**DOKUMEN PROYEK**

**12S4054 – PENAMBANGAN DATA**

**BPJS Case and Cost Prediction (Regression Problem)**

**using Decision Tree**



**Disusun oleh:**

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

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2023/204**

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc155390874)

[BAB 1 BUSINESS UNDERSTANDING 5](#_Toc155390875)

[1.1 Determine Business Objective 5](#_Toc155390876)

[1.2 Determine Project Goal 6](#_Toc155390877)

[1.3 Produce Project Plan 6](#_Toc155390878)

[BAB 2 DATA UNDERSTANDING 8](#_Toc155390879)

[2.1 Menelaah Data 8](#_Toc155390880)

[2.2 Validation Data 11](#_Toc155390881)

[BAB 3 DATA PREPARATION 16](#_Toc155390882)

[3.1 Data Selection 16](#_Toc155390883)

[3.2 Data Cleaning 17](#_Toc155390884)

[3.3 Data Construct 18](#_Toc155390885)

[BAB 4 MODELLING 19](#_Toc155390886)

[4.1 Building Test Scenario 19](#_Toc155390887)

[4.2 Build Model 20](#_Toc155390888)

[BAB 5 EVALUATION 22](#_Toc155390889)

[5.1 Evaluate Result 22](#_Toc155390890)

[5.1.1 Evaluate result case prediction 23](#_Toc155390891)

[5.1.2 Evaluate Result cost prediction 23](#_Toc155390892)

[BAB 6 DEPLOYMENT 25](#_Toc155390893)

[6.1 Plan Deployment 25](#_Toc155390894)

**DAFTAR TABEL**

[Table 1. Jadwal Pelaksanaan Proyek 4](#_Toc155390088)

[Table 2. Deskripsi Atribut 6](#_Toc155390089)

[Table 3. Summary Hasil Evaluasi 22](#_Toc155390090)

**DAFTAR GAMBAR**

[Gambar 1. Korelasi 15](#_Toc155399112)

[Gambar 2. Halaman Case Prediction 30](#_Toc155399113)

[Gambar 3. Cost Prediction 31](#_Toc155399114)

[Gambar 4. Homepage 31](#_Toc155399115)

# 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 di berbagai 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:

**Table 1. Jadwal Pelaksanaan Proyek**

|  |  |  |
| --- | --- | --- |
| **Tahapan** | **Waktu Pengerjaan** | **Kegiatan** |
| *Business Understanding* | 3 hari | Menentukan objektif bisnis, menentukan tujuan bisnis, dan membuat rencana 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 mendeskripsikan data dan memahami data yang akan digunakan dalam penelitian.

## 2.1 Menelaah 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.

**Table 2. Deskripsi Atribut**

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

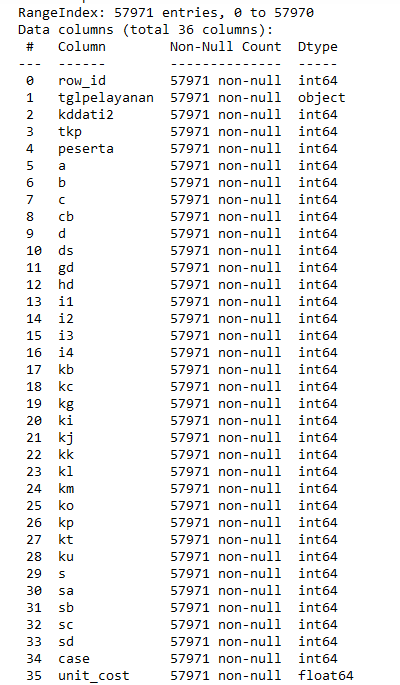
Berdasarkan dataset tersebut, akan dilakukan EDA (*Exploratory Data Analysis*) terhadap dataset *case\_cost\_prediction\_train.csv* untukmenganalisis karakteristik utamanya. Dalam pengerjaan proyek ini, tidak semua atribut dari dataset digunakan, melainkan hanya beberapa atribut yang relevan terhadap 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.

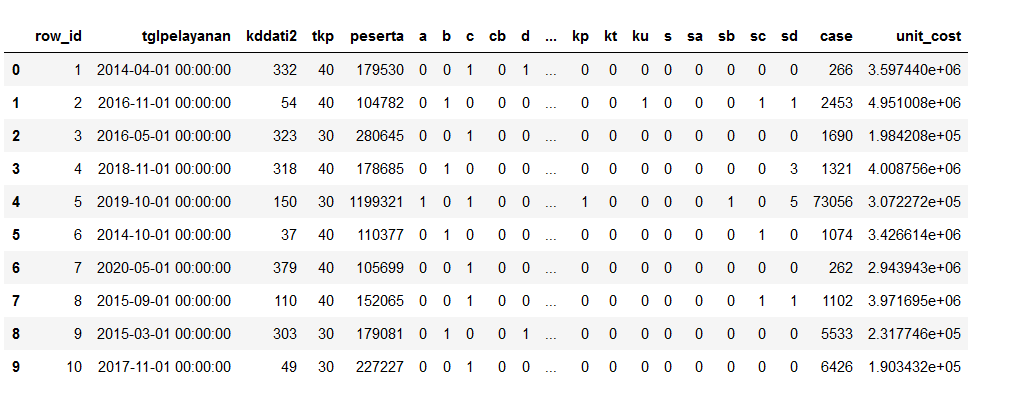
Berdasarkan hipotesis-hipotesis tersebut, atribut *kddati2*, peserta, *case*, dan *unit\_cost* memiliki pengaruh signifikan terhadap jumlah kasus dan biaya unit. Atribut-atribut ini dianggap relevan dan dapat digunakan dalam pengembangan model data mining untuk meramalkan jumlah kasus dan biaya unit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

* data.info()

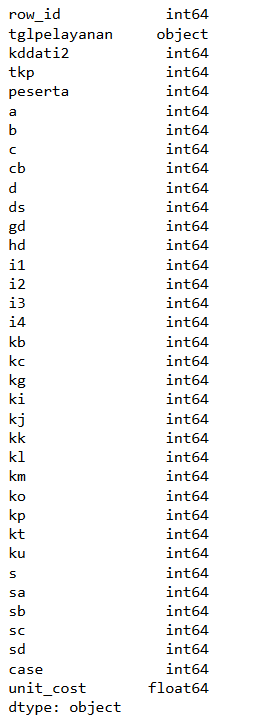
Untuk memberikan informasi ringkas mengenai dataset mengenai jumlah baris dan column, nama column, tipe data, dan jumlah nilai non-null.



