# Analisis Prediksi Resiko Diabetes Tahap Awal Menggunakan Algoritma *Naive Bayes*

**Muhtajuddin Danny[1)*)], Asep Muhidin[2)]**

[1)2)] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[*)]Correspondence Author: utat@pelitabangsa.ac.id, Cikarang, Indonesia
**DOI:** https://doi.org/10.37012/jtik.v9i2.2017

## Abstrak

Diabetes merupakan salah satu penyakit kronis yang diakibatkan adanya kelainan sekresi insulin pada kenaikan glukosa secara tidak teratur. Resiko penyakit stroke, penyakit jantung, kebutaan bahkan hingga resiko kematian merupakan penyakit komplikasi yang terjadi ketika adanya peningkatakn gula darah dalam tubuh pada penderita diabetes. Diabetes merupakan salah satu penyakit yang memiliki faktor resiko kematian yang tinggi. Deteksi dini penyakit diabetes perlu dilakukan sebagai upaya dalam menurukan tingkat kematian yang diakibatkan oleh faktor penyakit tersebut. Model yang diusulkan yaitu menerapkan algoritma *Naive Bayes* sebagai algoritma pengklasifikasi. Dataset yang dijadikan sebagai objek penelitian yaitu dataset *Early Stage Diabetes Risk Prediction* merupakan dataset terbuka yang bersumber dari *UCI Machine Learning*. Metode-metode yang digunakan dalam melakukan prediksi yaitu metode data mining. Data mining merupakan serangkaian tindakan untuk menemukan hubungan dari pola dan kecenderungan dari data yang disimpan. Desain alur sistem klasifikasi jenis pada penelitian ini, dimulai dari penentuan *Dataset*, *Loading* dan baca data, Analisis Eksplorasi Data, *Data Preprocessing*, membangun model data, evaluasi *Confusion Matrix*, dan *Hyperparameter Tuning*. Didapatkan nilai *True Positive* sebanyak 276, *True Negative* sebanyak 180, *False Positive* sebanyak 20 dan *False Negative* sebanyak 44. Nilai akurasi yang didapatkan dalam penelitian yaitu sebesar 87.88% dengan kategori *Good Classification* serta memiliki *error rate* yang rendah yaitu 12.12% termasuk kedalam kategori *Good Error Rate*. Hasil penelitian tersebut menunjukan bahwa algoritma *Naive Bayes* memiliki kinerja yang baik serta dapat dijadikan sebagai landasan dalam memprediksi risiko diabetes tahap awal.

**Kata Kunci:** Diabetes, *Data Mining*, Klasifikasi, *Naive Bayes*

## *Abstract*

*Diabetes is a chronic disease caused by abnormalities in insulin secretion due to irregular increases in glucose. The risk of stroke, heart disease, blindness and even the risk of death are complications that occur when there is an increase in blood sugar in the body of diabetes sufferers. Diabetes is a disease that has a high risk factor for death. Early detection of diabetes needs to be done as an effort to reduce the death rate caused by this disease. The proposed model applies the Naive Bayes algorithm as a classifier algorithm. The dataset used as the research object, namely the Early Stage Diabetes Risk Prediction dataset, is an open dataset sourced from UCI Machine Learning. The methods used to make predictions are data mining methods. Data mining is a series of actions to discover relationships from patterns and trends in stored data. The design of the type classification system flow in this research starts from determining the Dataset, Loading and reading data, Data Exploration Analysis, Data Preprocessing, building a data model, Confusion Matrix evaluation, and Hyperparameter Tuning. The True Positive value was 276, True Negative was 180, False Positive was 20 and False Negative was 44. The accuracy value obtained in the research was 87.88% in the Good Classification category and had a low error rate of 12.12%, which was included in the Good Error category. Rate. The results of this research show that the Naive Bayes algorithm has good performance and can be used as a basis for predicting the risk of early stage diabetes.*

***Keywords:*** *Diabetes, Data Mining, Classification, Naive Bayes*

# PENDAHULUAN

*American Diabetes Association* (ADA) melaporkan bahwa tiap 21 detik ada satu orang yang terkena diabetes. Prediksi sepuluh tahun yang lalu bahwa jumlah diabetes akan mencapai 350 juta pada tahun 2025, ternyata sudah jauh terlampaui. Lebih dari setengah populasi dunia yang menderita penyakit diabetes berada di Asia, terutama di India, China, Pakistan, dan Indonesia. Sementara itu suatu studi yang dilakukan di ibukota Saudi Arabia tahun 2012 melaporkan sebanyak 53% penduduknya memiliki resiko tinggi terhadap penyakit diabetes melitus (Yosmar, Amasdy, & Rahma, 2018).

Pasien didiagnosa menderita penyakit diabetes pada saat kadar glukosa darahnya melebihi nilai normal. Penyakit diabetes melitus adalah penyakit yang memiliki kompleksitas tinggi. Perawatan medis yang berkelanjutan sangat dibutuhkan guna menurunkan dampak komplikasi dengan pengecekan glikemik (Hana, 2020).

Banyaknya penderita diabetes dari tahun ke tahun semakin bertambah. Pasien diabetes di Indonesia sebesar 10 juta jiwa di tahun 2015. Merajuk pada data Federasi Diabetes Internasional, diprediksi penderita penyakit diabetes di Indonesia akan bertambah menjadi 16.2 juta pada tahun 2040. Guna menyikapi masalah ini, perlu adanya pendeteksian sejak dini penyakit diabetes. Deteksi sejak dini diharapkan dapat menurunkan resiko komplikasi pada pasien diabetes diwaktu mendatang. Guna menganalisa pasien pengidap penyakit diabetes sejak dini, pencatatan terhadap penyakit ini banyak dilakukan agar dapat dilakukan pencegahan. Salah satu metode pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan data mining (Hana, 2020).

