



**Muhammad  
Zaki Fuadi**

:

**Nomor Urut**

:

**17**

## DAFTAR ISI

BUKTI 1-ADS.....	2
1.    Kebutuhan Data.....	2
2.    Pengambilan Data.....	3
3.    Pengintegrasian Data.....	3
BUKTI 2-ADS.....	5
1.    Analisis Tipe dan Relasi Data.....	5
2.    Analisis Karakteristik Data.....	6
3.    Laporan Telaah Data.....	6
BUKTI 3-ADS.....	7
1.    Pengecekan Kelengkapan Data.....	7
2.    Rekomendasi Kelengkapan DATA.....	7
BUKTI 4-ADS.....	9
1.    Kriteria dan Teknik Pemilihan Data.....	9
2.    Attributes (Columns) dan Records (Row) Data.....	9
BUKTI 5-ADS.....	11
1.    Pembersihan Data Kotor.....	11
2.    Laporan dan Rekomendasi Hasil Pembersihan Data Kotor.....	11
BUKTI 6-ADS.....	13
1.    Analisis Teknik Transformasi Data.....	13
2.    Transformasi Data.....	13
3.    Dokumentasi Konstruksi Data.....	14
BUKTI 7-ADS.....	15
1.    Pelabelan Data.....	15
2.    Laporan Hasil Pelabelan Data.....	15
BUKTI 8-ADS.....	17
1.    Parameter Model.....	17
2.    Tools Pemodelan.....	17
BUKTI 9-ADS.....	19
1.    Penggunaan Model dengan Data Riil.....	19

2.	Penilaian Hasil Pemodelan	19
----	---------------------------	----

## BUKTI 1-ADS

Kode Unit	:	J.62DM100.004.1
Judul Unit	:	Mengumpulkan Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengumpulkan data untuk data science.

### Langkah Kerja:

- 1) Menentukan kebutuhan data
- 2) Mengambil data
- 3) Mengintegrasikan data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengubah teks
  - Aplikasi basis data
  - Tools pengambilan data

### 1. KEBUTUHAN DATA

---

#### Instruksi Kerja:

- Identifikasi kebutuhan data sesuai tujuan teknis data science
- Periksa ketersediaan data berdasarkan kebutuhan data sesuai aturan yang berlaku
- Tentukan volume data berdasarkan kebutuhan data sesuai tujuan teknis data science

#### 1) Tujuan teknis data science

Melakukan prediksi berdasarkan pengukuran diagnostik apakah seorang pasien menderita diabetes atau tidak.

#### 2) Kebutuhan data

Dataset bersumber dari "[https://github.com/arubhasy/dataset/blob/main/diabetes\\_kotor.csv](https://github.com/arubhasy/dataset/blob/main/diabetes_kotor.csv)". Dataset berisi kriteria-kriteria seseorang mengalami diabetes agar dapat membangun model klasifikasi apakah seseorang mengalami diabetes apa tidak.

#### 3) Ketersediaan data

Dataset memiliki 9 feature dan 769 row dengan rician sebagai berikut :

- a. Pregnant : Berapa kali hamil
- b. Glucose : Konsetrasi glukosa
- c. BloodPressure : Tekanan darah

- d. SkinThines : Ketebalan kulit
- e. Insulin : Insulin
- f. BMI : Indeks massa tubuh
- g. DiabetesPedigree : Fungsi selisih diabetes
- h. Age : Usia pasien
- i. Outcome : Label seorang apakah memiliki penyakit diabetes

#### 4) Volume data

Dataset memiliki 9 feature dan 769 row dengan size 27 kb.

## 2. PENGAMBILAN DATA

---

### Instruksi Kerja:

- Identifikasi metode dan tools pengambilan data sesuai tujuan teknis data science
- Tentukan tools pengambilan data sesuai tujuan teknis data science
- Siapkan tools pengambilan data sesuai tujuan teknis data science
- Jalankan proses pengambilan data sesuai dengan tools yang telah disiapkan

#### 1) Metode dan tools pengambilan data

Metode pengambilan data menggunakan download data melalui link "<https://github.com/arubhasy/dataset/tree/main>" dan tools digunakan melalui rapidminer

#### 2) Penyiapan tools pengambilan data

Import data melalui rapiminer, kemudian mencari file untuk dijadikan dataset.

#### 3) Proses pengambilan data

Dataset berformat .csv dapat langsung diimport data melalui rapidminer

## 3. PENGINTEGRASIAN DATA

---

### Instruksi Kerja:

- Periksa integritas data sesuai tujuan teknis data sciene
- Integrasikan data sesuai tujuan teknis data science

#### 1) Pemeriksaan integritas data

Integrasi data dapat dilihat ketika dataset sudah diproses pada rapidminer

Name	Type	Missing	Statistics			Filter (10 / 10 attributes)	Search for Attributes	
Insulin	Integer	374	Min	14	Max	846	Average	155.548
		374	Min	7	Max	99	Average	29.153
SkinThickness	Integer	227	Min	24	Max	122	Average	72.405
BloodPressure	Integer	35	Min	21	Max	81	Average	33.237
Age	Integer	22	Min	20	Max	671	Average	294.008
BMI	Integer	11	Min	44	Max	199	Average	121.687
Glucose	Integer	5	Min	1	Max	768	Average	384.500
No	Integer	0	Min		Max		Average	
			Min		Max		Average	

Showing attributes 1 - 10

Examples: 768 Special Attributes: 0 Regular Attributes: 10

## 2) Pengintegrasian data

Integrasi data dengan melakukan replacing missing value, menghapus outlier data, menghapus duplikat data.

## BUKTI 2-ADS

Kode Unit	:	J.62DMI00.005.1
Judul Unit	:	Menelaah Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menelaah data untuk data science.

