JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985

PREDIKSI PENYAKIT DIABETES BERDASARKAN PERBANDINGAN KLASIFIKASI METODE K-NEAREST NEIGHBOR, NAÏVE BAYES, DAN DECISION TREE MENGGUNAKAN RAPID MINER

Muhammad Rezanur Ardianto*[1)], Rushendra Rushendra[2)]
1. Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercubuana
2. Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercubuana

Article Info
Kata Kunci: Diabetes, *Decision Tree, Naïve Bayes, K-NN.*

Keywords: *Diabetes, Decision Tree, Naïve Bayes, K-NN.*

Article history:
Received 8 Oktober 2024
Revised 11 November 2024
Accepted 8 Desember 2024
Available online 15 March 2025

DOI :
https://doi.org/10.29100/jipi.v10i2.6079

* Corresponding author.
Muhammad Rezanur Ardianto
E-mail address:
41520110043@student.mercubuana.ac.id

ABSTRAK
Pada era digital seperti saat ini kegiatan manusia dipermudah dengan adanya teknologi tak terkecuali dalam bidang penjualan makanan dan minuman, namun dengan kemudahan tersebut mengakibatkan kesulitan masyarakat dalam melihat gizi dari makanan dan minuman yang mengakibatkan terjangkitnya penyakit Diabetes, akan tetapi penyakit tersebut banyak faktor yang dapat memengaruhinya. Oleh sebab itu penelitian ini dilakukan sebuah prediksi terjangkitnya penyakit Diabetes dengan melakukan perbandingan algoritma K-NN, *Naïve Bayes,* dan *Decision Tree.* Hasil dari perbandingan algoritma yang paling cocok pada kondisi *default* yaitu *Decision Tree* dengan tingkat akurasi 93, 60%, namun untuk menghindari *overfitting* dan *underfitting* perlu dilakukan sebuah optimisasi K *cross validation* pada K=5 sampai K=10, kemudian dilakukan optimisasi nilai Konstanta K pada algoritma K-NN dengan K=2, sehingga didapatkan hasil algoritma K-NN lebih cocok untuk prediksi penyakit diabetes dengan nilai akurasi 96.13%, *precision* 98,31%, *recall* 88.21%, dan F1 *score*-nya 93%.

ABSTRACT
In the digital era as it is today, human activities are facilitated by technology, including in the field of food and beverage sales, but with this convenience, it results in the difficulty of the community in seeing the nutrition of food and drinks which results in the outbreak of diabetes, but the disease has many factors that can affect it. Therefore, this study conducted a prediction of the outbreak of Diabetes disease by comparing the K-NN, Naïve Bayes, and Decision Tree algorithms. The results of the comparison of the most suitable algorithms in the default condition are Decision Tree with an accuracy level of 93, 60%, so to avoid overfitting and underfitting it is necessary to optimize K cross validation at K = 5 to K = 10, then optimize the value of the K Constant in the K-NN algorithm with K = 2, so that the results of the K-NN algorithm are more suitable for predicting diabetes with an accuracy value of 96.13%, precision *98,31%,* recall *88.21%, an F1* score-*nya 93%.*

I. PENDAHULUAN

ERA digital seperti saat ini kehidupan manusia dipermudah dengan adanya teknologi, teknologi tersebut lebih praktis dan juga lebih *modern.* Perkembangan digital yang terjadi di masyarakat meliputi; jaringan internet, perangkat keras, aplikasi, serta media sosial [1]. Berdasarkan perkembangan era digital tersebut muncullah sebuah terobosan atau strategi pemasaran penjualan seperti *e-commerce fashion*, *food and baverage*, dan kebutuhan masyrakat sehari hari yang dapat dengan mudah diakses melalui teknologi pada *smartphone* [2]. Namun dengan adanya teknologi tersebut mengakibatkan masyarakat kesulitan dalam memilah dan menentukan produk yang benar-benar aman bagi kesehatan masyarakat itu sendiri, salah satu contohnya adalah masyarakaat dapat kesulitan dalam menentukan kandungan gizi makanan yang akan dikonsumsi jika dilakukan pembelian secara *online,* hal tersebut dapat mengakibatkan masyarakat dapat dengan mudah terjangkit berbagai penyakit [3].

Salah satu penyebab terjangkitnya penyakit yang tidak memperhatikan komposisi dan kandungan dari makanan dan minuman adalah penyakit diabetes, menurut kementrian kesehatan pada tahun 2021 terdapat 19,5 juta orang yang terjangkit penyakit diabetes, angka tersebut kemungkinan akan terus meningkat hingga 2045 dapat mencapai

28,6 juta orang yang akan terjangkit, risiko seseorang terjangkit diabetes tidak hanya dari konsumsi gula berlebihan melainkan antara garam, gula, dan lemak harus seimbang. Makanan cepat saji yang dipesan melalui *online* tidak ada pencantuman nilai gizi, hal tersebut mengakibatkan penyakit diabetes melitus tipe 2 [4]. Berdasarkan dari pernyataan yang diungkapkan oleh kementrian kesehatan penyakit diabetes tidak hanya didasari oleh konsumsi gula berlebihan melainkan terdapat faktor lain. Oleh sebab itu perlu dilakukan penentuan faktor prediksi diabetes.

Pada penelitian sebelumnya terdapat penelitian-penelitian yang membahas terkait faktor atau risiko terjangkitnya penyakit diabetes. Penelitian pertama yang berjudul "Penerapan Algoritma *Naïve Bayes* Untuk Klasifikasi Penyakit Diabetes Mellitus" bahwa menggunakan 17 atribut klasifikasi dan menunjukan akurasi 90.20% untuk penentuan faktor terjangkitnya penyakit diabetes [5]. Penelitian kedua yang berjudul "Penerapan Algoritma *K-Nearest Neighbor* Untuk Klasifikasi Penyakit Diabetes Melitus" bahwa untuk menentukan faktor terjangkitnya penyakit diabetes menggunkan 8 atribut dan menghasilkan akurasi 93% [6]. Penelitian ketiga yang berjudul "*Clinical Decision Support System for Diabetic Patients by Predicting Type 2 Diabetes Using Machine Learning Algorithm*" bahwa pada penelitian tersebut menggunakan 9 atribut untuk menentukan seseorang terjangkit penyakit diabetes, dan hasil akurasinya menunjukkan angka 90% [7]. Berdasarkan penelitian sebelumnya terdapat penentuan risiko terjangkitnya diabetes menggunakan 1 algoritma saja, hal tersebut dapat mengakibatkan perbedaan dari segi akurasi dan juga banyaknya atribut yang digunakan, sedangkan pada penelitian ini akan melakukan komparasi terkait penentuan seseorang terjangkitnya penyakit diabetes menggunakan *algorithm* K-NN, *Naïve Bayes,* serta *Decision Tree*.

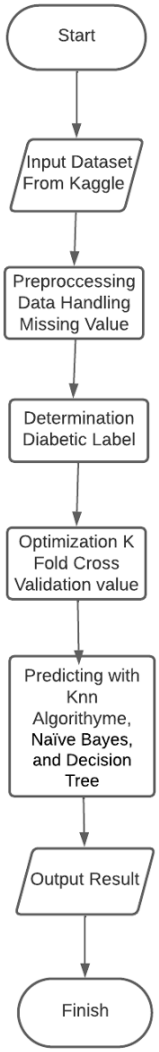
Penggunaan metode K-NN pada prediksi penyakit diabetes diyakini dapat menghasilkan akurasi yang cukup tinggi dalam memprediksi penyakit diabetes, apalagi pada data kesehatan yang tidak terstruktur, K-NN juga mudah untuk diimplementasikan pada prediksi penyakit diabetes, namun jika tidak diatur maka akan terjadi *overfitting* [8]. *Naïve Bayes* cenderung rendah dalam terjadinya kesalahan, namun algoritma tersebut kurang cocok dalam bidang kesehatan [9]. *Decision Tree* cenderung rendah dalam terjadi kesalahan prediksi, dan hasilnya mudah dipahami, namun *Decision Tree* juga rentan dalam terjadinya *overfitting* [10]. Mengambil dari *platform* Kaggle dapat mempermudah penelitian yang dilakukan oleh peneliti, selain itu juga *dataset* yang diperoleh sudah dilakukan pengelompokan berdasarkan topik yang diambil (prediksi diabetes) [11]. Pada penelitian ini lebih memilih *tools* RapidMiner daripada *tools* lainnya dikarenakan RapidMiner banyak menyediakan *Pre-Prcessing* data, selain itu juga fitur untuk membantu memproses, analisis data, visualisasi sangat berguna untuk menangani jenis data yang kompleks [12], dengan bantuan *tools, algorithm,* dan *platform* akan menghasilkan kesimpulan bahwa metode yang akurat untuk penentuan penyakit diabetes.

