92,8%.**



Gambar 8. AUC Model KNN

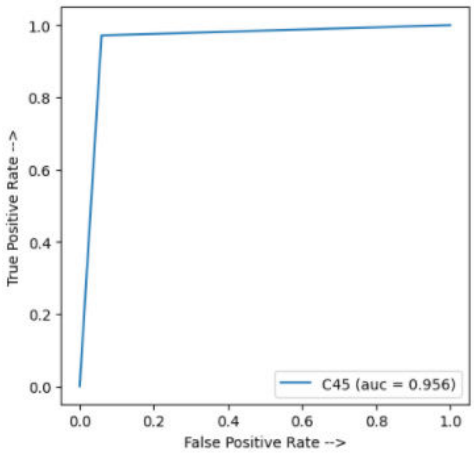
Dan untuk AUC dari model KNN adalah seperti pada gambar 7 yaitu 0.942.

Model C4.5



Gambar 9. Validasi Akurasi Model C4.5

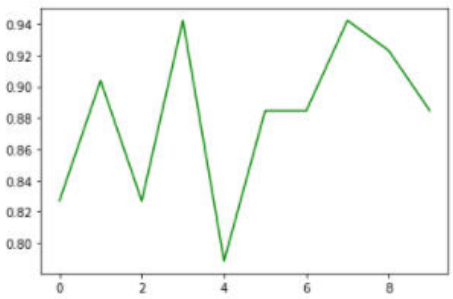
Dari 10 percobaan yang sudah dilakukan dihasilkan performa seperti gambar 8 dan untuk akurasi didapatkan dari rata-rata 10 percobaan yang dilakukan dan untuk akurasi model c4.5 adalah 96,7%.



Gambar 10. AUC Model C4.5

Dan untuk AUC dari model C4.5 aadalah seperti pada gambar 9 yaitu 0.956.

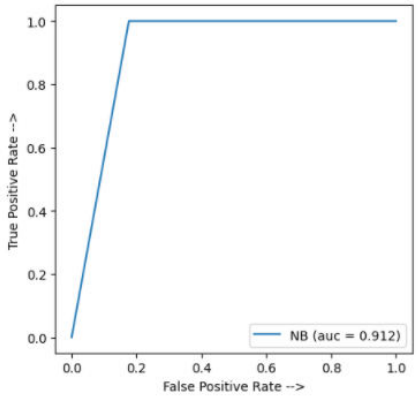
Model Naïve Bayes



Gambar 11. Validasi Akurasi Model Naive Bayes\

Dari 10 percobaan yang sudah dilakukan dihasilkan performa seperti gambar 10 dan untuk akurasi didapatkan dari rata-rata 10 percobaan yang dilakukan dan untuk akurasi model naïve bayes adalah 88,0%.

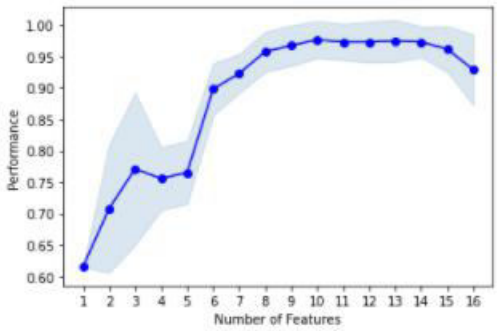
ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 1, Maret 2022



Gambar 12. AUC Model Naive Bayes

Dan untuk AUC dari model C4.5 aadalah seperti pada gambar 11 yaitu 0.912.

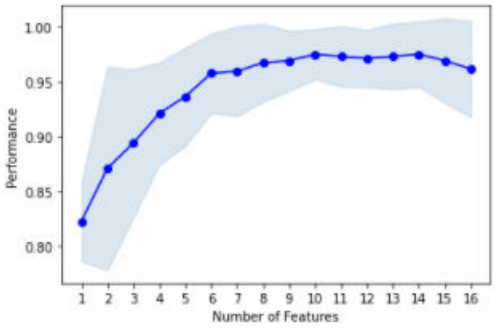
Model KNN + Backward Elimination



Gambar 13. Validasi Akurasi Model KNN + BE

Akurasi dari model KNN + Backward Elimination adalah 97,6% dengan 10 feature yang digunakan dan 6 feature dihilangkan. Dan AUC adalah sebesar 0,973.

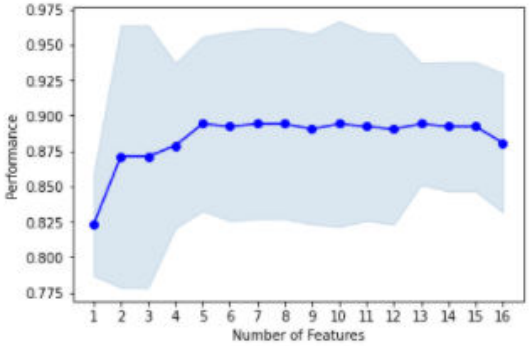
Model C4.5 + Backward Elimination



Gambar 14. Validasi Akurasi Model C4.5 + BE

Akurasi dari model C4.5 + Backward Elimination adalah 97,5% dengan 14 feature yang digunakan dan 2 feature dihilangkan. Dan AUC adalah sebesar 0,988.

