**DOKUMEN PROYEK**

**12S4054 – PENAMBANGAN DATA**

***Case and Cost Prediction (Regression Problem) using Decision Tree***



**Disusun oleh:**

|  |  |
| --- | --- |
| 12S20009 | Agnes Marpaung |
| 12S20021 | Sintia Lolita Silaen |
| 12S20040 | Esphi Hutabarat |

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2023/204**

# DAFTAR ISI

[DAFTAR ISI 2](#_30j0zll)

[BAB 1 4](#_1fob9te)

[BUSINESS UNDERSTANDING 4](#_3znysh7)

[1.1](#_2et92p0)  Determine Business Objective 4

[1.2](#_1t3h5sf)  Determine Project Goal 5

[1.3](#_4d34og8)  Produce Project Plan 5

[BAB 2 7](#_lnxbz9)

[DATA UNDERSTANDING 7](#_35nkun2)

[2.1](#_44sinio) Collecting Data 7

[2.2](#_2jxsxqh) Describe Data 7

[2.3](#_z337ya) Validation Data 7

[BAB 3 8](#_1y810tw)

[DATA PREPARATION 8](#_4i7ojhp)

[3.1 Data Selection 8](#_2xcytpi)

[3.2 Data Cleaning 8](#_1ci93xb)

[3.3 Data Construct 8](#_3whwml4)

[3.4 Labeling Data 8](#_2bn6wsx)

[BAB 4 9](#_3as4poj)

[MODELLING 9](#_1pxezwc)

[4.1 Build Model 9](#_49x2ik5)

[BAB 5 10](#_3o7alnk)

[EVALUATION 10](#_23ckvvd)

[5.1 Evaluate Result 10](#_ihv636)

[5.1.2 Evaluate result cost prediction 10](#_32hioqz)

[BAB 6 11](#_1hmsyys)

[DEPLOYMENT 11](#_41mghml)

[6.1 Plan Deployment 11](#_2grqrue)

# BAB 1

# BUSINESS UNDERSTANDING

Business understanding merupakan langkah pertama dalam metodologi CRISP-DM untuk melakukan prediksi jumlah kasus dan unit cost pada sebuah wilayah yang diakibatkan dari penambahan rumah sakit. Pada bab ini akan menjelaskan terkait tahapan data mining untuk meningkatkan pemahaman dalam menentukan objektif bisnis, menentukan tujuan bisnis, dan membuat rencana proyek.

## 1.1 Determine Business Objective

Dalam kehidupan sehari hari tentunya banyak pemanfaatan teknologi yang kita gunakan dan teknologi telah menjadi solusi utama dalam menyelesaikan berbagai masalah. Perkembangan teknologi yang pesat juga mempengaruhi berbagai sektor termasuk dalam pertumbuhan sektor kesehatan, khususnya dengan pertambahan jumlah rumah sakit diberbagai kota dan wilayah yang ada di Indonesia. Dengan penambahan ini, menimbulkan tantangan dalam mengelola data dan menghadapi berbagai situasi yang kompleks. Maka dari itu pengerjaan proyek ini bertujuan untuk melakukan prediksi jumlah kasus dan unit cost di suatu wilayah tertentu dengan menggunakan data BPJS Hackathon *(Case and Cost Predictio*n).

Pada proyek ini, dilakukan pengembangan model data mining yang mampu melakukan prediksi jumlah kasus dan unit cost yang timbul akibat penambahan jumlah rumah sakit di wilayah tertentu metode yang digunakan adalah Regression Problem dengan algoritma *Decision Tree Regression* yang merupakan salah satu teknik dari Data Mining untuk melakukan prediksi terhadap variabel kontinu atau nilai yang bersifat numerik yang terdiri dari variabel bebas(x) dan variabel tak bebas (y). Dalam konteks ini, algoritma *Decision Tree* akan mencoba memahami hubungan antara pertambahan rumah sakit dengan kenaikan jumlah kasus yang tercatat dan biaya yang dibutuhkan, memungkinkan analisis prediktif yang mendalam.