Penelitian ini bermaksud untuk menghasilkan metode yang paling tepat akurasinya dalam memprediksi penyakit diabet, sehingga dalam prediksi penyakit diabetes tidak akan terjadi kesalahan atau salah diagnosa. Menurut Siloam Hospital diabetes termasuk kategori penyakit kronis dan berbahaya, apalagi jika sudah terjadinya komplikasi [13]. Jika penelitian ini tidak dilakukan maka penyakit diabetes akan sulit untuk diprediksi secara akurat dan tepat, selain itu juga jika tidak dapat diprediksi secara akurat maka angka masyarakat yang terjangkit penyakit diabetes maka akan benar-benar meningkat di angka 28,6 juta pada tahun 2045 [4].

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan metodologi penelitian agar dapat dilakukan dengan terarah dan tahapan pengerjaannya jelas, hal tersebut dapat dilihat pada Gambar 1.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985



Gambar 1. Metodologi Penelitian

A. *Dataset*

Dataset merupakan sebuah kumpulan data yang akan dikelola oleh peneliti untuk menyelesaikan penelitiannya [14]. Pada dataset ini peneliti menggunakan 953 data dengan 18 atribut sebagai faktor-faktor prediksi penyakit diabetes. Berikut penjelasan dari masing-masing atribut; *Age* untuk pengelompokkan usia dari pasien terdiagnosa, *Gender* untuk data jenis kelamin dari pasien, *Family_Diabetes* untuk pasien yang memiliki keluarga riwayat diabetes, *highBP* untuk pengidap darah tinggi, *PhysicallyActive* untuk indikasi aktivitas atau kegiataan per hari, *BMI* untuk penjelasan berat badan, *Smoking* untuk indikasi perokok atau tidak, *Alcohol* untuk indikasi konsumsi alkohol, *Sleep* untuk penjelasan jam tidur dalam 1 hari, *SoundSleep* untuk indikasi dari seberapa lama untuk tidur dengan nyenyak, *RegularMedicine* untuk mengetahui apakah pasien tersebut mengkonsumsi obat, *JunkFood* untuk mengetahui seberapa sering konsumsi makanan cepat saji, *Stress* untuk mengukur tingkat stres dari pasien, *BPLevel* untuk mengetahui tekanan darah dari pasien, *Pregancies* untuk mengetahui pasien tersebut sedang hamil atau tidak, *Pdiabetes* untuk mengetahui apakah memiliki riwayat diabetes, *UriationFreq* untuk mengetahui apakah buang air kecil dalam jumlah banyak atau sedikit, *dan Diabetic* untuk menentukna apakahpasien terjangkit diabetes atau tidak.

B. *Preproccessing Data*

Preprocessing data adalah tahap penting untuk melakukan analisis data, yang melibatkan persiapan dan prubahan pada data mentah menjadi lebih terstruktur hingga siap dilakukan analisis. *Fase* untuk menjaga kualitas sertta keakuratan data dan memastikan data siap digunakan secara efektif dalam *tools* [15]. *Cleansing data* adalah proses untuk menghilangkan atau membersihkan data dari yang tidak terpakai atau data yang salah. Peneliti menggunakan *dataset* dari Kaggle dimana pada *dataset* tersebut sudah dikelompokan, sehingga sudah dapat dilakukan penelitian, akan tetapi pada penelitian ini peneliti menemukan beberapa data yang kosong atau belum terisi, sehingga perlu

dilakukan *tools handling missing value* yang berfungsi untuk mengisi data yang kosong [16].

C. *Diabetic Label*

Diabetic Label adalah proses pelabelan yang dilakukan oleh peneliti untuk menghasilkan target dari penelitian yang akan dilakukan oleh peneliti, pelabelan dilakukan berdasarkan kolom yang dapat mengidentifikasi dari kolom-kolom yang lain atau kategori data yang lainnya, misalnya pada penelitian ini untuk menentukan pasien yang terjangkit penyakit diabetes, namun ada beberapa kategori yang mendukung terjangkitnya penyakit tersebut contohnya usia, konsumsi makanan, dll [17].

D. *K-Fold Cross Validation*

K-Fold Cross Validation dapat digunakan evaluasi dari tahap suatu algoritma dengan membagi data *sampling* secara random kemudian menggabungkannya sebanyak dari nilai K. Tujuan pengujiannya adalah untuk menemukan model dengan nilai akurasi terbaik [18]. Pada penelitian ini peneliti menggunakan K-*Fold cross validation* dengan langkh sebagai berikut: mempersiapkan data latih dan data uji, membagi data menjadi K=5 sampai K=10, menerpkan algorotma yang dugunakan, dan dilakukan perhitungan akurasi dari K=5 sampai dengan K=10.

E. *Algorithm Clasification*

Algorithm Clasification adalah metode yang digunakan untuk mengelompokkan data ke dalam kelas atau kategori berdasarkan atribut atau sifatnya [19]. *algorithm* yang diterapkan pada penelitian ini meliputi *K-Nearest Neighbor, Naïve Bayes,* serta metode *Decision Tree,* berikut penjelasan algoritma yang digunakan pada penelitian ini: *K-Nearest Neighbor* merupakan *algorithm* yang digunakan unutk mendeteksi serta pembanding dari kondisi sebelumnya dengan yagng baru. memahami terkait data yang dievaluasi dengan data yang dikenal sebagai K-NN [19]. K-NN dapat menghasilkan akurasi yang cukup tinggi dalam memprediksi penyakit diabetes, apalagi pada data kesehatan yang tidak terstruktur, K-NN juga mudah diimplementasikan, namun jika tidak diatur maka akan terjadi *overfitting.* Untuk menghindari terjadinya *overfitting* peneliti melakukan *tunning* pada *dataset* yaitu dengan mengubah nilai K dengan cara menentukan jumlah tetangga terdekat dalam memprediksi kelas [8].

$$dis = \sqrt{\sum_{i=0}^{n}(X_1 - X_2)^2} \quad (1)$$

Keterangan :
Dis = *distance*
i = *variable*
n = dimensi data
X1 = data sampel
X2 = data uji

Naïve bayes adalah salah satu teknik pengajaran mesin yang menggunakan perhitungan probabilitas. Algoritma ini menggunakan asumsi bahwa kelas-kelas tidak saling tergantung (independen) [20]. *Naïve Bayes* cenderung rendah dalam terjadinya kesalahan, namun algoritma tersebut kurang cocok dalam bidang kesehatan [9].

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q}(P(Xi|Y))}{P(X)} \quad (2)$$

Keterangan:
P(Y|X) = *probability* pada data X dengan *class* Y
P(Y) = *probability* pada *class* Y
P(X) = *probability* pada X
$P(Y)\prod_{i=1}^{q}\big(P(Xi|Y)\big)$ = *probability* bebas pada semua vektor X

Decision tree merupakan algoritma dari klasifikasi yang populer karena mudah dipahami orang dan termasuk pada metode klasifikasi data *mining*. *Decision tree* merupakan gagasan *flowchart* struktur *tree*, di mana setiap *node* mewakili atribut, serta cabang menunjukkan hasil pengujian atau hasil nilai atribut, sedangkan daunny menunjukkan kelas [21]. *Decision Tree* cenderung rendah dalam terjadi kesalahan prediksi, dan hasilnya mudah dipahami, namun *Decision Tree* juga rentan dalam terjadinya *overfitting,* untuk menghindari hal tersebut Dengan mengubah kriteria pemilihan node, dengan begitu model dapat lebih menyesuaikan dengan data yang tersedia [10]

$$Gain\ (S,A) = Entrophy\ (s) - \sum_{i=1}^{n} \frac{Si}{S} * Entrophy \quad (3)$$

$$Entrophy\ (s) = \sum_{i=1}^{n} -pi * Log_2\ pi \quad (4)$$

Keterangan :
S = Himpunan *Case*
A = *Atribute*
n = Jumlah banyaknya *Atribute* A
Si = Jumlah *Case* pada banyaknya ke-i
S = Jumlah *Case* pada S
Pi = Proporsi Si pada S