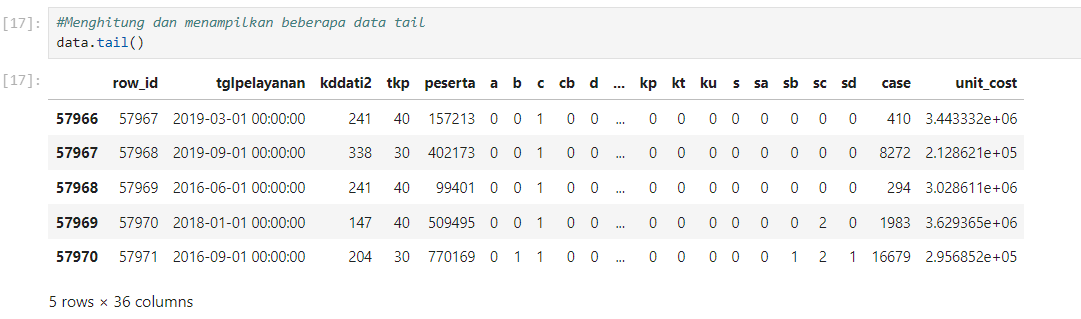
* data.head (10)



* data.dtypes



* data.tail()

**

* data.describe()



## 2.2 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*

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

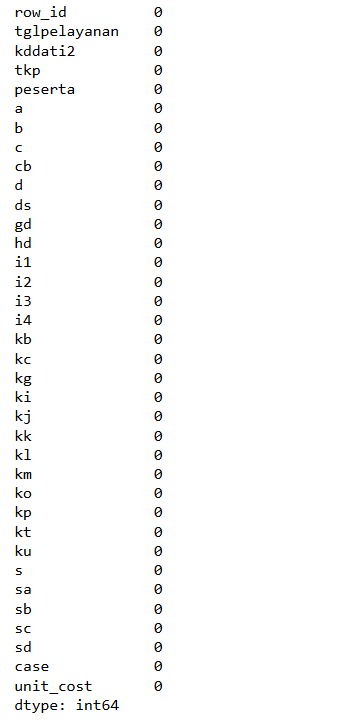
* Atribut *case*

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

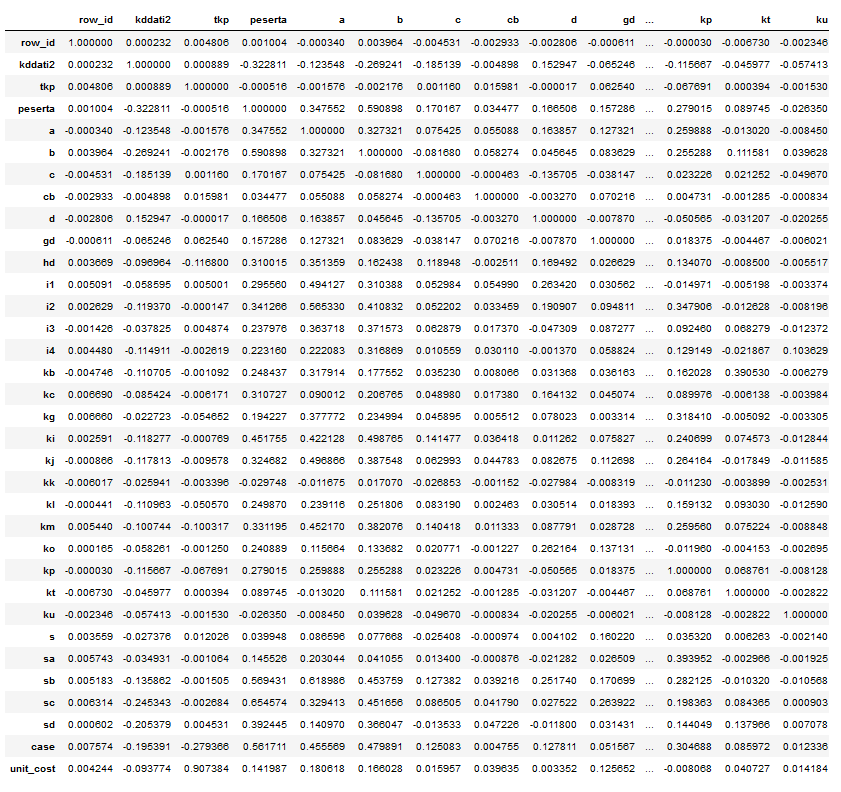
* Atribut *unit\_cost*

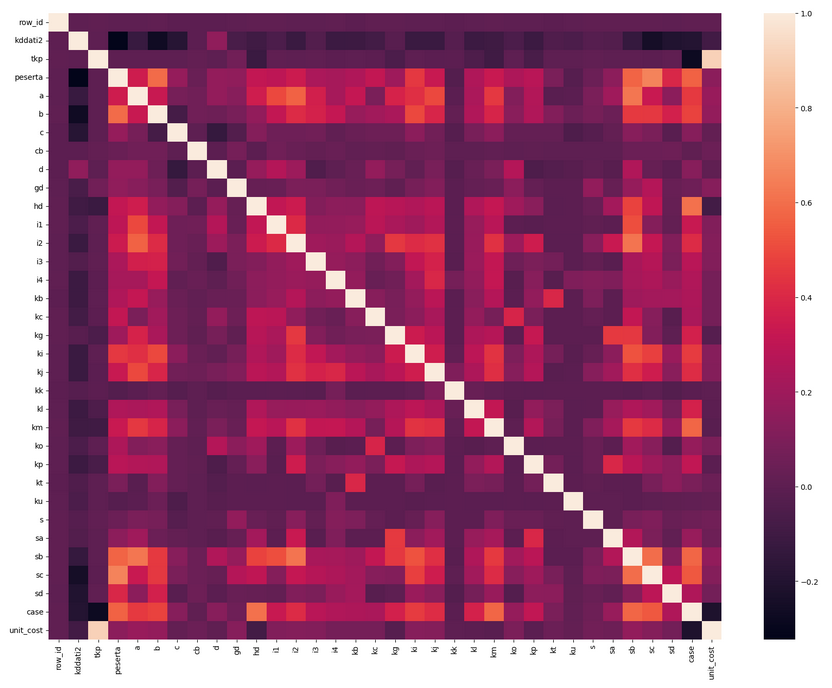
|  |  |
| --- | --- |
|  |  |

* Memeriksa Missing Value

**

* Memeriksa Corellation

**

**

**Gambar 1. Korelasi**

# BAB 3 DATA PREPARATION

Data preparation adalah tahap dalam proses analisis data yang dilakukan untuk menghasilkan data yang berkualitas baik. Tahapan dalam data preparation meliputi data *Selection*, data *Cleaning*, data *Construct*, dan Labeling Data. Data preparation dilakukan setelah pengumpulan data awal pada fase business understanding dan sebelum analisis data dilakukan. Tahapan data preparation meliputi beberapa proses, seperti ekstraksi, transformasi, pembersihan, standarisasi, integrasi, validasi, *formatting*, dan *summarization*.

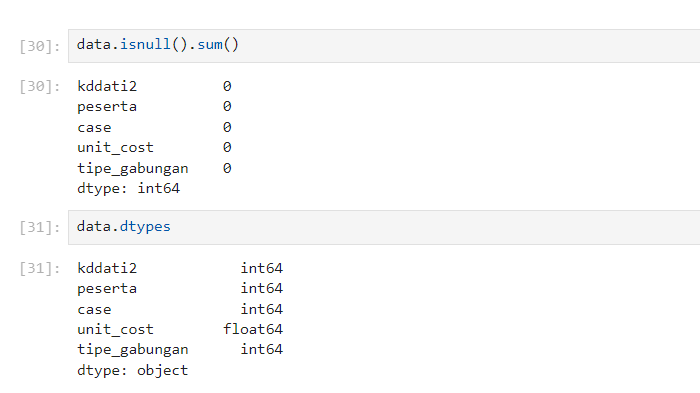
## 3.1 Data Selection

Tahapan ini melibatkan proses pemilihan atribut atau kolom yang relevan dalam sebuah dataset. Atribut yang tidak diperlukan akan dihapus. Atribut yang tidak digunakan diidentifikasi dan dimasukkan ke dalam variabel bernama 'to\_drop', kemudian dilakukan penghapusan kolom sesuai dengan daftar atribut yang terdapat dalam 'to\_drop'.