Model Naïve Bayes + Backward Elimination



Gambar 15. Validasi Akurasi Model Naïve Bayes+ BE

Akurasi dari model Naïve Bayes + Backward Elimination adalah 89,4% dengan 13 feature yang digunakan dan 3 feature dihilangkan. Dan AUC adalah sebesar 0,958.

Uji T

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-1.829983	2	two-sided	0.208743	[-0.08, 0.03]	0.525561	1.037	0.087612

Gambar 16. Hasil Uji T

Dari gambar 15 dapat dikatahui bahwa nilai t hitung adalah -1,829 jika dibandingkan dengan nilai t tabel yaitu 4,302 maka nilai t hitung lebih kecil dari t tabel sehingga H0 diterima. Artinya tidak ada perbedaan signifikan akurasi sebelum dan sesudah dioptimasi.

6. Deployment

Dari permasalahan yang sudah dipaparkan dari tahap business understanding yaitu tingginya angka kematian dan kasus diabetes, maka dari itu penelitian ini dibuat untuk mencari algoritma terbaik untuk deteksi dini penyakit diabetes.
Setelah dilakukan pengujian didapatkan hasil seperti pada tabel 3 berikut :

Tabel 3. Keseluruhan Performa Model

Model	Sebelum		Sesudah	
	Akurasi (wt%)	AUC	Akurasi (wt%)	AUC
KNN	92,8%	0,942	97,6%	0,973
Naïve Bayes	88,0%	0,912	89,4%	0,958
C4.5	96,7%	0,956	97,5%	0,988

Dilihat dari tabel diatas adalah hasil akurasi dan nilai auc dari keseluruhan model dapat disimpulkan bahwasanya optimasi menggunakan metode backward elimination dapat meningkatkan nilai akurasi maupun

nilai auc dari ketiga algoritma yang diuji.Model dengan nilai akurasi tertinggi adalah model KNN yang dioptimasi dan Model dengan nilai auc tertinggi adalah model C4.5 yang dikombinasikan dengan backward elimination.

Dari tujuan bisnis yang sudah dipaparkan pada tahap pemahaman bisnis, dapat disimpulkan bahwasanya tujuan dapat dicapai.

KESIMPULAN

Dari penelitian yang sudah dilakukan, dihasilkan akurasi dan nilai auc dari masing-masing model, untuk model algoritma KNN mendapatkan akurasi sebesar 0,928 dan nilai auc 0,942, model Naïve Bayes mendapatkan akurasi 0,880 dan nilai auc 0,912, model C45 mendapatkan akurasi 0,967 dan nilai auc 0,956 dan setelah dioptimasi menggunakan metode backward elimination adalah model KNN + backward elimination mendapatkan akurasi 0,976 dan nilai auc 0,973, model Naïve Bayes + backward elimination mendapatkan akurasi 0,894 dan nilai auc 0,958, model C4.5 + backward elimination mendapatkan akurasi 0,975 dan nilai auc 0,988. Dari hasil akurasi dan nilai auc yang dihasilkan model dapat disimpulkan metode backward elimination dapat mengoptimalkan performa dari algoritma data mining dalam penelitian ini adalah algoritma KNN, Naïve Bayes, dan C4.5 dan dari nilai auc yang dihasilkan setelah dioptimasi ketiga algoritma masuk kedalam kategori excellent classification. Dari uji paired t-test yang dilakukan, didapati bahwa t hitung adalah -1,829 yang dimana lebih kecil dari t tabel yaitu 4,302 sehingga h0 diterima dan dapat disimpulkan tidak ada perbedaan yang signifikan dari akurasi sebelum dan sesudah dilakukan optimasi, namun hal tersebut tidak menjadi masalah dikarenakan dalam hal kesehatan peningkatan akurasi sekecil apapun sangat berpengaruh pada diagnosis yang dihasilkan nantinya. Dari performa yang dihasilkan dari model yang sudah diuji, dapat disimpulkan bahwasanya algoritma terbaik dari segi akurasi adalah algoritma KNN yang sudah dioptimasi dengan nilai akurasi 0,976 dan dari segi nilai auc yang dihasilkan adalah algoritma C4.5 yang sudah dioptimasi dengan nilai auc 0,988.