Data Mining merupakan konsep untuk mengeksplorasi informasi dari sekumpulan data. Proses ini menggunakan matematika, statistik, dan machine learning untuk mengekstrak pengetahuan. Data BPJS Hackathon merupakan data kategorikal yang memerlukan analisis sebelum membentuk model, dengan tujuan meningkatkan performa data dan mengidentifikasi faktor-faktor yang berpotensi menyebabkan fraud atau non fraud. Proyek ini memerlukan analisis data untuk mencapai prediksi jumlah kasus dan biaya, dimana analisis dilakukan sebelum data dijadikan dasar pembentukan model. Model tersebut akan diimplementasikan menggunakan metode regresi dengan algoritma *Decision Tree* yang telah dipilih sebelumnya.

## 1.2 Determine Project Goal

Pada Case and Cost Prediction bertujuan untuk mengembangkan model data mining yang mampu melakukan prediksi jumlah kasus dan unit cost di sebuah daerah sebagai hasil dari penambahan rumah sakit berdasarkan data BPJS Hackathon.

## 1.3 Produce Project Plan

Tahap perencanaan yang dilakukan untuk mencapai tujuan pengerjaan adalah sebagai berikut:

**Tabel 1. Jadwal Pelaksanaan Proyek**

|  |  |  |
| --- | --- | --- |
| **Tahapan** | **Waktu Pengerjaan** | **Kegiatan** |
| *Business Understanding* | 3 hari | Menentukan objektif bisnis, menentukan tujuan bisnis, dan membuat rencan a proyek. |
| *Data Understanding* | 4 hari | Mengumpulkan data, menelaah data, memvalidasi data |
| *Data Preparation* | 3 hari | Memilih data, membersihkan data, mengkonstruksi data, menentukan label data, dan mengintegrasikan data. |
| *Modeling* | 3 hari | Membangun skenario pengujian dan membangun model. |
| *Evaluation* | 3 hari | Melakukan evaluasi hasil pemodelan dan melakukan review terhadap proses pemodelan. |
| *Deployment* | 4 hari | Membuat rencana deployment model, Monitoring and Maintenance rencana deployment model dan meninjau proyek. |

# BAB 2

# DATA UNDERSTANDING

Dalam tahapan data understanding yang merupakan tahapan pemahaman terhadap data yang akan digunakan, tahapan ini dimulai dari mengumpulkan data, mendeskripsikan data dan memahami data yang akan digunakan dalam penelitian.

## 2.1 Collecting Data

Collecting data atau pengumpulan data merupakan tahap awal dalam menemukan data yang akan digunakan dalam penelitian. Pada penelitian ini dataset yang akan digunakan dalam memprediksi jumlah kasus dan unit cost pada sebuah daerah yang diakibatkan adanya penambahan Rumah Sakit kerja sama, berdasarkan dataset *train* yaitu data case\_cost\_prediction\_train.csv.

## 2.2 Describe Data

Dataset yang digunakan untuk memprediksi jumlah kasus dan biaya unit di suatu daerah akibat penambahan Rumah Sakit kerja sama adalah *case\_cost\_prediction\_train.csv.* Dataset ini terdiri dari 57.971 observasi dan memiliki 36 variabel. Berikut adalah tabel yang membahas atribut dalam dataset.

|  |  |  |
| --- | --- | --- |
| No. | Atribut | Deskripsi |
| 1 | *row\_id* | ID dari setiap data |
| 2 | *tgl\_pelayanan* | periode bulan pelayanan di rumah sakit |
| 3 | *kddati2* | kode kabupaten/kota |
| 4 | *tkp* | tingkat pelayanan; 30:rawat jalan; 40:rawat inap |
| 5 | *peserta* | jumlah peserta akhir pada kabupaten/kota periode tersebut |
| 6 | *a,b,c,...,sd* | tipe rumah sakit yang melayani peserta JKN-KIS |
| 7 | *case* | jumlah kunjungan rumah sakit |
| 8 | *unit\_cost* | jumlah biaya pelayanan rumah sakit |

