

# Uas Data Science

Deteksi Dini Diabetes Menggunakan  
Naive Bayes Classifier dengan  
Laplacian Smoothing

18410100002

APRIANTO

18410100228


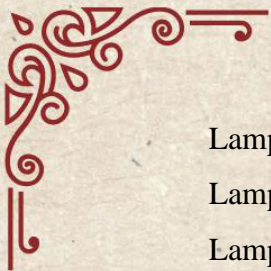
KRISTIN ANGELINA



## DAFTAR ISI

	Halaman
DAFTAR ISI.....	1
1. LATAR BELAKANG .....	3
2. TUJUAN.....	4
3. LANDASAN TEORI.....	5
3.1. Diabetes.....	5
3.2. Machine Learning .....	6
3.3. <i>Naïve Bayes Classifier</i> .....	7
3.4. <i>Laplacian Smoothing</i> .....	8
3.5. <i>Dataset Training</i> .....	8
3.6. <i>Dataset Testing</i> .....	8
3.7. <i>Confusion matrix</i> .....	9
3.8. Akurasi.....	9
3.9. Presisi.....	10
3.10. <i>Recall</i> .....	10
3.11. <i>F-1 Score</i> .....	11
4. CARA KERJA <i>MACHINE LEARNING</i> .....	12
5. CONTOH PERHITUNGAN MANUAL.....	14
5.1. Sebelum diterapkan <i>Laplacian Smoothing</i> .....	14
5.1.1. Kelas Hasil Ya.....	14
5.1.2. Kelas Hasil Tidak .....	16
5.1.3. Hasil Klasifikasi .....	17
5.2. Setelah diterapkan <i>Laplacian Smoothing</i> .....	17
5.2.1. Kelas Hasil Ya.....	19
5.2.2. Kelas Hasil Tidak .....	19
5.2.3. Hasil Klasifikasi .....	20
5.3. Kesimpulan .....	21
6. PENJELASAN KOLOM DATASET .....	22
7. ANALISA HASIL PERHITUNGAN .....	23
8. DAFTAR PUSTAKA.....	29
LAMPIRAN .....	31





Lampiran 1 - Dataset Diabetes .....	31
Lampiran 2 - Hasil Kinerja Klasifikasi .....	33
Lampiran 3 - Dataset Gejala.....	37



## 1. LATAR BELAKANG

Diabetes atau *Diabetes Melitus* merupakan salah satu penyakit yang ditandai dengan gejala lebihnya kadar gula darah diatas ambang normal yaitu sama atau lebih dari 200 mg/dl dan melebihi kadar gula darah puasa yang diatas atau sama dengan 126 mg/dl menurut Misnadiarly dalam Hestiana. Penyakit diabetes ini merupakan salah satu penyakit berbahaya karena sering tidak disadari oleh penderitanya dan saat diketahui sudah terjadi komplikasi dalam dirinya, dari situ diabetes ini juga sering dikenal sebagai *silent killer* (Hestiana, 2017).

Berdasarkan interpretasi data yang dilaporkan oleh International Diabetes Federation (IDF) dalam IDF Diabetes Atlas edisi 9 tahun 2019, Jumlah penyandang diabetes dunia pada 2019 adalah sebanyak 463 juta penderita dan akan di prediksi melonjak naik di tahun 2045 dengan jumlah 700 juta penderita (atau meningkat sebanyak 51%). Masih dalam data yang sama, Indonesia menduduki peringkat ketujuh negara dengan penyandang diabetes terbesar di dunia dari 211 negara yang terdata oleh IDF dengan jumlah penyandang sebesar 10,7 juta orang. Sedangkan untuk kawasan Pasifik Barat, Indonesia menduduki peringkat kedua dibawah negara China dari 36 negara yang terdata dengan tingkat prevalensi sebesar 6,2% dari jumlah populasi orang dewasa sebesar 172,2 juta jiwa pada 2019 (International Diabetes Federation, 2019).

Faktor-faktor yang memiliki andil dalam penyakit diabetes antara lain adalah faktor genetis atau keturunan dari keluarga, faktor usia, faktor gaya hidup (terkait pada pola makan, aktivitas sehari-hari) dan faktor riwayat terkena penyakit diabetes gestasional pada wanita saat hamil (Sunur, 2020). Sebagian besar penderita diabetes tidak menyadari bahwa dirinya berisiko atau sudah terdiagnosa mengalami diabetes, hal ini disebabkan oleh minimnya pengetahuan mengenai gejala-gejala yang terjadi membuat seakan tidak terjadi apa-apa. Menurut data yang disadur dari laman suara.com yang berasal dari Riset Kesehatan Dasar Badan Penelitian dan Pengembangan Kesehatan (Riskesdas Litbangkes) 2018 dan Konsensus Perkumpulan Endokrinologi Indonesia (PERKENI) 2015, 75% dari total penyandang diabetes di Indonesia belum menyadari bahwa dirinya menyandang diabetes. 25% sisanya sudah menyadari mereka menyandang diabetes dengan 17% pasien menjalani terapi diabetes dan 8% sisanya tidak menjalankan terapi (Rossa & Halidi, 2019).



Kondisi tersebut dapat ditangani ketika penyandang diabetes lebih dini mengetahui dirinya terkena diabetes atau setidaknya memahami dirinya berisiko atau tidak dibandingkan dengan pengobatan pasca diagnosis terkena diabetes yang sudah terlambat dan umumnya sudah disertai beragam komplikasi.

Kecerdasan buatan atau *Artificial Intelligence* (AI) disini dapat berperan dalam membantu deteksi dini adanya penyakit diabetes dengan menggunakan *Machine Learning* (ML) atau pembelajaran mesin. ML merupakan cabang dari AI dan dapat didefinisikan sebagai proses pemecahan masalah praktis dengan mengumpulkan dataset dan membangun model statistik dari data set tersebut, *Machine learning* mampu beradaptasi dengan data baru secara mandiri untuk menghasilkan keputusan yang andal dari perhitungan sebelumnya (Zohuri & Rahmani, 2019).

Machine learning dalam hal ini dapat digunakan untuk membantu seseorang mengenali lebih dini mengenai risiko penyakit diabetes. Klasifikasi merupakan salah satu jenis subset dalam machine learning yang dapat digunakan untuk mengklasifikasikan suatu hal kedalam salah satu jenis kategori. Dalam hal ini, machine learning dapat digunakan untuk melakukan klasifikasi terhadap gejala yang diberikan apakah gejala-gejala tersebut masuk kedalam penyakit diabetes atau tidak.

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

## 2. TUJUAN

Tujuan dari penulisan laporan ini adalah:

- ✗ Diharapkan pembaca atau seseorang mampu melakukan pengecekan dini terhadap terkait penyakit diabetes secara mandiri menggunakan bantuan Machine Learning.
- ✗ Menghemat pengeluaran pembaca karena pengecekan dini yang disediakan sama sekali tidak membutuhkan biaya dan tidak perlu melakukan pengecekan secara medis.



### 3. LANDASAN TEORI

#### 3.1. Diabetes

*Diabetes melitus* atau juga dikenal sebagai penyakit kencing manis yaitu penyakit yang disebabkan karena terganggunya metabolisme karbohidrat, protein dan lemak yang umumnya ditandai dengan peningkatan kadar glukosa dalam darah akibat kelainan sekresi insulin atau menurunnya kerja insulin yang menyebabkan gangguan pada kinerja metabolisme, kegagalan pada berbagai organ terutama pada organ mata, ginjal, saraf, jantung dan pembuluh darah (Wisudanti dalam Wahyuni et al., 2019). Diabetes sendiri merupakan salah satu penyakit yang berpotensi paling banyak menimbulkan komplikasi (memicu timbulnya penyakit lain) karena berkaitan dengan tingginya kadar gula darah sehingga berdampak pada rusaknya pembuluh darah, saraf dan struktur internal lainnya. Komplikasi ini dapat timbul jika penderita diabetes tidak ditangani dengan baik (Hayat, 2016).

Menurut infografis yang disajikan dalam laman Direktorat Pencegahan dan Pengendalian Penyakit Tidak Menular (P2PTM) Kementerian Kesehatan 4 Republik Indonesia (Kemenkes RI), penyakit diabetes dibedakan menjadi empat jenis diantaranya adalah *Diabetes Melitus* tipe 1 yang disebabkan tidak adanya produksi insulin dalam tubuh sama sekali, *Diabetes Melitus* tipe 2 yang disebabkan tidak efektifnya kerja insulin, *Diabetes Mellitus Gestasional* yang terjadi saat masa kehamilan dan *Diabetes Melitus* tipe lainnya yang disebabkan oleh penyakit lain, penggunaan obat ataupun yang lainnya (P2PTM Kemenkes RI, 2018)

Menurut (Tandra, 2017) Beberapa gejala pada penderita baru *diabetes mellitus* terdiri dari (1) Banyak kencing, (2) Sering merasakan haus, (3) Berat badan menurun, (4) Pandangan mata kabur, (5) Luka yang sulit sembuh, (6) Sering merasa kesemutan, (7) Merasa lemas, (8) Kulit terasa kering dan gatal, dan (9) Memiliki keturunan diabetes. Menurut (Tandra, 2017) Beberapa gejala pada penderita baru *diabetes mellitus* terdiri dari (1) Banyak kencing, (2) Sering merasakan haus, (3) Berat badan menurun, (4) Pandangan mata kabur, (5) Luka yang sulit sembuh, (6) Sering merasa kesemutan, (7) Merasa lemas, (8) Kulit terasa kering dan gatal, dan (9) Memiliki keturunan diabetes.

