Mata Kuliah - Pembelajaran Mesin

Nama Kelompok : [dikosongi]

Anggota :

202010370311153 – Ivon Viqro Dini

202010370311168 – Ria Wulandari

Step 1 : Data Understanding and Preprocessing

(DM) Diabetes Mellitus merupakan penyakit kronis yang ditandai dengan ketidakmampuan tubuh dalam melakukan metabolisme karbohidrat, lemak, dan protein sehingga terjadinya peningkatan kadar gula darah (hiperglikemia) yang disebabkan karena menurunnya kadar hormon insulin. Penyakit diabetes bisa terjadi karena faktor keturunan (genetika) dan pengaruh lingkungan gaya hidup yang tidak sehat. Dalam mengukur kadar gula darah diperlukan tes hemoglobin A1c untuk diagnosa dan mengontrol kondisi penderita diabetes. Adapun tujuan yang dilakukan dalam penelitian ini, yaitu membuat analisa model prediksi untuk klasifikasi penderita diabetes menggunakan R Shiny dan mengevaluasi hasil kinerja klasifikasi dari metode support vector machine. Terdapat banyak cara dalam mendiagnosa penyakit diabetes, berikut salah satu algoritma machine learning yang digunakan dalam kasus klasifikasi pada penelitian ini, yaitu support vector machine (SVM).

- ## Real-World Data

  Dataset yang digunakan dalam tahap ini yaitu dari kaggle.com wanita keturunan indian pima yang mempunyai 9 atribut dan 1 label, 9 atribut ini terdiri dari Gender, Age, hypertension, Heart_disease, Smoking_history, bmi, hbA1c_level, blood_glucose level, diabetes. Data ini berupa tabular berjumlah

- Url dataset : https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/code

Tabel 1. Dataset yang digunakan.

Atribut	Keterangan
Gender	identitas pria / wanita
age	Umur pasien
hypertension	Tekanan darah
heart_disease	Penderita diabetes memiliki resiko yang lebih tinggi untuk mengalami penyakit kardiovaskular
smoking_history	Riwayat merokok pada penderita diabetes
bmi	Indeks massa badan weight in kg/height in
hbA1c_level	Rata-Rata kadar glukosa darah selama 2 hingga 3 bulan terakhir
blood_glucose level	Level glukosa darah dalam diabetes
diabetes	Penderita penyakit diabetes

- **Data Pre-Processing**

a. Load Dataset

Dataset yang digunakan diambil dari situs Kaggle terdapat 100000 data dengan target kelas (Gender, Age, hypertension, Heart_disease, Smoking_history, bmi, hbA1c_level, blood_glucose level, diabetes)

b. Data Mining

Perlu diketahui tahapan yang dilakukan pada data mining ini yaitu seleksi data lalu pemilihan data sebelum lanjut ke tahapan penggalian informasi dimulai. berikutnya preprocessing dimana tahapan ini memiliki 3 step diantaranya data cleaning, data selection, transformasi data. Data mining merupakan proses pengumpulan sebuah informasi penting pada suatu data yang berukuran besar dan dapat diartikan juga sebagai teknik penambahan data untuk melakukan suatu analisa dengan teknik penyaringan informasi secara lebih akurat.

c. Forward Selection

Suatu metode yang bertujuan untuk menambah variabel yang akan ditentukan pada nilai tertentu. fungsi yang digunakan untuk mengklasifikasikan antar fitur adalah garis sedangkan fungsi yang digunakan untuk mengklasifikasikan fitur dalam 3-D disebut

bidang, begitu pula fungsi yang mengklasifikasikan titik dalam dimensi yang lebih tinggi disebut hyperplane. Namun hasil akurasi yang diperoleh masih belum cukup dikategorikan baik dan tidak digeneralisasikan dengan baik. Oleh karena itu perlu meningkatkan akurasi ke dimensi yang lebih tinggi, hal ini disebut sebagai fungsi kernel.

d. Pemodelan (SVM)

Algoritma yang digunakan adalah Support Vektor Machine (SVM) untuk melakukan klasifikasi. hasil dari klasifikasi ini menghasilkan model atau fungsi yang menjelaskan perbedaan konsep kelas data dan untuk tahapan pemodelan svm ini dengan bertujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space. untuk tingkat akurasi pada model svm ini sangat bergantung terhadap fungsi kernel dan parameter yang digunakan.