### Langkah Kerja:

- 1) Menganalisis tipe dan relasi data
- 2) Menganalisis karakteristik data
- 3) Membuat laporan telaah data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Tools pengolahan data
  - Tools pembuat grafik

## 1. ANALISIS TIPE DAN RELASI DATA

### Instruksi Kerja:

- Identifikasi tipe data yang terkumpul sesuai tujuan teknis
- Uraikan nilai atribut data yang terkumpul sesuai dengan batasan konteks bisnisnya
- Identifikasi relasi antar data yang terkumpul sesuai dengan tujuan teknis

### 1) Analisis tipe data

✓ Pregnancies	Real	0	Min 0	Max 1	Average 0.228
✓ Glucose	Real	0	Min 0	Max 1	Average 0.499
✓ BloodPressure	Real	0	Min 0	Max 1	Average 0.494
✓ SkinThickness	Real	0	Min 0	Max 1	Average 0.240
✓ Insulin	Real	0	Min 0	Max 1	Average 0.237
✓ BMI	Real	0	Min 0	Max 1	Average 0.420
✓ DiabetesPedigreeFunction	Real	0	Min 0	Max 1	Average 0.281

- Pregnant : Rasio
- Glucose : Interval
- BloodPressure : Interval

- SkinThines : Interval
- Insulin : Interval
- BMI : Interval
- DiabetesPedigree : Interval
- Age : Interval
- Outcome : Binominal

## 2) Analisis relasi data

Relasi data dapat dilihat dengan operator correlation matrix

Attribu...	Pregna...	Glucose	Blood...	SkinTh...	Insulin	BMI	Diabet...	Age	Outcome	No
Pregnan...	1	0.139	0.207	0.084	0.085	-0.003	0.007	0.540	-0.226	-0.046
Glucose	0.139	1	0.223	0.183	0.406	0.154	0.115	0.266	-0.495	0.021
BloodPr...	0.207	0.223	1	0.193	0.079	0.185	0.036	0.324	-0.175	0.024
SkinThic...	0.084	0.183	0.193	1	0.160	0.328	0.039	0.126	-0.211	0.025
Insulin	0.085	0.406	0.079	0.160	1	0.084	0.049	0.148	-0.233	0.023
BMI	-0.003	0.154	0.185	0.328	0.084	1	0.032	-0.005	-0.146	0.038
Diabete...	0.007	0.115	0.036	0.039	0.049	0.032	1	0.046	-0.197	-0.034
Age	0.540	0.266	0.324	0.126	0.148	-0.005	0.046	1	-0.237	0.004
Outcome	-0.226	-0.495	-0.175	-0.211	-0.233	-0.146	-0.197	-0.237	1	0.038
No	-0.046	0.021	0.024	0.025	0.023	0.038	-0.034	0.004	0.038	1

Terlihat feature No terhadap label, memiliki nilai 0,038 sehingga perlu difilter featurenya, beberapa feature memiliki rata-rata nilai terhadap label sekitar 0,2.

## 2. ANALISIS KARAKTERISTIK DATA

### Instruksi Kerja:

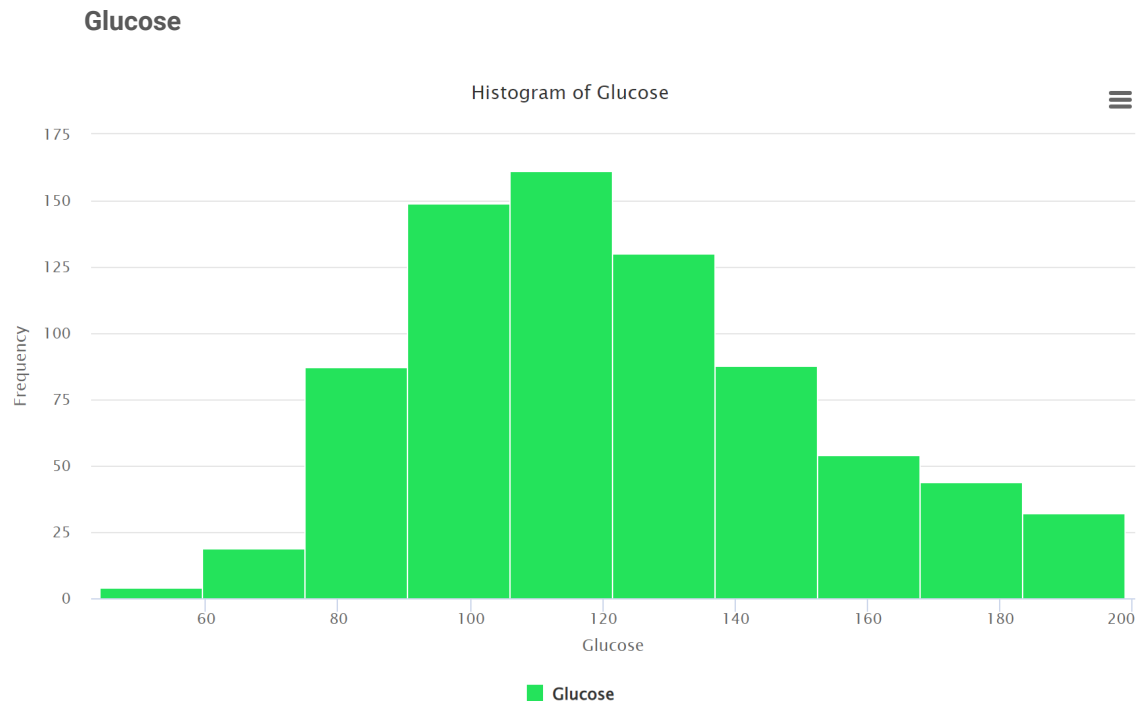
- Sajikan karakteristik data yang terkumpul dengan deskripsi statistik dasar
- Sajikan karakteristik data yang terkumpul dengan visualisasi grafik
- Analisis karakteristik data dari hasil penyajian data untuk telaah data