Berdasarkan data yang terkumpul dari IDF Diabetes Atlas edisi ke-9 tahun 2019, jumlah penderita diabetes terbesar adalah diabetes dengan tipe 1 yaitu sebanyak 8.483



jiwa dari 10.681 jiwa atau jika di prosentasekan sebanyak 79% (*International Diabetes Federation*, 2019). Gejala yang ditampakkan oleh diabetes melitus tipe 1 dan tipe 2 adalah sama, hal yang membedakannya adalah rentang waktu kemunculan gejala-gejalanya. Pada *diabetes melitus* tipe 1 gejala muncul dan berkembang dalam waktu beberapa minggu dan tampak lebih jelas dibandingkan *diabetes melitus* tipe 2 yang gejala awalnya tidak tampak jelas namun perlahan, bahkan sering terjadi kasus penderita *diabetes melitus* menyadari bahwa dirinya terkena saat sudah mengalami komplikasi yang serius (Sunur, 2020).

### **3.2. Machine Learning**

*Machine learning* (ML) atau pembelajaran mesin merupakan bagian atau subset dari kecerdasan buatan atau *Artificial Intelligence* (AI) yang merupakan simulasi dari kecerdasan manusia oleh sistem komputer. ML sendiri adalah kombinasi dari ilmu komputer dan statistik, ilmu komputer berfokus pada pemecahan masalah dan identifikasi apakah masalah dapat diselesaikan pada semua tahapan. Sedangkan statistik pada sisi pemodelan data, hipotesis dan mengukur keandalan (Pavithra Devi & Jayanthi, 2018). ML belajar dari pengalaman di masa lalu untuk meningkatkan kinerja di masa depan. *Machine learning* menggunakan data untuk belajar secara mandiri dengan memodifikasi atau meningkatkan algoritma tanpa adanya keterlibatan manusia. Algoritma dikembangkan dari berbagai disiplin ilmu untuk mengukur tingkat akurasi serta kemampuan dan kecepatan penyesuaian.

*Machine learning* sudah banyak diaplikasikan dalam berbagai bidang, tak terkecuali bidang medis. Bidang medis merupakan salah satu bidang yang berpotensi untuk dikembangkan dengan kemampuan machine learning. Dalam penelitian yang dilakukan oleh (Battineni et al., 2020) menyebutkan bahwa 5 ML dalam dunia medis dapat digunakan untuk mengekstrak pengetahuan medis, menemukan ide-ide baru untuk praktisi dan spesialis, mendiagnosis secara mandiri berbagai penyakit berdasarkan peraturan klinis, meningkatkan data medis, mengurangi fluktuasi tarif pasien, menghemat biaya medis serta mengurangi angka kematian yang disebabkan oleh penyakit kronis melalui deteksi dini dan perawatan yang efektif. Dalam penelitian ini, ML digunakan untuk mendiagnosis penyakit diabetes dalam diri seseorang berdasarkan gejala-gejala yang dirasakannya.



Pembelajaran dari *machine learning* umumnya dibagi menjadi empat algoritma atau kategori utama yaitu *supervised*, *semi-supervised*, *unsupervised* dan *reinforced type*. Dalam penelitian ini, algoritma atau kategori yang digunakan untuk mendiagnosis penyakit diabetes adalah algoritma *Supervised*. *Supervised learning* yaitu algoritma ML yang dikenal sebagai pembelajaran yang diawasi, karena pengguna perlu untuk menyediakan input dan output yang diinginkan (data historis) serta menemukan cara untuk menghasilkan output yang diinginkan berdasarkan input yang diberikan. Algoritma ini dapat memberikan output untuk input yang belum pernah dihasilkan sebelumnya tanpa bantuan manusia (Zohuri & Rahmani, 2019).

Dalam penelitian ini, algoritma *supervised learning* digunakan dengan menyajikan data historis penderita diabetes untuk selanjutnya dilakukan pembelajaran (*data training*) untuk dapat mengklasifikasikan suatu input yang akan diberikan berupa keputusan bahwa seseorang mengidap penyakit diabetes atau tidak.

### 3.3. Naïve Bayes Classifier

*Naïve Bayes* merupakan salah satu metode klasifikasi dalam *Machine Learning* yang dapat memprediksi keanggotaan kelas kepada kelas tertentu dengan menggunakan probabilitas. Dasar dari metode *Naïve Bayes* adalah Teorama *Bayes* yang dikembangkan Thomas Bayes yang awalnya digunakan dalam teori probabilitas dan keputusan selama abad ke-18 (Han et al., 2012). Klasifikasi dari metode *Naïve Bayes* dapat didasarkan pada rumus berikut ini.

$$P(C_i | X) = \frac{P(X|C_i)P(C_i)}{P(X)} \dots (3.1)$$

*Naïve Bayes* melakukan klasifikasi dengan mencari probabilitas tertinggi dari kelas yang diujikan (probabilitas posterior, diwakili  $(C_i|X)$ ), dalam penelitian ini berkorelasi dengan potensi terkena diabetes rendah, sedang atau tinggi.  $(X)$  merupakan probabilitas dari prediktor yaitu seberapa besar kemungkinan data uji akan masuk kedalam kelas tertentu (kelas rendah, sedang atau tinggi dalam potensi diabetes).  $(C_i|)$  merupakan probabilitas prior yaitu probabilitas kelas terhadap target atribut, dikorelasikan dengan probabilitas laki-laki (kelas target) dalam jenis kelamin yang ada (atribut target).  $(X|C_i|)$  merupakan probabilitas dari kelas target terhadap kelas prediktor,



dapat dikorelasikan dengan probabilitas diabetes rendah (kelas prediktor) pada jenis kelamin laki-laki (kelas dalam target jenis kelamin).

### 3.4. Laplacian Smoothing

*Laplacian smoothing* atau laplace smoothing merupakan salah satu metode atau algoritma pemulusan (*smoothing*) tertua yang digunakan. Menurut (Kilimci & Ganiz, 2015) *laplacian smoothing* merupakan metode yang berpengaruh untuk mencegah masalah probabilitas nol. Metode *laplacian smoothing* juga dikenal sebagai *add-one smoothing* yaitu menambahkan angka 1 pada setiap frekuensi token yang didapat (Listiowarni & Setyaningsih, 2018). Bentuk perhitungan dengan *laplacian smoothing* dapat dilihat pada rumus berikut ini.

$$P(W_i|class) = \frac{\text{freq}(W_i, \text{class}) + 1}{N_{\text{class}} + V_{\text{class}}} \dots (3.2)$$

*Laplacian Smoothing* menambahkan nilai 1 pada setiap frekuensi kelas dan menambahkan V class yaitu jumlah atribut pada kelas tertentu untuk menghasilkan probabilitas tanpa hasil nol. Pada penelitian ini, *laplacian smoothing* digunakan untuk menghindari nilai nol pada probabilitas yang dihasilkan oleh metode *naïve bayes*. Nilai probabilitas nol akan berpengaruh terhadap hasil klasifikasi karena mengkalkulasi beberapa probabilitas untuk mendapatkan nilai akhir yang menentukan nilai probabilitas.

### 3.5. Dataset Training

*Dataset training* adalah kumpulan data yang berisi nilai-nilai dari kedua komponen sebelumnya dan digunakan untuk melatih model dalam mengenali *class* yang cocok/sesuai, berdasarkan prediktor yang tersedia. Dalam pembuatan laporan ini, *dataset training* digunakan dalam perhitungan manual dengan membandingkan hasil klasifikasi dari aplikasi dan algoritma *Naïve Bayes* yang bertujuan untuk menguji kerja aplikasi yang dikembangkan.

### 3.6. Dataset Testing

*Dataset testing* adalah kumpulan data yang bersifat baru dan akan digunakan dalam pengklasifikasian oleh *classifier* model yang telah dibangun sehingga hasil akurasi klasifikasi (*model performance*) dapat dievaluasi.