F. Output Result

Pada tahap ini merupakan tahap akhir, *fase* ini adalah akan menkomparasikan hasil dari ketiga algoritma yang telah digunakan untuk memprediksi penyakit diabetes. Berdasarkan dari ketiga hasil tersebut menghasilkan tingkat akurasi yang paling tinggi, sehingga paling cocok untuk memprediksi penyakit diabetes. Peneliti juga melihat hasil berdasarkan nilai *Recall* dan *Precision. Recall* adalah rasio dari prediksi benar positif dengan jumlah data yang benar positif keseluruhan. *Precision* adalah rasio prediksi benar positif dengan jumlah data yang dipredikси benar positif, sedangkan *F1 Score* merupakan perbandingan dari rata-rata dari bobot presisi dan *recall* [22].

$$Recall = (True\ Positive)/(True\ Positive + False\ Negative) \quad (5)$$

$$Precission = (True\ Positive)/(True\ Positive + False\ Positive) \quad (6)$$

$$F1\ Score = 2 * (Recall * Precission)/(Recall + Precission) \quad (7)$$

III. HASIL DAN PEMBAHASAN

Penelitian ini akan memberikan hasil serta pembahasan yang telah diteliti terkait permasalahan dari pembelian makanan dan minuman secara *online*, hal tersebut mengakibatkan beberapa terjangkitnya penyakit salah satunya adalah diabetes, namun penyakit diabetes tidak hanya diakibatkan dari faktor konsumsi gula saja melainkan banuak faktor-faktor lain yang dapat mempengruhinya. Oleh sebab itu perlu dilakukan prediksi diagnosa penyakit diabetes, akan tetapi dalam memprediksi pasien terkena diabetes terdapat banyak metode yang digunakan, maka dari itu akan dilakukan metode mana yang paling akurat untuk digunakan dalam melakukan prediski penyakit Diabetes.

A. Dataset

Dataset yang digunakan adalah 953 data dengan 18 atribut namun dalam penyajiannya peneliti mengelompokkan kedalam 2 tabel agar dapat terbaca dengan jelas, pada tabel pertama terdapat kategori; *Age* untuk pengelompokkan usia dari pasien terdiagnosa, *Gender* untuk data jenis kelamin dari pasien, *Family_Diabetes* untuk pasien yang memiliki keluarga riwayat diabetes, *highBP* untuk pengidap darah tinggi, *PhysicallyActive* untuk indikasi aktivitas atau kegiataan per hari, *BMI* untuk penjelasan berat badan, *Smoking* untuk indikasi perokok atau tidak, *Alcohol* untuk indikasi konsumsi alkohol, *Sleep* untuk penjelasan jam tidur dalam 1 hari. Hal tersebut ditunjukkan pada Tabel 1.

TABEL I.
1-9 DATASET PREDIKSI PASIEN TERDIAGNOSA DIABETES BERDASARKAN PENGUMPULAN DATA PADA KAGGLE.

No	*Age*	*Gender*	*Family Dibetes*	*High BP*	*Physical Active*	*BMI*	*Smoking*	*Alcohol*	*Sleep*
1	50-59	*Male*	*no*	*yes*	*one hr or more*	39	*no*	*no*	8
2	50-59	*Male*	*no*	*yes*	*less than half an hr*	28	*no*	*no*	8
3	40-49	*Male*	*no*	*no*	*one hr or more*	24	*no*	*no*	6
4	50-59	*Male*	*no*	*no*	*one hr or more*	23	*no*	*no*	8
5	40-49	*Male*	*no*	*no*	*less than half an hr*	27	*no*	*no*	8
953	60 *or older*	*Female*	*yes*	*yes*	*one hr or more*	30	*no*	*no*	7

Pada Tabel 2 terdapat kolom lanjutan sebagai berikut: *SoundSleep* untuk indikasi dari seberapa lama untuk tidur dengan nyenyak, *RegularMedicine* untuk mengetahui apakah pasien tersebut mengkonsumsi obat, *JunkFood* untuk mengetahui seberapa sering konsumsi makanan cepat saji, *Stress* untuk mengukur tingkat stres dari pasien, *BPLevel*

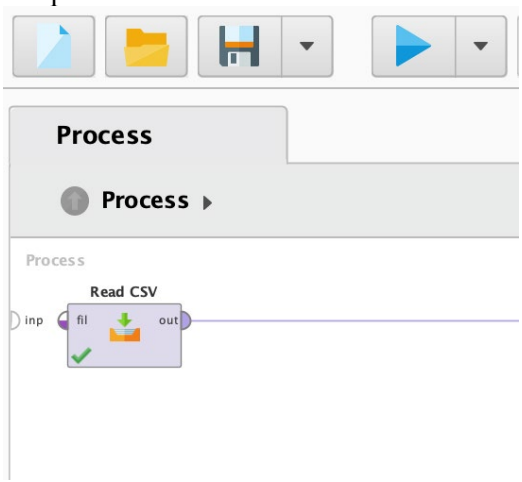
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985

untuk mengetahui tekanan darah dari pasien, *Pregancies* untuk mengetahui pasien tersebut sedang hamil atau tidak, *Pdiabetes* untuk mengetahui apakah memiliki riwayat diabetes, *UriationFreq* untuk mengetahui apakah buang air kecil dalam jumlah banyak atau sedikit, *dan Diabetic* untuk menentukna apakahpasien terjangkit diabetes atau tidak. Dataset ditunjukkan pada Tabel 2.

TABEL II.
10-18 DATASET PREDIKSI PASIEN TERDIAGNOSA DIABETES BERDASARKAN PENGUMPULAN DATA PADA KAGGLE.

No	*Sound Sleep*	*Regular Medicine*	*Junk Food*	*Stress*	*BPLevel*	*Pregancies*	*Pdiabetes*	*Uriation Freq*	*Diabetic*
1	6	*no*	*occasionally*	*Some-times*	*high*	0	0	*not much*	*no*
2	6	*yes*	*very often*	*Some-times*	*normal*	0	0	*not much*	*no*
3	6	*no*	*occasionally*	*Some-times*	*normal*	0	0	*not much*	*no*
4	6	*no*	*occasionally*	*Some-times*	*normal*	0	0	*not much*	*no*
5	8	*no*	*occasionally*	*Some-times*	*normal*	0	0	*not much*	*no*
953	4	*yes*	*occasionally*	*Some-times*	*high*	2	0	*quite often*	*yes*

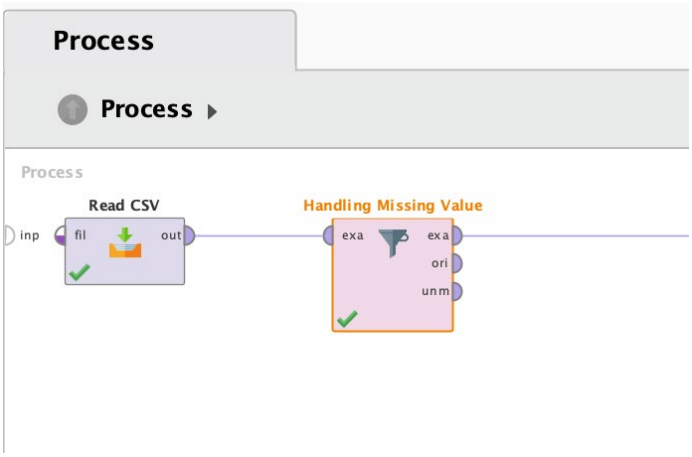
Dataset yang telah dikumpulkan kemudian dilakukan sebuah proses lanjutan pada *tools* RapidMiner, pada RapidMiner menggunakan operator Read CSV, operator tersebut digunakan untuk men-ekstrak *file* Excel kedalam RapidMiner. Proses dapat ditunjukkan pada Gambar 2.



Gambar 2. Proses ekstrak *file* Excel kedalam RapidMiner.

