TUGAS BESAR ANALISIS DAN SAINS DATA

Nama : Ahmad Muslikh
Kelas/NIM : B/60200122067
Judul dataset : Diabetes Mellitus Disease
Link dataset : https://www.kaggle.com/datasets/prosperchuks/health-dataset

---

1. Business Understanding

Deskripsi permasalahan:

Di era modern saat ini, diabetes merupakan salah satu penyakit tidak menular yang menjadi perhatian serius dalam dunia kesehatan. Pola hidup yang tidak sehat, seperti konsumsi makanan tinggi gula, kurangnya aktivitas fisik, dan stres, telah menyebabkan peningkatan jumlah penderita diabetes secara signifikan di berbagai belahan dunia, termasuk Indonesia. Penyakit ini dapat menimbulkan komplikasi serius seperti gangguan jantung, gagal ginjal, bahkan kematian apabila tidak ditangani secara tepat dan dini.

Salah satu tantangan dalam penanganan penyakit diabetes adalah proses diagnosis yang cepat dan akurat. Umumnya diagnosis dilakukan melalui pemeriksaan laboratorium yang membutuhkan waktu, biaya, dan akses ke fasilitas medis. Oleh karena itu, dibutuhkan solusi berbasis teknologi yang dapat membantu dalam proses identifikasi atau klasifikasi risiko diabetes dengan lebih efisien.

Dengan memanfaatkan teknik data mining, khususnya klasifikasi, kita dapat membangun model prediktif untuk mengidentifikasi apakah seseorang berisiko terkena diabetes berdasarkan atribut-atribut tertentu seperti kadar glukosa, tekanan darah, indeks massa tubuh, dan lain-lain. Model ini diharapkan dapat membantu tenaga medis atau masyarakat umum dalam melakukan deteksi dini secara digital terhadap risiko penyakit diabetes, sehingga tindakan pencegahan atau pengobatan dapat dilakukan lebih cepat dan tepat sasaran.

Tugas analisis: Klasifikasi

2. Data Understanding

Deskripsi data:

Dataset yang digunakan dalam proyek ini bertujuan untuk membangun model klasifikasi guna mendeteksi kemungkinan seseorang mengidap penyakit diabetes berdasarkan beberapa indikator gaya hidup, kesehatan fisik, serta riwayat penyakit.

Dataset ini terdiri dari 18 kolom attribute dan 70.692 kolom, Dimana setiap baris merepsentasikan data satu individu.

```
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   70692 non-null  float64
 1   Sex                   70692 non-null  float64
 2   HighChol              70692 non-null  float64
 3   CholCheck             70692 non-null  float64
 4   BMI                   70692 non-null  float64
 5   Smoker                70692 non-null  float64
 6   HeartDiseaseorAttack  70692 non-null  float64
 7   PhysActivity          70692 non-null  float64
 8   Fruits                70692 non-null  float64
 9   Veggies               70692 non-null  float64
 10  HvyAlcoholConsump     70692 non-null  float64
 11  GenHlth               70692 non-null  float64
 12  MentHlth              70692 non-null  float64
 13  PhysHlth              70692 non-null  float64
 14  DiffWalk              70692 non-null  float64
 15  Stroke                70692 non-null  float64
 16  HighBP                70692 non-null  float64
 17  Diabetes              70692 non-null  float64
dtypes: float64(18)
```

No	Nama Atribut	Deksripsi	Tipe Data	ΣMissing Value
1	Age	Usia responden dalam kategori usia 1–13 (representasi kelompok umur)	Numerik	
2	Sex	Jenis Kelamin (0 = laki laki, 1= perempuan)	Binominal	
3	HighCol	Memiliki kadar highChol (0 = Ya, 1 = Tidak)	Binominal	
4	CholCheck	Pernah melakukan pemeriksaan kolestrol (0 = Ya, 1 = Tidak)	Binominal	
5	BMI	Indeks massa tubuh (Body Mass Index)	Numerik	
6	Smoker	Merokok dalam 30 hari terakhir (1 = Ya, 0 = Tidak)	Binominal	
7	HeartDiseaseorAttack	Riwayat penyakit jantung atau serangan	Binominal	