Tabel 2. tipe kernel yang digunakan

Nama Kernel	Fungsi
Linear	$K(x,y) = (x^T \text{ y})$
Gussian	$K(x,y) = exp\{-\frac{\lvert x-y \rvert^2}{2y^2}\}$
Polynomial	$K(x,y) = yx^T \text{y} + \text{coef . 0})^d$

e. Confusion Matrix

Merupakan metode pengukuran untuk mencari suatu masalah klasifikasi yang dilakukan oleh machine learning dengan berupa dua kelas atau lebih. ada 4 istilah yang merupakan representasi hasil dari klasifikasi pada confusion matrix yaitu (TP) true positive merupakan data bersifat positive terdeteksi benar, (TN) true negative bersifat negative terdeteksi dengan benar, (FP) false positive data bersifat negative namun terdeteksi menjadi data bersifat positive, (FN) false negative kebalikan dari true positive data yang bersifat positive namun terdeteksi data negative.

Tabel 3. Confusion Matrix

Classification	Actual +	Actual -
prediksi +	True Positive (TP)	False Positive (FP)
prediksi -	False Negative (FN)	True Negative (TN)

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

Gambar 1. Confusion Matrix

- ## Exploratory Data Analysis (EDA)

EDA (Exploratory Data Analysis) adalah suatu pendekatan analisis data yang digunakan untuk memahami karakteristik dasar dari dataset sebelum menerapkan model statistik atau machine learning. Tujuan utama dari EDA adalah untuk merangkum, meringkas, dan menggali informasi dari data secara visual dan deskriptif**.**

1. **Deskribsi Statistik**

   deskripsi statistik adalah langkah awal yang krusial dalam analisis data, yang melibatkan penggunaan metrik-metrik statistik untuk merangkum, mengorganisir, dan menyajikan data dengan tujuan memahami karakteristik utama dari dataset.

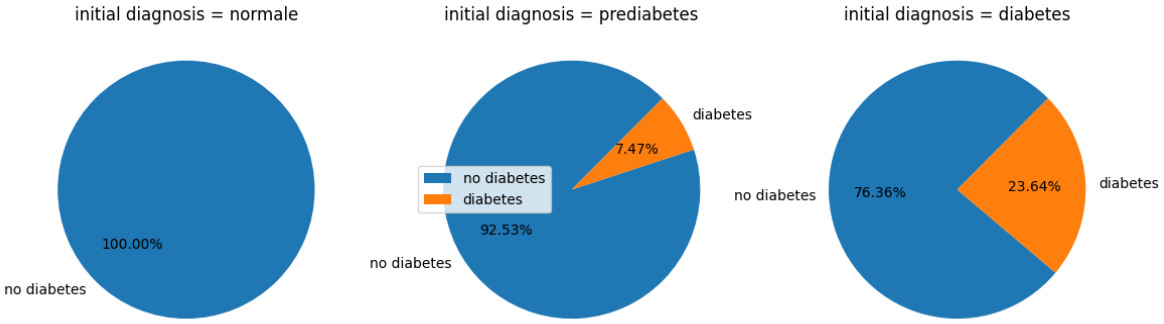
	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes	smoking_history_num
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.00000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000	0.27465
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883	1.40306
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000	-1.00000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000	-1.00000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000	0.00000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000	1.00000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000	4.00000

Gambar 2. Deskripsi Statistik dari Prediksi Diabetes

2. **Visualisasi Data**

   Visualisasi data dalam konteks prediksi diabetes adalah sarana yang efektif untuk membawa pemahaman yang lebih baik tentang hubungan antar variabel, distribusi data, dan pola-pola penting. Bahasa visual melibatkan berbagai jenis grafik dan diagram yang dapat memberikan wawasan langsung kepada pemirsa. Visualisasi dalam prediksi diabetes menggunakan elemen-elemen seperti warna, skala, dan label untuk memberikan dimensi tambahan dan memudahkan pemirsa dalam menginterpretasi dan memahami informasi yang

terkandung dalam visualisasi. Dengan menggunakan bahasa visual ini, kita dapat merangkum kompleksitas data dan membuatnya lebih dapat diakses oleh berbagai pemangku kepentingan.Berikut adalah beberapa visualisasi yang telah dilakukan.