B. *Preprocessing Data*

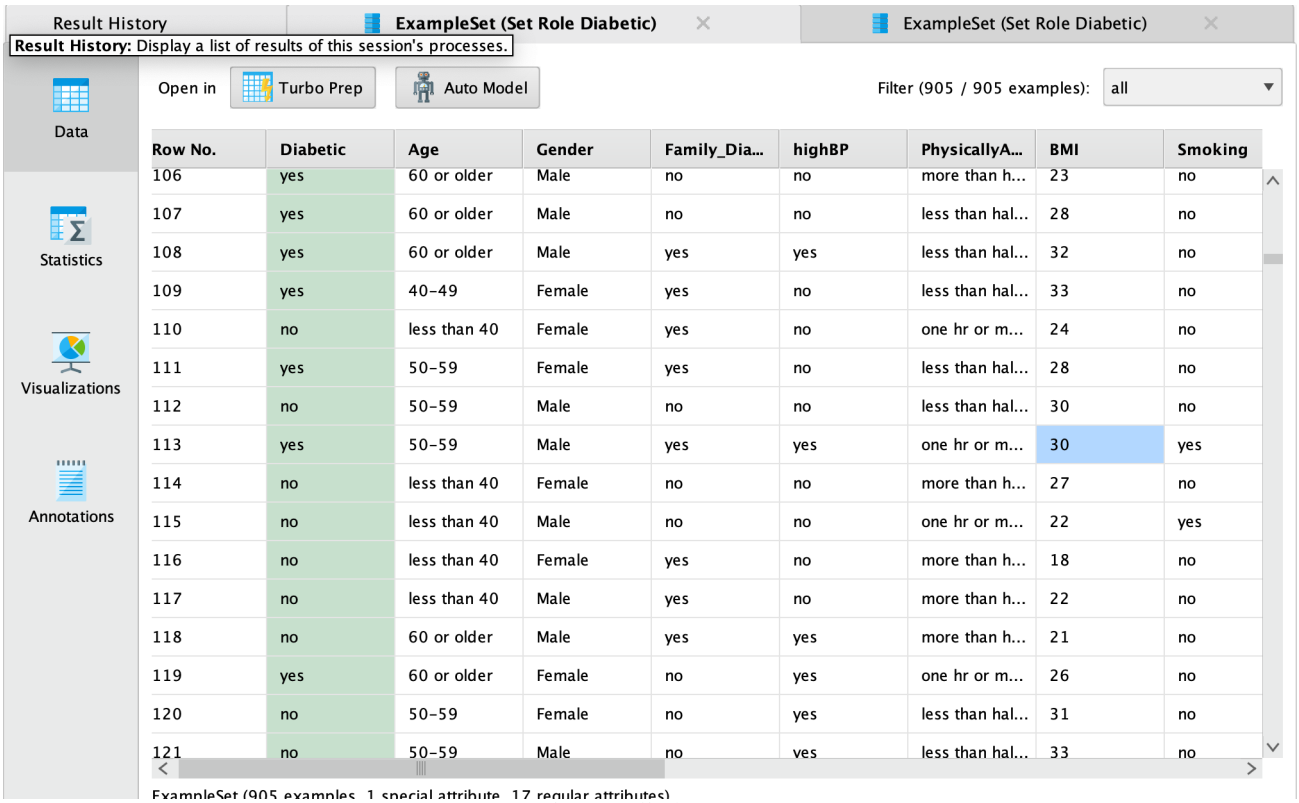
Preprocessing Data pada penelitian ini diperlukan sebuah pembersihan data *missing value,* dalam melakukan pembersihan tersebut menggunakan operator pada RapidMiner yang bernama "*Handling Missing Value"*. Pada *dataset* yang digunakan peneliti terdapat beberapa kesalahan atau data yang kosong sehingga peneliti menggunakan *missing value* untuk mengisi data kosong tersebut agar *dataset* yang digunakan terstruktur dan selaras dengan data-data laiinnya, hal tersebut mengakibatkan *dataset* yang digunakan dapat dilakukan penelitian untuk prediksi penyakit diabetes. Oleh karena itu *d*apat ditunjukan pada Gambar 3.



Gambar 3. *Handling Missing Value* digunakan untuk *Cleansing Dataset*

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985

Hasil dari proses *Handling Missing Value* pada RapidMiner menunjukan bahwa baris ke 115 dan kolom BMI sudah terisi, dapat ditunjukka pada Gambar 4.

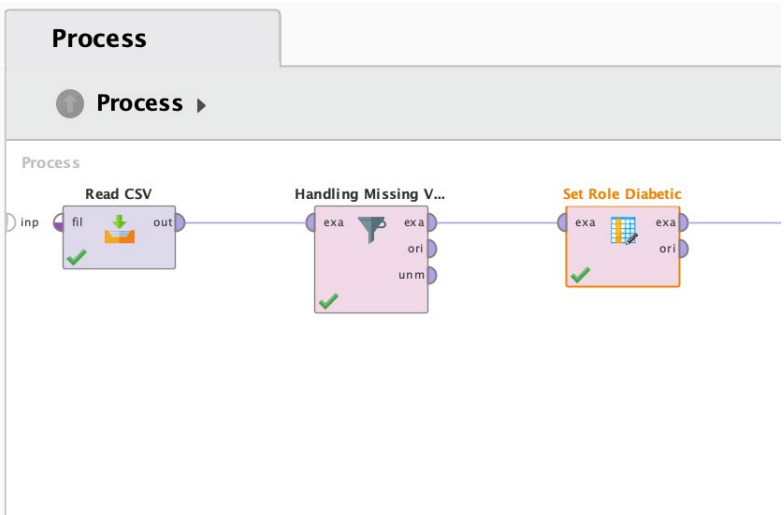


Row No.	Diabetic	Age	Gender	Family_Dia...	highBP	PhysicallyA...	BMI	Smoking
106	yes	60 or older	Male	no	no	more than h...	23	no
107	yes	60 or older	Male	no	no	less than hal...	28	no
108	yes	60 or older	Male	yes	yes	less than hal...	32	no
109	yes	40-49	Female	yes	no	less than hal...	33	no
110	no	less than 40	Female	yes	no	one hr or m...	24	no
111	yes	50-59	Female	yes	no	less than hal...	28	no
112	no	50-59	Male	no	no	less than hal...	30	no
113	yes	50-59	Male	yes	yes	one hr or m...	30	yes
114	no	less than 40	Female	no	no	more than h...	27	no
115	no	less than 40	Male	no	no	one hr or m...	22	yes
116	no	less than 40	Female	yes	no	more than h...	18	no
117	no	less than 40	Male	yes	no	more than h...	22	no
118	no	60 or older	Male	yes	yes	more than h...	21	no
119	yes	60 or older	Female	no	yes	one hr or m...	26	no
120	no	50-59	Female	no	yes	less than hal...	31	no
121	no	50-59	Male	no	yes	less than hal...	33	no

Gambar 4. Hasil *Cleansing Dataset*

C. *Diabetic Label*

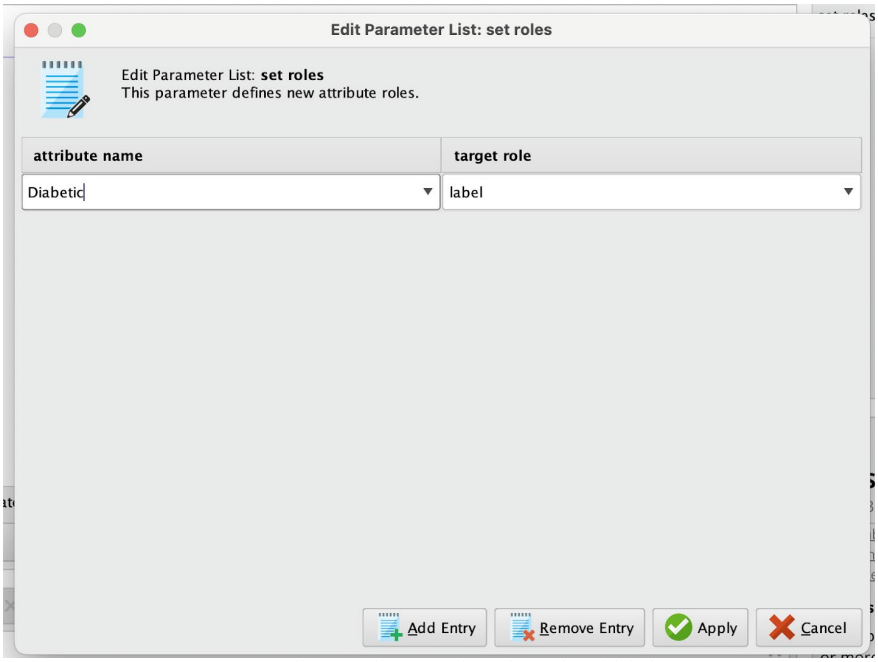
Diabetic Label adalah sebuah tahapan yang dilakukan untuk pelabelan dari *Dataset* yang telah digunakan, pada tahap ini menggunakan proses operator *set role*, dapat ditunjukkan pada Gambar 5.