UCAPAN TERIMAKASIH

Persembahan tugas akhir dan rasa terimakasih kepada :

1. Allah SWT Pencipta alam semesta yang memberiku hidup yang berkah dan rizkinya.
2. Kedua orang tua yaitu bapak dan ibu tercinta yang selalu memberikan dukungannya dan pengorbanannya yang tak terkira harganya, doa serta kasih sayangnya tak akan pernah ananda lupakan sepanjang hidup.
3. Untuk seluruh keluargaku dan saudara dimanapun berada yang secara tidak langsung memberikan dorongan dan semangat yang tak pernah putus sehingga tugas akhir ini berjalan lancer.
4. Teruntuk sahabat dan kawan seperjuangan yang selalu memberikan semangat semoga kita sukses bersama di masa depan kelak.

REFERENSI

[1] W. D. Septiani dan M. Marlina, “Comparison Of Decision Tree, Naïve Bayes, And Neural Network Algorithm For Early Detection Of Diabetes,” *Pilar Nusa Mandiri J. Comput. Inf. Syst.*, vol. 17, no. 1, Art. no. 1, Mar 2021, doi: 10.33480/pilar.v17i1.2213.

[2] L. Lenny and F. Fridalina, “Faktor-Faktor yang Berhubungan dengan Kepatuhan Berobat Jalan Pasien Diabetes Mellitus Tipe II,” Jurnal Ilmu Kesehatan Masyarakat, vol. 7, no. 02, pp. 85–93, Jul. 2018, doi: 10.33221/jikm.v7i02.110.

[3] Yunitasari, H. S. Hopipah, dan R. Mayasari, “Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritme k-nearest neighbor (k-NN) and Naive Bayes,” *Technomedia J.*, vol. 6, no. 1, hlm. 99–110, Jul 2021, doi: 10.33050/tmj.v6i1.1531.

[4] B. Betrisandi, “KLASIFIKASI NASABAH ASURANSI JIWA MENGGUNAKAN ALGORITMA NAIVE BAYES BERBASIS BACKWARD ELIMINATION,” *Ilk. J. Ilm.*, vol. 9, no. 1, Art. no. 1, Apr 2017, doi: 10.33096/ilkom.v9i1.116.96-101.

[5] A. A. Abdillah, “Optimasi Linear Sampling dan Information Gain pada Algoritma Decision Tree untuk Diagnosis Penyakit Diabetes,” *MULTINETICS*, vol. 7, no. 1, Art. no. 1, Mei 2021.

[6] Sutrisno, “Classification of Diabetes Particle Swarm Optimization K-Nearest Neighbor,” Bachelor thesis, Universitas Bumigora, 2021.

[7] D. Pramadhana, “Klasifikasi Penyakit Diabetes Menggunakan Metode CFS Dan ROS dengan Algoritma J48 Berbasis Adaboost,” *Edumatic J. Pendidik. Inform.*, vol. 5, no. 1, hlm. 89–98, Jun 2021, doi: 10.29408/edumatic.v5i1.3336.

[8] A. Muhidin dan M. Casdi2, “OPTIMASI ALGORITMA NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION

(PSO) DAN STRATIFIED UNTUK MENINGKATKAN AKURASI PREDIKSI PENYAKIT DIABETES," *J. SIGMA*, vol. 10, no. 1, hlm. 151–157, Jun 2019.

[9] P. Arsi dan O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *J. Inform. J. Pengemb. IT*, vol. 3, hlm. 290–294, Okt 2018, doi: 10.30591/jpit.v3i3.1008.

[10] A. T. Kharroubi dan H. M. Darwish, "Diabetes mellitus: The epidemic of the century," *World J. Diabetes*, vol. 6, no. 6, hlm. 850–867, Jun 2015, doi: 10.4239/wjd.v6.i6.850.

[11] D. T. Larose, Data mining methods and models. Hoboken, Nj: Wiley-Interscience, 2006.

[12] J. Han and M. Kamber, Data mining : concepts and techniques. Haryana, India ; Burlington, Ma: Elsevier, 2012.

[13] M. Ary dan D. Rismiati, "Ukuran Akurasi Klasifikasi Penyakit Mesothelioma Menggunakan Algoritma K-Nearest Neighbor dan Backward Elimination," *SATIN - Sains Dan Teknol. Inf.*, vol. 5, hlm. 11–18, Jun 2019, doi: 10.33372/stn.v5i1.444.

[14] I. Indrayanti, D. Sugianti, dan M. A. A. Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-Tech*, vol. 12, no. 2, Art. no. 2, 2017, doi: 10.47775/ictech.v12i2.3.

[15] R. Sulaehani, "PREDIKSI KEPUTUSAN KLIEN TELEMARKETING UNTUK DEPOSITO PADA BANK MENGGUNAKAN ALGORITMA NAIVE BAYES BERBASIS BACKWARD ELIMINATION," *Ilk. J. Ilm.*, vol. 8, no. 3, Art. no. 3, Des 2016, doi: 10.33096/ilkom.v8i3.83.182-189.