LAPORAN PROYEK DATA MINING

Case Sample BPJS 2015 – 2021

Using K-Nearest Neighbor (KNN)

Oleh:

12S20001	**Marcelino Manalu**
12S20026	**Mastuari Octafina Sirumapea**
12S20038	**Arni Febryarti Sitorus**

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

2023/2024

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

Bab 1

Business Understanding

Business understanding bertujuan untuk menjawab permasalahan bisnis mengenai tugas analitis yang perlu diselesaikan. Fokus permasalahan mencakup klasifikasi pasien ke dalam kelompok tertentu berdasarkan pola kunjungan yang sering atau jarang pasien menggunakan kartu BPJS pada saat berkunjung ke Rumah Sakit. Tujuannya adalah menghasilkan sebuah model data mining yang dapat meningkatkan perencanaan sumber daya dan kapasitas fasilitas kesehatan, mengoptimalkan alokasi tenaga medis dan non-medis berdasarkan kebutuhan aktual pasien dan merancang program pelayanan yang lebih terarah dan sesuai dengan kebutuhan populasi pasien.

1.1 Determine Business Objective

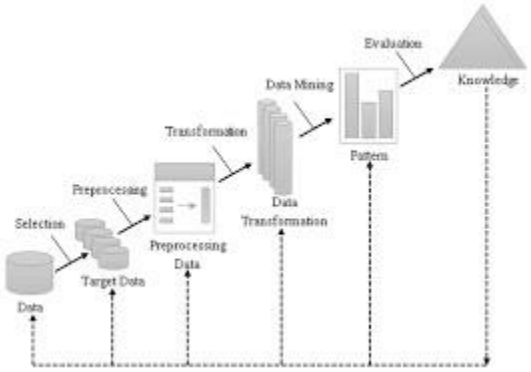
Sistem pelayanan kesehatan merupakan aspek krusial dalam memberikan perawatan yang berkualitas kepada masyarakat. Dalam dinamika layanan kesehatan saat ini, pemahaman mendalam terhadap pola pelayanan pasien menjadi esensial. Organisasi kesehatan, termasuk BPJS, berhadapan dengan tantangan mengelola sumber daya secara efisien, meningkatkan pengalaman pasien, dan memastikan bahwa pelayanan kesehatan sesuai dengan kebutuhan populasi.

Penyedia layanan kesehatan perlu memiliki pemahaman yang lebih baik tentang pola kunjungan pasien, jenis layanan yang paling sering dibutuhkan, dan faktor-faktor yang mempengaruhi keputusan pasien. Oleh karena itu, analisis mendalam terhadap data pelayanan kesehatan menjadi suatu kebutuhan strategis.

1.2 Determining Data Mining Goal

Dalam konteks analisis pola pelayanan kesehatan, tujuan utama penggunaan model K-Nearest Neighbors (KNN) adalah mengklasifikasikan pasien ke dalam kelompok-kelompok tertentu berdasarkan pola kunjungan yang sering atau jarang. Dengan memanfaatkan algoritma KNN, kita dapat membangun model yang mampu mengidentifikasi kemiripan pola kunjungan antar-pasien. Hal ini memungkinkan pembagian pasien ke dalam kelompok-kelompok yang mencerminkan tingkat frekuensi kunjungan, yang dapat berkisar dari tinggi hingga rendah.

Salah satu teknik yang dapat diterapkan dalam pelaksanaan eksplorasi data adalah Knowledge Discovery in Databases (KDD). KDD merupakan suatu proses yang bertujuan untuk menggali informasi dari dataset, mengidentifikasi pola-pola yang muncul, dan menerapkan algoritma untuk mendeteksi serta memvisualisasikan pola tersebut agar dapat dipahami secara efektif oleh pengguna. Proses KDD melibatkan beberapa langkah kunci, meliputi pemilihan data, pra-pemrosesan data, transformasi data, penambangan data, dan evaluasi model. Ilustrasi yang diberikan juga menggambarkan serangkaian tahapan yang dilalui selama proses KDD.



Langkah awal melibatkan proses pembersihan dan seleksi data untuk memastikan kegunaan data. Langkah berikutnya adalah integrasi data, di mana data dengan karakteristik serupa digabungkan menjadi satu set data. Proses Transformasi data melibatkan penerapan prosedur penambangan data pada data yang telah terpilih. Pada tahap penambangan data, berbagai teknik digunakan untuk menghasilkan data yang sesuai dengan tujuan bisnis yang diinginkan. Evaluasi model melibatkan pengukuran terhadap hasil yang telah dicapai, dengan mempertimbangkan apakah hasil tersebut memenuhi standar yang diinginkan.

1.3 Produce Project Plan

Pada bagian ini menjelaskan aktivitas, detail, dan durasi perencanaan proyek yang akan dilakukan.

Aktivitas	Detail	Durasi
Pemilihan Kasus dan Algoritma	Pemilihan Kasus	2 hari
	Penentuan Algoritma	2 hari

Business Understanding	Menentukan Objektif Bisnis	5 hari
	Menentukan Tujuan Bisnis	5 hari
	Membuat Rencana Proyek	2 hari
Data Understanding	Mengumpulkan Data	2 hari
	Menelaah Data	2 hari
	Memvalidasi Data	2 hari
Data Preparation	Memilah Data	2 hari
	Membersihkan data	2 hari
	Mengkonstruksi Data	2 hari
	Menentukan Label Data	4 hari
Modeling	Mengintegrasikan Data	5 hari
	Membangun Skenario Pengujian	5 hari
Deployment	Melakukan Deployment Model	5 hari
	Membuat laporan akhir proyek	2 hari

Tools yang diterapkan dalam pelaksanaan proyek mencakup penggunaan bahasa pemrograman Python dalam tahap analisis hingga pembangunan model. Pelaksanaannya dilakukan melalui platform berbasis notebook, baik itu menggunakan Google Colab atau Jupyter Notebook.