Diabetes tipe 1 disebabkan karena tubuh mengalami kerusakan, ketika tubuh menghasilkan insulin, tubuh tersebut tidak mampu menggunakan insulin sebagaimana mestinya. Diabetes tipe 2 menghasilkan kelas dari resistensi insulin dimana sel-sel tidak mampu menggunakan insulin dalam proporsi yang tepat. Deteksi dini penyakit diabetes perlu dilakukan sebagai upaya dalam menurukan tingkat kematian yang diakibatkan oleh faktor penyakit tersebut (Fernanda, Ratnawati, & Adikara, 2017).

Berdasarkan tingginya tingkat penderita penyakit ini maka pencegahan dan prediksi menjadi faktor penting dalam penelitian saat ini. Metode-metode yang digunakan dalam melakukan prediksi yaitu metode data mining. Data mining merupakan serangkaian tindakan untuk menemukan hubungan dari pola dan kecenderungan dari data yang disimpan

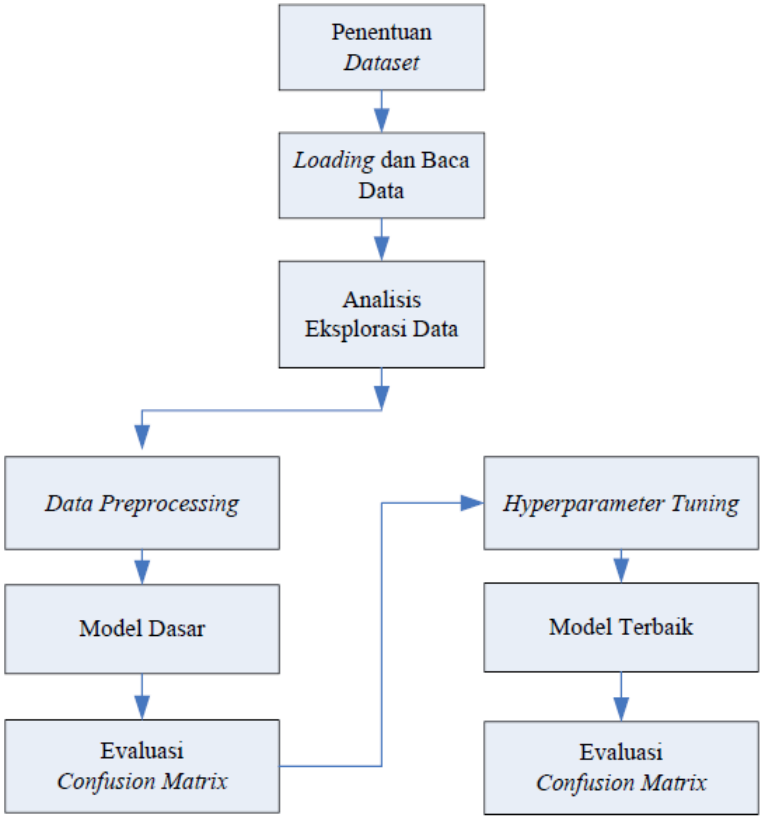
(Fernanda, Ratnawati, & Adikara, 2017). Beberapa metode yang di lakukan untuk melakukan prediksi seperti Algoritma C45, SVM, Naïve Bayes, Random Forest dll. Dari beberapa metode data mining, algoritma C45 merupakan algoritma yang relatif mudah di pahami dimana keputusan digambarkan dalam bentuk pohon keputusan (Fauzi et al., 2020). Beberapa literatur yang relevan berdasarkan metode algoritma C45 atau studi kasus penyakit diabetes diantaranya: (Aris, 2019) menggunakan metode C.45 untuk Identifikasi Penyakit Diabetes Melitus menghasilkan suatu pohon keputusan yang mencantumkan atribut-atribut seperti jenis kelamin, gula darah, usia, proko dan tipe gula darah lainnya. Penelitian (Noviandi, 2018) penggunaan metode C45 dalam prediksi penyakit diabetes dengan tingkat akurasi 70.32% (Yusnaeni & Widiarina, 2022).

Dalam upaya mengidentifikasi kategori penyakit diabetes sebagai langkah awal deteksi dini, penerapan teknik data mining sangat dibutuhkan karena teknik tersebut akan sangat membantu dalam proses identifikasi. Salah satu peran data mining tersebut yaitu metode kalsifikasi. Beberapa algoritma metode klasifikasi banyak digunakan peneliti khusunya dalam mendeteksi dini resiko penyakit diabetes. Algoritma *Modified K-Nearest Neighbor* (MKNN) diujikan pada dataset lokal di RSUD Kota Mataram dengan hasil akurasi 93,33% dan masuk dalam kategori *Excellent Classification* (Fernanda, Ratnawati, & Adikara, 2017). Pemilihan algoritma klasifikasi yang tepat sangat penting dalam mendeteksi dini penyakit diabetes. Melakukan pengujian beberapa algoritma sangat membantu dalam pemilihan model terbaik, akurasi kinerja algoritma terbaik akan dipilih nantinya sebagai model prediksi. Pada penelitian ini *Naive Bayes* akan dipilih sebagai algoritma pengklasifikasi karena memiliki beberapa kelebihan diantaranya yaitu cepat dalam proses perhitungan, algoritma sederhana yang memiliki tingkat akurasi yang tinggi (Wijanarto & Puspitasari, 2019). Sedangkan dataset yang dijadikan sebagai objek penelitian yaitu dataset *Early Stage Diabetes Risk Prediction* merupakan dataset terbuka yang bersumber dari *UCI Machine Learning* (Wijanarto & Puspitasari, 2019).

Walaupun kelebihan metode Naïve Bayes sangat sederhana, efisien dan hanya memerlukan komputasi matematika yang tidak terlalu kompleks, namun kelemahan teknik ini yaitu memerlukan pengetahuan awal untuk mengambil suatu keputusan, tingkat keberhasilan metode ini sangat tergantung pada pengetahuan awal yang diberikan (Wijanarto & Puspitasari, 2019).