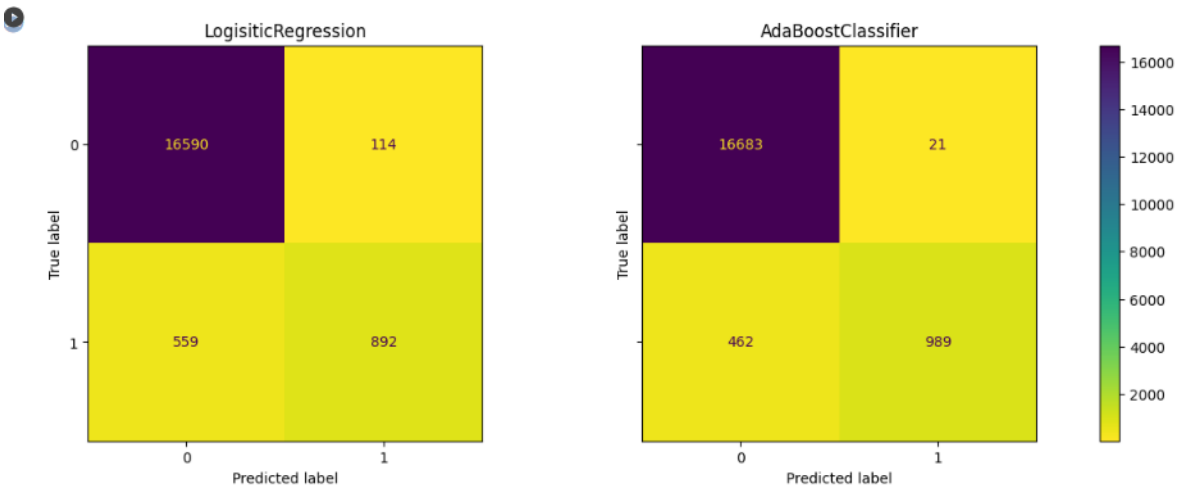


Gambar 3. Visualisasi hbA1C

Kita dapat melihat bahwa nilai HbA1c_level berpengaruh terhadap prediksi menderita diabetes atau tidak, ketika nilai HbA1c_level berada pada nilai normal, kita melihat bahwa tidak ada catatan dengan diabetes seiring dengan peningkatan nilai HbA1c_level, jumlah catatan penderita diabetes meningkat ## khususnya bila >= 6,5.

- **Model Building & Training**

  Model Building & Training adalah tahap penting dalam pengembangan model machine learning. Proses ini melibatkan pembuatan dan pelatihan model untuk dapat memahami pola atau hubungan dalam data dan membuat prediksi atau klasifikasi.



	Model	Precision_Neg	Precision_Pos	Recall_Neg	Recall_Pos	F1_Score_Neg	F1_Score_Pos	Accuracy
0	LogisiticRegression	0.967	0.887	0.993	0.615	0.980	0.726	0.963
1	AdaBoostClassifier	0.973	0.979	0.999	0.682	0.986	0.804	0.973

Gambar 4. Model Training

Kita dapat melihat bahwa AdaBoost Classifier berkinerja lebih baik daripada Regresi Logisitik yang menunjukkan skor lebih tinggi di semua skor metrik. Namun, Pengklasifikasi AdaBoost memberikan prediksi yang buruk pada label Positif dan kami memperkirakan 31,8% penderita diabetes tidak menderita diabetes. Kesalahan dalam jumlah besar ini bisa berbahaya jika kita menggunakan model prediksi diabetes kita. Jadi ketidakseimbangan label dapat sangat memengaruhi hasil algoritme pembelajaran mesin kami. Ketidak seimbangan ini menyebabkan algoritma dengan Akurasi Tinggi tetapi tanpa keterampilan hal ini terjadi karena dataset yang tidak seimbang

- ..