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

## 3.2 Data Cleaning

Data cleaning adalah proses dalam menganalisis data yang bertujuan untuk mengidentifikasi, mengoreksi, dan menghapus kesalahan atau ketidaksesuaian dalam dataset. Data cleaning sangat penting dalam proses data mining karena dapat meningkatkan kualitas data, meningkatkan akurasi, dan meningkatkan keandalan hasil analisis. Tahapan dalam data cleaning meliputi mendeteksi kesalahan atau data yang rusak, memperbaiki atau menghapus data yang tidak diperlukan atau yang duplikat, dan memastikan konsistensi dan keakuratan data.



## 3.3 Data Construct

Data yang tersedia memiliki nilai boolean yang menunjukkan bahwa data tersebut tidak memiliki nilai kosong atau missing value, sehingga tidak perlu menggunakan fungsi dropna() untuk menghapusnya. Selanjutnya, perlu dilakukan konstruksi data sebelum proses pemodelan. Konstruksi data adalah bagian dari transformasi data yang mencakup representasi fitur, penentuan korelasi, dan integrasi data. Representasi fitur digunakan untuk mengurangi kompleksitas data, meningkatkan akurasi prediksi, dan memilih fitur yang paling berpengaruh.

Untuk memprediksi biaya, digunakan atribut 'kddati2', 'peserta', 'tkp', dan 'tipe\_gabungan'. Sedangkan untuk memprediksi kasus, digunakan atribut 'kddati2', 'peserta', 'unit\_cost', 'tipe\_gabungan', dan 'tkp'. 'Tipe Gabungan' adalah jumlah dari semua jenis rumah sakit yang melayani peserta JKN-KIS.

