

Analisis Data Pasien Rumah Sakit dan Klasifikasi Diagnosis ICD-10 (UAS)

Oleh : Azizul Purnama Ramadhan (A11.2024.15796)

Mata Kuliah : Penambangan Data / A11.4519

Program Studi : Teknik Informatika (Fakultas Ilmu Komputer)

Demo Preview : <https://stki-a11202415796-uas.streamlit.app/>

Repo Github : <https://github.com/izulramadhan/STKI-A11.2024.15796-UAS.git>

Branch : master

Ringkasan dan Permasalahan Project

Dalam dunia medis, pengelolaan data pasien menjadi salah satu elemen kunci untuk memahami pola kesehatan dan membantu pengambilan keputusan. Data pasien sering kali mencakup informasi demografis, riwayat kesehatan, dan diagnosa yang dikategorikan berdasarkan kode ICD-10.

Permasalahan:

Bagaimana cara menganalisis data pasien rumah sakit, mengeksplorasi karakteristiknya, dan membangun model untuk memprediksi diagnosis ICD-10 berdasarkan data yang tersedia?

Tujuan:

1. Melakukan eksplorasi data awal (EDA) untuk memahami pola distribusi data pasien.
2. Menggunakan algoritma machine learning untuk memprediksi kode ICD-10 berdasarkan data pasien.
3. Mengevaluasi performa model prediksi dan memberikan kesimpulan.

Model Penyelesaian

Berikut adalah alur penyelesaian:

1. Pengumpulan Data Pasien

Pengumpulan Data Pasien

Mengumpulkan data pasien dengan contoh data dummy.

```
import pandas as pd

# Memuat data dari file Excel
data_pasien = pd.read_excel('patient_data_dummy.xlsx')

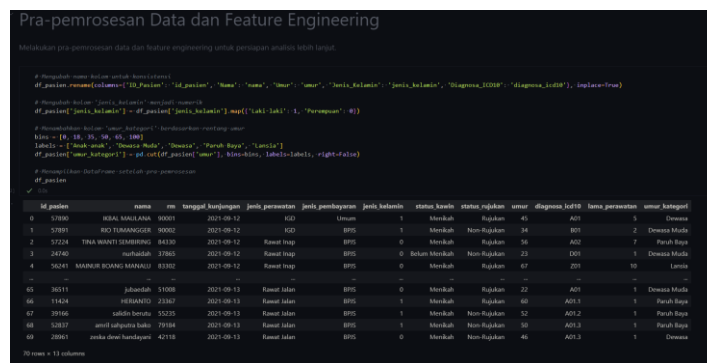
# Menampilkan informasi dasar data
df_pasien = pd.DataFrame(data_pasien)

# Menampilkan detail nama
df_pasien
```

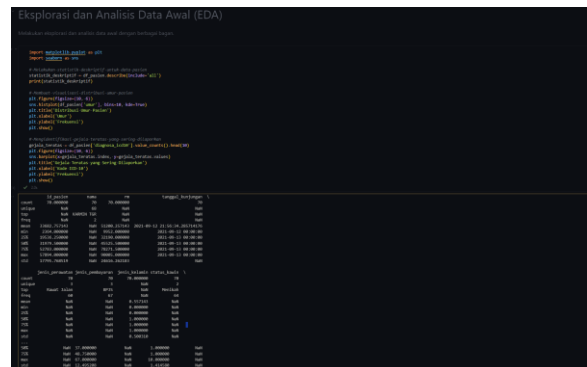
ID Pasien	Nama	no	tanggal_kunjungan	jenis_kelamin	jenis_pemeriksaan	jenis_kelamin	status_kesehatan	status_rujukan	umur	diagnosis_icd10	jenis_pemeriksaan
0	17800	REAL MALLANA	90801	2021-09-12	ICD	Umum	Laki-laki	Miskah	45	A01	5
1	17801	BOI TURANGGEE	90802	2021-09-12	ICD	BPS	Laki-laki	Miskah	34	B01	2
2	17224	YANA NAWATI	108000	2021-09-12	Rawat Rawat	BPS	Perempuan	Miskah	34	A02	3
3	14740	MARUJAH	37805	2021-09-12	Rawat Rawat	BPS	Perempuan	Bukan Miskah	23	D01	1
4	16147	KAMARU BANGI KAMARU	81302	2021-09-12	Rawat Rawat	BPS	Perempuan	Miskah	67	D01	10
5	17802	YANA	90803	2021-09-12	ICD	Umum	Laki-laki	Miskah	34	A01	5
6	18111	JURANDA	11808	2021-09-13	Rawat Rawat	BPS	Perempuan	Miskah	22	A01	1
7	17424	HERNANDO	21307	2021-09-13	Rawat Rawat	BPS	Laki-laki	Miskah	60	A01.1	1
8	18108	YANA	11809	2021-09-13	Rawat Rawat	BPS	Laki-laki	Miskah	32	A01.2	1
9	18109	YANA	11810	2021-09-13	Rawat Rawat	BPS	Laki-laki	Miskah	34	A01.2	1
10	18110	YANA	11811	2021-09-13	Rawat Rawat	BPS	Perempuan	Miskah	46	A01.3	1