# METODE

Penelitian ini menggunakan Klasifikasi *Gaussian Naive Bayes* yaitu teknik pembelajaran mesin klasik yang dapat digunakan untuk memprediksi nilai diskrit ketika semua variabel prediktornya berupa numerik. Desain alur sistem klasifikasi jenis pada penelitian ini, dimulai dari Penentuan *Dataset*, *Loading* dan Baca Data, Analisis Eksplorasi Data, *Data Preprocessing*, Membangun Model data, Evaluasi *Confusion Matrix*, *Hyperparameter Tuning* (Erlin, Marlim, Junadhi, L., & Agustina, 2022), dan output identifikasi jenis ditunjukkan pada Gambar 1 (Erlin, Marlim, Junadhi, L., & Agustina, 2022).




**Gambar 1.** Tahapan Penelitian

Dataset penyakit diabetes diambil melalui data kaggle. Dataset tersebut akan diolah melalui proses pengolahan data dan digunakan untuk mengidentifikasi penyebab utama penyakit diabetes (Sitanggang, Nicholas, Wilson, V. Sinaga, & Simanjuntak, 2022). *Dataset* yang digunakan dalam penelitian ini berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* sebagai bagian dari *Pima Indians Diabetes Database*. Dataset terdiri

atas beberapa variabel prediktor medis (independen) dan satu variabel target (dependen), yaitu target (hasil), seperti diperlihatkan pada Tabel 1.

**Tabel 1.** Variabel Yang Digunakan Dalam Prediksi Diabetes

| No. | Variabel | Keterangan |
|---|---|---|
| 1 | *Pregnancies* | Kehamilan: berapa kali pasien hamil |
| 2 | *Glucose* | Konsentrasi glukosa plasma selama dua jam dalam tes toleransi glukosa oral |
| 3 | *BloodPressure* | Tekanan darah: tekanan darah diastolik (mmHg) |
| 4 | *SkinThickness* | Ketebalan lipatan kulit trisep (mm) |
| 5 | *Insulin* | Insulin serum dua jam (µlU/mL) |
| 6 | BMI | Indeks massa tubuh (kg/m2) |
| 7 | *DiabetesPedigreeFunction*/DPF | Fungsi yang menilai kemungkinan diabetes berdasarkan riwayat keluarga |
| 8 | *Age* | Usia di tahun ini |
| 9 | *Outcome/Target* | Hasil: variabel kelas (0 jika nondiabetes, 1 jika diabetes) |

Proses *Loading* dan baca data mengubah data mentah menjadi format yang mudah dipahami. Proses ini dilakukan karena data mentah sering kali dalam bentuk format yang tidak beraturan. Tujuannya agar bisa dijadikan sumber informasi melalui sekumpulan data yang bisa diteruskan untuk diolah datanya (Sitanggang, Nicholas, Wilson, V. Sinaga, & Simanjuntak, 2022). *Dataset* dalam format *.csv* dimuat ke dalam variabel independen. Terdapat 768 data pasien yang semuanya perempuan berusia 21 tahun ke atas, yang terdiri atas sembilan variabel dengan rincian delapan variabel independen, yaitu *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin,* BMI, *Diabetes Pedigree Function* / DPF, dan *Age*; dan satu variabel dependen, yaitu Target. Hasil pemeriksaan pada *dataset* menggunakan *data.head()* memperlihatkan adanya beberapa variabel bernilai 0 yang mengindikasikan nilai yang hilang.

**Tabel 2.** Kelompok Umur oleh Departemen Kesehanan RI (2009)

| No | Kriteria | Umur (Tahun) |
|---|---|---|
| 1 | Masa Balita | 0 – 5 |
| 2 | Masa Kanak-Kanak | 6 - 11 |
| 3 | Masa Remaja Awal | 12 - 16 |
| 4 | Masa Remaja Akhir | 17 - 25 |
| 5 | Masa Dewasa Awal | 26 - 35 |
| 6 | Masa Dewasa Akhir | 36 - 45 |
| 7 | Masa Lansia Awal | 46 - 55 |
| 8 | Masa Lansia Akhir | 56 - 65 |
| 9 | Masa Manula | 65+ |

Normalisasi kriteria kelompok umur atribut age atau umur dapat dilihat pada tabel 3.

**Tabel 3.** Kelompok Umur oleh Departemen Kesehanan RI (2009)

| No | Kriteria | Umur (Tahun) |
|---|---|---|
| 1 | Remaja / *Adolescents* | 16 - 25 |
| 2 | Dewasa / *Adults* | 26 - 45 |
| 3 | Lansia / *Elderly* | 46 - 65 |
| 4 | Manula / *Seniors* | 65+ |

Setelah dilakukan normalisasi, atribut *age* atau umur pada dataset dibagi menjadi 4 kelompok. Kelompok pertama yaitu kriteria remaja dengan rentang umur 16 – 25 tahun, kelompok kedua yaitu kriteria dewasa dengan rentang umur 26 – 45 tahun, kelompok ke tiga yaitu kriteria lansia dengan rentang umur 46 – 65tahun dan terakhir kelompok dengan kriteria manula yaitu mereka yang memiliki umur diatas 65 tahun.

Setelah tahap Loading dan Baca Data, proses selanjutnya adalah *data exploration*. Data exploration adalah tahap yang bertujuan untuk memahami data. Pada proses eksplorasi ini kumpulan dataset yang telah didapatkan melalui situs Kaggle, dilakukan preprocessing dengan melihat data duplikat dan memeriksa *missing value*. Analisis eksplorasi data bertujuan menganalisis dataset yang digunakan untuk meringkas karakteristik utama dataset tersebut menggunakan bantuan statistika dan mempresentasikannya melalui teknik visual. Pada tahap ini, data diperiksa sebelum dibangunnya model, sehingga didapatkan wawasan maksimal dari dataset yang dimiliki (Erlin, Marlim, Junadhi, L., & Agustina, 2022).

Pada tahap *Data Preprocessing*, pengecekan dilakukan terhadap nilai data yang hilang karena dataset bisa saja memuat data yang tidak lengkap. Nilai data yang hilang digantikan dengan nilai median dari setiap variabel, sehingga setiap data pada variabel dataset memiliki nilai yang lengkap. Pada tahap ini juga dilakukan pengecekan terhadap data yang tidak seimbang. Penanganan terhadap data yang tidak seimbang dilakukan menggunakan *synthetic minority over-sampling technique* (SMOTE). Teknik ini digunakan untuk meningkatkan jumlah kelas minoritas melalui sampel data sintesis dengan tetap mempertahankan jumlah kelas mayoritas (Erlin, Marlim, Junadhi, L., & Agustina, 2022).