# BAB 4 MODELLING

Pada bab sebelumnya, persiapan data telah dilakukan untuk mempersiapkan model yang akan dikembangkan. Selanjutnya pada bab ini akan mengeksplorasi pembangunan model Regresi Prediksi dengan menerapkan algoritma *Decision Tree Regressor.* Adapun tujuan dibangunnya model ini pada konteks BPJS Hackathon adalah untuk memprediksi jumlah kasus dan biaya unit di suatu daerah sebagai hasil dari penambahan Rumah Sakit pada wilayah tersebut. Selanjutnya regresi mempunyai tujuan untuk melakukan proses prediksi nilai yang bersifat kontinu, dan ada berbagai macam algoritma yang dapat digunakan untuk prediksi kasus ini, salah satunya adalah *Long Short-Term Memory*. Adapun alasan kelompok kami memilih untuk menggunakan Decision Tree Regression yaitu karena kemampuannya dalam mengingat informasi dalam jangka waktu yang panjang sambil menghapus informasi yang sudah tidak relevan. Harapan kami menggunakan algoritma ini supaya lebih memudahkan pemahaman dan interpretasi kinerjanya dalam menyelesaikan proyek Data Mining ini. Oleh karena itu, tahap-tahap yang akan dilakukan dalam pembuatan model ini disajikan sebagai berikut.

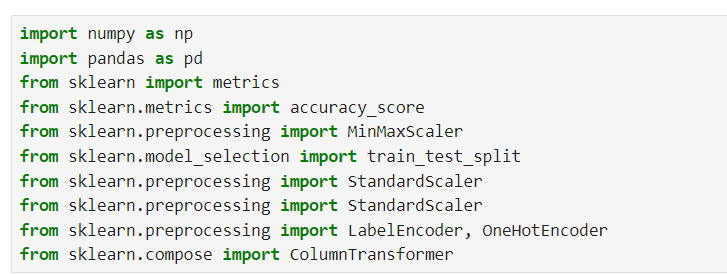
## 4.1 Building Test Scenario

Dalam membangun *test scenario* untuk prediksi *case and cost* *(Regression Problem)* menggunakan *Decision Tree*. Data yang ditambahkan dalam melakukan pemodelan yaitu pemilihan variabel yang relevan seperti tglpelayanan, kddati2 (kode wilayah), peserta, case, dan unit\_cost yang kemudian dibagi menjadi data pelatihan dan uji. Selanjutnya, dengan menerapkan algoritma *Decision Tree*, model yang telah dilatih bertujuan untuk memahami hubungan antara variabel input dan output, yaitu jumlah kasus dan biaya pelayanan rumah sakit. Proses evaluasi dilakukan dengan menggunakan data uji untuk mengukur seberapa akurat model *Decision Tree* dalam memprediksi kasus dan biaya, memberikan gambaran yang lebih baik dalam perencanaan dan pengelolaan rumah sakit dalam mengantisipasi kasus dan estimasi biaya pelayanan.

## 4.2 Build Model

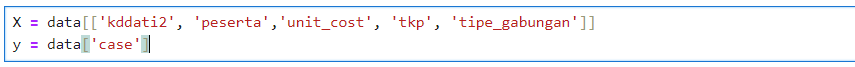
Dalam proses pengembangan model, dataset yang telah disiapkan melalui tahap *preprocessing*, seperti yang telah dibahas pada bab sebelumnya. Adapun *library* yang diperlukan dalam pembuatan model ini adalah sebagai berikut. Pertama sekali, digunakan *library* 'sklearn' untuk membangun model *Decision Tree Regression.* Kemudian model ini akan diuji pada dataset BPJS yang telah disiapkan. *Library* ini memungkinkan implementasi model regresi pohon keputusan serta proses pelatihan dan pengujian model pada data yang telah disiapkan. Dibawah ini adalah proses dalam pembangunan model yang bertujuan untuk mencari nilai dari *case and cost.*

* Dibutuhkan library sklearn untuk membangun model *Decision Tree Regression*. Selanjutnya, model akan dijadikan sebagai bahan latihan dan pengujian menggunakan dataset dari BPJS.

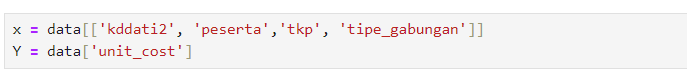


* Selanjutnya mendefinisikan fitur yang digunakan untuk melakukan prediksi.

Berikut merupakan fitur yang digunakan dalam kasus ‘*case’*

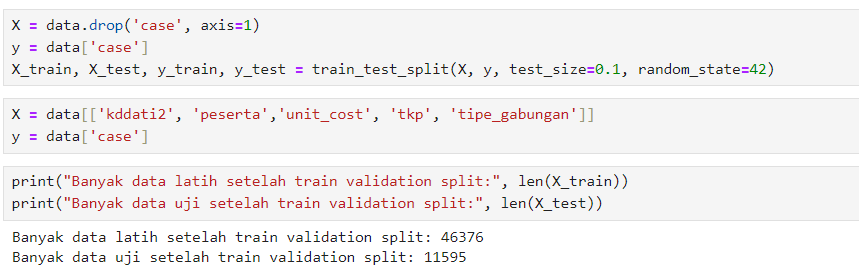


Berikut merupakan fitur yang digunakan dalam kasus ‘*cost’*

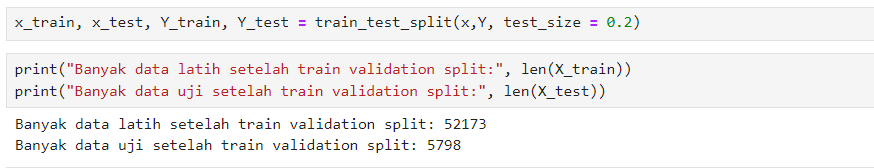
**

* Setelah fitur didefinisikan, langkah selanjutnya adalah membagi data menjadi data training dan data test.

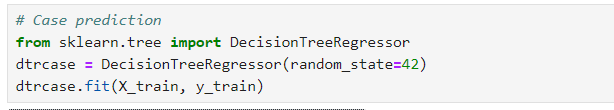
Untuk Cost Prediction menggunakan rasio 90:10



Untuk case menggunakan rasio 80:20



* Setelah membagi data, langkah selanjutnya adalah menambahkan lapisan pada model Decision Tree Regression. Hal Ini melibatkan inisialisasi model Sequential yang akan mengimplementasikan jaringan saraf Decision Tree Regression. Proses ini termasuk menambahkan lapisan normalisasi batch dan lapisan keluaran padat. Setelah itu, akan dicetak untuk mendapatkan gambaran model yang akan dibuat. Dengan demikian, kita akan membuat model regresi Decision Tree-nya.