### 3.7. Confusion matrix

*Confusion matrix* merupakan tabel pencatat hasil kerja klasifikasi. *Confusion matrix* melakukan pengujian untuk memperkirakan objek yang benar dan salah. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi (Aprilia et al., 2016). Tabel berikut adalah contoh tabel *confusion matrix* yang menunjukkan klasifikasi dua kelas.

		Kelas yang Terprediksi	
		Positif	Negatif
Kelas yang Sesungguhnya	Positif	True positive count (TP)	False negative count (FN)
	Negatif	False positive count (FP)	True negative count (TN)

Penjelasan dari tabel di atas adalah sebagai berikut:

- ✘ *True Positives* (TP) adalah *record* yang diprediksikan sebagai positif atau benar dan kenyataan dari *record* tersebut adalah positif atau benar,
- ✘ *False Positives* (FP) adalah *record* yang diklasifikasikan sebagai negatif atau salah dan kenyataan dari *record* tersebut adalah positif atau benar,
- ✘ *False Negatives* (FN) adalah *record* yang diklasifikasikan sebagai positif atau benar dan kenyataan dari *record* tersebut adalah negatif atau salah,
- ✘ *True Negatives* (TN) adalah *record* yang diklasifikasikan sebagai negatif atau salah dan kenyataan dari *record* tersebut adalah negatif atau salah.

Dari 4 poin di atas dapat digambarkan bahwa nilai prediksi adalah keluaran dari program dimana nilainya Positif dan Negatif dan Nilai Aktual adalah nilai sebenarnya dimana nilainya Benar dan Salah (Anggreany, 2021). Data uji yang dimasukkan ke dalam *confusion matrix*, akan dihitung nilai-nilai *recall*, *precision* dan *accuracy*.

### 3.8. Akurasi

Akurasi menggambarkan tingkat keakuratan model dalam pengklasifikasian yang telah dihitung dengan menggunakan *Confusion matrix*. Dengan kata lain, Akurasi menjelaskan seberapa banyak data aktual yang benar diklasifikasikan oleh sistem dengan



ketentuan jumlah data yang benar diklasifikasikan sistem dibagi jumlah data keseluruhan (Arthana, 2019). Dalam laporan ini, akurasi menjawab pertanyaan “Berapa persen pasien yang benar diprediksi terkena diabetes dan tidak terkena diabetes dari keseluruhan pasien?”. Rumus dari akurasi adalah hasil penjumlahan antara *True Positive* dan *True Negative* dibagi dengan hasil penjumlahan antara *True Positive*, *False Positive*, *False Negative*, dan *True Negative*. Penulisannya adalah sebagai berikut:

$$Akurasi = \frac{(TP+TN)}{(TP+FP+FN+TN)} \dots (3.3)$$

### 3.9. Presisi

Presisi menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model melalui perhitungan *Confusion matrix* (Arthana, 2019). Dalam laporan ini, presisi menjawab pertanyaan “Berapa persen pasien yang benar terkena diabetes dari keseluruhan pasien yang diprediksi terkena diabetes?”. Rumus dari presisi adalah nilai *True Positive* dibagi dengan hasil penjumlahan antara *True Positive* dan *False Positive*, penulisannya adalah sebagai berikut:

$$Presisi = \frac{(TP)}{(TP+FP)} \dots (3.4)$$

### 3.10. Recall

*Recall*, atau *sensitivity* menggambarkan tingkat keberhasilan sistem dalam menemukan/menghasilkan prediksi yang sesuai dengan *class* sebenarnya (Arthana, 2019). Dalam laporan ini, *recall* menjawab pertanyaan “Berapa persen pasien yang diprediksi terkena diabetes dibandingkan dengan keseluruhan pasien yang sebenarnya terkena diabetes?”. Rumus dari *recall* adalah nilai *True Positive* dibagi dengan hasil penjumlahan antara *True Positive* dan *False Negative*, penulisannya adalah sebagai berikut:

$$Recall = \frac{(TP)}{(TP+FN)} \dots (3.5)$$



### 3.11. F-1 Score

F-1 Score menggambarkan perbandingan rata-rata presisi dan *recall* yang dibobotkan. Akurasi digunakan sebagai acuan performansi algoritma jika dataset memiliki jumlah data *False Negatif* dan *False Positif* yang sangat mendekati (*symmetric*) (Arthana, 2019). Namun jika jumlahnya tidak mendekati, maka disinilah F1 Score digunakan dalam perhitungan sebagai acuan. Rumus dari F-1 Score adalah:

$$F - 1 \text{ Score} = \frac{(2 * \text{Recall} * \text{Presisi})}{(\text{Recall} + \text{Presisi})} \dots (3.6)$$



#### 4. CARA KERJA MACHINE LEARNING

*Naïve Bayes Classifier* (NBC) merupakan salah satu metode yang digunakan untuk klasifikasi dalam ruang lingkup *Machine Learning*. Cara kerja NBC adalah melakukan perhitungan probabilitas untuk memprediksi peluang dimasa depan berdasarkan pengalaman sebelumnya (Informatikalogi, 2021). Perhitungan probabilitas yang dimaksud adalah menghitung nilai probabilitas setiap kelas dari perbandingan data yang akan diklasifikasi dengan data yang digunakan sebagai *dataset testing* sesuai jumlah kelas yang ada. Kelas yang dimaksudkan adalah variabel yang digunakan untuk melakukan klasifikasi, dalam studi kasus ini yang dimaksud kelas adalah gejala-gejala diabetes yang digunakan. *Dataset testing* yang dimaksud adalah kumpulan data yang akan digunakan sebagai acuan klasifikasi terhadap input data gejala baru.

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

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

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

Untuk hasil prediksi ya, dari dataset akan dihitung berapa buah data yang memenuhi kondisi `usia=50-60; turun_bb=ya; keturunan=ya; hasil=ya`; terhadap jumlah *dataset*. Hasil



dari prediksi “ya” terhadap gejala tersebut dapat misalkan sebagai  $\frac{a}{n}$  dengan a sebagai jumlah data yang sesuai dan n sebagai jumlah dataset keseluruhan. Hal yang sama dilakukan untuk hasil prediksi “tidak” yaitu dengan mencari probabilitas terhadap data yang memenuhi kondisi **usia=50-60; turun\_bb=ya;keturunan=ya;hasil=tidak**. Hasil prediksi “tidak” terhadap kondisi tersebut dapat dimisalkan sebagai  $\frac{b}{n}$  dengan b sebagai jumlah data yang sesuai dengan kondisi dan n sebagai jumlah dataset keseluruhan. Dari perhitungan probabilitas terhadap kelas hasil “ya” dan “tidak”, kemudian akan diambil nilai tertinggi dari keduanya sebagai hasil klasifikasi. Hasil probabilitas tertinggi menandakan bahwa sebuah kondisi pada hasil tertentu akan condong kearah hasil tersebut.

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

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

Untuk mendapatkan pemahaman lebih lanjut mengenai perhitungan probabilitas untuk mengklasifikasikan apakah seseorang menderita diabetes dapat dilihat pada bagian contoh perhitungan [berikut](#).



## 5. CONTOH PERHITUNGAN MANUAL

Pada bagian sebelumnya telah dijelaskan mengenai bagaimana cara kerja *Naïve Bayes classifier* untuk melakukan klasifikasi. Pada bagian ini akan dijelaskan mengenai contoh perhitungan manual dari metode *Naïve Bayes classifier*. Contoh perhitungan pada bagian ini akan menggunakan dataset pada lampiran [berikut](#) untuk sumber data dan bagian dibawah ini sebagai input klasifikasi atau permasalahan yang akan diselesaikan menggunakan perhitungan.

Kode Kelas	Kelas Gejala	Kelas Hasil
A	Range Usia	50-60 Tahun
B	Jenis Kelamin	Wanita
C	Banyak Kencing?	Ya
D	Turun Berat Badan?	Tidak
E	Luka Sukar Sembuh?	Tidak
F	Sering Kesemutan?	Ya
G	Sering merasa Lemas?	Ya
H	Kulit Gatal-gatal?	Ya
I	Riwayat / Keturunan Diabetes?	Ya
Y	Terdeteksi Diabetes?	Ya / Tidak ?

Berdasarkan data input gejala diatas, klasifikasi dilakukan dengan melakukan perhitungan terhadap sejumlah 9 kelas gejala pada masing-masing kelas hasil ya dan tidak. Probabilitas tertinggi dari kelas hasil ya / tidak akan menjadi hasil dari klasifikasi.

$$P(C_i | X) = \frac{P(X|C_i)P(C_i)}{P(X)} \dots (5.1)$$

### 5.1. Sebelum diterapkan Laplacian Smoothing

Dibawah ini merupakan contoh perhitungan klasifikasi *Naïve Bayes* dengan melakukan perhitungan masing-masing untuk kelas hasil ya dan tidak. Hasil perhitungan tertinggi pada probabilitas kelas hasil ya atau tidak akan menjadi hasil klasifikasi.