Gambar 5. Pelabelan *Dataset*

Pada proses pelabelan menggunakan *set role*, selanjutnya dilakukan sebuah pemilihan *atributr* yang nantinya akan digunakan sebagai kelas, pada penelitian ini menggunakan kelas *Diabetic* sebagai label dari *Dataset* diabetes. Tahap pelabelan diabetes ini salah satu digunakan untuk proses prediksi diabetes dengan cara menentukan kolom sebagai kelas *dataset*-nya, pada penelitian ini menggunakan kolom diabetes untuk menentukan apakah pasien tersebut terkena penyakit diabetes atau tidak, jika tidak dilakukan maka penelitian ini tidak akan dapat menghasilkan prediksi penyakit diabetes. Dapat ditunjjukan pada Gambar 6.

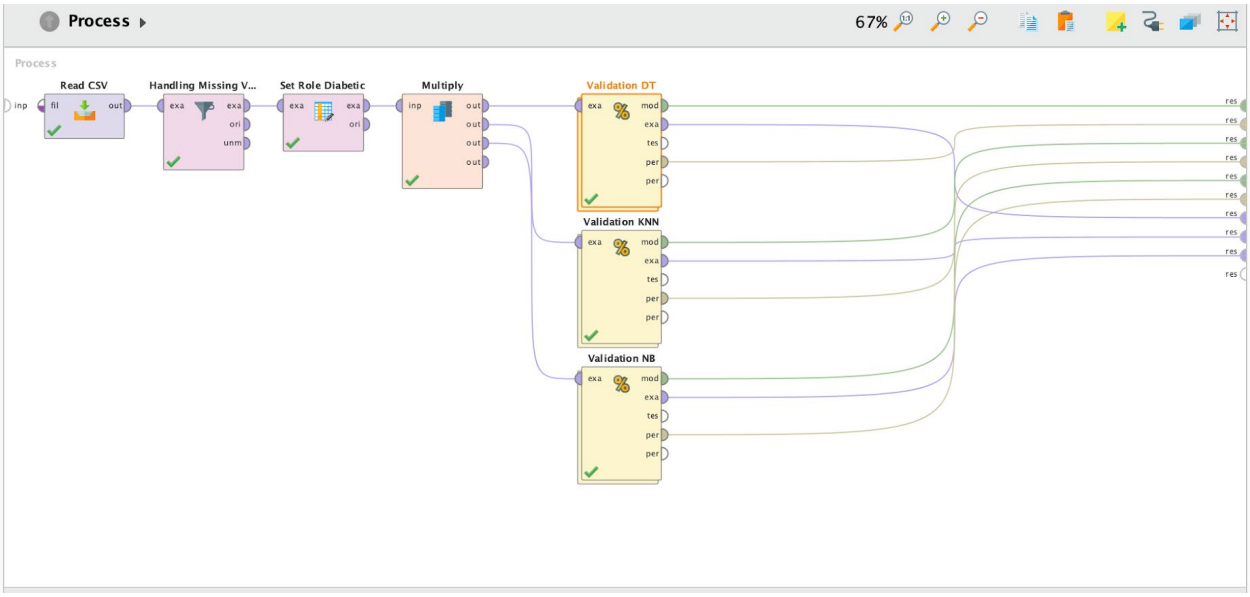
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985



Gambar 6. Kelas yang digunakan sebagai label

D. *K-Fold Cross Validation*

Proses ini dilakukan sebuah pemisah antara data latih dengan data *test* menggunakan operator *K-Fold Cross Validation* pada setiap *algorithm* yang diterapkan, selain itu juga menggunakan operator *multiply* untuk membandingkan hasil dari masing-masing metode yang digunakan, peneliti menggunakan K=5 sampai dengan K=10 untuk menghindari terjadinya *underfitting* dan *overfitting,* serta dapat mengurangi kebiasan dari data yang digunakan, sehingga dengan menggunakan K=5 sampai K=10 dapat meningkatkan akurasi dari hasil yang diperoleh [23]. Dapat ditujjukan pada Gambar 7.

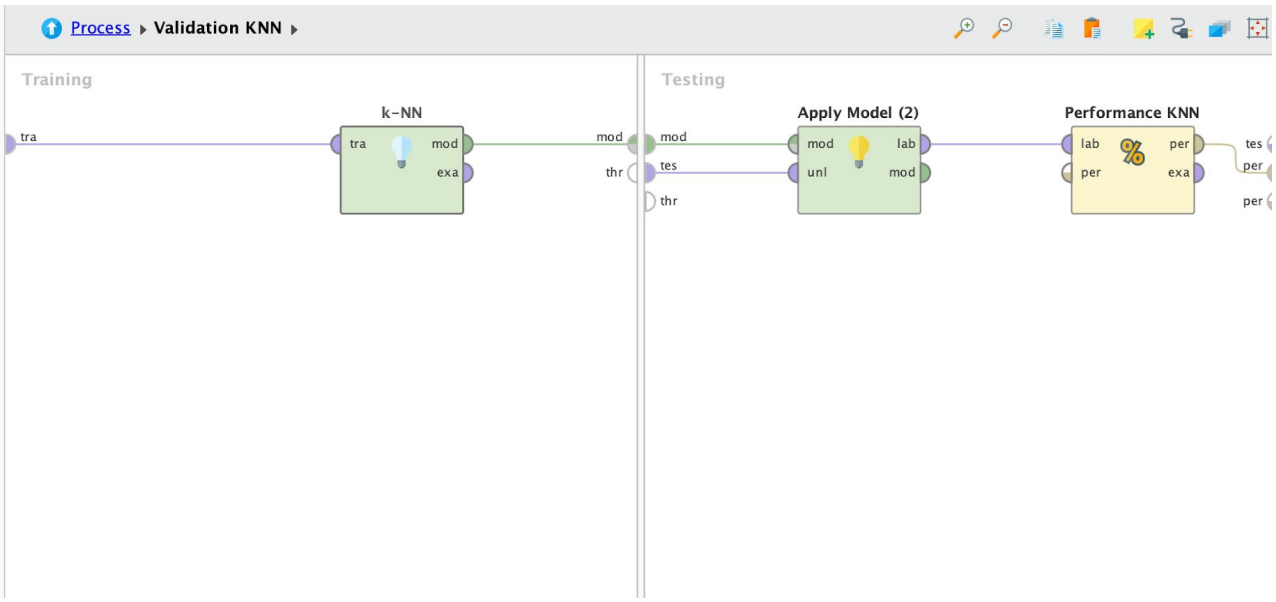


Gambar 7. Operator *K-Fold Cross* untuk pengujian *Dataset*

E. *Algorithm K-NN, Naïve Bayes, Decision Tree*

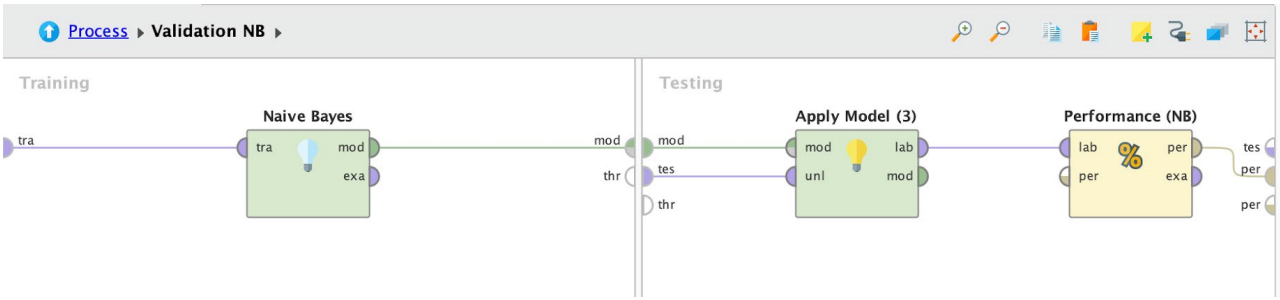
Proses ini dilakukan untuk penerapan algoritma yang diterapkan pada penelitian ini melalui *tools* RapidMiner. Penerapan metode K-NN dan juga beberapa operator tambahan seperti *Apply Model* serta *Performance,* hasil dari performa pada metode K-NN dilihat berdasarkan akurasi, *recall,* dan, *precision.* dapat ditunjukkan pada Gambar 8.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985