### 1) Analisis karakteristik data dengan deskripsi statistik dasar

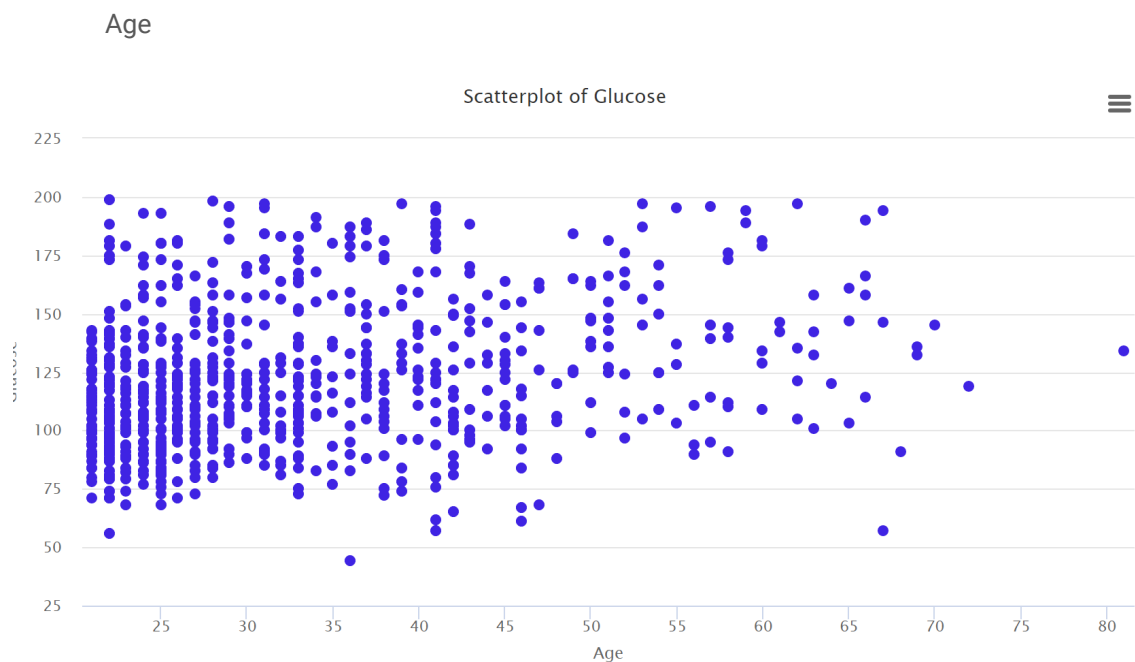
Feature	Min	Max	Average
Pregnancies	0	17	3,84
Glucose	44	199	121,6
BloodPressure	24	122	72,3
SkinThickness	7	99	29,1
Insulin	14	846	155,7
BMI	20	617	294
Age	1	2328	431,8
DiabetesPedigreeFunction	21	81	33,2



## 2) Analisis karakteristik data dengan visualisasi grafik



Pada visualisasi feature glucose memiliki distribusi normal



Pada visualisasi feature age memiliki outlier pada nilai 80, dilihat ada jarak yang jauh terhadap persebaran data yang lain.

### 3. LAPORAN TELAAH DATA


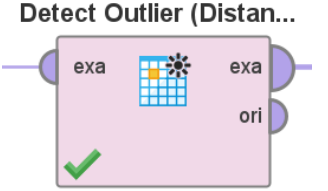
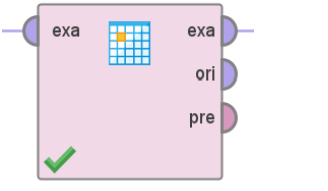
#### Instruksi Kerja:

- Dokumentasikan hasil analisis dalam bentuk laporan sesuai dengan tujuan teknis
- Susun hipotesis berdasar hasil analisis sesuai tujuan teknis data science

#### Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) menganalisis tipe dan relasi data; dan (2) menganalisis karakteristik data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat laporan telaah data; dapat diabaikan.

#### 1) Dokumentasi hasil analisis

Solusi	Operator
Pada bagian <b>Missing Value</b> dapat diganti dengan menggunakan data nilai rata-rata pada setiap feature sesuai dengan jenis data.	
Memfilter nilai outlier dengan menghapus nilai tersebut menggunakan operator <b>detect outlier</b>	
Feature yang masih belum berdistribusi normal seperti <b>Age, Diabetesdegree, BMI, Insulin</b> dilakukan dengan menormalisasi dengan rentang 0 sampai satu	

### 3) Hipotesis (bila ada)

Ada feature yang masih membuat model belum optimal sehingga perlu dilakukan pengecekan korelasi menggunakan correlation matrix

Attribu...	Pregna...	Glucose	Blood...	SkinTh...	Insulin	BMI	Diabet...	Age	Outcome
Pregnan...	1	0.139	0.207	0.084	0.085	-0.003	0.007	0.540	-0.226
Glucose	0.139	1	0.223	0.183	0.406	0.154	0.115	0.266	-0.495
BloodPr...	0.207	0.223	1	0.193	0.079	0.185	0.036	0.324	-0.175
SkinThic...	0.084	0.183	0.193	1	0.160	0.328	0.039	0.126	-0.211
Insulin	0.085	0.406	0.079	0.160	1	0.084	0.049	0.148	-0.233
BMI	-0.003	0.154	0.185	0.328	0.084	1	0.032	-0.005	-0.146
Diabete...	0.007	0.115	0.036	0.039	0.049	0.032	1	0.046	-0.197
Age	0.540	0.266	0.324	0.126	0.148	-0.005	0.046	1	-0.237
Outcome	-0.226	-0.495	-0.175	-0.211	-0.233	-0.146	-0.197	-0.237	1

Namun, ternyata beberapa fitur memiliki nilai yang serupa sehingga tidak ada feature yang didrop.