#### 5.1.1. Kelas Hasil Ya

$$X1 = P(hasil = ya) = \frac{54}{100} = 0.54 \dots (5.2)$$



Pada persamaan (5.2) diatas dapat dilihat bahwa probabilitas kelas hasil ya pada seluruh dataset adalah 0.54 yang akan akan dikalikan dengan probabilitas kelas gejala yang lain. Persamaan diatas pada formula *Naïve Bayes classifier* dikenal sebagai  $P(C_i)$ , dapat dilihat pada persamaan (1.1).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \dots (5.3)$$

Pada persamaan 5.3 diatas digunakan untuk menghitung probabilitas dari suatu kelas gejala yang sesuai dengan kondisi kelas hasil ya atau tidak. Probabilitas dari dua kondisi tersebut disubstitusi kedalam bentuk persamaan 5.3. Makna dari  $P(A \cap B)$  adalah irisan dimana dari dataset akan dicari sejumlah data yang memenuhi 2 kondisi, sedangkan  $P(B)$  merupakan probabilitas dari kondisi B. Implementasi dari persamaan 5.3 dapat dilihat pada perhitungan berikut ini.

$$A1 = P(Usia = 50 - 60 | hasil = ya) \dots (5.4)$$

$$A1 = \frac{P(Usia=50-60 \cap hasil=ya)}{P(hasil=ya)} \dots (5.5)$$

$$A1 = \frac{19}{54} = 0.352 \dots (5.6)$$

Formula 3.4 diatas menjelaskan perhitungan dari persamaan 4.3 dimana  $P(A \cap B)$  variabel A menggambarkan kelas gejala usia dan variabel B menggambarkan kelas hasil ya. Pada perhitungan 4.5 meggambarkan bahwa pembilang merupakan irisan dari kedua kondisi dimana pada dataset dicari jumlah data yang cocok dengan kondisi memiliki usia 50-60 tahun dan hasil klasifikasinya ya. Dari *dataset*, jumlah data yang memenuhi kondisi tersebut ada 19 baris data. Sedangkan pada penyebutnya adalah mencari jumlah data yang memiliki hasil klasifikasi ya, dapat dilihat kembali pada persamaan 4.2 bahwa jumlah data yang memenuhi kondisi tersebut ada sebanyak 54 baris data. Sehingga probabilitas untuk kelas gejala usia tersebut yang memiliki hasil klasifikasi ya adalah 0.35. selanjutnya perhitungan yang sama dilakukan untuk 8 kelas gejala lainnya pada kelas hasil ya lalu dilanjutkan mencari 9 kelas gejala lainnya pada kelas hasil tidak.

$$B1 = P(Jenis Kelamin = wanita | hasil = ya) = \frac{22}{54} = 0.407$$

$$C1 = P(Banyak Kencing = ya | hasil = ya) = \frac{36}{54} = 0.667$$



$$D1 = P(\text{Turun BB} = \text{tidak} | \text{hasil} = \text{ya}) = \frac{22}{54} = 0.407$$

$$E1 = P(\text{Luka Sukar} = \text{tidak} | \text{hasil} = \text{ya}) = \frac{20}{54} = 0.37$$

$$F1 = P(\text{Kesemutan} = \text{ya} | \text{hasil} = \text{ya}) = \frac{32}{54} = 0.593$$

$$G1 = P(\text{Lemas} = \text{ya} | \text{hasil} = \text{ya}) = \frac{36}{54} = 0.667$$

$$H1 = P(\text{Kulit Gatal} = \text{ya} | \text{hasil} = \text{ya}) = \frac{31}{54} = 0.574$$

$$I1 = P(\text{Keturunan} = \text{ya} | \text{hasil} = \text{ya}) = \frac{39}{54} = 0.722$$

Beberapa perhitungan dibawah sudah dilakukan untuk mencari probabilitas pada setiap kelas gejala. Selanjutnya pada probabilitas  $X1$  hingga  $I1$  dilakukan perkalian untuk menghasilkan nilai probabilitas akhir yang akan digunakan sebagai acuan dalam menentukan gejala-gejala yang diinputkan memiliki kecenderungan hasil ya atau tidak. Perhitungan dapat dilihat dibawah ini.

$$Y1 = X1 \times A1 \times B1 \times C1 \times D1 \times E1 \times F1 \times G1 \times H1 \times I1$$

$$Y1 = (0.54)(0.352)(0.407)(0.667)(0.407)(0.37)(0.593)(0.667)(0.574)(0.722)$$

$$Y1 = 0.00127$$

### 5.1.2. Kelas Hasil Tidak

$$X2 = P(\text{hasil} = \text{tidak}) = \frac{46}{100} = 0.46$$

$$A2 = P(\text{Usia} = 50 - 60 | \text{hasil} = \text{tidak}) = \frac{10}{46} = 0.217$$

$$B2 = P(\text{Jenis Kelamin} = \text{wanita} | \text{hasil} = \text{tidak}) = \frac{22}{46} = 0.478$$

$$C2 = P(\text{Banyak Kencing} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{20}{46} = 0.435$$

$$D2 = P(\text{Turun BB} = \text{tidak} | \text{hasil} = \text{tidak}) = \frac{25}{46} = 0.543$$



$$E2 = P(\text{Luka Sukar} = \text{tidak} | \text{hasil} = \text{tidak}) = \frac{21}{46} = 0.457$$

$$F2 = P(\text{Kesemutan} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{20}{46} = 0.435$$

$$G2 = P(\text{Lemas} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{17}{46} = 0.37$$

$$H2 = P(\text{Kulit Gatal} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{12}{46} = 0.261$$

$$I2 = P(\text{Keturunan} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{13}{46} = 0.282$$

$$Y2 = X2 \times A2 \times B2 \times C2 \times D2 \times E2 \times F2 \times G2 \times H2 \times I2$$

$$Y2 = (0.46)(0.217)(0.487)(0.435)(0.543)(0.457)(0.435)(0.37)(0.261)(0.283)$$

$$Y2 = 0.00006$$

### 5.1.3. Hasil Klasifikasi

Dapat dilihat pada perhitungan diatas menunjukkan perhitungan probabilitas untuk kelas hasil tidak. Selanjutnya akan diambil nilai tertinggi antara hasil probabilitas kelas ya (variabel Y1) dan hasil probabilitas kelas tidak (variabel Y2). Hasil probabilitas antar kedua variabel Y1 dan Y2 jika dibandingkan dapat dilihat pada tabel berikut ini.

Y1	Y2	Hasil (Terbesar)
0.00127	0.00006	Y1 / "Ya"

Dari tabel diatas dapat terlihat jika variabel Y1 memiliki nilai probabilitas dibandingkan dengan variabel Y2 sehingga dapat diambil kesimpulan bahwa dari kumpulan gejala yang diinputkan akan **terklasifikasi kedalam hasil Ya (Pengidap Penyakit Diabetes)**.

### 5.2. Setelah diterapkan Laplacian Smoothing

Pada bagian sebelumnya telah dijelaskan mengenai perhitungan manual dari metode *Naïve Bayes classifier*. Perhitungan sebelumnya dilakukan menggunakan metode *Naïve Bayes* secara normal, belum diterapkan pemulusan menggunakan *laplacian smoothing* atau dikenal juga dengan sebutan *laplace*. Pada bagian ini akan dijelaskan mengenai perhitungan menggunakan *laplacian smoothing* untuk menghindari nilai 0 pada



probabilitas yang dihitung. Sebagai contoh dapat dilihat pada persamaan 4.6, pada persamaan ditemukan hasil  $\frac{19}{54}$  dimana pembilang merupakan jumlah data yang cocok dengan kondisi yang diinputkan. Jika tidak menggunakan *laplace*, perhitungan tersebut berpotensi mengembalikan nilai 0 pada pembilang yang akan mempengaruhi hasil probabilitas. Jika tidak ada data yang sesuai dengan kondisi yang sesuai, maka probabilitasnya adalah  $\frac{0}{n}$  dan hasil dari probabilitas adalah 0.

Hasil klasifikasi tidak dapat ditentukan jika hasil probabilitas akhir (variabel Y) memiliki nilai 0 pada kedua hasil ya dan tidak. Peran *laplace* membantu untuk menghindari nilai 0 dengan menambahkan nilai 1 pada pembilang dan nilai x pada penyebut. Nilai x adalah jumlah atribut unik yang ada pada setiap kelas gejala. Untuk memberikan gambaran mengenai nilai x dapat dilihat pada tabel berikut ini (kolom jumlah atribut).