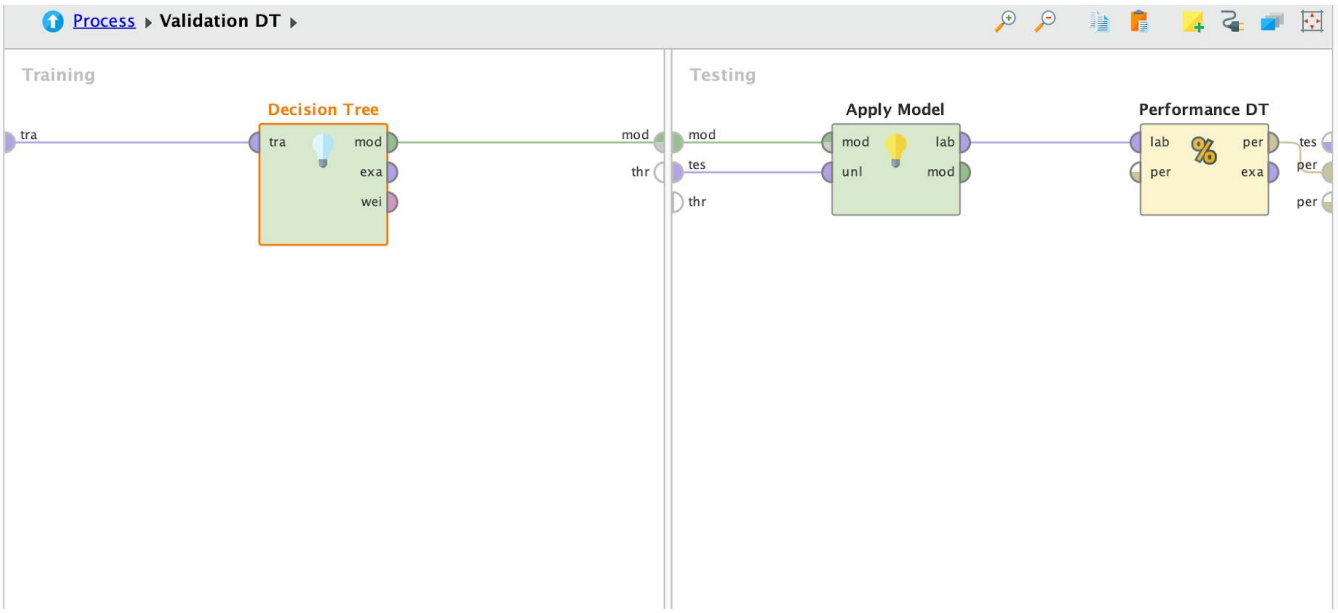
Gambar 8. Penerapan Algoritma K-NN pada RapidMiner

Algoritma *Naïve Bayes* pada RapidMiner dengan bantuan operator *Apply Model* dan juga *Performance,* hasil dari performa pada metode K-NN dilihat berdasarkan akurasi, *recall,* dan, *precision.* Dapat ditunjukkan pada Gambar 9.



Gambar 9. Penerapan Algoritma *Naïve Bayes* pada RapidMiner

Algoritma *Decision Tree* pada RapidMiner menggunakan operator *Apply Model* serta *Performance,* hasil dari performa pada metode K-NN dilihat berdasarkan akurasi, *recall,* dan, *precision.* Dapat diltunjukkan pada Gambar 10.

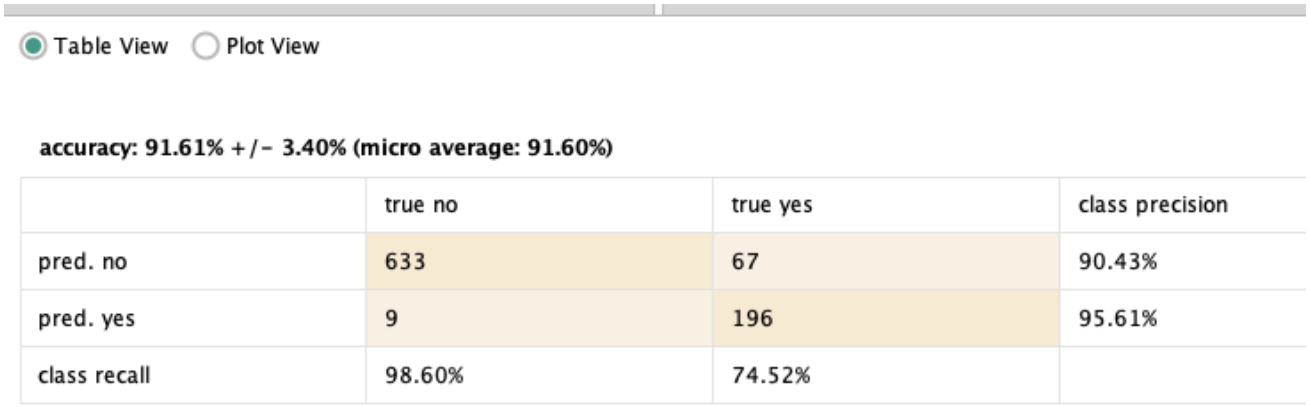


Gambar 10. Penerapan Algoritma *Decision Tree* pada RapidMiner

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985

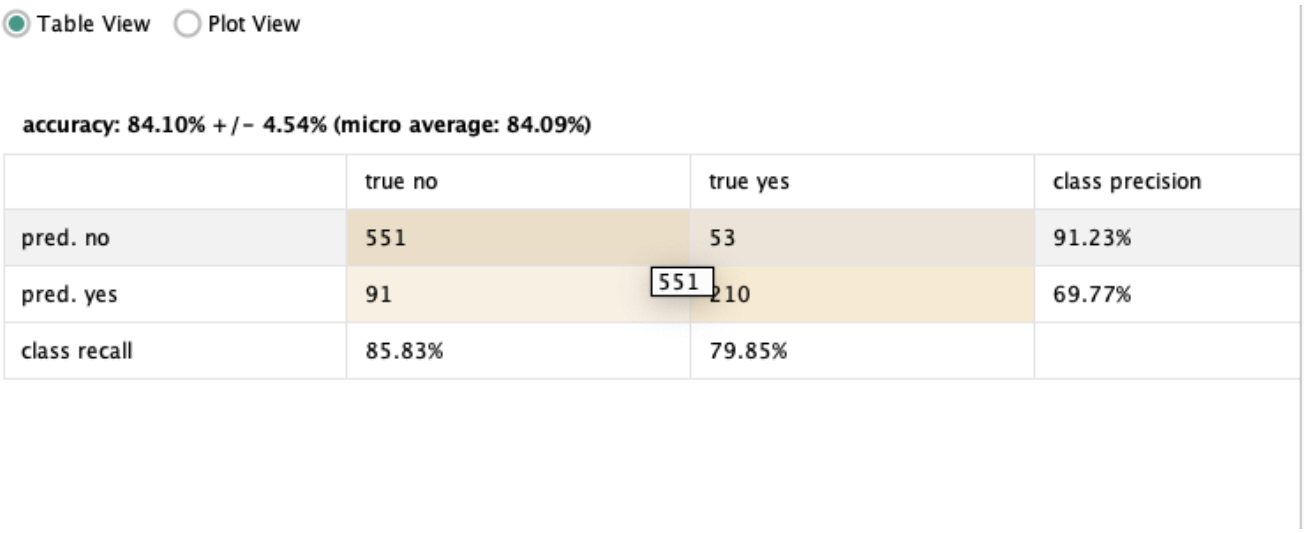
F. Output Result

Proses ini adalah hasil dari ketiga algoritma yang digunakan dalam melakukan prediksi penyakit Diabetes mengunakan *algortihm K-Nearest Neighbor* pada *tools* RapidMiner menghasilkan nilai akurasi 91.61%, dengan nilai *precision*-nya 95.61%, sedangkan nilai *recall*-nya adalah 74.52%, serta F1 *score*-nya adalah 97,7%. berikut hasil dari algoritma K-NN ditunjukan Gambar 11.