## BUKTI 3-ADS

Kode Unit	:	J.62DM100.006.1
Judul Unit	:	Memvalidasi Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memvalidasi data untuk data science.

### Langkah Kerja:

- 1) Melakukan pengecekan kelengkapan data
- 2) Membuat rekomendasi kelengkapan data

### Peralatan dan Perlengkapan:


- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengubah teks

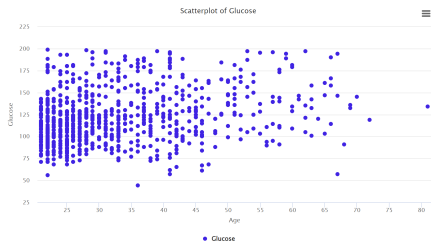
## 1. PENGECEKAN KELENGKAPAN DATA

### Instruksi Kerja:

- Sajikan penilaian kualitas data dari hasil telaah sesuai tujuan teknis data science
- Sajikan penilaian tingkat kecukupan data dari hasil telaah sesuai tujuan teknis data science

### 1) Penilaian kualitas data

Captured	Temuan Masalah
	Ditemukan <b>Missing Value</b> pada seluruh feature. Total <b>Missing Value</b> sebanyak 374 values di feature Insulin. Beberapa feature juga memiliki <b>Missing Value</b> .

<table><tr><td>8</td><td>125</td><td>96</td></tr><tr><td>8</td><td>125</td><td>96</td></tr><tr><td>8</td><td>125</td><td>96</td></tr></table>	8	125	96	8	125	96	8	125	96	Terdapat ada 10 baris <b>duplikat</b>
8	125	96								
8	125	96								
8	125	96								
	Terdapat <b>outlier</b> data pada beberapa feature									

## 2) Penilaian tingkat kecukupan data


Setelah ditelaah ada 374 data missing value dan ada data 10 duplikat baris dan ada beberapa nilai outlier di dataset, sehingga ini perlu dilakukan cleansing data agar membangun model yang optimal.

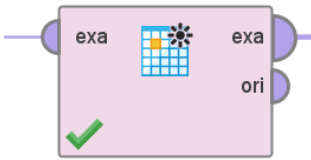
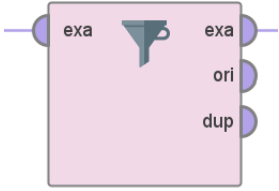
## 2. REKOMENDASI KELENGKAPAN DATA

### Instruksi Kerja:

- Susun rekomendasi hasil penilaian kualitas sesuai tujuan teknis data science
- Susun rekomendasi hasil penilaian kecukupan data sesuai tujuan teknis data science

### 1) Rekomendasi hasil penilaian kualitas data

Captured	Solusi
	Pada bagian <b>Missing Value</b> dapat diganti dengan menggunakan data nilai rata-rata pada setiap feature sesuai dengan jenis data

<p><b>Detect Outlier (Distan...</b></p> 	<p>Memfilter nilai outlier dengan menghapus nilai tersebut menggunakan operator <b>detect outlier</b></p>
<p><b>Remove Duplicates</b></p> 	<p>Ada beberapa baris yang memiliki nilai yang sama sehingga dilakukan <b>remove duplicate</b></p>

### 3) Rekomendasi hasil penilaian tingkat kecukupan data

Setelah dilakukan cleansing data, 758 row data, serta 9 feature. Hal ini cukup menjadi bahan untuk prediksi terhadap seseorang memiliki penyakit diabetes, sehingga nanti ketika proses split data, data training cukup untuk membangun model prediksi.

## BUKTI 4-ADS

<b>Kode Unit</b>	:	J.62DMI00.007.1
<b>Judul Unit</b>	:	Menentukan Objek Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilah dan memilih data yang sesuai permintaan atau kebutuhan.

### Langkah Kerja:

- 1) Memutuskan kriteria dan teknik pemilihan data
- 2) Menentukan attributes (columns) dan records (row) data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi spreadsheet
  - Aplikasi notepad plus
  - Aplikasi SQL (Structured Query Language)

## 1. KRITERIA DAN TEKNIK PEMILIHAN DATA

### Instruksi Kerja:

- Identifikasi kriteria pemilihan data sesuai dengan tujuan teknis dan aturan yang berlaku
- Tetapkan teknik pemilihan data sesuai dengan kriteria pemilihan data

### 1) Kriteria pemilihan data

Attribu...	Pregna...	Glucose	Blood...	SkinTh...	Insulin	BMI	Diabet...	Age	No	Outcome
Pregnan...	1	0.139	0.207	0.084	0.085	-0.003	0.007	0.540	-0.046	-0.226
Glucose	0.139	1	0.223	0.183	0.406	0.154	0.115	0.266	0.021	-0.495
BloodPr...	0.207	0.223	1	0.193	0.079	0.185	0.036	0.324	0.024	-0.175
SkinThic...	0.084	0.183	0.193	1	0.160	0.328	0.039	0.126	0.025	-0.211
Insulin	0.085	0.406	0.079	0.160	1	0.084	0.049	0.148	0.023	-0.233
BMI	-0.003	0.154	0.185	0.328	0.084	1	0.032	-0.005	0.038	-0.146
Diabete...	0.007	0.115	0.036	0.039	0.049	0.032	1	0.046	-0.034	-0.197
Age	0.540	0.266	0.324	0.126	0.148	-0.005	0.046	1	0.004	-0.237
No	-0.046	0.021	0.024	0.025	0.023	0.038	-0.034	0.004	1	0.038
Outcome	-0.226	-0.495	-0.175	-0.211	-0.233	-0.146	-0.197	-0.237	0.038	1

Terlihat pada feature "no" memiliki nilai korelasi terhadap nilai label sebesar 0,038, sehingga drop feature "no" tersebut. Namun beberapa memiliki korelasi negatif, artinya jika glukosa tinggi maka kemungkinan tidak memiliki penyakit diabetes

## 2) Teknik pemilihan data

Dipilih feature predictor yang tidak mendekati 0, sehingga feature yang digunakan yaitu : Pregnant, Glucose, Bloodpredcit, SkinThic, Insulin, BMI, DiabetesDiagree, dan Age

## 2. ATTRIBUTES (COLUMNS) DAN RECORDS (ROW) DATA

---

### Instruksi Kerja:

- Identifikasi attributes (columns) data sesuai dengan kriteria pemilihan data
- Identifikasi records (row) data sesuai dengan kriteria pemilihan data

### 1) Attributes (columns) data

Feature Predictor : Pregnant, Glucose, Bloodpredcit, SkinThic, Insulin, BMI, DiabetesDiagree, dan Age.