Kelas Gejala	Atribut	Jumlah Atribut
Range Usia	{ 20-40; 40-50; 50-60 }	3
Jenis Kelamin	{ Pria; Wanita }	2
Banyak Kencing?	{ Ya; Tidak }	2
Turun Berat Badan?	{ Ya; Tidak }	2
Luka Sukar Sembuh?	{ Ya; Tidak }	2
Sering Kesemutan?	{ Ya; Tidak }	2
Sering merasa Lemas?	{ Ya; Tidak }	2
Kulit Gatal-gatal?	{ Ya; Tidak }	2
Riwayat / Keturunan Diabetes?	{ Ya; Tidak }	2
Terdeteksi Diabetes?	{ Ya; Tidak }	2

Untuk menggunakan *laplace*, probabilitas pada pembilang ditambahkan nilai 1 (*default laplacian smoothing*) dan pada penyebut ditambahkan nilai x dimana x adalah jumlah atribut pada kelas gejala. Karena pada persamaan 4.6 adalah kondisi  $P(Usia = 50 - 60 | hasil = ya)$ , maka nilai x adalah jumlah atribut pada kelas usia. Pada tabel diatas dapat dijadikan sebagai acuan bahwa jumlah atribut kelas usia adalah 3, sehingga bentuk *laplace* seperti tabel dibawah ini.

Sebelum Laplace	Setelah Laplace
$\frac{19}{54}$	$\frac{19 + 1}{54 + x} = \frac{19 + 1}{54 + 3} = \frac{20}{57}$



### 5.2.1. Kelas Hasil Ya

$$X1 = P(\text{hasil} = \text{ya}) = \frac{54 + 1}{100 + 2} = \frac{55}{102} = 0.539$$

$$A1 = P(\text{Usia} = 50 - 60 | \text{hasil} = \text{ya}) = \frac{19 + 1}{54 + 3} = \frac{20}{57} = 0.351$$

$$B1 = P(\text{Jenis Kelamin} = \text{wanita} | \text{hasil} = \text{ya}) = \frac{22 + 1}{54 + 2} = \frac{23}{56} = 0.411$$

$$C1 = P(\text{Banyak Kencing} = \text{ya} | \text{hasil} = \text{ya}) = \frac{36 + 1}{54 + 2} = \frac{37}{56} = 0.661$$

$$D1 = P(\text{Turun BB} = \text{tidak} | \text{hasil} = \text{ya}) = \frac{22 + 1}{54 + 2} = \frac{23}{56} = 0.411$$

$$E1 = P(\text{Luka Sukar} = \text{tidak} | \text{hasil} = \text{ya}) = \frac{20 + 1}{54 + 2} = \frac{21}{56} = 0.375$$

$$F1 = P(\text{Kesemutan} = \text{ya} | \text{hasil} = \text{ya}) = \frac{32 + 1}{54 + 2} = \frac{33}{56} = 0.589$$

$$G1 = P(\text{Lemas} = \text{ya} | \text{hasil} = \text{ya}) = \frac{36 + 1}{54 + 2} = \frac{37}{56} = 0.661$$

$$H1 = P(\text{Kulit Gatal} = \text{ya} | \text{hasil} = \text{ya}) = \frac{31 + 1}{54 + 2} = \frac{32}{56} = 0.571$$

$$I1 = P(\text{Keturunan} = \text{ya} | \text{hasil} = \text{ya}) = \frac{39 + 1}{54 + 2} = \frac{40}{56} = 0.714$$

$$Y1 = X1 \times A1 \times B1 \times C1 \times D1 \times E1 \times F1 \times G1 \times H1 \times I1$$

$$Y1 = (0.539)(0.351)(0.411)(0.661)(0.411)(0.375)(0.589)(0.661)(0.571)(0.714)$$

$$Y1 = 0.00126$$

### 5.2.2. Kelas Hasil Tidak

$$X2 = P(\text{hasil} = \text{tidak}) = \frac{46 + 1}{100 + 2} = \frac{47}{102} = 0.461$$

$$A2 = P(\text{Usia} = 50 - 60 | \text{hasil} = \text{tidak}) = \frac{10 + 1}{46 + 3} = \frac{11}{49} = 0.224$$



$$B2 = P(\text{Jenis Kelamin} = \text{wanita} | \text{hasil} = \text{tidak}) = \frac{22 + 1}{46 + 2} = \frac{23}{48} = 0.479$$

$$C2 = P(\text{Banyak Kencing} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{20 + 1}{46 + 2} = \frac{21}{48} = 0.435$$

$$D2 = P(\text{Turun BB} = \text{tidak} | \text{hasil} = \text{tidak}) = \frac{25 + 1}{46 + 2} = \frac{26}{48} = 0.542$$

$$E2 = P(\text{Luka Sukar} = \text{tidak} | \text{hasil} = \text{tidak}) = \frac{21 + 1}{46 + 2} = \frac{22}{48} = 0.458$$

$$F2 = P(\text{Kesemutan} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{20 + 1}{46 + 2} = \frac{21}{48} = 0.438$$

$$G2 = P(\text{Lemas} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{17 + 1}{46 + 2} = \frac{18}{48} = 0.375$$

$$H2 = P(\text{Kulit Gatal} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{12 + 1}{46 + 2} = \frac{13}{48} = 0.271$$

$$I2 = P(\text{Keturunan} = \text{ya} | \text{hasil} = \text{tidak}) = \frac{13 + 1}{46 + 2} = \frac{14}{48} = 0.292$$

$$Y2 = X2 \times A2 \times B2 \times C2 \times D2 \times E2 \times F2 \times G2 \times H2 \times I2$$

$$Y2 = (0.461)(0.224)(0.479)(0.438)(0.542)(0.458)(0.438)(0.375)(0.271)(0.292)$$

$$Y2 = 0.00007$$

### 5.2.3. Hasil Klasifikasi

Dapat dilihat pada perhitungan diatas menunjukkan perhitungan probabilitas untuk kelas hasil tidak. Selanjutnya akan diambil nilai tertinggi antara hasil probabilitas kelas ya (variabel Y1) dan hasil probabilitas kelas tidak (variabel Y2). Hasil probabilitas antar kedua variabel Y1 dan Y2 jika dibandingkan dapat dilihat pada tabel berikut ini.

Y1	Y2	Hasil (Terbesar)
0.00126	0.00007	Y1 / "Ya"

Dari tabel diatas dapat terlihat jika variabel Y1 memiliki nilai probabilitas dibandingkan dengan variabel Y2 sehingga dapat diambil kesimpulan bahwa dari kumpulan gejala yang diinputkan akan **terklasifikasi kedalam hasil Ya (Pengidap Penyakit Diabetes).**



### 5.3. Kesimpulan

Berdasarkan perhitungan yang sudah dilakukan pada variabel Y1 untuk kelas hasil ya dan variabel Y2 untuk kelas hasil tidak dengan perhitungan tanpa *laplace* dan dengan menggunakan *laplace*. Hasil dari perhitungan dapat dilihat pada tabel perbandingan dibawah ini.

	Y1	Y2
Tanpa Laplace	0.00127	0.00006
Dengan Laplace	0.00126	0.00007

Dapat terlihat pada tabel diatas bahwa nilai antara perhitungan tanpa *laplace* dan dengan *laplace* tetap mengklasifikasikan hasil ke variabel Y1 yaitu hasil ya. Namun dapat dilihat bahwa terdapat perbedaan sedikit nilai pada nilai sebelum dan setelah *laplace*. Perbedaan nilai tersebut diakibatkan karena penambahan nilai 1 dan nilai x. Dengan melakukan perhitungan menggunakan *laplacian smoothing*, kinerja dari *Naïve Bayes classifier* dapat dioptimalkan untuk menghindari nilai 0 (terutama saat jumlah data sedikit).



## 6. PENJELASAN KOLOM DATASET

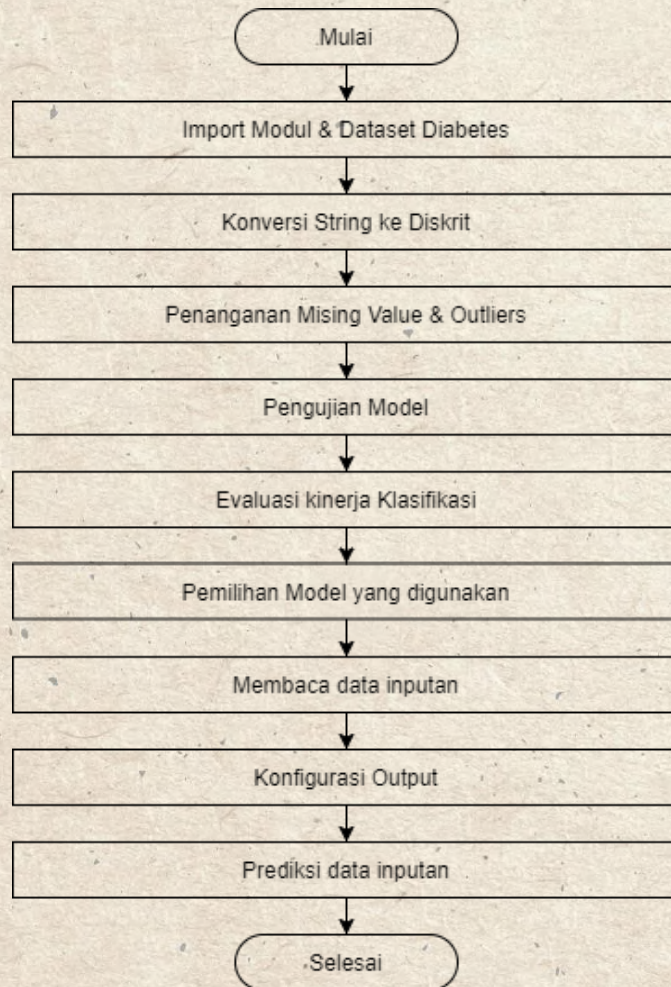
Pada studi kasus deteksi dini menggunakan machine learning dengan metode *Naïve Bayes classifier* ini, kolom yang digunakan pada dataset mengacu pada beberapa gejala dari penderita diabetes (Tandra, 2017). Kolom dataset yang digunakan sebagai parameter inputan ada 9, dimana 8 kolom sebagai gejala diabetes dan 1 kolom sebagai kolom hasil klasifikasi. Adapun kolom yang digunakan dalam dataset dijelaskan dalam tabel berikut ini.