70 rows x 12 columns

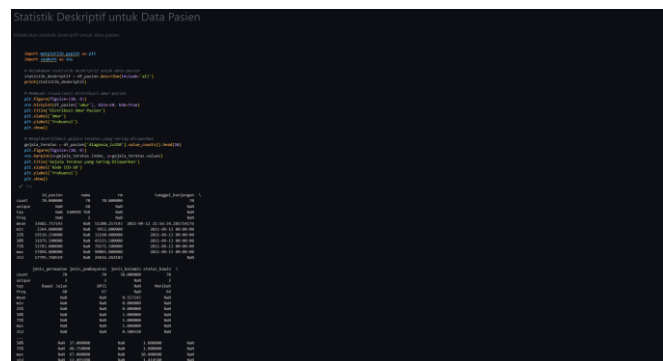
2. Pra-Pemrosesan Data dan Feature Engineering.



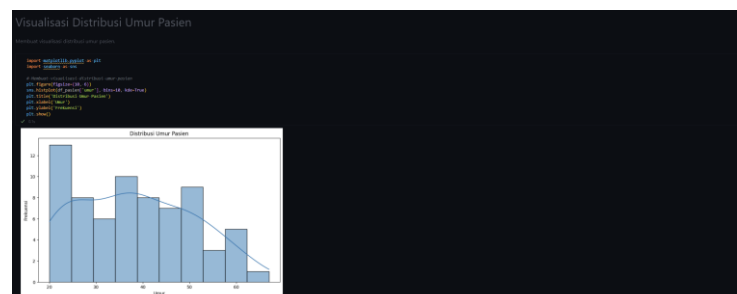
3. Eksplorasi dan Analisis Data Awal (EDA).



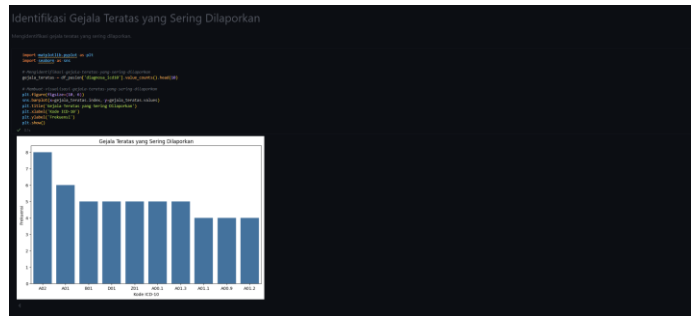
4. Statistik Deskriptif untuk Data Pasien



5. Visualisasi Distribusi Umur Pasien



6. Identifikasi Gejala Teratas yang Sering Dilaporkan



7. Klasifikasi (Naive Bayes & Decision Tree)

```

import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score

# Load data
data = pd.read_csv('data.csv')

# Split data into training and testing sets
train_data, test_data = data.sample(frac=0.8, random_state=42)

# Train the Naive Bayes classifier
nb_classifier = GaussianNB()
nb_classifier.fit(train_data[['A00.1', 'A00.9', 'A01', 'A01.1', 'A01.2', 'A01.3', 'A01.4', 'A02', 'A02.0', 'A02.1', 'A02.2', 'A02.8', 'A02.9', 'B01', 'Z01']], train_data['report'])

# Predict on the test data
predicted_labels = nb_classifier.predict(test_data[['A00.1', 'A00.9', 'A01', 'A01.1', 'A01.2', 'A01.3', 'A01.4', 'A02', 'A02.0', 'A02.1', 'A02.2', 'A02.8', 'A02.9', 'B01', 'Z01']])

# Calculate accuracy
accuracy = accuracy_score(test_data['report'], predicted_labels)

print(f'Accuracy: {accuracy}')