Feature Target : Outcome

### 3) Records (row) data

Row No.	Pregnancies	Glucose	BloodPres...	SkinThickn...	Insulin	BMI	DiabetesPe...	Age	Outcome
1	0.353	0.671	0.490	0.304	0.242	0.485	0.424	0.483	Diabetes
2	0.059	0.265	0.429	0.239	0.242	0.378	0.237	0.167	Normal
3	0.471	0.897	0.408	0.239	0.242	0.327	0.455	0.183	Diabetes
4	0.059	0.290	0.429	0.174	0.137	0.401	0.113	0	Normal
5	0.294	0.465	0.510	0.239	0.242	0.363	0.136	0.150	Normal
6	0.176	0.219	0.265	0.272	0.126	0.017	0.167	0.083	Diabetes
7	0.588	0.458	0.490	0.239	0.242	0.512	0.090	0.133	Normal
8	0.118	0.987	0.469	0.413	0.903	0.438	0.106	0.533	Diabetes
9	0.471	0.523	0.735	0.239	0.242	0.421	0.157	0.550	Diabetes
10	0.235	0.426	0.694	0.239	0.242	0.547	0.129	0.150	Normal
11	0.588	0.800	0.510	0.239	0.242	0.028	0.363	0.217	Diabetes

ExampleSet (758 examples, 0 special attributes, 9 regular attributes)

Memiliki 758 rows. records ini akan dibagi menjadi 2 melalui proses split data yaitu data training dan data testing dengan proporsi lebih banyak data training untuk membangun model dan data testing untuk menguji peforma model.



## BUKTI 5-ADS

<b>Kode Unit</b>	:	J.62DM100.008.1
<b>Judul Unit</b>	:	Membersihkan Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membersihkan data yang sesuai permintaan atau kebutuhan.

### Langkah Kerja:

- 1) Melakukan pembersihan data yang kotor
- 2) Membuat laporan dan rekomendasi hasil membersihkan data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi spreadsheet
  - Aplikasi text editor
  - Aplikasi SQL (Structured Query Language)

## 1. PEMBERSIHAN DATA KOTOR

### Instruksi Kerja:

- Tentukan strategi pembersihan data berdasarkan hasil telaah data
- Koreksi data yang kotor berdasarkan strategi pembersihan data

#### 1) Strategi pembersihan data

- Menyesuaikan penggantian type data
- Mengecek missing value serta mengganti nilai dengan rata-rata jika berdistribusi normal, atau dengan median jika nilai feature berdistribusi tidak normal, atau jika memiliki jumlah data missing value yang banyak maka bisa dilakukan drop feature tersebut.
- Mengecek duplikat baris, jika ada maka dilakukan proses duplikat agar baris tidak banyak dengan kriteria yang sama,
- Mengecek outlier, bisa dilihat dari boxplot atau scatterplot untuk melihat sebaran data, jika ada maka akan dilakukan handling outlier, bisa menggunakan metode IQR atau difiltering data.
- Mengecek data yang mungkin tidak konsisten, seperti pada jenis kelamin : (Wanita. Pria) namun ada data seperti (W, P, Laki-laki, Perempuan), ini perlu dilakukan keseragaman.
- Normalisasi data jika ada data yang tidak berdistribusi normal atau transformasi data jika rentang data itu terlalu jauh, sehingga mungkin saja memperlambat proses cleansing.

#### 2) Koreksi data kotor

Kesalahan	Koreksi
Masih ada missing value	Melakukan replacing missing value
Ada rentang nilai minimum dengan maksimum	Melakukan normalisasi
Ada feature yang tidak berkorelasi kuat dengan label target	Dilakukan drop feature
Ada beberapa baris yang memiliki nilai yang sama sehingga dilakukan <b>remove duplicate</b>	Dilakukan Remove duplikat

## 2. LAPORAN DAN REKOMENDASI HASIL PEMBERSIHAN DATA KOTOR

### Instruksi Kerja:

- Deskripsikan masalah dan teknis koreksi data sesuai dengan kondisi data dan strategi pembersihan data
- Lakukan evaluasi berdasarkan analisis koreksi yang telah dilakukan
- Dokumentasikan evaluasi proses dan hasil pembersihan data kotor

### 1) Masalah dan teknis koreksi data

Masalah teknis pada dataset ini yaitu feature-feature prediktor memiliki korelasi negatif ke label target, sehingga sulit untuk menentukan intepetasi data, seperti glukosa memiliki nilai tinggi maka seseorang itu memiliki penyakit diabetes, padahal glukosa adalah gula.

### 3) Evaluasi berdasarkan analisis koreksi yang telah dilakukan

Evaluasi karena pada model memiliki nilai akurasi sebesar 79,70%, sehingga dilakukan kembali correlation matrix untuk mengecek korelasi antar fitur target, dan melakukan drop salah satu feature namun hasilnya tidak meningkatkan nilai akurasi model.