* Selanjutnya mendefinisikan objek y\_pred dengan untuk memprediksi hasil model regresi yang akan dibangun.



# BAB 5 EVALUATION

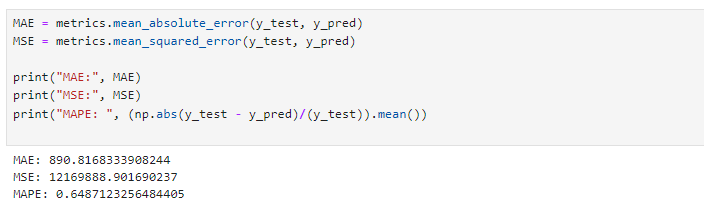
Pada bab *Evaluation* akan dijelaskan tentang bagaimana evaluasi yang diperoleh terhadap model *Decision tree* dalam memprediksi case and cost dalam sebuah daerah akibat penambahan Rumah Sakit yang terjadi dalam suatu wilayah. Evaluasi ini dilakukan dengan tujuan agar dapat mengetahui bagaimana hasil pada tahap modelling terhadap tujuan yang ingin dicapai seperti yang telah dijelaskan dalam business understanding.

## 5.1 Evaluate Result

Evaluasi terhadap model dalam memprediksi Case and Cost dalam suatu wilayah sebagai dampak dari penambahan Rumah Sakit dapat dilaksanakan dengan berbagai metode, salah satunya adalah melalui penerapan algoritma Decision Tree Regression. Dalam menilai kinerja model yang menggunakan Decision Tree Regression dalam memprediksi Case and Cost, dapat dilakukan dengan menggunakan beberapa metrik evaluasi seperti mean absolute error (MAE). MAE digunakan untuk mengukur sejauh mana perbedaan antara nilai prediksi dan nilai sebenarnya. Selain itu, metrik evaluasi yang digunakan adalah Mean Absolute Percentage Error (MAPE), yang sering digunakan dalam analisis deret waktu. Kedua metrik ini memberikan informasi tentang tingkat kesalahan dalam prediksi model, berguna untuk mengevaluasi kemampuan model dalam memprediksi jumlah kasus dan biaya unit di suatu wilayah akibat penambahan Rumah Sakit. Oleh karena itu, Decision Tree menjadi pilihan yang tepat untuk memprediksi jumlah kasus dan biaya unit di wilayah tersebut.

### 5.1.1 Evaluate result case prediction

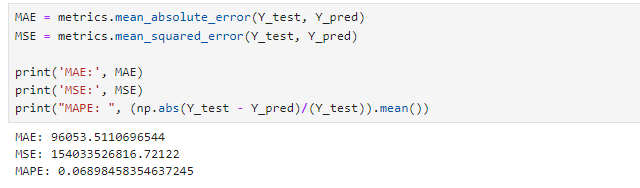
Pada bagian ini dilakukan evaluate result pada *case* prediction agar mengetahui bagaimana kinerja model yang dibangun.



Dari cuplikan gambar di atas dapat dilihat bahwa MAE: 890.8168333908244, MSE: 12169888.901690237 dan MAPE : MAPE: 0.6487123256484405.

### 5.1.2 Evaluate Result cost prediction

Pada bagian ini dilakukan evaluate result pada *cose* prediction agar mengetahui bagaimana kinerja model yang dibangun.



Dari cuplikan gambar di atas dapat dilihat bahwa MAE: 96053.5110696544, MSE: 154033526816.72122 dan MAPE : MAPE: 0.06898458354637245.

Dari hasil evaluasi yang diperoleh dapat dibuat summary untuk Case and Cost prediction akibat penambahan rumah sakit.

**Table 3. Summary Hasil Evaluasi**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Menggunakan Seleksi Fitur** | | | | |
|  | Key Performance Index | Tujuan Teknis | Hasil Model | Kondisi |
| Case | MAE | < 900 | 890.81 | Terpenuhi |
| MAPE | < 90% | 0.6487 | Terpenuhi |
| Cost | MAE | < 97000 | 96053.51 | Terpenuhi |
| MAPE | < 70% | 0.0689 | Terpenuhi |