Berdasarkan dataset tersebut, akan dilakukan EDA (*Exploratory Data Analysis*) terhadap dataset *case\_cost\_prediction\_train.csv* untukmenganalisis karakteristik utamanya. Dalam pengerjaan proyek ini, tidak semua atribut dari dataset digunakan, melainkan hanya beberapa atribut yang relevan terhadap tujuan penelitian. Oleh karena itu, atribut yang paling sesuai untuk memprediksi jumlah kasus dan biaya unit adalah atribut *case*, atribut *unit\_cost,* dan beberapa atribut lain yang relevan. Berikut adalah beberapa hipotesis terkait atribut dalam dataset yang akan digunakan.

* atribut *kddati2* digunakan untuk mengetahui kasus per kabupaten/kota berdasarkan kode yang telah ditetapkan.
* atribut *peserta* digunakan untuk mengetahui jumlah peserta.
* atribut *case* digunakan dalam mengetahui kasus kunjungan ke rumah sakit.
* atribut *unit\_cost* digunakan untuk mengetahui biaya pelayanan rumah sakit.

Berdasarkan hipotesis-hipotesis tersebut, atribut *kddati2*, peserta, *case*, dan *unit\_cost* memiliki pengaruh signifikan terhadap jumlah kasus dan biaya unit. Atribut-atribut ini dianggap relevan dan dapat digunakan dalam pengembangan model data mining untuk meramalkan jumlah kasus dan biaya unit di suatu daerah sebagai dampak dari penambahan Rumah Sakit kerja sama sesuai dengan tujuan proyek.

## 2.3 Validation Data

Pada sub bab ini akan dijelaskan tahap validasi data untuk melakukan pemeriksaan menyeluruh terhadap keakuratan, kelengkapan, dan kualitas sumber data sebelum digunakan untuk analisis atau pengolahan lebih lanjut. Tahapan ini melibatkan identifikasi dan penanganan potensi kesalahan seperti nilai yang hilang (missing value) atau noise dalam data. Pengecekan memungkinkan untuk membersihkan dan menormalkan data agar menjadi konsisten, lengkap, dan akurat, memastikan bahwa informasi yang diolah dapat diandalkan dalam pengambilan keputusan dan analisis mendatang.