#### 4) Dokumentasi evaluasi proses dan hasil pembersihan

Attributes	Pregna...	Glucose	Blood...	SkinTh...	Insulin	BMI	Diabet...	Outcome
Pregnancies	1	0.139	0.207	0.084	0.085	-0.003	0.007	-0.226
Glucose	0.139	1	0.223	0.183	0.406	0.154	0.115	-0.495
BloodPressure	0.207	0.223	1	0.193	0.079	0.185	0.036	-0.175
SkinThickness	0.084	0.183	0.193	1	0.160	0.328	0.039	-0.211
Insulin	0.085	0.406	0.079	0.160	1	0.084	0.049	-0.233
BMI	-0.003	0.154	0.185	0.328	0.084	1	0.032	-0.146
DiabetesPed...	0.007	0.115	0.036	0.039	0.049	0.032	1	-0.197
Outcome	-0.226	-0.495	-0.175	-0.211	-0.233	-0.146	-0.197	1

Disini melihat korelasi feature prediktor dengan feature target, setelah dilakukan drop feature pada salah satu feature yaitu feature "age". Namun tidak ada memenuhi asumsi, dengan meningkatnya akurasi model, namun ternyata sebaliknya, sehingga proses drop feature tidak dilakukan

Row No.	Pregnancies	Glucose	BloodPres...	SkinThickn...	Insulin	BMI	DiabetesPe...	Age	Outcome
1	0.353	0.671	0.490	0.304	0.242	0.485	0.424	0.483	Diabetes
2	0.059	0.265	0.429	0.239	0.242	0.378	0.237	0.167	Normal
3	0.471	0.897	0.408	0.239	0.242	0.327	0.455	0.183	Diabetes
4	0.059	0.290	0.429	0.174	0.137	0.401	0.113	0	Normal
5	0.294	0.465	0.510	0.239	0.242	0.363	0.136	0.150	Normal
6	0.176	0.219	0.265	0.272	0.126	0.017	0.167	0.083	Diabetes
7	0.588	0.458	0.490	0.239	0.242	0.512	0.090	0.133	Normal
8	0.118	0.987	0.469	0.413	0.903	0.438	0.106	0.533	Diabetes
9	0.471	0.523	0.735	0.239	0.242	0.421	0.157	0.550	Diabetes
10	0.235	0.426	0.694	0.239	0.242	0.547	0.129	0.150	Normal
11	0.588	0.800	0.510	0.239	0.242	0.028	0.363	0.217	Diabetes

ExampleSet (758 examples, 0 special attributes, 9 regular attributes)

## BUKTI 6-ADS

<b>Kode Unit</b>	:	J.62DM100.009.1
<b>Judul Unit</b>	:	Mengkonstruksi Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengkonstruksi data untuk proyek data science.

### Langkah Kerja:

- 1) Menganalisis teknik transformasi data
- 2) Melakukan transformasi data
- 3) Membuat dokumentasi konstruksi data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Tools pengolah kata

## 1. ANALISIS TEKNIK TRANSFORMASI DATA

---

### Instruksi Kerja:

- Lakukan analisis data untuk menentukan representasi fitur data awal
- Lakukan analisis representasi fitur data awal untuk menentukan teknik rekayasa fitur yang diperlukan untuk pembangunan model data science

#### 1) Representasi fitur data awal

Dilihat dari correlation matrix, sehingga ditentukan featurenya seperti Pregnant, Glucose, Bloodpredcit, SkinThic, Insulin, BMI, DiabetesDiagree, dan Age.

#### 2) Teknik rekayasa fitur data

Identifikasi fitur yang memiliki korelasi terhadap feature label, pemilihan fitur yang relevan, mengidentifikasi ketergantungan linear.

## 2. TRANSFORMASI DATA

---

### Instruksi Kerja:

- Lakukan transformasi untuk mendapatkan fitur data awal
- Lakukan rekayasa fitur data untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science

## 1) Transformasi data

Tidak ada tranformasi data, namun melakukan normalisasi data.

## 2) Rekayasa fitur data data

Normalisasi dengan method range normalization degan range 0 sampai satu

## 3. DOKUMENTASI KONSTRUKSI DATA

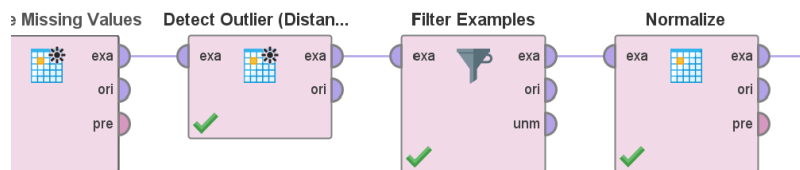
### Instruksi Kerja:

- Jabarkan teknis transformasi data dalam bentuk tertulis
- Tuangkan hasil transformasi data dan rekomendasi hasil transformasi dalam bentuk tertulis

### Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) mengalisis teknik transformasi data; dan (2) melakukan transformasi data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat dokumentasi konstruksi data; dapat diabaikan.

## 1) Dokumentasi teknis transformasi data



Berikut pipeline proyek, dimana disana melakukan normalisasi dengan hasil membuat rentang nilai numerikal mempunyai rentang dari 0 sampai 1.

## 2) Dokumentasi hasil dan rekomendasi transformasi data

Pregnancies	Glucose	BloodPres...	SkinThickn...	Insulin	BMI	DiabetesPe...	Age
0.353	0.671	0.490	0.304	0.242	0.485	0.424	0.483
0.059	0.265	0.429	0.239	0.242	0.378	0.237	0.167
0.471	0.897	0.408	0.239	0.242	0.327	0.455	0.183
0.059	0.290	0.429	0.174	0.137	0.401	0.113	0
0.294	0.465	0.510	0.239	0.242	0.363	0.136	0.150
0.176	0.219	0.265	0.272	0.126	0.017	0.167	0.083
0.588	0.458	0.490	0.239	0.242	0.512	0.090	0.133
0.118	0.987	0.469	0.413	0.903	0.438	0.106	0.533
0.471	0.523	0.735	0.239	0.242	0.421	0.157	0.550
0.235	0.426	0.694	0.239	0.242	0.547	0.129	0.150
0.588	0.800	0.510	0.239	0.242	0.028	0.363	0.217
0.588	0.613	0.571	0.239	0.242	0.386	0.976	0.600
0.294	0.787	0.490	0.130	0.275	0.366	0.397	0.500
0.412	0.361	0.490	0.239	0.242	0.015	0.327	0.183

Rekomendasi normalisasi menggunakan normalisasi menggunakan metode range tranformation dengan range 0 sampai 1.