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



## 7. ANALISA HASIL PERHITUNGAN

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



Pertama dilakukan import data yang akan digunakan untuk training model klasifikasi beserta modul yang akan digunakan untuk membantu proses klasifikasi. Modul *Naïve Bayes* yang digunakan berasal dari Library Scikit-learn dengan modul *Multinomial Naïve Bayes*. Sebenarnya pada *library* scikit-learn modul klasifikasi dengan *Naïve Bayes* memiliki beberapa jenis seperti *Gaussian Classifier*, *Bernoulli Classifier* dan *Multinomial Classifier*. *Gaussian Classifier* cocok diterapkan terhadap data berjenis kontinu yang memiliki distribusi normal. *Bernoulli Classifier* cocok digunakan untuk



data yang bersifat binary (memiliki nilai benar/salah atau 0/1). *Multinomial classifier* cocok digunakan terhadap data yang memiliki beberapa atribut seperti usia memiliki 3 atribut (Packt, 2018). Berdasarkan tiga jenis model klasifikasi naïve bayes, model yang paling cocok digunakan dalam kasus ini adalah model klasifikasi *Multinomial NB* dimana dapat menampung beberapa atribut dan tidak terkait dengan statistik persebaran data (distribusi data).

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

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

```

NILAI SETIAP pclass
-----
usia
-----
1. 20-40
2. 40-50
3. 50-60

-----
jkel
-----
1. wanita
2. pria

-----
banyak_kencing
-----
1. ya
2. tidak

-----
turun_bb
-----
1. ya
2. tidak
  
```

Pengecekan outliers pada setiap kelas

```

-----
luka_sukar
-----
1. ya
2. tidak

-----
kesemutan
-----
1. ya
2. tidak

-----
lemas
-----
1. ya
2. tidak

-----
kulit_gatal
-----
1. tidak
2. ya
  
```

```

-----
keturunan
-----
1. ya
2. tidak

-----
hasil
-----
1. ya
2. tidak
  
```

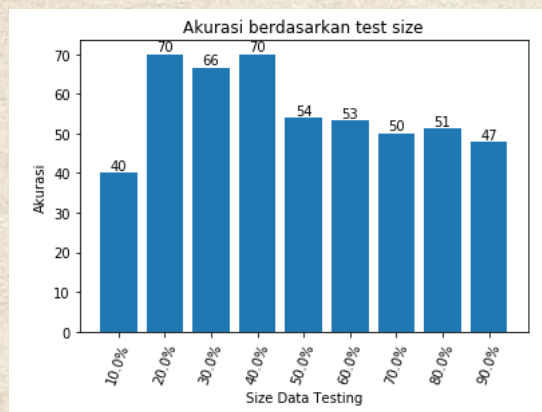
```

Out[6]: usia      0
        jkel      0
        banyak_kencing  0
        turun_bb   0
        luka_sukar   0
        kesemutan   0
        lemas       0
        kulit_gatal  0
        keturunan   0
        hasil       0
        dtype: int64
  
```

Pengecekan *missing values*



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



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

Kinerja dari hasil klasifikasi pada setiap *test size* dapat dilihat pada [lampiran kinerja](#) klasifikasi untuk melihat detail dari klasifikasi yang berisi mengenai informasi akurasi, *confusion matrix*, presisi, *recall* dan *f-1 score*.

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



	usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan
5	20-40	pria	ya	ya	ya	ya	tidak	tidak	ya
HASIL : TERDETEKSI DIABETES									

	usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan
6	40-50	pria	ya	ya	ya	ya	ya	tidak	tidak
HASIL : TIDAK TERDETEKSI DIABETES									

Berikutnya akan dilakukan analisa dari hasil perhitungan manual dengan hasil perhitungan melalui python 3. Perhitungan manual direpresentasikan dengan perhitungan menggunakan website SiDeDi (Sistem Informasi Deteksi Diabetes) yang merujuk pada penelitian terkait Penggunaan Metode *Naïve Bayes* Classifier dengan Laplacian Smoothing yang pernah dilakukan dan hasil-nya sudah verifikasi dengan perhitungan manual (Aprianto et al., 2021),(Aprianto et al., 2020). Perhitungan diwakili dengan penggunaan website terkait untuk menyingkat waktu(Aprianto et al., 2020). Perhitungan diwakili dengan penggunaan website terkait untuk menyingkat waktu.

Analisa dilakukan dengan melakukan perbandingan hasil antara perhitungan manual (diwakili dengan menggunakan website SiDeDi) dengan perhitungan melalui python 3 sebanyak 10 baris data gejala sebagai inputan yang dapat dilihat pada lampiran [dataset gejala](#).

Data ke-	Hasil Klasifikasi		Hasil
	Manual	Python	
1	Ya	Ya	Sama
2	Tidak	Ya	Berbeda
3	Ya	Ya	Sama
4	Tidak	Ya	Berbeda

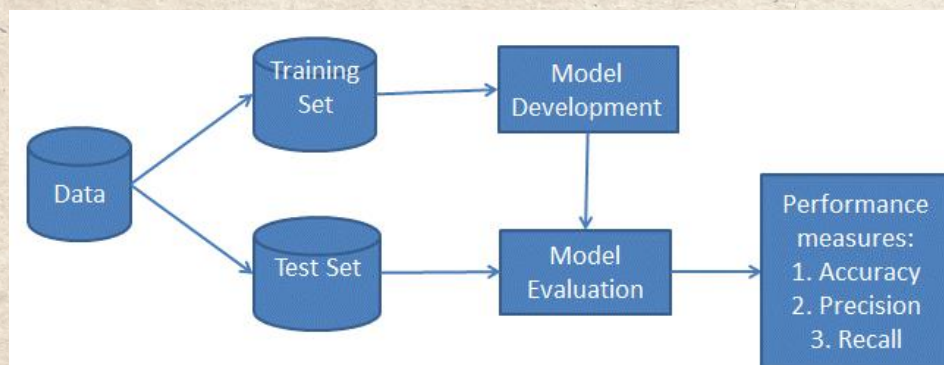


5	Ya	Ya	Sama
6	Ya	Ya	Sama
7	Ya	Tidak	Berbeda
8	Tidak	Tidak	Sama
9	Ya	Tidak	Berbeda
10	Ya	Ya	Sama
<b>Kesamaan</b>			<b>60%</b>

Berdasarkan perbandingan diatas bahwa terdapat cukup banyak perbedaan antara hasil prediksi perhitungan manual dan perhitungan dengan menggunakan python. Hal tersebut karena disebabkan oleh beberapa hal yang dapat terangkum sebagai berikut.

1. Pembagian dataset kedalam *training set* dan *testing set*

Perbedaan utama dari hasil klasifikasi terletak pada test size. Dalam perhitungan yang dihasilkan oleh website SiDeDi mengacu pada perhitungan manual dimana perhitungan melibatkan seluruh dari dataset sebagai data acuan klasifikasi, berbeda dengan penggunaan modul *Multinomial Naïve Bayes* pada python yang terdapat pembagian antara 100% data kedalam data training dan data testing dimana dalam proses klasifikasi akan menggunakan data testing. Data training digunakan dalam proses pembelajaran mesin yang dilakukan oleh metode *Naïve Bayes classifier*. Model klasifikasi yang digunakan dalam python dapat dilihat pada gambar berikut ini (mengacu pada Navlani, 2018).