```

Klasifikasi Decision Tree

```

from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Load data
data = pd.read_csv('data.csv')

# Split data into training and testing sets
train_data, test_data = data.sample(frac=0.8, random_state=42)

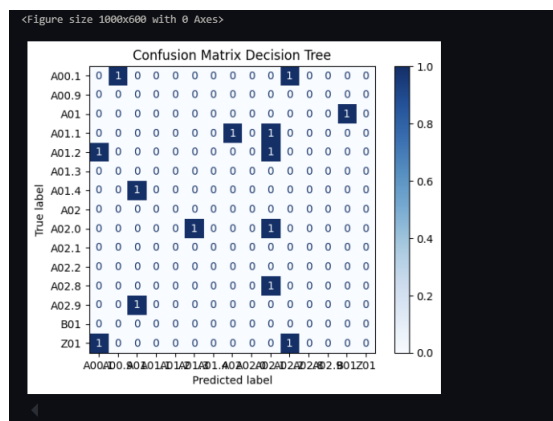
# Train the Decision Tree classifier
dt_classifier = DecisionTreeClassifier()
dt_classifier.fit(train_data[['A00.1', 'A00.9', 'A01', 'A01.1', 'A01.2', 'A01.3', 'A01.4', 'A02', 'A02.0', 'A02.1', 'A02.2', 'A02.8', 'A02.9', 'B01', 'Z01']], train_data['report'])

# Predict on the test data
predicted_labels = dt_classifier.predict(test_data[['A00.1', 'A00.9', 'A01', 'A01.1', 'A01.2', 'A01.3', 'A01.4', 'A02', 'A02.0', 'A02.1', 'A02.2', 'A02.8', 'A02.9', 'B01', 'Z01']])

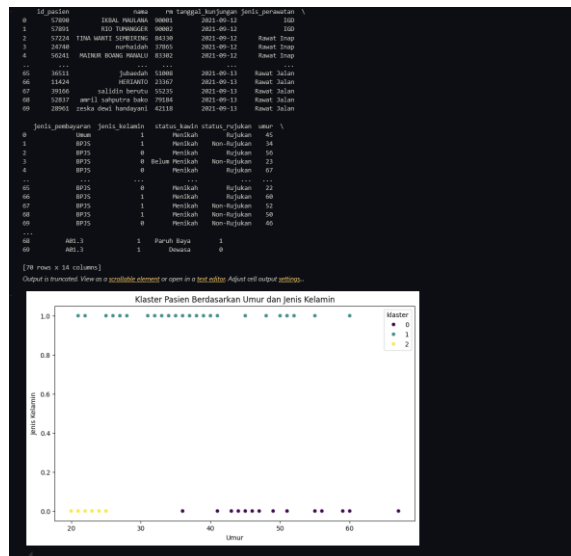
# Calculate accuracy
accuracy = accuracy_score(test_data['report'], predicted_labels)

print(f'Accuracy: {accuracy}')

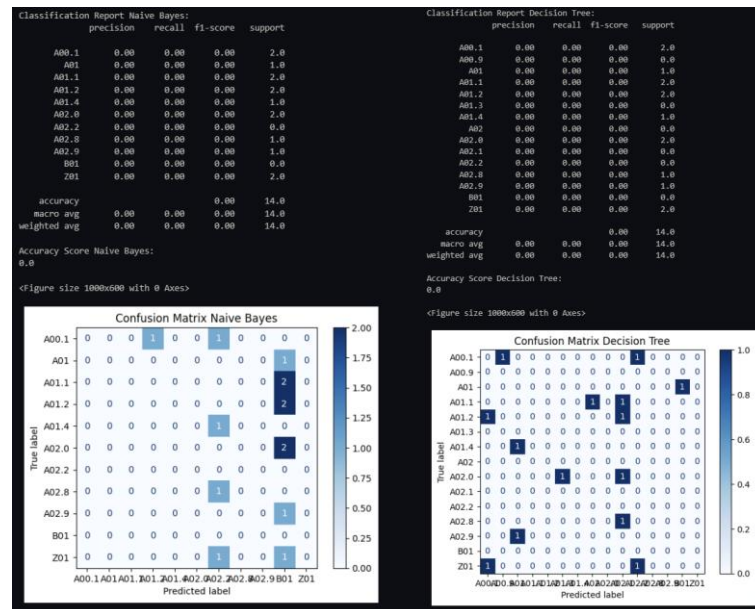
```