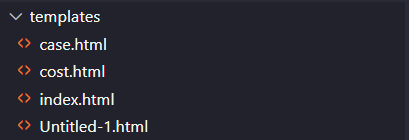
# BAB 6 DEPLOYMENT

## 6.1 Plan Deployment

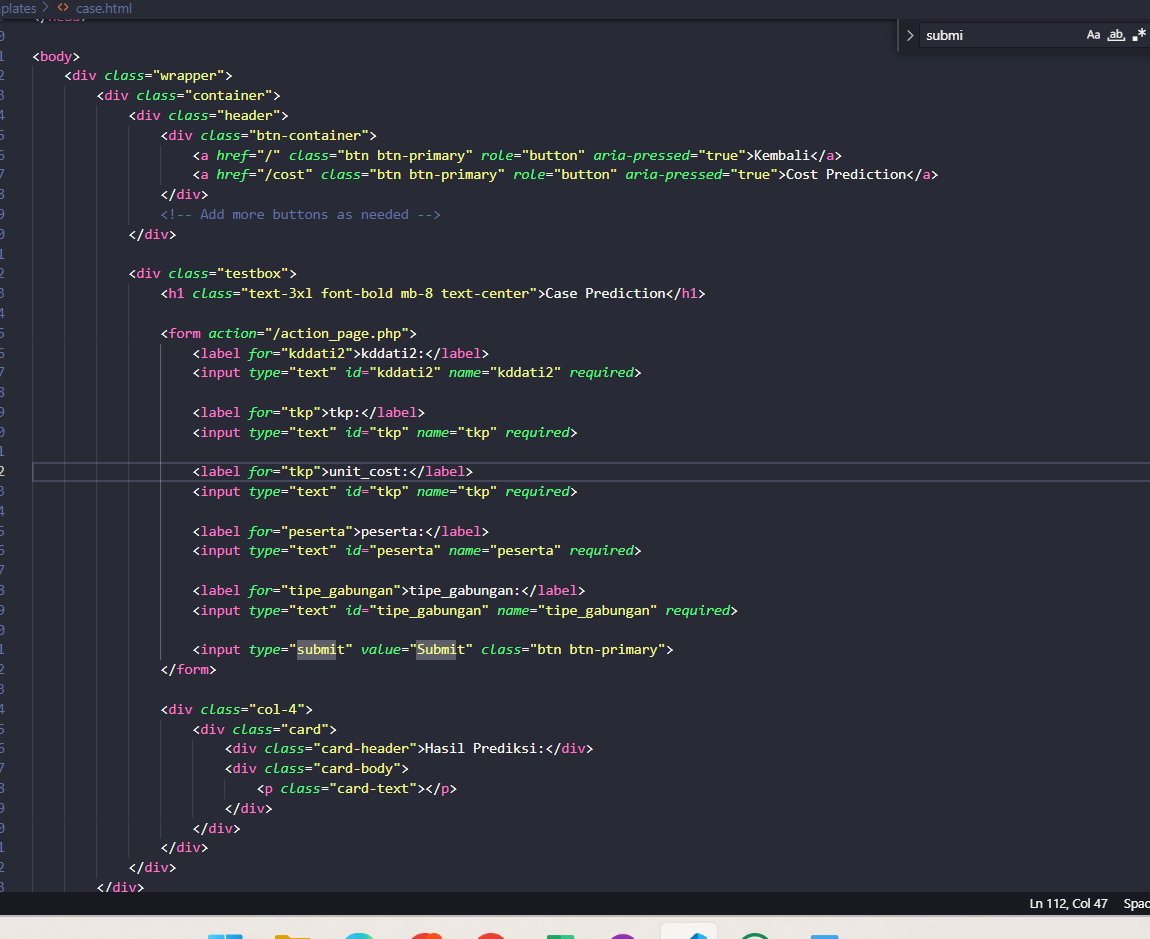
Dalam tahap perencanaan deployment, model yang terbentuk dari proses pemodelan akan diterapkan sesuai dengan kebutuhan data mining yang diinginkan. Langkah-langkah yang kami ambil dalam deployment yaitu :

1. Mempersiapkan interface dan Script Code

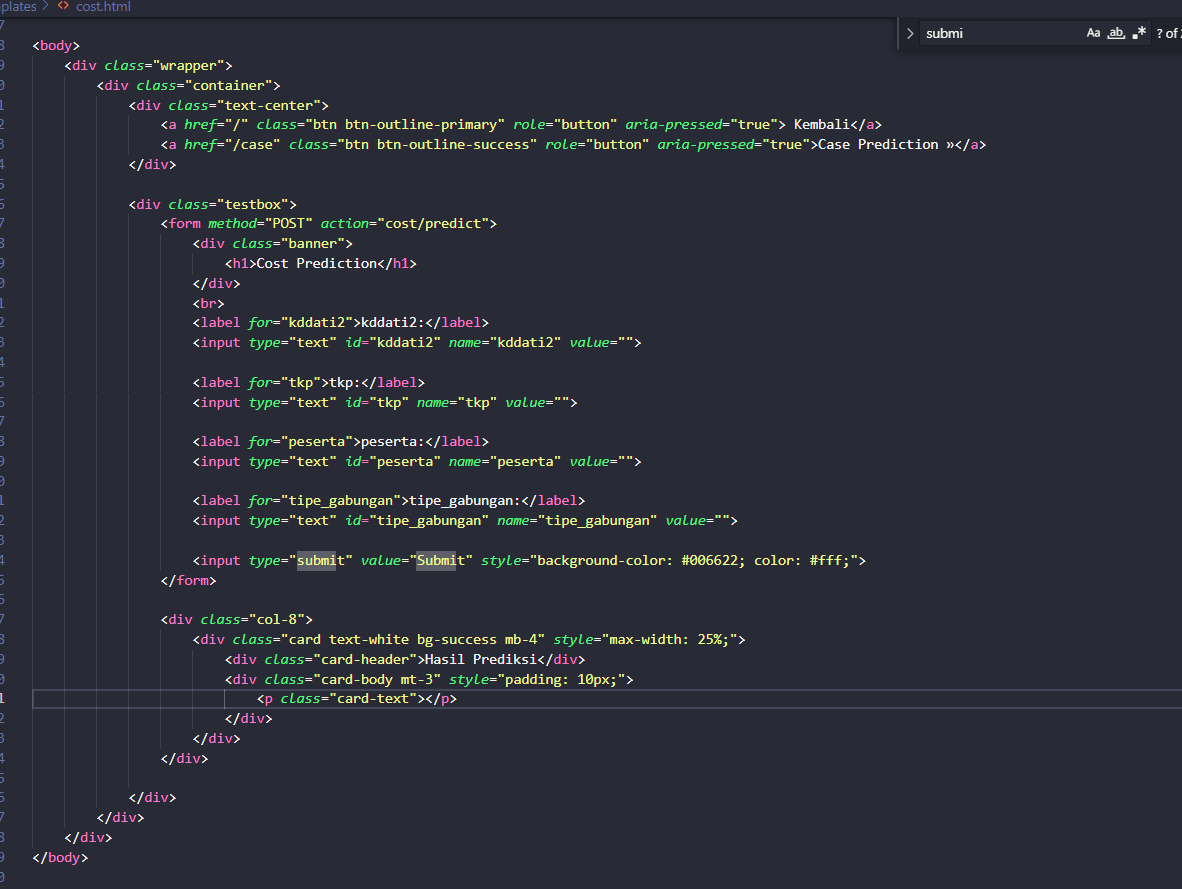
Pada tahap ini, persiapan dilakukan sebelum implementasi antarmuka (interface) dengan merancang kerangka desain HTML. Langkah ini mencakup pembuatan struktur dasar halaman web dan penyusunan elemen-elemen antarmuka. Tujuannya adalah agar proses implementasi antarmuka dengan HTML dapat dilakukan secara terstruktur. Desain antarmuka disimpan dalam folder