2. Perbedaan random state

Berkaitan dengan poin 1 yaitu dataset yang dijadikan acuan klasifikasi oleh python terbagi kedalam 2 kategori *training set* dan *testing set*. Dalam studi kasus ini digunakan sampel untuk test sebesar 40% dan 60% sisanya untuk *training set*. Saat melakukan pemanggilan model *Multinomial Naïve Bayes* dibutuhkan beberapa parameter saat inisiasi seperti berikut ini.



```
In [22]: 1 # alpha 1 -> Laplacian smoothing
2 model = MultinomialNB(alpha=1.0)
3
4 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =0.4, random_state = 44)
5
6 # training data
7 model.fit(X_train,y_train)

Out[22]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

Pada gambar diatas dapat terlihat bahwa terdapat parameter `random_state` yang digunakan untuk melakukan inisiasi terhadap 40% dari data mana saja yang akan dimasukan kedalam *testing set* sebagai acuan klasifikasi. Default dari random set adalah 44 dan ini dapat diganti. Jika random size berganti maka dampak yang terjadi adalah perubahan pada data mana saja yang dipilih untuk testing size dan dapat menghasilkan perbedaan pada hasil klasifikasi. Namun karena pada perhitungan manual data yang digunakan adalah 100% data untuk testing, maka random state tidak berpengaruh.



## 8. DAFTAR PUSTAKA

- Anggreany, M. S. (2021). *Confusion matrix*.  
<https://socs.binus.ac.id/2020/11/01/confusion-matrix/>
- Aprianto, Prasetyo, R. E., & Sahartian, O. (2021). SiDedi (Sistem Informasi Deteksi Diabetes): Sistem Pendukung Keputusan Deteksi Dini Diabetes. *JSIKA*.  
<https://drive.google.com/file/d/1Ulb1ytcyasTnpDu6U3PXiaf0f-GSm47Q/view>
- Aprianto, Young, C., & Sari, R. (2020). *PKM-KC: SiDedi (Sistem Informasi Deteksi Diabetes): Aplikasi Android Deteksi Dini Diabetes*.  
[https://drive.google.com/file/d/1UioahT0e3HuTjQnhpDrK9\\_1HPPJh4fN5/view](https://drive.google.com/file/d/1UioahT0e3HuTjQnhpDrK9_1HPPJh4fN5/view)
- Aprilia, R., Muludi, K., & Aristoteles. (2016). Pemetaan Sebaran Asal Siswa Dan Klasifikasi Jarak Asal Siswa Sma Negeri Di Kabupaten Pringsewu Menggunakan Metode Naïve Bayes. *Jurnal Komputasi Ilmu Komputer Unila*, 4(2), 52–66.
- Arthana, R. (2019). *Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning*. <https://rey1024.medium.com/mengenal-accuracy-precision-recall-dan-specificity-septa-yang-diprioritaskan-b79ff4d77de8>
- Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*, 10(21). <https://doi.org/10.3390/jpm10020021>
- Han, J., Kamber, M., & Pei, J. (2012). Data mining: Data mining concepts and techniques. In *The Morgan Kaufmann Series in Data Management Systems* (3rd ed.). Elsevier. <https://doi.org/10.1109/ICMIRA.2013.45>
- Hayat, C. (2016). Identifikasi Dini Penyakit Diabetes Melitus Menggunakan Expert System Builder Early Identification of Diabetes Mellitus Disease Using Expert System Builder. *Jurnal Teknik Dan Ilmu Komputer*, 5(20), 431–445.
- Hestiana, D. W. (2017). Journal of Health Education. *Journal of Health Education*, 2(2), 138–145. <https://doi.org/10.1080/10556699.1994.10603001>
- Informatikalogi. (2021). *Algoritma Naive Bayes*. <https://informatikalogi.com/algoritma-naive-bayes/>
- International Diabetes Federation. (2019). Global Diabetes Data Report 2010-2045. *Journal IDF*. <https://diabetesatlas.org/data/en/world/>
- Kilimci, Z. H., & Ganiz, M. C. (2015). Evaluation of Classification Models for Language Processing. *2015 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*. <https://ieeexplore.ieee.org/abstract/document/7276787/figures#figures>
- Listiowarni, I., & Setyaningsih, E. R. (2018). Analisis Kinerja Smoothing pada Naive Bayes untuk Pengkategorian Soal Ujian. *Jurnal Teknologi Dan Manajemen Informatika*, 4(2).
- Navlani, A. (2018). *Naive Bayes Classification using Scikit-learn*.



<https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>

P2PTM Kemenkes RI. (2018). *Lindungi Keluarga Dari Diabetes*.  
<http://p2ptm.kemkes.go.id/post/lindungi-keluarga-dari-diabetes>

Packt. (2018). *Implementing 3 Naive Bayes classifiers in scikit-learn*.  
<https://hub.packtpub.com/implementing-3-naive-bayes-classifiers-in-scikit-learn/>

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

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

Sunur, I. C. (2020). *Mengenal Perbedaan Diabetes Tipe 1 dan Tipe 2*.  
<https://Www.Alodokter.Com/Mengenal-Perbedaan-Diabetes-Tipe-1-Dan-Tipe-2>.  
<https://www.alodokter.com/mengenal-perbedaan-diabetes-tipe-1-dan-tipe-2>

Tandra, H. (2017). *Segala Sesuatu yang Harus Anda Ketahui Tentang Diabetes*. Gramedia.

Wahyuni, R., Ma'ruf, A., & Mulyono, E. (2019). Hubungan Pola Makan Terhadap Kadar Gula Darah Penderita Diabetes Mellitus. *Jurnal Medika Karya Ilmiah Kesehatan*, 4(2). <http://jurnal.stikeswhs.ac.id/index.php/medika>

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



## LAMPIRAN

### Lampiran 1 - Dataset Diabetes

usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan	hasil
20-40	wanita	ya	ya	ya	ya	ya	tidak	ya	ya
40-50	wanita	tidak	ya	tidak	tidak	ya	ya	tidak	tidak
20-40	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
50-60	wanita	ya	tidak	ya	ya	tidak	ya	ya	ya
40-50	pria	ya	ya	tidak	ya	ya	ya	tidak	ya
20-40	pria	ya	tidak	tidak	tidak	tidak	ya	ya	tidak
50-60	wanita	tidak	ya	ya	tidak	ya	ya	tidak	ya
50-60	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
20-40	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
50-60	pria	ya	ya	tidak	ya	tidak	ya	ya	ya
40-50	pria	tidak	tidak	ya	tidak	ya	ya	ya	ya
50-60	wanita	ya	tidak	tidak	ya	tidak	ya	ya	ya
40-50	wanita	tidak	ya	ya	ya	tidak	tidak	tidak	tidak
50-60	wanita	tidak	tidak	ya	tidak	tidak	tidak	ya	tidak
20-40	pria	ya	ya	tidak	ya	tidak	ya	tidak	ya
40-50	wanita	ya	tidak	ya	tidak	tidak	tidak	tidak	tidak
20-40	pria	tidak	ya	ya	tidak	ya	tidak	ya	ya
40-50	wanita	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
50-60	pria	ya	ya	tidak	ya	tidak	ya	ya	ya
20-40	pria	ya	tidak	ya	tidak	ya	tidak	ya	ya
40-50	wanita	ya	ya	tidak	ya	tidak	ya	tidak	ya
40-50	pria	tidak	ya	ya	tidak	ya	tidak	ya	ya
40-50	wanita	ya	tidak	tidak	ya	tidak	ya	tidak	tidak
20-40	pria	tidak	ya	tidak	tidak	ya	tidak	ya	tidak
40-50	wanita	ya	tidak	ya	tidak	ya	ya	tidak	ya
50-60	pria	tidak	ya	tidak	tidak	tidak	tidak	ya	tidak
40-50	pria	ya	tidak	tidak	ya	tidak	ya	tidak	tidak
20-40	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
20-40	wanita	tidak	ya	tidak	ya	tidak	ya	ya	ya
40-50	pria	ya	ya	ya	ya	tidak	tidak	tidak	ya
40-50	pria	tidak	ya	tidak	ya	ya	ya	tidak	ya
20-40	pria	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
40-50	wanita	ya	ya	ya	ya	tidak	tidak	ya	ya
20-40	pria	tidak	ya	ya	tidak	ya	tidak	ya	ya
40-50	wanita	ya	ya	ya	ya	tidak	tidak	ya	ya
20-40	pria	tidak	ya	ya	tidak	ya	tidak	ya	ya
40-50	wanita	ya	tidak	tidak	ya	tidak	tidak	ya	ya
50-60	wanita	tidak	tidak	ya	tidak	ya	tidak	ya	ya
40-50	pria	ya	ya	ya	tidak	ya	tidak	tidak	ya
20-40	pria	ya	tidak	tidak	ya	tidak	tidak	ya	tidak
40-50	wanita	tidak	ya	ya	tidak	ya	tidak	tidak	tidak
20-40	pria	ya	tidak	ya	tidak	ya	tidak	ya	ya
50-60	wanita	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
40-50	pria	ya	tidak	ya	tidak	ya	ya	ya	ya
40-50	pria	ya	tidak	ya	ya	ya	ya	tidak	ya
50-60	wanita	ya	tidak	ya	ya	tidak	tidak	ya	ya
20-40	pria	tidak	ya	tidak	ya	tidak	tidak	tidak	tidak
40-50	pria	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
50-60	wanita	tidak	ya	ya	ya	ya	tidak	ya	ya
50-60	pria	tidak	ya	tidak	ya	tidak	tidak	ya	tidak
20-40	pria	ya	tidak	ya	tidak	ya	ya	ya	ya
20-40	pria	ya	ya	tidak	ya	tidak	tidak	tidak	tidak
20-40	wanita	ya	tidak	ya	tidak	tidak	ya	tidak	tidak
40-50	pria	ya	ya	tidak	ya	ya	ya	tidak	ya
50-60	wanita	tidak	ya	ya	ya	ya	tidak	ya	ya
50-60	pria	tidak	ya	tidak	ya	tidak	tidak	ya	tidak
20-40	wanita	ya	tidak	ya	tidak	ya	tidak	ya	ya
20-40	pria	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
20-40	wanita	ya	tidak	ya	tidak	tidak	ya	tidak	tidak
40-50	pria	ya	ya	tidak	ya	ya	ya	tidak	ya



usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan	hasil
50-60	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
20-40	wanita	ya	ya	tidak	ya	tidak	ya	ya	ya
40-50	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
40-50	pria	tidak	tidak	tidak	ya	ya	ya	ya	ya
20-40	pria	ya	ya	ya	ya	tidak	tidak	ya	ya
40-50	pria	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
20-40	wanita	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
50-60	wanita	ya	tidak	ya	ya	ya	ya	tidak	ya
40-50	pria	tidak	ya	ya	ya	ya	tidak	ya	ya
40-50	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
20-40	pria	tidak	ya	tidak	ya	ya	tidak	ya	ya
20-40	wanita	tidak	ya	ya	tidak	tidak	tidak	tidak	tidak
40-50	pria	tidak	tidak	tidak	ya	ya	tidak	ya	tidak
50-60	wanita	ya	ya	tidak	tidak	tidak	ya	ya	ya
50-60	pria	ya	tidak	ya	tidak	ya	tidak	ya	ya
20-40	pria	tidak	ya	tidak	ya	tidak	tidak	tidak	tidak
20-40	wanita	ya	tidak	ya	ya	tidak	ya	ya	ya
50-60	pria	tidak	ya	ya	tidak	tidak	tidak	tidak	tidak
40-50	pria	tidak	tidak	ya	tidak	ya	tidak	ya	tidak
50-60	pria	ya	tidak	ya	ya	tidak	ya	tidak	ya
20-40	wanita	ya	ya	tidak	ya	ya	tidak	ya	ya
50-60	wanita	ya	tidak	ya	tidak	ya	ya	ya	ya
20-40	pria	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
40-50	wanita	ya	tidak	ya	ya	tidak	tidak	tidak	tidak
40-50	pria	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
50-60	pria	ya	tidak	ya	tidak	ya	ya	ya	ya
40-50	pria	ya	ya	tidak	ya	tidak	ya	ya	ya
40-50	wanita	tidak	ya	ya	ya	tidak	ya	tidak	ya
50-60	wanita	ya	tidak	ya	tidak	tidak	tidak	tidak	tidak
20-40	pria	tidak	ya	tidak	ya	tidak	ya	ya	ya
40-50	pria	ya	tidak	ya	ya	tidak	tidak	tidak	tidak
20-40	pria	tidak	ya	tidak	ya	ya	tidak	ya	ya
50-60	wanita	ya	tidak	ya	tidak	ya	tidak	tidak	tidak
20-40	pria	tidak	ya	tidak	ya	ya	ya	ya	ya
50-60	pria	ya	tidak	ya	tidak	ya	tidak	ya	ya
40-50	wanita	tidak	ya	tidak	ya	tidak	ya	tidak	tidak
50-60	wanita	ya	tidak	ya	tidak	ya	tidak	ya	ya



## Lampiran 2 - Hasil Kinerja Klasifikasi

<----- Sample test size 10.0% ----->

Akurasi

=====  
40.0%

*Confusion matrix*

=====  
[[ 1 6]  
 [ 0 3]]

Classification Report

=====

	precision	recall	f1-score	support
0	1.00	0.14	0.25	7
1	0.33	1.00	0.50	3
accuracy			0.40	10
macro avg	0.67	0.57	0.38	10
weighted avg	0.80	0.40	0.33	10

<----- Sample test size 20.0% ----->

Akurasi

=====  
70.0%

*Confusion matrix*

=====  
[[ 4 6]  
 [ 0 10]]

Classification Report

=====

	precision	recall	f1-score	support
0	1.00	0.40	0.57	10
1	0.62	1.00	0.77	10
accuracy			0.70	20
macro avg	0.81	0.70	0.67	20
weighted avg	0.81	0.70	0.67	20

<----- Sample test size 30.0% ----->

Akurasi

=====  
66.67%

*Confusion matrix*

=====  
[[ 5 10]  
 [ 0 15]]



### Classification Report

	precision	recall	f1-score	support
0	1.00	0.33	0.50	15
1	0.60	1.00	0.75	15
<i>accuracy</i>			0.67	30
macro avg	0.80	0.67	0.62	30
weighted avg	0.80	0.67	0.62	30

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

Akurasi

70.0%

Confusion matrix