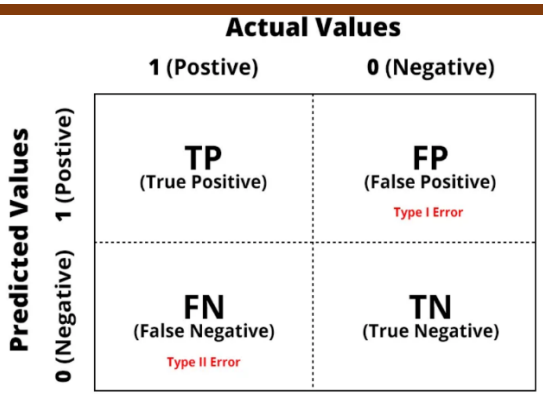
*Logistic regression* memodelkan hubungan antara variabel respons kategori dan covariate. Secara khusus, ada kombinasi linier dari variabel independen dengan log-peluang

probabilitas suatu peristiwa. *Logistic regression* merupakan model linier yang lebih cocok untuk masalah klasifikasi dibandingkan penggunaannya untuk regresi. *Logistic regression* juga dikenal dalam literatur sebagai regresi logit, klasifikasi entropi maksimum (MaxEnt), atau pengklasifikasi log-linear. Dalam logit, probabilitas yang menggambarkan kemungkinan hasil dari percobaan tunggal dimodelkan menggunakan fungsi logistik.

Berdasarkan tipe data yang bersifat nominal ataupun kategorikal serta setiap variabel bisa dianggap independen maka dipilih teknik klasifikasi dengan model Naïve Bayes. Selain itu jumlah data tidak terlalu banyak sehingga lebih sederhana, mudah, dan cepat diimplementasikan (Suryadewiansyah & Tju, 2022).

Pada tahapan *modelling,* alur penelitian merupakan model yang diusulkan dalam penelitian. Dataset yang telah melalui tahapan sebelumnya yaitu tahapan *data preparation* dengan melakukan normalisasi data, kemudian dibagi menjadi dua bagian yaitu *data training* sebanyak 90% dan *data testing* sebanyak 10% dari dataset. Algoritma pengklasifikasi *Naive Bayes* diterapkan pada *data training* untuk melatih data tersebut sehingga dapat meprediksi risiko diabetes tahap awal pada *data testing* yang merupakan data yang disiapkan untuk menguji kekauratan dalam memprediksi. Proses tersebut dilakukan berulang sebanyak 10 kali yang merupakan teknik *k-fold cross validation* dimana k adalah jumlah iterasi atau perulangan.

Model yang diusulkan, selanjutnya dievaluasi dengan menggunakan *confussion matrix* untuk mengetahui nilai akurasi dan tingkat kesalahan. Model klasifikasi yang dihasilkan akan diuji. Dalam penelitian ini, *Confusion Matrix* digunakan selama pengujian. *Precision*, *recall*, dan *accuracy* membentuk *confusion matrix*. Tingkat kesamaan antara nilai aktual dan nilai yang diharapkan diukur dengan menggunakan akurasi. Sedangkan keakuratan informasi yang diminta pengguna dan respons yang diberikan sistem digunakan untuk mengukur nilai presisi. Tingkat keberhasilan sistem dalam memulihkan informasi ditunjukkan dengan nilai *recall* (Oktafini & Rianto, 2023).




**Gambar 2.** *Confusion Matrix*

Pada gambar 2 terlihat bahwa terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada *confusion matrix*. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN).

Sedangkan untuk menghitung tingkat error, diperlukan persamaan *error rate* atau *misclasification rate* yang merupakan persentase jumlah *record* data yang diklasifikasikan diprediksi secara salah oleh algoritma. Untuk menghitung *error rate* atau *misclasification rate* digunakan persamaan berikut:

$$Error\ rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}$$

Untuk menentukan kinerja sebuah model diperlukan satu ukuran untuk mewakili kinerja dari setiap model. *Area under the curve* (AUC) adalah suatu daerah di bawah *receiver operating characteristic* (ROC). *Receiver operating characteristic* (ROC) merupakan kurva yang dihasilkan dari tarik ulur antara sensitivitas dan spesifisitas pada berbagai titik potong. Nilai AUC secara teoritis berada di antara 0 dan 1. Nilai AUC memberikan gambaran tentang keseluruhan pengukuran atas kesesuaian dari model yang digunakan. Semakin besar *area under curve* maka semakin baik variabel yang diteliti dalam memprediksi kejadian. Untuk menghitung nilai AUC, maka digunakan persamaan berikut:

$$AUC = \frac{(sensitivity + specificity)}{2}$$

Sebuah pedoman umum untuk mengklasifikasikan keakuratan pengujian diagnostik menggunakan AUC dapat dilihat pada sistem tradisional (Gorunescu, 2011) seperti yang tertera pada tabel 4 berikut:

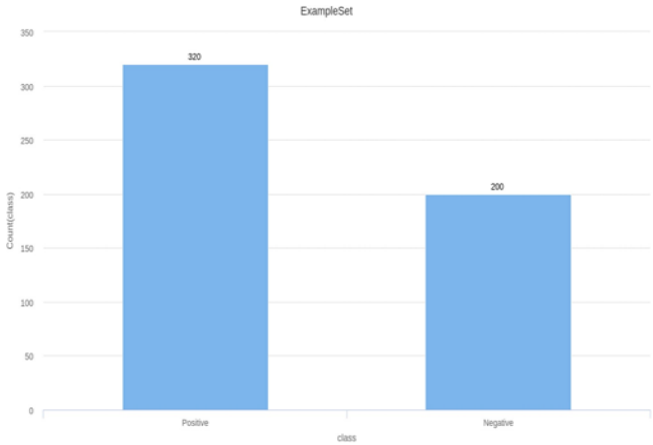
**Tabel 4.** Klasifikasi keakuratan pengujian diagnostic

| Nilai AUC | Klasifikasi |
|---|---|
| 0.9 - 1 | *Excellent Classification* |
| 0.8 – 0.9 | *Good Classification* |
| 0.7 – 0.8 | *Fair Classification* |
| 0.6 – 0.7 | *Poor Clasification* |
| < 0.6 | *Failure* |

Tahapan ***Hyperparameter Tuning*** yaitu tahap deployment, dimana pada tahap ini dilakukan perencanaan untuk menerapkan model yang telah melalui proses evaluasi. Model yang telah dievaluasi dan telah mencapai tujuan dari sebuah bisnis bisa dimanfaatkan untuk mempermudah proses bisnis tersebut.