		jantung (1 = Ya, 0 = Tidak)		
8	PhysActivity	Melakukan aktivitas fisik dalam 30 hari terakhir (1 = Ya, 0 = Tidak)	Binominal	
9	Fruits	Mengonsumsi buah minimal sekali sehari (1 = Ya, 0 = Tidak)	Binominal	
10	Veggies	Mengonsumsi sayur minimal sekali sehari (1 = Ya, 0 = Tidak)	Binominal	
11	HvyAlcoholConsump	Konsumsi alkohol berat (1 = Ya, 0 = Tidak)	Binominal	
12	GenHealth	Penilaian kesehatan umum (1 = Sangat Baik, 5 = Sangat Buruk)	Ordinal	
13	Diffwalk	Kesulitan berjalan atau naik tangga (1 = Ya, 0 = Tidak)	Binominal	
14	Stroke	Pernah mengalami stroke (1 = Ya, 0 = Tidak)	Binominal	
15	HighBP	Memiliki tekanan darah tinggi (1 = Ya, 0 = Tidak)	Binominal	
16	Menthealth	Jumlah hari tidak sehat secara mental dalam 30 hari terakhir	Numerik	
17	PhysHlth	Jumlah hari tidak sehat secara fisik dalam 30 hari terakhir	Numerik	
18	Diabetes	**LABEL –** Status diabetes (0 = tidak, 1 = diabetes)	Binominal	

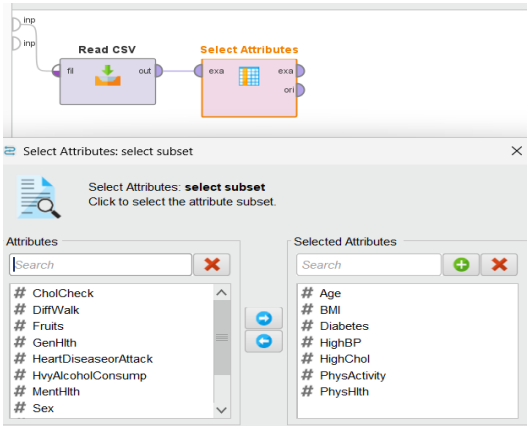
3. Data Preparation

Deskripsi:

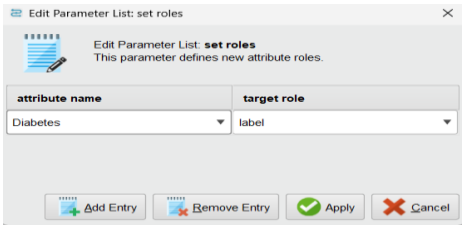
Pada tahap ini, dilakukan serangkaian proses untuk membersihkan, memfilter, dan menyiapkan data agar siap digunakan dalam proses modeling. Proses ini sangat penting agar algoritma klasifikasi dapat bekerja secara optimal dan hasil yang diperoleh lebih akurat serta dapat diandalkan.

Berdasarkan hasil eksplorasi awal terhadap dataset yang terdiri dari 70.692 baris dan 18 atribut, tidak ditemukan nilai yang hilang (missing value) dalam dataset ini, sehingga tidak diperlukan proses imputasi data.

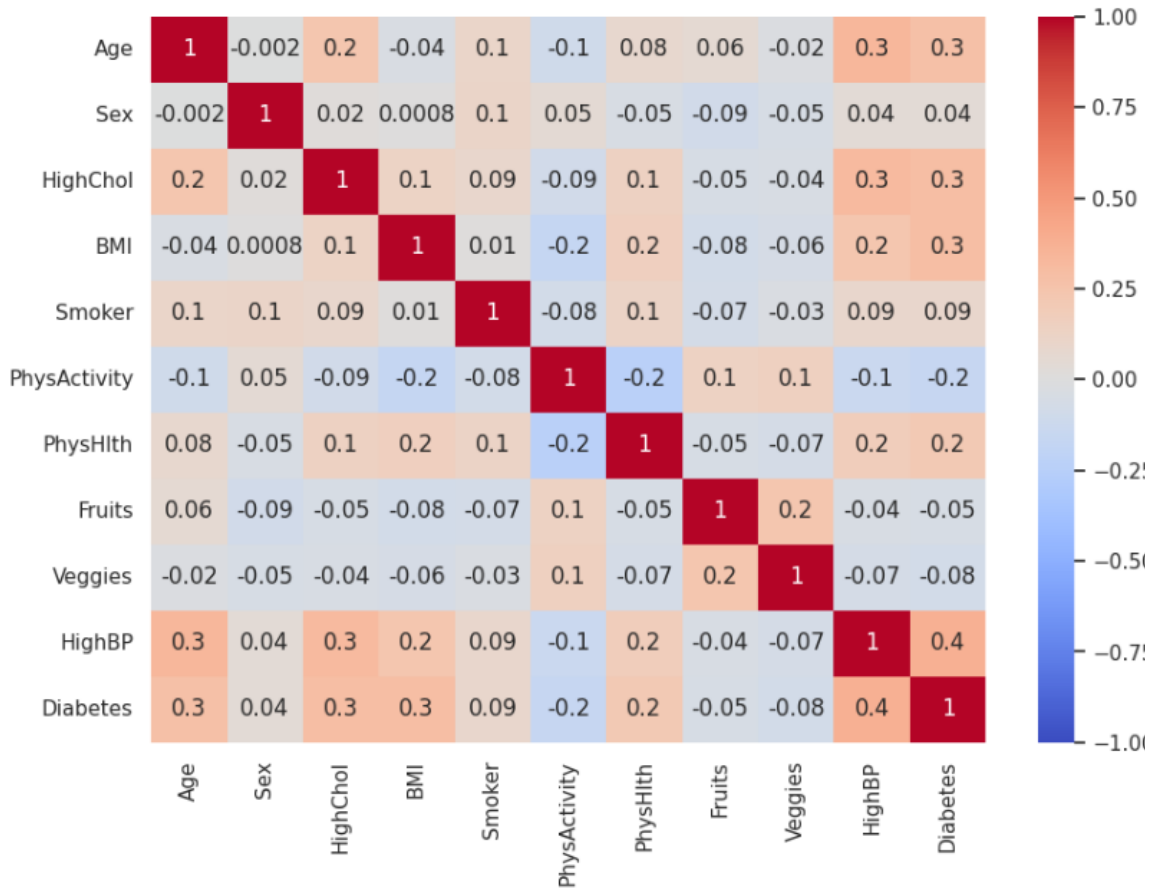
Namun tidak semua attribute digunakan dalam proses klasifikasi ini, gunakan operator selec attribute