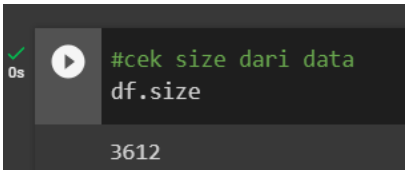
BAB 2

Data Understanding

Data understanding merupakan tahapan yang dilakukan setelah pemahaman awal mengenai business understanding sudah dipenuhi. Data understanding bertujuan untuk mendapatkan gambaran data secara utuh. Pada bagian ini berisi data awal dan analisis deskripsi data.

2.1 Collect Initial Data

Pengumpulan data awal perlu dilakukan untuk mengenali data yang dimiliki dan mengetahui jumlah data, misalnya jumlah baris dan kolom. Data yang digunakan dalam kasus case and cost prediction (regression problem) adalah data BPJS Kesehatan 2015-2021 yang berasal dari Hackathon. Ini menggunakan format data yang sudah terstruktur, yaitu 2022 Metadata Data Sampel BPJS Kesehatan. Jumlah data yang digunakan adalah 3612



Gambar. Ukuran jumlah data pada dataframe

2.2 Describe Analysis

Pada describe analysis melakukan pendefinisian label data yang digunakan untuk memastikan bahwa data yang digunakan merupakan dataset yang seimbang. Dataset pengujian yang digunakan untuk mengklasifikasian pasien berdasarkan pola kunjungan, terdiri dari 7 variabel dari 516 observasi data.

gambar. Bentuk data

Berikut adalah variabel yang dimiliki pada file data pengujian yang digunakan.

```
[20] df.columns

Index(['kode_prov', 'nama_provinsi', 'kode_kabupaten/kota',
       'nama_kabupaten/kota', 'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6'],
      dtype='object')
```

gambar. Jumlah atribut pada dataset

Langkah selanjutnya melibatkan Exploratory Data Analysis (EDA) untuk memberikan pemahaman mendalam terhadap data, menemukan konteks dalam data, mengidentifikasi hubungan antar variabel, dan merumuskan hipotesis untuk membangun model prediksi dalam kasus yang diberikan. Proses ini dimulai dengan mengidentifikasi jenis setiap atribut, yang pertama-tama dilakukan dengan menggunakan kode sebagai berikut.

```
#melihat informasi yang terdapat pada dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 516 entries, 0 to 515
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   kode_prov            35 non-null     float64
 1   nama_provinsi        35 non-null     object 
 2   kode_kabupaten/kota  515 non-null    float64
 3   nama_kabupaten/kota  515 non-null    object 
 4   Unnamed: 4           0 non-null      float64
 5   Unnamed: 5           0 non-null      float64
 6   Unnamed: 6           0 non-null      float64
dtypes: float64(5), object(2)
memory usage: 28.3+ KB
```

Gambar. melihat tipe atribut

Diperoleh informasi mengenai atribut dan tipe atribut.

Tabel. Informasi mengenai atribut, tipe atribut dan keterangan atribut

No.	Nama Atribut	Tipe Atribut	Deskripsi
1	kode_prov	float64	
2	nama_provinsi	object	

3	kode_kabupaten/kota	float64	
4	nama_kabupaten/kota	object	

2.3 Data Validation

Proses validasi data bertujuan untuk mengevaluasi akurasi dan kualitas sumber data sebelum data tersebut diolah atau digunakan. Validasi dilakukan dengan melakukan pengecekan terhadap adanya gangguan (noise) pada sumber data. Jika terdeteksi adanya gangguan, langkah selanjutnya adalah membersihkan data dengan maksud menciptakan data yang konsisten, lengkap, dan akurat. Oleh karena itu, dilakukan pemeriksaan terhadap atribut utama yang akan digunakan dalam dataset.

- atribut kode_prov, nama_provinsi, kode_kabupaten/kota, nama_kabupaten/kota.

```
df = df[['kode_prov', 'nama_provinsi', 'kode_kabupaten/kota', 'nama_kabupaten/kota']]
df.head()
```

	kode_prov	nama_provinsi	kode_kabupaten/kota	nama_kabupaten/kota
0	11.0	ACEH	1101.0	SIMEULUE
1	NaN	NaN	1102.0	ACEH SINGKIL
2	NaN	NaN	1103.0	ACEH SELATAN
3	NaN	NaN	1104.0	ACEH TENGGARA
4	NaN	NaN	1105.0	ACEH TIMUR

BAB 3. Data Preparation

Data preparation dapat dilakukan dengan melakukan data cleaning, data integration, data transformation dan data reduction. Tahapan ini dilakukan setelah pengumpulan data awal yang telah dilakukan yaitu pada tahap business understanding, lalu proses persiapan data, pemilihan variabel yang akan dianalisis dan pembersihan data.

3.1 Package

3.2 Dataset Description

3.3 Data Selection

3.4 Data Cleaning

3.5 Data Construct

3.6 Data Labeling

3.7 Data Integration

BAB 4. Modeling

BAB 5. MODEL EVALUATION

Tahap Evaluation akan menjelaskan bagaimana evaluasi terhadap model yang dipilih untuk memprediksi kunjungan dan biaya pada rumah sakit.

5.1 Mengevaluasi Hasil Pemodelan