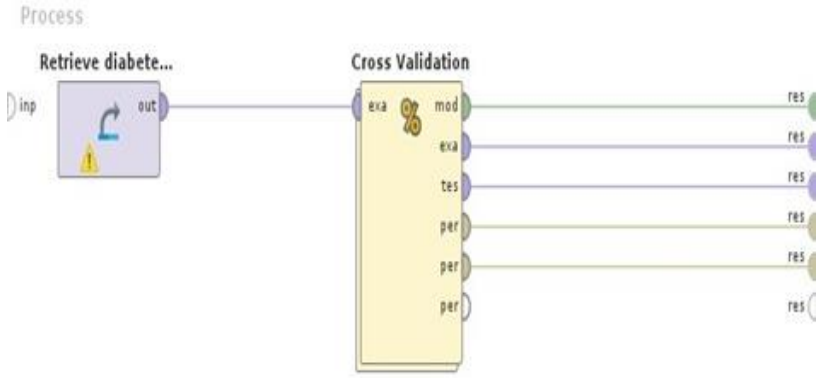
## HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian yaitu *Early stage diabetes risk prediction dataset* diperoleh dari website UCI *Machine Learning Repository*. Data yang digunakan adalah sebanyak 520 instances, memiliki 16 atribut prediktor dan 1 atribut kelas. Atribut kelas yang digunakan adalah *class*, atribut kelas *class* memiliki nilai *Positive* dan *Negative* dimana *Positive* diinterpretasikan sebagai *Positive Diabetes* dan *Negative* diinterpretasikan sebagai *Negative Diabetes*. Jumlah nilai *Positive* pada class adalah 320 dan jumlah nilai *Negative* pada class adalah 200.




**Gambar 3.** Jumlah Nilai *Class* pada *Dataset Rapidminer*

Rapidminer digunakan sebagai aplikasi pendukung dalam pemrosesan data dengan teknik *data mining*. Dalam prosesnya, data yang digunakan merupakan data bersih yang tidak memiliki masalah sehingga data tidak perlu dilakukan *preprocessing* dan dapat langsung digunakan dalam penelitian. Model yang diusulkan dalam penelitian ini yaitu menerapkan algoritma *naive bayes* sebagai algoritma pengklasifikasi dalam deteksi dini penyakit diabetes yang kemudian dilakukan perhitungan kinerja model dengan perhitungan akurasi dan *error rate*.

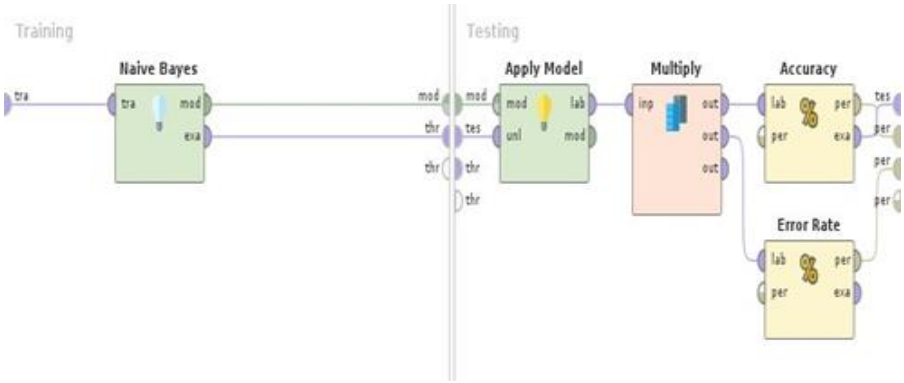



**Gambar 4.** Proses *Retrieve* Data dan *Cross Validation RapidMiner*

Langkah pertama yang dilakukan dalam memulai pengujian model menggunakan aplikasi RapidMiner yaitu dengan mengimport dataset ke dalam aplikasi. Pada proses mengimport dataset, atribut class diberikan role sebagai label dengan tujuan membedakan atribut prediktor dengan atribut kelas. Setelah dataset siap, gunakan operator *retrieve* untuk membaca dataset yang akan dilakukan pengujian. RapidMiner mendukung beberapa format atau ekstensi file salah satunya adalah ekstensi CSV. Format atau ekstensi yang digunakan pada dataset adalah *Comma Separated Values* (CSV). Operator selanjutnya yang digunakan yaitu operator *Cross Validation* dimana operator tersebut digunakan sebagai operator iterasi pengujian yaitu 10 kali pengujian.

Pembagian dataset dilakukan pada proses *cross validation*, dimana persentase pembagian dataset pada penelitian ini yaitu 90% dijadikan sebagai data *training* dan 10% dijadikan sebagai data *testing*. Operator *Naive Bayes* merupakan operator yang digunakan sebagai model algoritma pelatihan terhadap data *training* dalam penelitian. Sedangkan pada bagian *testing* digunakan beberapa operator yaitu operator *Apply Model* sebagai operator penerapan model terhadap *exampleset*, operator *Perfomance Accuracy* sebagai operator yang berguna untuk evaluasi kinerja seperti perhitungan akurasi, *Perfomance Error Rate*

sebagai operator yang berguna untuk evaluasi kinerja statistik dari tugas klasifikasi seperti perhitungan tingkat error. Gambar 5 menunjukan proses data *splitting cross validation.*




**Gambar 5.** Proses *Data Splitting Cross Validation RapidMiner*

Perhitungan secara manual dilakukan untuk menguji dan menentukan nilai akurasi dan nilai *error rate* kemudian dibandingkan dengan hasil keluaran dari aplikasi.