Selanjutnya yaitu penetapan attribute diabetes sebagai label dengan menggunakan operator SetRole



Selanjutnya yaitu Pembagian dataset menjadi data training (80%) dan data testing (20%) dengan menggunakan operator split data



Melihat hubungan setiap attribute terhadap dengan attribute yang lainnya, disini nilai yang indikator semakin merah semakin kuat hubungan positifnya dan jika indikator semakin biru maka hubungan kuat hubungan negatifnya, dan jika nilainya mendekati 0 maka tidak memiliki hubungan, oleh karena itu hapus attribute yang tidak memiliki hubungan. Untuk mendapatkan accuracy yang lebih baik.

4. Modelling

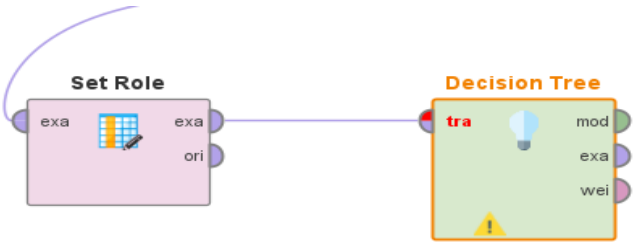
Algoritma: Decision Tree

Keterangan:

Dataset yang digunakan mengandung 2 tipe data yaitu kategorikal dan numerik, sehingga Decision Tree dianggap pilihan yang tepat. Decision tree merupakan metode klasifikasi yang menggunakan struktur pohon untuk memisahkan data berdasarkan atribut-atribut pada data pelatihan. Setiap node pada pohon keputusan membagi data menjadi dua atau lebih kelompok berdasarkan nilai atribut tertentu.

Decision Tree dipilih karena memiliki beberapa keunggulan yang sangat sesuai dengan karakteristik dataset ini, yaitu:

1. Mampu menangani atribut numerik dan kategorikal tanpa perlu melakukan normalisasi atau transformasi lanjutan.
2. Tidak sensitif terhadap skala data, sehingga tidak diperlukan preprocessing seperti scaling atau standardisasi.
3. Dapat menangani data dengan banyak variabel dan melakukan feature selection secara otomatis.
4. Mudah dipahami dan diinterpretasikan, karena hasilnya berupa struktur pohon logis yang menjelaskan proses pengambilan keputusan secara visual.




```
#Data splitting

y = (dm['Diabetes']).astype(int)
X = dm.loc[:, dm.columns != 'stroke']  # everything except "stroke"
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Split dataset menjadi 2, menjadi data testing dan data train,