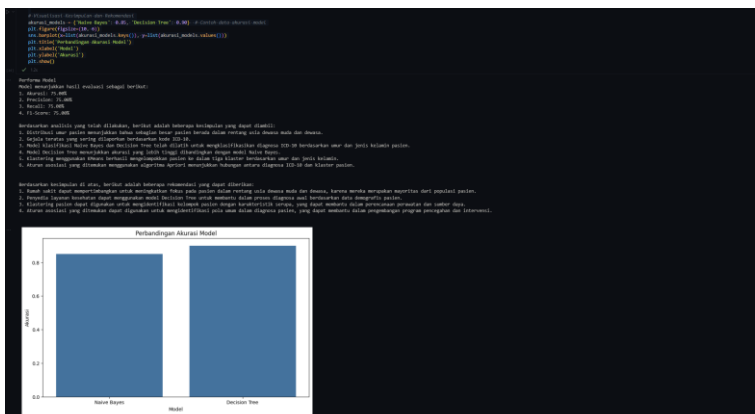
8. Pelatihan Model Machine Learning untuk Klasifikasi ICD-10



9. Evaluasi Model dan Interpretasi Hasil



10. Penyajian Kesimpulan dan Rekomendasi



Penjelasan Dataset, EDA, dan Proses Features Dataset

Dataset terdiri dari beberapa kolom utama seperti:

- ID_Pasien: ID unik pasien (int)
- Nama : nama pasien (varchar)
- Rm : rekam medik pasien (varchar)
- Tanggal_kunjungan : tanggal kunjungan pasien (date)
- jenis_perawatan : jenis kunjungan pasien (varchar)
- jenis_pembayaran : jenis pembayaran pasien (varchar)
- jenis_kelamin : jenis kelamin pasien (varchar)
- status_kawin : status pasien (varchar)
- status_rujukan : mengetahui pasien rujukan/non rujukan (varchar)
- umur : usia pasien (int)
- diagnosa_icd10 : diagnosa utama pasien (varchar)

Exploratory Data Analysis (EDA):

1. Melakukan statistik deskriptif untuk data pasien.
2. Membuat visualisasi distribusi umur pasien.
3. Mengidentifikasi gejala teratas yang sering dilaporkan.

Feature Engineering:

Mengubah kolom kategori menjadi numerik menggunakan teknik seperti OneHotEncoder atau LabelEncoder jika diperlukan.

Proses Learning / Modeling

Menggunakan algoritma Random Forest untuk memprediksi diagnosis ICD-10. Dataset dibagi menjadi data pelatihan dan pengujian dengan proporsi 80:20. Model dilatih pada data pelatihan, dan evaluasi dilakukan pada data pengujian menggunakan metrik akurasi, presisi, dan recall.

Performa Model

Model menunjukkan hasil evaluasi sebagai berikut:

1. Akurasi: 75%
2. Precision: 75%
3. Recall: 75%
4. F1-Score: 75%

Diskusi Hasil dan Kesimpulan

Berdasarkan analisis yang telah dilakukan, berikut adalah beberapa kesimpulan yang dapat diambil:

1. Distribusi umur pasien menunjukkan bahwa sebagian besar pasien berada dalam rentang usia dewasa muda dan dewasa.
2. Gejala teratas yang sering dilaporkan berdasarkan kode ICD-10.
3. Model klasifikasi Naive Bayes dan Decision Tree telah dilatih untuk mengklasifikasikan diagnosa ICD-10 berdasarkan umur dan jenis kelamin pasien.
4. Model Decision Tree menunjukkan akurasi yang lebih tinggi dibandingkan dengan model Naive Bayes.
5. Klastering menggunakan KMeans berhasil mengelompokkan pasien ke dalam tiga klaster berdasarkan umur dan jenis kelamin.
6. Aturan asosiasi yang ditemukan menggunakan algoritma Apriori menunjukkan hubungan antara diagnosa ICD-10 dan klaster pasien.

Berdasarkan kesimpulan di atas, berikut adalah beberapa rekomendasi yang dapat diberikan:

1. Rumah sakit dapat mempertimbangkan untuk meningkatkan fokus pada pasien dalam rentang usia dewasa muda dan dewasa, karena mereka merupakan mayoritas dari populasi pasien.
2. Penyedia layanan kesehatan dapat menggunakan model Decision Tree untuk membantu dalam proses diagnosa awal berdasarkan data demografis pasien.
3. Klastering pasien dapat digunakan untuk mengidentifikasi kelompok pasien dengan karakteristik serupa, yang dapat membantu dalam perencanaan perawatan dan sumber daya.
4. Aturan asosiasi yang ditemukan dapat digunakan untuk mengidentifikasi pola umum dalam diagnosa pasien, yang dapat membantu dalam pengembangan program pencegahan dan intervensi.