```
[[ 7 12]
 [ 0 21]]
```

### Classification Report

	precision	recall	f1-score	support
0	1.00	0.37	0.54	19
1	0.64	1.00	0.78	21
<i>accuracy</i>			0.70	40
macro avg	0.82	0.68	0.66	40
weighted avg	0.81	0.70	0.66	40

<----- Sample test size 50.0% ----->

Akurasi

54.0%

Confusion matrix

```
[[ 5 18]
 [ 5 22]]
```

### Classification Report

	precision	recall	f1-score	support
0	0.50	0.22	0.30	23
1	0.55	0.81	0.66	27
<i>accuracy</i>			0.54	50
macro avg	0.53	0.52	0.48	50
weighted avg	0.53	0.54	0.49	50



<----- Sample test size 60.0% ----->

Akurasi

=====

53.33%

*Confusion matrix*

=====

```
[[ 8 20]
 [ 8 24]]
```

Classification Report

=====

	precision	recall	f1-score	support
0	0.50	0.29	0.36	28
1	0.55	0.75	0.63	32
accuracy			0.53	60
macro avg	0.52	0.52	0.50	60
weighted avg	0.52	0.53	0.51	60

<----- Sample test size 70.0% ----->

Akurasi

=====

50.0%

*Confusion matrix*

=====

```
[[ 9 23]
 [12 26]]
```

Classification Report

=====

	precision	recall	f1-score	support
0	0.43	0.28	0.34	32
1	0.53	0.68	0.60	38
accuracy			0.50	70
macro avg	0.48	0.48	0.47	70
weighted avg	0.48	0.50	0.48	70

<----- Sample test size 80.0% ----->

Akurasi

=====

51.25%

*Confusion matrix*

=====

```
[[ 4 35]
 [ 4 37]]
```



# Classification Report

	precision	recall	f1-score	support
0	0.50	0.10	0.17	39
1	0.51	0.90	0.65	41
<i>accuracy</i>			0.51	80
macro avg	0.51	0.50	0.41	80
weighted avg	0.51	0.51	0.42	80

<----- Sample test size 90.0% ----->

## Akurasi

47.78%

## Confusion matrix

```
[[11 31]
 [16 32]]
```

# Classification Report

	precision	recall	f1-score	support
0	0.41	0.26	0.32	42
1	0.51	0.67	0.58	48
<i>accuracy</i>			0.48	90
macro avg	0.46	0.46	0.45	90
weighted avg	0.46	0.48	0.46	90



### Lampiran 3 - Dataset Gejala

usia	jkel	banyak_kencing	turun_bb	luka_sukar	kesemutan	lemas	kulit_gatal	keturunan
50-60	wanita	ya	tidak	tidak	ya	ya	ya	ya
20-40	wanita	tidak	tidak	tidak	tidak	tidak	ya	tidak
40-50	wanita	tidak	ya	ya	tidak	ya	tidak	ya
20-40	wanita	tidak	tidak	tidak	tidak	tidak	tidak	tidak
20-40	pria	tidak	tidak	ya	tidak	ya	tidak	ya
20-40	pria	ya	ya	ya	ya	tidak	tidak	ya
40-50	pria	ya	ya	ya	ya	ya	tidak	tidak
40-50	pria	tidak	ya	ya	tidak	ya	tidak	tidak
50-60	pria	tidak	ya	ya	ya	tidak	tidak	ya