* Atribut *kddati2*

|  |  |
| --- | --- |
|  |  |

* Atribut *peserta*

|  |  |
| --- | --- |
|  |  |

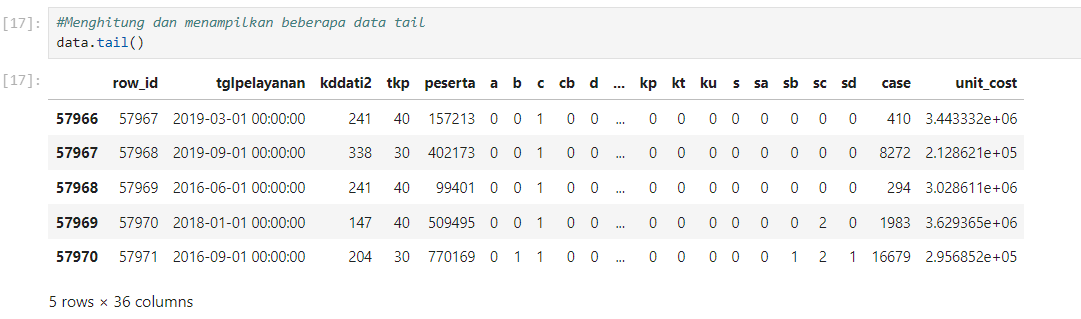
* Atribut *case*

|  |  |
| --- | --- |
|  |  |

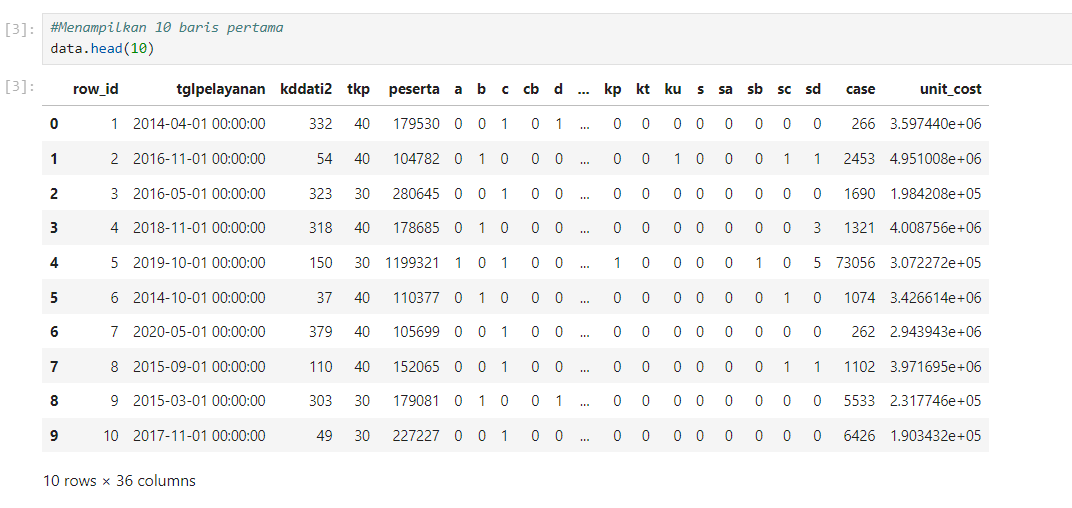
* Atribut *unit\_cost*

|  |  |
| --- | --- |
|  |  |

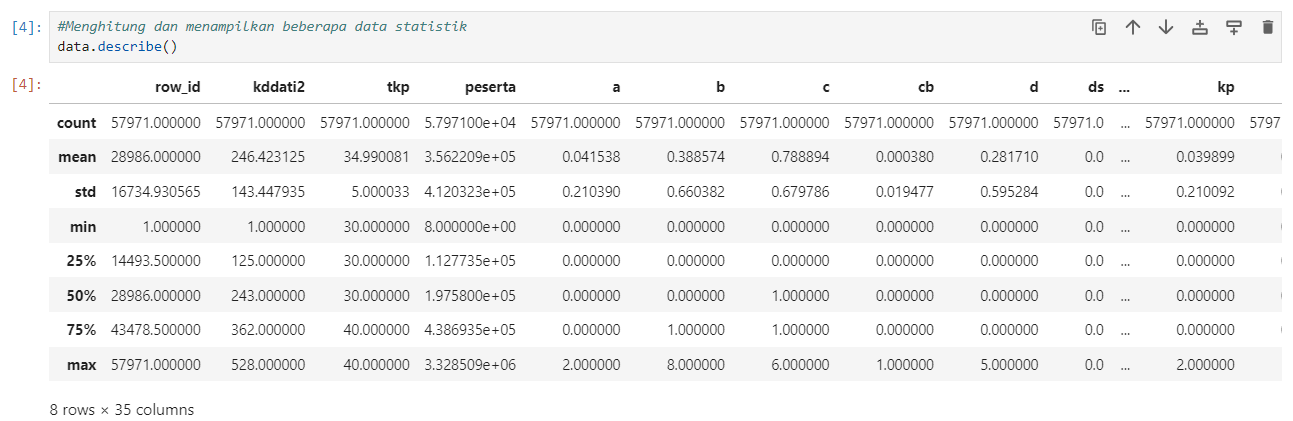
* Data *tail*

**

* Data *head*

**

* *Data Describe*

**

# BAB 3

# DATA PREPARATION

## 3.1 Data Selection

## 3.2 Data Cleaning

## 3.3 Data Construct

## 3.4 Labeling Data

# BAB 4

# MODELLING

## 4.1 Build Model

# BAB 5

# EVALUATION

## 5.1 Evaluate Result

### 5.1.2 Evaluate result cost prediction

# BAB 6

# DEPLOYMENT

## 6.1 Plan Deployment