Gambar 11. Hasil prediksi penyakit diabetes menggunakan algoritma *K-Nearest Neighbor*

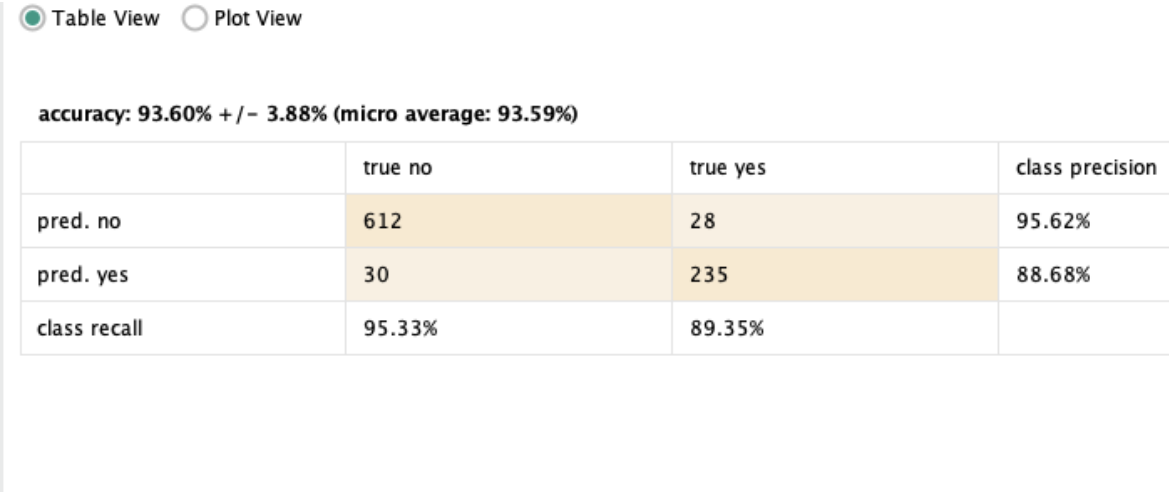
Pada penerapan *algoritm Naïve Bayes* dengan dataset 953 data serta 18 atribut menghasilkan nilai akurasi 84.10%, dengan nilai *precision*-nya 69.77%, sedangkan nilai *recall*-nya adalah 79.85%, serta F1 *score*-nya adalah 88,7%%. berikut hasil dari algoritma *Naïve Bayes* ditunjukkan melalui Gambar 12.



Gambar 12. Hasil prediksi penyakit diabetes menggunakan algoritma *Naïve Bayes*

Pada penerapan *Decision Tree* dengan dataset 953 data serta 18 atribut menghasilkan nilai akurasi 93.60%, dengan nilai *precision*-nya 88.68%, sedangkan nilai *recall*-nya adalah 89.35%, serta F1 *score*-nya adalah 94,3%. Berikut hasil dari algoritma *Decision Tree* dapat ditunjukkan pada Gambar 13.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 973-985



Gambar 13 Hasil prediksi penyakit diabetes menggunakan algoritma *Decision Tree*

Hasil pada penelitian ini menunjukan bahwa akurasi tertinggi adalah dari algoritma *Decision Tree,* namun untuk menghindari terjadinya hasil yang *underfitting* dan *overfitting* maka pada penelitian ini dilakukan sebuah optimisasi *K-Cross Validation* pada K=5 sampai dengan K=10, untuk mempersingkat hasil penelitian, peneliti telah meringkas hasil dari ketiga algoritma hal tersebut dapat diperhatikkan pada Tabel 3.

TABEL III.
HASIL K=5 SAMPAI K=10 PADA KETIGA ALGORITMA

Hasil Akurasi untuk nilai K fold Validation (K=5)					**Hasil Akurasi untuk nilai K fold Validation (K=6)**				
Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*	Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*
Decision Tree	92,93	89,33	85,93	*87,5%*	*Decision Tree*	91,16	83,27	87,07	85%
Naïve Bayes	86,14	69,93	78,71	*74%*	*Naïve Bayes*	84,2	70,13	79,47	74,5%
K-Nearest Neighbor	89,17	95,08	66,16	*78%*	*K-Nearest Neighbor*	89,17	91,88	68,82	78,7%
Hasil Akurasi untuk nilai K fold Validation (K=7)					**Hasil Akurasi untuk nilai K fold Validation (K=8)**				
Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*	Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*
Decision Tree	93,59	89,58	88,21	*88,9%*	*Decision Tree*	95,36	94,02	89,73	92%
Naïve Bayes	85,67	69,54	79,85	*74%*	*Naïve Bayes*	85,81	69	78,71	74%
K-Nearest Neighbor	90,16	94,39	70,34	*81%*	*K-Nearest Neighbor*	89,06	92,27	68,06	78,3%
Hasil Akurasi untuk nilai K fold Validation (K=9)					**Hasil Akurasi untuk nilai K fold Validation (K=10)**				
Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*	Nama Algo-ritma	Akurasi (%)	*Precision %*	*Recall %*	*F1 Score*
Decision Tree	92,15	93,12	87,45	*90,1%*	*Decision Tree*	93,6	88,68	89,35	89%
Naïve Bayes	83,76	69,33	79,09	*74%*	*Naïve Bayes*	84,10	69,77	79,85	74,5%
K-Nearest Neighbor	91,06	95,07	73,38	*83%*	*K-Nearest Neighbor*	91,61	95,61	74,52	84%

Menurut hasil dari optimisasi K=5 sampai K=10 menunjukan bahwa *decision tree* hasilnya tidak jauh berbeda dnegan K-NN oleh sebab itu dilakukan *tunning* nilai K pada algoritma K-NN dengan pengoptimlan nilai K=2, sehingga memperoleh akurasi yang tinggi sebesar 96.13%. Oleh sebab itu algoritma yang paling cocok untuk melakukan prediksi penyakit diabetes adalah algoritma *K-Nearest Neighbor* dengan nilai akurasi 96.13%, *precision* 98,31%, *recall* 88,21% serta F1 *score*-nya 93%. sedangkan pada *decision tree* dan juga Naïve bayes tidak mencapai nilai persentase yang di dapat dari algoritma K-NN. Penelitian sebelumnya dengan K-NN pada penyakit diabetes menghasilkan tingkat akurasi 93% dengan presisi 100% dan *recall* 60% [6]. Penelitian K-NN pada penyakit diabetes kedua menghasilkan akurasi 72% tanpa analisa presisi dan *recall* [8]. Penelitian ketiga K-NN pada penyakit diabetes menghasilkan tingkat akurasi 66,7% [25]. Agar menghasilkan akurasi serta hasil evaluasi yang lebih baik dari penelitian ini, peneliti merekomendasikan untuk melakukan penelitian terbaru dengan dataset lebih banyak dan juga atribut yang digunakan lebih kompleks lagi, sehingga akan lebih akurat untuk menentukan hasil prediksi diabetes. Dapat diperhatikan pada Gambar 14.