Tabel 5 menunjukan nilai *confusion matrix* menggunakan 10-*fold cross validation*.

**Tabel 5.** *Confusion Matrix 10-fold Cross Validation*

| ***Class*** | **True. Positive** | **True. Negative** |
|---|---|---|
| **Pred. Positive** | 277 **(*TP*)** | 20 **(*FP*)** |
| **Pred. Negative** | 43 **(*FN*)** | 180 **(*TN*)** |

$$Accuracy=\frac{(TP+TN)}{(TP+TN+FP+FN)}=\frac{(277+180)}{(277+180+20+43)}=\frac{457}{520}=0.8788=87.88\,\%$$

$$Error\ rate=\frac{(FP+FN)}{(TP+TN+FP+FN)}=\frac{(20+43)}{(277+180+20+43)}=\frac{(63)}{(520)}=0.1212=12.12\,\%$$

$$Sensitivity/Recall=\frac{TP}{(TP+FN)}=\frac{277}{(277+43)}=\frac{277}{320}=0.8656=86.56\,\%$$

$$Precision=\frac{TP}{(TP+FP)}=\frac{277}{(277+20)}=\frac{277}{297}=0.9327=93.27\,\%$$

$$Specificity=\frac{TN}{(TN+FP)}=\frac{180}{(180+20)}=\frac{180}{200}=0.9000=90.00\,\%$$

$$Prevalence=\frac{(TP+FN)}{(TP+FP+FN+TN)}=\frac{(277+43)}{(277+20+43+180)}=\frac{320}{520}=0.6154=61.54\,\%$$

$$PPV = \frac{(sensitivity * prevalence)}{((sensitivity * prevalence) + (1 - specificity) * (1 - prevalence))}$$

$$\frac{(0.8656 * 0.6154)}{((0.8656 * 0.6154) + (1 - 0.9000) * (1 - 0.6154))} = \frac{0.5327}{0.5712} = 0.9327 = 93.27\ \%$$

$$NPV = \frac{(specificity * (1 - prevalence))}{(((1 - sensitivity) * prevalence) + ((specificity) * (1 - prevalence)))}$$

$$= \frac{(0.900 * (1 - 0.6154))}{(((1 - 0.8656) * 0.6154) + ((0.900) * (1 - 0.6154)))} = \frac{0.3462}{0.4288}$$

$$= 0.8072 = 80.772\%$$

Hasil perhitungan akurasi dan *error rate* dari aplikasi RapidMiner menunjukan kesesuaian dengan perhitungan manual. Perhitungan manual untuk mencari nilai akurasi dilakukan dengan persamaan kemudian dibandingkan dengan hasil perhitungan akurasi dari aplikasi RapidMiner yang ditunjukan pada tabel 5. Sementara untuk perhitungan *error rate*, persamaan digunakan dalam perhitungan manualnya yang kemudian dibandingkan dengan hasil keluaran aplikasi RapidMiner untuk nilai *error rate* tersebut. Dari perbandingan perhitungan nilai akurasi dan *error rate*, dapat disimpulkan bahwa antara perhitungan manual dan aplikasi RapidMiner untuk kedua nilai tersebut tidak terdapat kesalahan. Nilai akurasi dan *error rate* merupakan nilai dari hasil perhitungan model algoritma naive *bayes* dengan teknik validasi yaitu *10-cross validation* yang di implementasikan pada *Early stage diabetes risk prediction dataset* yang bersumber dari "UCI *Machine Learning Repository*". Tabel 6 menunjukan ringkasan hasil pengujian dalam penelitian.

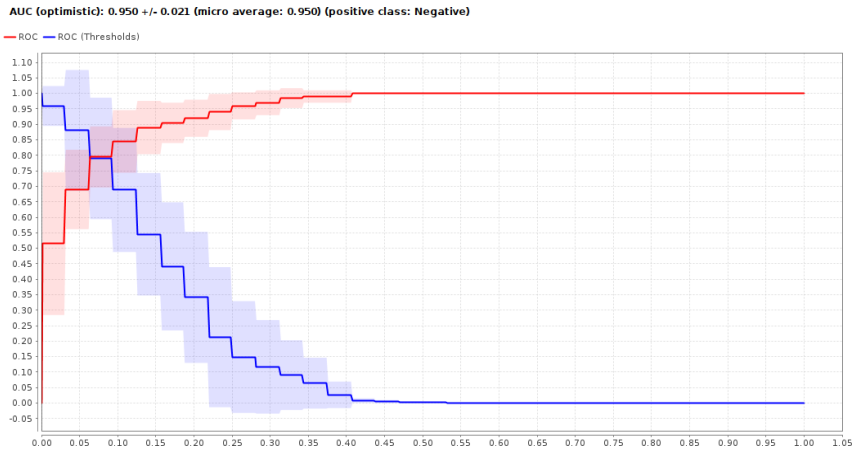
**Tabel 6.** Hasil pengujian

| **Model** | **TP** | **TN** | **FP** | **FN** | ***Accuracy*** | ***Error Rate*** |
|---|---|---|---|---|---|---|
| *Naive Bayes* | 276 | 180 | 20 | 44 | **87.88 %** | **12.12 %** |

Pada tabel 6, menunjukan ringkasan dari hasil pengujian. Dimana model klasifikasi yang digunakan yaitu algoritma *naive bayes* dengan teknik validasi yaitu *10-cross validation*. Dari model tersebut didapatkan nilai *True Positive* sebanyak 276, *True Negative* sebanyak 180, *False Positive* sebanyak 20 dan *False Negative* sebanyak 44. Nilai akurasi