## BUKTI 7-ADS

<b>Kode Unit</b>	:	J.62DM100.010.1
<b>Judul Unit</b>	:	Menentukan Label Data

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menentukan label data untuk pembangunan model data science.

### Langkah Kerja:

- 1) Melakukan pelabelan data
- 2) Membuat laporan hasil pelabelan data

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi pelabelan data

## 1. PELABELAN DATA

---

### Instruksi Kerja:

- Uraikan kesesuaian antara analisis hasil pelabelan data sejenis yang sudah ada dengan Standard Operating Procedure (SOP) pelabelan
- Lakukan pelabelan data sesuai dengan SOP pelabelan

#### 1) SOP pelabelan data (bila ada)

SOP pelabelan data pada dataset ini dilakukan setelah proses set role, pelabelan ini dilakukan untuk membagi data training dan data testing. Data training menjadi bahan model untuk belajar dan data testing menjadi bahan untuk menakar peforma model yang dibangun.

## 2) Proses pelabelan data

Proses pelabelan data dengan memilih feature target berdasarkan rumusan masalah yang sudah ditentukan kemudian, dibagi menjadi proporsi 80:20, 90:10, 60:40 serta 70:30. Tahap ini dapat mempengaruhi dalam membangun model. Jadi, outputnya menjadi 2 yaitu data training dan data testing, sesuai dengan proporsi yang ditentukan. Data training untuk melatih model serta data testing untuk melakukan pengujian performa model.

## 2. LAPORAN HASIL PELABELAN DATA

---

### Instruksi Kerja:

- Uraikan statistik hasil pelabelan pada laporan
- Uraikan evaluasi proses pelabelan pada laporan

### 1) Statistik hasil pelabelan

Model	Proporsi : Akurasi model
Random Forest	60 : 40 = 82,60% 70:30 = 80,89% 80:20 = 79,54%
Decision Tree	60:40 = 76,43% 70:30 = 80,79% 80:20 = 77,72%

### 3) Evaluasi proses pelabelan

Evaluasi proses pelabelan, semakin sedikit proporsi data training, maka akurasi model semakin meningkat. Hal ini perlu dikaji, padahal semakin banyak data, semakin banyak belajar, sehingga perlu dicroscheck kembali, proses pemodelan. Namun pada case ini memilih proporsi data 80:20, namun dilakukan analisa lebih lanjut, seperti melakukan hyperparameter tuning, mungkin saja parameter model belum optimal.

## BUKTI 8-ADS

<b>Kode Unit</b>	:	J.62DMI00.013.1
<b>Judul Unit</b>	:	Membangun Model

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membangun model.

### Langkah Kerja:

- 1) Menyiapkan parameter model
- 2) Menggunakan tools pemodelan

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer dan peralatannya
  - Perangkat lunak data science di antaranya: rapid miner, weka, atau development untuk bahasa pemrograman tertentu seperti Python atau R.
- Perlengkapan
  - Dokumen best practices kriteria dan evaluasi penilaian

## 1. PARAMETER MODEL

### Instruksi Kerja:

- Identifikasi parameter-parameter yang sesuai dengan model
- Tetapkan nilai toleransi parameter evaluasi pengujian sesuai dengan tujuan teknis

#### 1) Parameter-parameter model

Model	Parameter
Random Forest	<b>number of tree</b> : Jumlah pohon keputusan yang akan digunakan <b>criterion</b> : mengukur kualitas split pada setiap node dalam pohon keputusan <b>max_depth</b> : Kedalaman maksimum dari setiap pohon keputusan <b>min_samples_split</b> : Jumlah minimum sampel <b>min_samples_leaf</b> : Jumlah minimum sampel yang diperlukan di setiap leaf node <b>class_weight</b> : menyeimbangkan kelas target yang tidak seimbang dengan memberikan bobot yang berbeda



	pada kelas-kelas tertentu
Decision Tree	<b>Criteria</b> : mengukur kualitas split <b>Max_depth</b> : kedalaman maksimum pohon keputusan. <b>min samples split</b> : menentukan jumlah minimal sampel membagu node menjadi 2 <b>min sample leaf</b> : menentukan jumlah mimumum sampel berbentuk daun <b>max feature</b> : menentukan jumlah fitur maksimum yang dipertimbangkan saat mencari split terbaik.

## 2) Nilai toleransi parameter evaluasi pengujian

Matriks evaluasi pada masing-masing model dapat dilihat pada confusion matriks seperti,

- akurasi : proporsi prediksi benar dari keseluruhan prediksi
- precision : proporsi prediksi positif yang benar dari kesesluruhan positif
- recall : proporsi prediksi aktual yang diprediksi secara posiitif
- f1 score : rata-rata dari precision dan recall

## 2. TOOLS PEMODELAN

---

### Instruksi Kerja:

- Identifikasi tools untuk membuat model sesuai dengan tujuan teknis data science
- Bangun algoritma untuk teknik pemodelan yang ditentukan menggunakan tools yang dipilih
- Eksekusi algoritma pemodelan sesuai dengan skenario pengujian dan tools untuk membuat model yang telah ditetapkan
- Optimasi parameter model algoritma untuk menghasilkan nilai parameter evaluasi yang sesuai dengan skenario pengujian