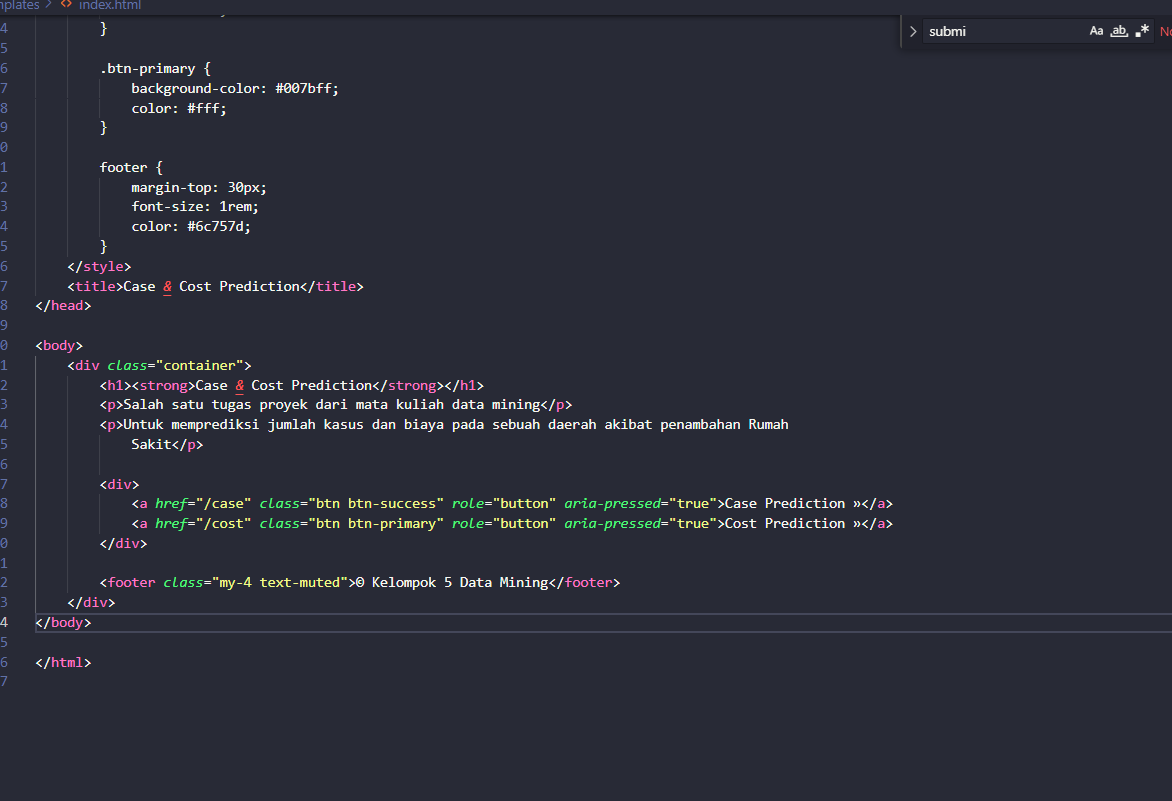
1. Membuat Interface HTML
2. Case Prediction