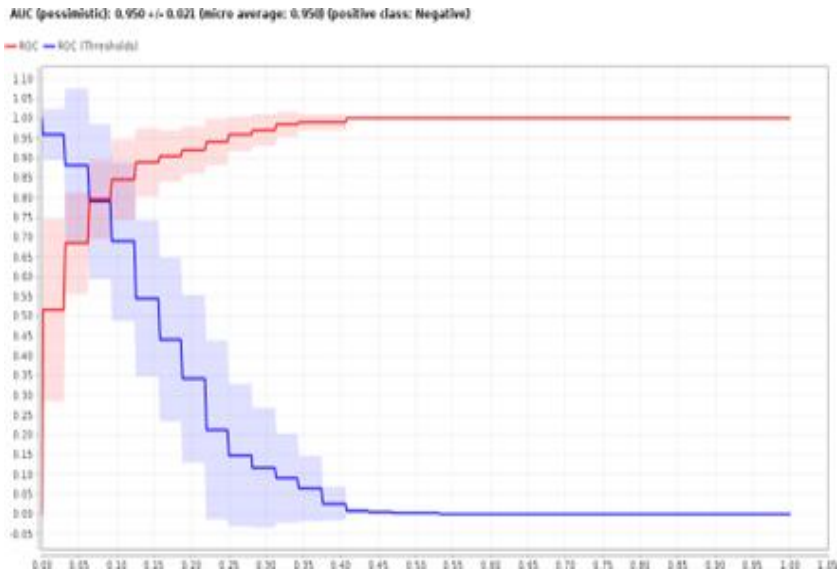
yang didapat yaitu sebesar 87.88%, nilai tersebut didapat dari persamaan. Angka 12.12% merupakan nilai *Error Rate* yang dihitung menggunakan persamaan.

Nilai *Area Under Curve* (AUC) merupakan nilai yang menunjukan tingkat dari keakuratan model *empiric*.



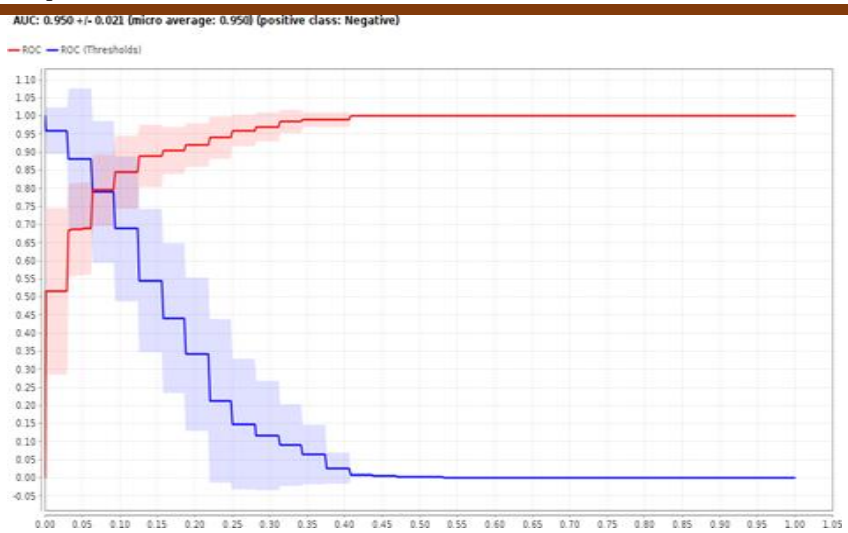

**Gambar 6.** AUC *Optimistic Dengan Algoritma Naive Bayes RapidMiner*

Gambar 6 menunjukan gambar grafik AUC *Optimistic* yang merupakan hasil keluaran dari aplikasi RapidMiner. AUC *Optimistic* memplot contoh positif sebelum contoh negatif.



**Gambar 7.** AUC *Pessimistic Dengan Algoritma Naive Bayes RapidMine5*

Gambar 7 menunjukan gambar grafik AUC *Pessimistic*, AUC *Pessimistic* memplot contoh negatif sebelum memplot contoh positif.




**Gambar 8.** AUC Dengan Algoritma *Naive Bayes RapidMiner*

Sedangkan gambar 8 merupakan grrafik *Area Under Curve* (AUC) versi normal yang menghitung luas dengan mengambil rata-rata AUC *Optimistic* dan AUC *Pessimistic*.

Pada bagian ini akan dibahas mengenai hasil pengukuran kinerja model, pembahasan tersebut dilakukan untuk menjawab rumusan masalah yaitu seberapa tinggi akurasi algoritma *naive bayes* terhadap *dataset Early Stage Diabetes Risk Prediction* ? Seberapa rendah *error rate a*lgoritma *naive bayes* terhadap *dataset Early Stage Diabetes Risk Prediction* ? Sebagai jawaban dari rumusan masalah tersebut, tabel 7 menunjukan hasil pengujian akurasi dan *error rate* model pengklasifikasian *naive bayes*.

**Tabel 7.** Hasil pengujian akurasi dan *error rate*

| NO | *Naive Bayes* | Nilai (%) | Keterangan |
|---|---|---|---|
| 1 | *Accuracy* | 87.88 % | *Good Classification* |
| 2 | *Error Rate* | 12.12 % | *Good Error Rate* |

Tabel hasil pengujian akurasi dan *error rate* pada tabel 7 merupakan hasil analisa model *naive bayes* dengan teknik *10-Fold Cross Validation*. Pada penelitian ini, nilai akurasi didapatkan yaitu sebesar 87.88%. Nilai akurasi merupakan jumlah relatif dari contoh yang diklasifikasikan dengan benar atau dengan kata lain persentase prediksi yang benar. Nilai tersebut termasuk dalam kategori *Good Classification* berdasarkan tabel klasifikasi keakuratan pengujian diagnostic. Dengan kata lain, model yang diusulkan dalam penelitian yang menghasilkan nilai akurasi tersebut dapat direkomendasikan sebagai dasar untuk memprediksi penyakit diabetes sejak dini.