### 1) Tools untuk membuat model

Tools untuk membuat model, pada case ini yaitu menggunakan decision tree dan random forest, masing model-model tersebut digunakan melalui operator "Apply model" dengan bahan testing dari hasil operator "Split data"

### 2) Algoritma untuk teknik pemodelan

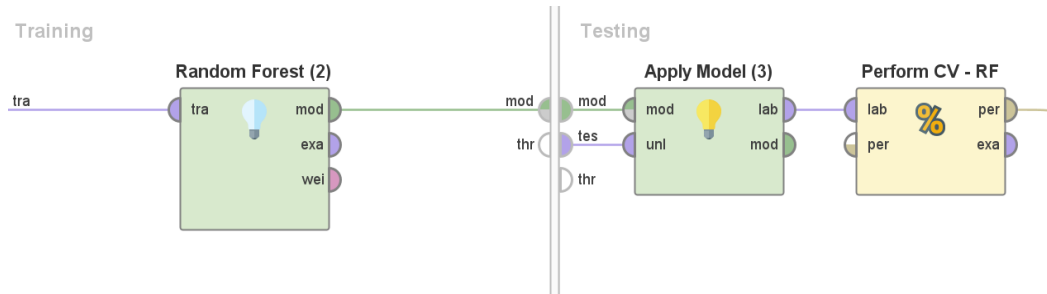
Algoritma yang digunakan pada case ini perbandingan antara decision tree dan random forest.

### 3) Eksekusi algoritma pemodelan

Eksekusi algoritma pemodelan melalui operator “Apply model” untuk dilakukan pengujian peforma model menggunakan data testing, dengan hasil akurasi prediksi dalam menebak label target.

### 4) Optimasi parameter model algoritma

Optimasi parameter model algoritma, mengasumsikan metode “cross validation” untuk meningkatkan akurasi model, namun ketika uji coba, hasil outputnya menurunkan peforma model, sehingga cross validation hanya menjadi uji coba saja. Selain itu, dilakukan untuk menentukan parameter-parameter model untuk mendapatkan peforma terbaik. Salah satu cara yaitu melakukan hyperparameter tuning untuk mendapatk parameter-parameter model yang optimal. Namun, pada rapidminer dapat dilakukan secara manual, berbeda ketika digunakan membangun menggunakan python.



## BUKTI 9-ADS

<b>Kode Unit</b>	:	J.62DM100.014.1
<b>Judul Unit</b>	:	Mengevaluasi Hasil Pemodelan

### Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengevaluasi hasil pemodelan.

### Langkah Kerja:

- 1) Menggunakan model dengan data riil
- 2) Menilai hasil pemodelan

### Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Tools untuk mengeksekusi model
  - Tools untuk pengumpulan data riil

## 1. PENGGUNAAN MODEL DENGAN DATA RIIL

---

### Instruksi Kerja:

- Kumpulkan data baru untuk evaluasi pemodelan sesuai kebutuhan yang mengacu kepada parameter evaluasi
- Uji model dengan menggunakan data riil yang telah dikumpulkan

#### 1) Data untuk evaluasi pemodelan

Data untuk evaluasi pemodelan yaitu menggunakan data testing, hasil dari proses operator "Split data", sehingga data tersebut dapat mengetahui hasil dari prediksi.

#### 2) Pengujian model

Pengujian model dilakukan menggunakan operator "performance" khusus pada klasifikasi. Hasil tersebut mengeluarkan output berupa confusion matrix. Confusion matrix dapat mengukur akurasi serta presisi prediksi hasil dari model.

## 2. PENILAIAN HASIL PEMODELAN

---

### Instruksi Kerja:

- Nilai keluaran pengujian model berdasarkan metrik kesuksesan
- Dokumentasikan hasil penilaian sesuai standar yang berlaku

### 1) Penilaian hasil/keluaran pengujian model berdasarkan metrik kesuksesan

Dari model paling optimal pada kasus ini menggunakan model random forest yang memiliki akurasi sebesar 82,80% dengan ketentuan, precision positif : 96% precision negatif : 79%, recall positif : 51,90%, serta recall negatif: 98,99%.

- Precision diabetes : semua data yang diprediksi sebagai diabetes oleh model 96%, yaitu benar-benar diabetes
- Precision normal : semua data yang diprediksi sebagai normal oleh model 79%, yaitu benar-benar normal
- recall diabetes : model berhasil menemukan 51,90% dari semua data diabetes sebenarnya
- Recall normal : model berhasil menemukan 98,99% dari semua data normal sebenarnya.

Dari informasi sebelumnya, bahwa model cenderung lebih akurat memprediksi kelas normal, tetapi memiliki kekurangan dalam prediksi diabetes.

### 3) Dokumentasi hasil penilaian

accuracy: 82.60%

	true Diabetes	true Normal	class precision
pred. Diabetes	82	3	96.47%
pred. Normal	76	293	79.40%
class recall	51.90%	98.99%	

Pada confusion matrix, bahwa model prediksi diabetes dengan data aktual diabetes sebesar 82 buah, model menebak benar pada prediksi normal dengan data aktual normal sebesar 293 buah, selanjutnya, model menebak salah pada data aktual normal namun prediksi model diabetes sebesar 3 buah, selain itu model menebak salah pada aktual diabetes namun prediksi model normal sebesar 76 buah.

**Kesimpulannya** bahwa model yang dibangun memiliki kinerja yang baik dalam memprediksi kelas Normal, namun perlu lebih ditingkatkan dalam mendeteksi dan memprediksi kelas Diabetes dengan lebih baik. Beberapa analisa lebih lanjut akan dilakukan seperti meningkatkan jumlah data, menggunakan model lebih kompleks seperti gradient boosting atau yang lainnya, melakukan hyperparameter tuning untuk menentukan parameter-parameter model yang membuat model lebih optimal.