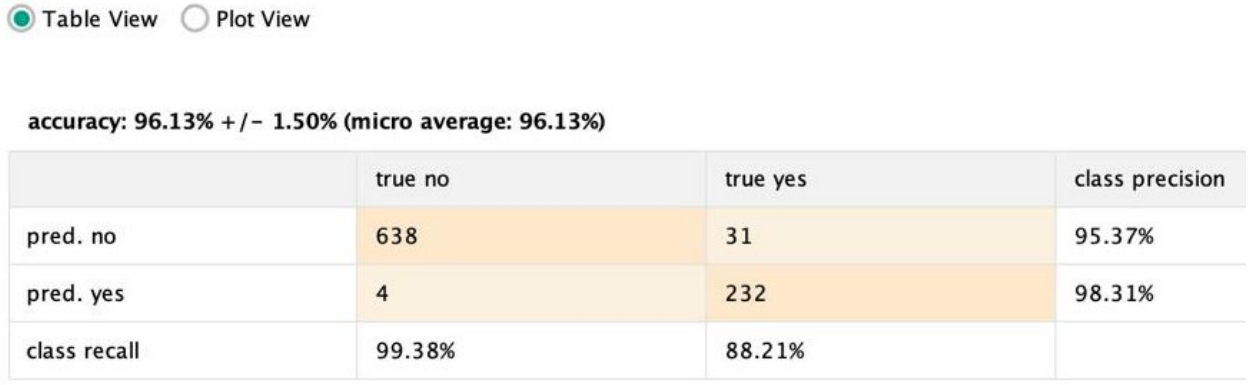


Table View Plot View

accuracy: 96.13% +/- 1.50% (micro average: 96.13%)

	true no	true yes	class precision
pred. no	638	31	95.37%
pred. yes	4	232	98.31%
class recall	99.38%	88.21%	

Gambar 14. Hasil *Tunning* K=2 dengan nilai akurasi tertinggi yaitu 96.13%

IV. KESIMPULAN

Penelitian ini memperoleh hasil terkait prediksi penyakit Diabetes dengan perbandingan ketiga Algoritma yaitu: *K-Nearest Neighbor, Naïve Bayes,* serta *Decision Tree* pada *tools* RapidMiner disimpulkan bahwa *algorithm K-Nearest Neighbor* sangat cocok untuk digunakan dalam melakukan prediksi penyakit Diabetes, hal tersebut disebabkan karena akurasi *K-Nearest Neighbor* menghasilkan nilai 96.13%, *precision* 98,31%, *recall* 88.21%, dan F1 *score*-nya 93%, sedangkan dari algoritma *Decision Tree* adalah 93.6%, *precision* 88,68%, *recall* 89.35%, dan *F1 Score* 89% serta paling tidak cocok adalah *Naïve Bayes* akurasinya adalah 84.09% *precision* 69,77%, *recall* 79.85%, dan *F1 Score 74,5%.*

DAFTAR PUSTAKA

[1] Astuti, "PERAN PERKEMBANGAN TEKNOLOGI DIGITAL TERHADAP STRATEGI PEMASARAN DAN DISTRIBUSI UMKM KOTA MAKASSAR," *Indonesian Journal of Business and Management,* vol. 6, no. (1), pp. 175-180, 2023.

[2] E. A. Sahrul, "PERAN E-COMMERCE, MEDIA SOSIAL DAN DIGITAL TRANSFORMATION UNTUK PENINGKATAN KINERJA BISNIS UMKM," *Jurnal Muara Ilmu Ekonomi dan Bisnis ,* vol. 7, no. (2), pp. 286-299, 2023.

[3] R. D. Rasdiyatno, "TRANSFORMASI INDUSTRI MAKANAN DAN MINUMAN MENUJU INOVASI NUTRISI DAN KEAMANAN PANGAN," *Jurnal Sains dan Teknologi ,* vol. 2, no. (4), pp. 80-89, 2024.

[4] Rokom, "Saatnya Mengatur Si Manis," Sehat Negeriku, 10 01 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/blog/20240110/5344736/saatnya-mengatur-si-manis/#:~:text=Menurut%20IDF%2C%20Indonesia%20menduduki%20peringkat,merupakan%20ibu%20dari%20segala%20penyakit. [Accessed 03 07 2024].

[5] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk KlasifikasiPenyakit Diabetes Mellitus," *Jurnal Sistem Komputer dan Kecerdasan Buatan,* vol. 4, no. (1), pp. 15-221, 2020.

[6] H. A. D. Fasnuari, "APPLICATION OF K-NEAREST NEIGHBOR ALGORITHM FOR CLASSIFICATION OF DIABETES MELLITUS CASE STUDY : RESIDENTS OF JATITENGAH VILLAGE," *Jurnal Ilmiah Teknik Informatika,* vol. 16, no. (2), p. 133–142, 2022.

[7] R. Islam, "Clinical Decision Support System for Diabetic Patients byPredicting Type 2 Diabetes Using Machine Learning Algorithms," *National Library of Medicine,* vol. 23, no. (1), pp. 1-11, 2023.

[8] C. Susanto, "Sistem Pakar Prediksi Penyakit Diabetes Menggunakan Metode K-NN Berbasis Android," *COGITO SMART JOURNAL,* vol. 8, no. (2), pp. 359-370, 2022.

[9] C. A. Rahayu, "PREDIKSI PENDERITA DIABETES MENGGUNAKAN METODE NAIVE BAYES," *(Jurnal Informatika dan Teknik Elektro Terapan,* vol. 11, no. (3), pp. 261-266, 2023.

[10] F. M. Hana, "Penerapan Algoritma ID3 Decision Tree Pada Klasifikasi Penyakit Diabetes," *Jurnal Dinamika Informatika,* vol. 12, no. (2), pp. 1-15, 2023.

[11] A. F. Riany, "PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT JANTUNG KORONER MENGGUNAKAN ALGORITMA NAÏVE BAYES," *MDP STUDENT CONFERENCE,* vol. 2, no. (1), pp. 297-305, 2023.

[12] A. H. Yuanti, "Analisis Pengaruh Covid-19 Terhadap Kesehatan Mental Dengan Visualisasi Data Rapidminer," *Gudang Jurnal Multidisiplin Ilmu,* vol. 2, no. (1), p. 183–187, 2024.

[13] S. E. Wijayaningrum, "Diabetes - Causes, Types, Symptoms, and Treatments," Siloam Hospitals, 23 06 2024. [Online]. Available: https://www.siloamhospitals.com/en/informasi-siloam/artikel/diabetes. [Accessed 03 07 2024].

[14] M. Kurniawan, "Pemodelan Dataset Tambang Terbuka pada PT. United Tractors Semen Gresikdengan Metode Artificial Neural Network," *PROMINE,* vol. 12, no. (1), pp. 1-6, 2024.

[15] A. Wahyu, "Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa," *Jurnal Edukasi dan Penelitian Informatika,* vol. 8, no. (1), pp. 175-179, 2022.

[16] J. M. A. S. Dachi, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam,* vol. 2, no. (2), pp. 87-103, 2023.

[17] E. Poerwandono, "Penerapan Data Mining Untuk Penilaian Kinerja Karya Di PT Riksa Dinar Djaya Menggunakan Metode Naive Bayes Classification," *Jurnal Sains dan Teknologi,* vol. 5, no. (1), pp. 336-340, 2023.

[18] Nurainun, "Penerapan Algoritma Naïve Bayes Classifier Dalam Klasifikasi Status Gizi Balita dengan Pengujian K-Fold Cross Validation," *Journal of Computer System and Informatics (JoSYC),* vol. 4, no. (3), pp. 578-586, 2023.

[19] A. P. Rushendra, "IMPLEMENTATION OF LOAD BALANCING WITH PER CONNECTION CLASSIFIER AND FAILOVER AND UTILIZATION OF TELEGRAM BOT (CASE STUDY : PT TUJUH MEDIA ANGKASA)," *Jurnal Teknik Informatika ,* vol. 5, no. (1), pp. 273-282, 2024.

[20] E. Novianto, "KLASIFIKASI ALGORITMA K-NEAREST NEIGHBOR, NAIVE BAYES, DECISION TREE UNTUK PREDIKSI STATUS KELULUSAN MAHASISWA S1," *Jurnal Teknologi dan Sistem Informasi Univrab,* vol. 8, no. (2), pp. 146-154, 2023.

[21] M. Afriansyah, "Optimasi Algoritma Naïve BayesUntuk Klasifikasi Buah Apel Berdasarkan Fitur Warna RGB," *BULLETIN OF COMPUTER SCIENCE RESEARCH,* vol. 3, no. (3), pp. 242-249, 2023.

[22] D. Septhya, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM: Indonesian Journal of Machine Learning and Computer Science,* vol. 3, no. (1), pp. 15-19, 2023.

[23] R. Kurniawan, "Klasifikasi Tingkat Kematangan Buah Sawit Berbasis Deep Learning dengan Menggunakan Arsitektur Yolov5," *JOURNAL OF INFORMATION SYSTEM RESEARCH (JOSH),* vol. 5, no. (1), pp. 302-309, 2023.

[24] H. Hafid, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada data kasus Covid-19 di Indonesia," *Journal of Mathematics, Computations, and Statistics,* vol. 6, no. (2), pp. 161-168, 2023.

[25] A. Asmarani, "Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes," *Jurnal Informatika Dan Rekayasa Komputer,* vol. 2, no. (2), pp. 231-239, 2022.