Dalam mengklasifikasi, model yang diusulkan dalam penelitian yaitu analisa model *naive* bayes dengan teknik *10-Fold Cross Validation* juga memiliki tingkat kesalahan yang rendah dalam mengklasifikasi. Nilai *error rate* yang dihasilkan yaitu sebesar 12.12%, nilai tersebut sangat rendah dan termasuk kedalam kategori *Good Error Rate*. Nilai *error rate* atau tingkat kesalahan merupakan jumlah relatif dari contoh yang salah diklasifikasikan atau dengan kata lain persentase prediksi yang salah. Kesalahan dalam mengklasifikasi bisa terjadi karena beberapa faktor diantaranya adalah proses normalisasi pada tahapan *data preprocessing*. Proses normalisasi yang dilakukan terhadap dataset pada penelitian ini adalah memberikan pengkategorian pada atribut *age* atau umur yang memiliki nilai numerik menjadi 4 kategori *polynominal* yaitu *adolescents, adults, ederly* dan *seniors.* Terlalu banyak kategori dalam satu atribut jika tidak distandarkan dengan atribut lain akan menyebabkan kesalahan dalam mengklasifikasi.

## KESIMPULAN DAN REKOMENDASI

Penelitian yang dilakukan dalam memprediksi risiko diabetes tahap awal dimulai dengan menentukan algoritma *naive bayes* sebagai algoritma klasifikasi. Analisis data dilakukan sebelum proses penerapan model, tujuannya untuk memastikan kualitas data apakah data yang digunakan merupakan data kotor yang perlu dilakukan pembersihan, terdapat *missing value* dan hal-hal lain yang mungkin dapat mempengaruhi kinerja model yang diusulkan atau bahkan data tersebut merupakan data bersih sehingga tidak perlu dilakukan tahap *data preprocessing.*

Model yang diusulkan pada penelitian yaitu penerapan algoritma *naive bayes* sebagai algoritma pengklasifikasi terhadap dataset *Early Stage Diabetes Risk Prediction* memiliki hasil yang baik. Model yang diusulkan mampu menghasilkan nilai akurasi yaitu sebesar 87.88%. Hasil tersebut menunjukan bahwa dalam memprediksi risiko diabetes tahap awal, model yang diusulkan dapat dijadikan rekomendasi karena masuk kedalam kategori *Good Classification*. Dalam mengklasifikasi, tingkat kesalahan atau *error rate* dalam mengklasifikasi juga dapat dijadikan sebagai landasan evaluasi kinerja model. Semakin rendah tingkat kesalahan klasifikasi maka semakin bagus model tersebut. Penerapan algoritma klasifikasi *Naive Bayes* dalam memprediksi risiko diabetes tahap awal memiliki tingkat kesalahan klasifikasi yang rendah yaitu sebesar 12.12%.

# REFERENSI

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., & Putra, A. D. (2022, Juni). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma regresi Logistik. *Journal of Machine Learning and Artificial Intelligence, 1*(2), 107-114.

Erlin, Marlim, Y. N., Junadhi, L., S., & Agustina, N. (2022, Mei). Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 11*(2), 88-96. doi:10.22146/jnteti.v11i2.3586

Fernanda, S. I., Ratnawati, D. E., & Adikara, P. P. (2017, Juni). Identifikasi Penyakit Dibetes Melitus Menggunakan Metode Modified K-Nearest Neighbor (MKNN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 1*(6), 507-513.

Hana, F. M. (2020, Oktober). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan), 4*(1), 32-39. doi:10.47970/siskom-kb.v4i1.173

Jackins, V, Vimal, S, Kaliappan, M, & Lee, MY (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of …, Springer, https://doi.org/10.1007/s11227-020-03481-x

Oktafini, R., & Rianto, R. (2023, Agustus). Perbandingan Algoritma Support Vector Machine (SVM) dan Decission Tree untu Sistem Rekomendasi Tempat Wisata. *Jurnal Nasional Teknologi dan Sistem Informasi, 9*(2), 113-121. doi:10.25077/TEKNOSI.v9i2.2023.113-121

Saritas, MM, & Yasar, A (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International journal of intelligent systems and …, ijisae.org, https://ijisae.org/IJISAE/article/view/934

Sitanggang, D., Nicholas, N., Wilson, V. Sinaga, A. R., & Simanjuntak, A. D. (2022, Desember). Implementasi Data Mining untuk Memprediksi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor dan Logistic Regression. Jurnal Teknik Informasi dan Komputer (Tekinkom), 5(2), 493. doi:10.37600/tekinkom.v5i2.698

Stephens, CR, Huerta, HF, & Linares, AR (2018). When is the Naive Bayes approximation not so naive?. Machine Learning, Springer, https://doi.org/10.1007/s10994-017-5658-0

Suryadewiansyah, M. K., & Tju, T. E. (2022, Agustus). Naive Bayes dan Confusion Matrix untuk Efisiensi Analisa Intrusion Detection System Alert. Jurnal Nasional Teknologi dan Sistem Informasi, 8(2), 81-88. doi:10.25077/TEKNOSI.v8i2.2022.81-88

Wilson, AJ, Lakeland, BS, Wilson, TJ, & ... (2023). A naive Bayes classifier for identifying Class II YSOs. Monthly Notices of the …, academic.oup.com, https://academic.oup.com/mnras/article-abstract/521/1/354/7009217

Wijanarto, W., & Puspitasari, R. (2019, September). Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus. Eksplora Informatika, 9(1), 50-59. doi:10.30864/eksplora.v9i1.257

Yang, FJ (2018). An implementation of naive bayes classifier. 2018 International conference on computational …, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/8947658/

Yosmar, R., Amasdy, D., & Rahma, F. (2018, Agustus). Survei Resiko Penyakit Diabetes Melitus Terhadap Masyarakat Kota Padang. Jurnal Sains Farmasi & Kliis, 5(2), 134-141. doi:10.25077/jsfk.5.2.134-141.2018

Yusnaeni, W., & Widiarina, W. (2022, Januari). Penerapan Algoritma C4.5 dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes). Jurnal Teknik Komputer, 8(1), 56-60. doi:10.31294/jtk.v8i1.11566

Zhang, H, Jiang, L, & Yu, L (2021). Attribute and instance weighted naive Bayes. Pattern Recognition, Elsevier, https://www.sciencedirect.com/science/article/pii/S0031320320304775