```
X_train.shape
```

```
(49484, 7)
```

```
X_test.shape
```

```
(21208, 7)
```

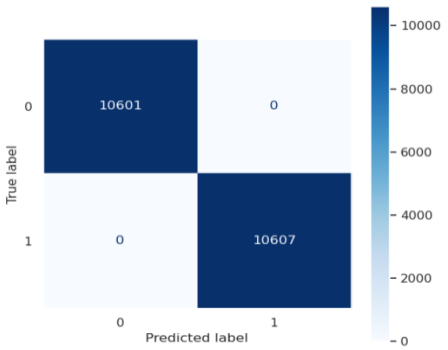
5. Evaluation

Deskripsi:

Evaluasi dilakukan untuk mengukur performa model Decision Tree dalam mengklasifikasikan status diabetes berdasarkan data uji. Dalam proyek ini, metode evaluasi yang digunakan adalah pembagian dataset menjadi dua bagian, yaitu 80% untuk data pelatihan (training set) dan 20% untuk data pengujian (testing set), menggunakan operator Split Data pada RapidMiner. Model dilatih menggunakan data latih dan kemudian diterapkan pada data uji untuk melihat sejauh mana kemampuan model dalam menggeneralisasi terhadap data baru.

Beberapa metrik evaluasi yang digunakan antara lain accuracy, precision, recall, dan confusion matrix. Accuracy digunakan untuk melihat seberapa besar proporsi prediksi yang benar dari keseluruhan data uji. Precision mengukur ketepatan model dalam memprediksi kelas positif (yaitu penderita diabetes), sedangkan recall mengukur sejauh mana model mampu menemukan seluruh kasus diabetes yang sebenarnya. Selain itu, confusion matrix digunakan untuk melihat performa model secara lebih rinci melalui jumlah true positive, true negative, false positive, dan false negative.

Secara umum, model Decision Tree menunjukkan performa yang baik dalam proses klasifikasi. Nilai-nilai evaluasi yang diperoleh menunjukkan bahwa model memiliki tingkat ketepatan dan sensitivitas yang seimbang, sehingga cocok digunakan untuk kasus klasifikasi kesehatan seperti deteksi dini diabetes. Hal ini penting karena dalam konteks medis, kesalahan prediksi—baik false positive maupun false negative—dapat berdampak pada keputusan penanganan pasien.



Hasil evaluasi model klasifikasi menggunakan algoritma Decision Tree ditunjukkan melalui confusion matrix yang menggambarkan performa prediksi terhadap dua kelas, yaitu kelas 0 (tidak diabetes) dan kelas 1 (diabetes). Berdasarkan confusion matrix tersebut, model berhasil mengklasifikasikan seluruh data uji dengan sangat baik, tanpa satu pun kesalahan prediksi. Tercatat sebanyak 10.601 data yang benar-benar tidak mengidap diabetes berhasil diklasifikasikan dengan tepat (true negative), dan sebanyak 10.607 data yang benar-benar mengidap diabetes juga berhasil diprediksi dengan benar (true positive). Tidak terdapat kasus false positive maupun false negative, yang berarti model tidak pernah salah dalam memprediksi status diabetes.

Decision Trees - Method 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

6. Deployment

Deskripsi:

Berdasarkan hasil evaluasi model klasifikasi menggunakan algoritma Decision Tree yang menunjukkan nilai sempurna pada seluruh metrik evaluasi, termasuk akurasi, precision, recall, dan F1-score yang semuanya bernilai 1.000000, maka model ini dinyatakan sangat layak untuk di-deploy. Tahap deployment dilakukan dengan tujuan mengimplementasikan model ke dalam sistem operasional atau lingkungan nyata, agar dapat dimanfaatkan langsung oleh pengguna untuk memprediksi status diabetes berdasarkan data kesehatan dan gaya hidup yang dimiliki.

Model yang telah dibangun dapat diintegrasikan ke dalam aplikasi berbasis web, desktop, atau mobile untuk mendukung skrining dini penyakit diabetes. Dengan memasukkan data seperti usia, indeks massa tubuh (BMI), tekanan darah, kebiasaan merokok, aktivitas fisik, dan riwayat penyakit, sistem dapat secara otomatis memberikan hasil prediksi kepada pengguna. Hasil prediksi ini dapat dimanfaatkan oleh tenaga medis, penyedia layanan kesehatan, atau bahkan masyarakat umum untuk mengambil keputusan dini terkait potensi risiko diabetes.

Namun, sebelum diimplementasikan secara luas, diperlukan pengujian tambahan pada data yang benar-benar baru (real-world data) untuk memastikan kemampuan generalisasi model. Selain itu, aspek non-teknis seperti keamanan data pengguna, privasi, serta antarmuka pengguna (user interface) juga perlu diperhatikan agar model ini tidak hanya kuat dari sisi akurasi, tetapi juga mudah digunakan dan aman secara operasional. Dengan demikian, deployment model ini diharapkan dapat menjadi kontribusi nyata dalam mendukung upaya deteksi dan pencegahan penyakit diabetes secara lebih cepat dan efisien.