1. Cost Prediction



1. Index Html

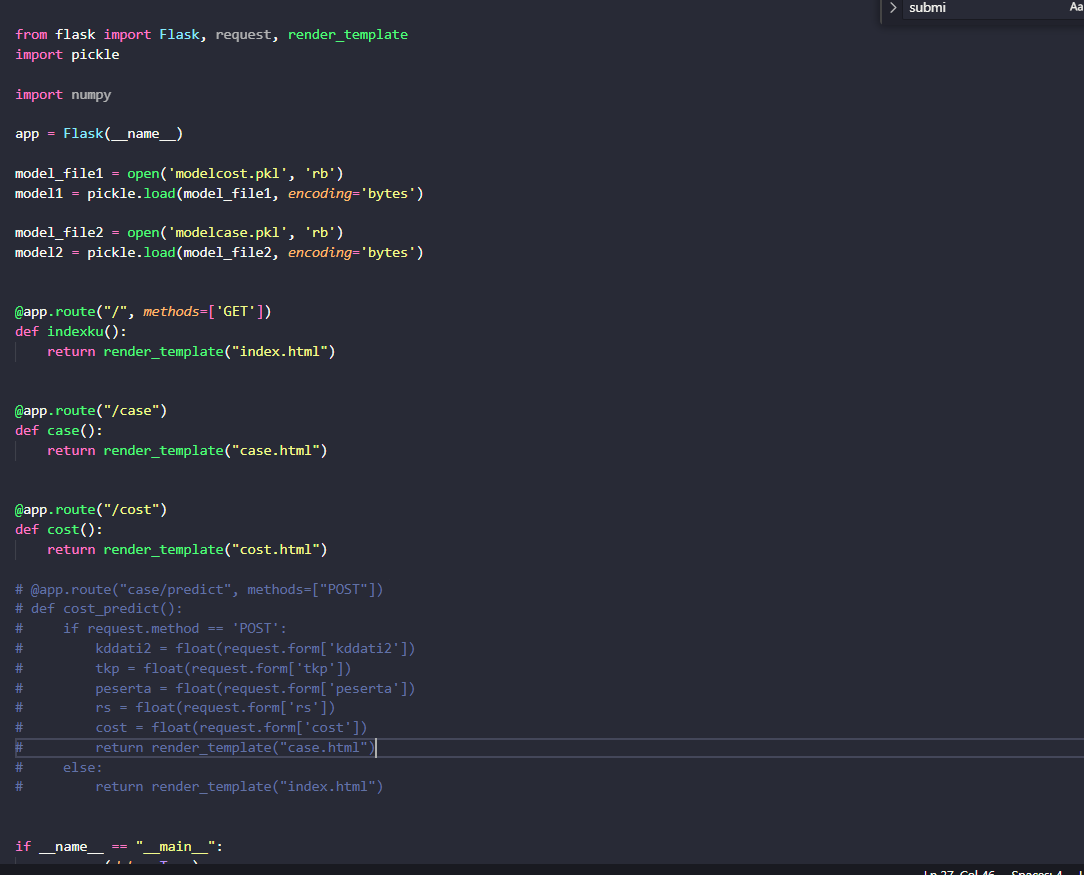


1. Mempersiapkan Script dan Code

Pada proyek ini, instalasi Flask dilakukan untuk membantu menyediakan kerangka kerja web. Flask merupakan sebuah web framework yang ditulis menggunakan bahasa Python dan termasuk dalam kategori microframework. Dengan menggunakan Flask dan bahasa Python, dapat membuat situs web yang terstruktur dan dapat mengatur perilaku situs dengan lebih mudah.



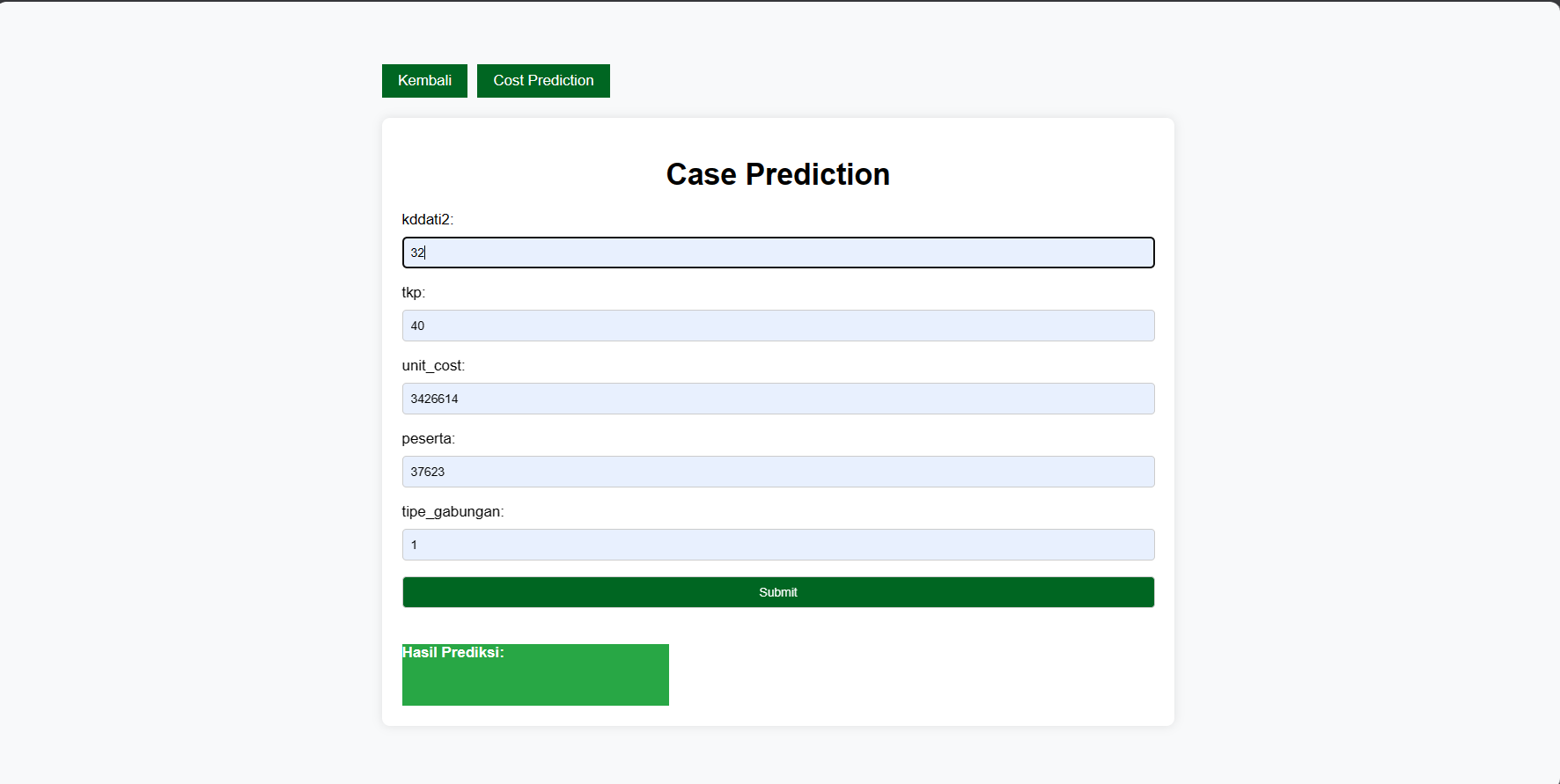
1. Create App.py



## 6.2 Web Aplication

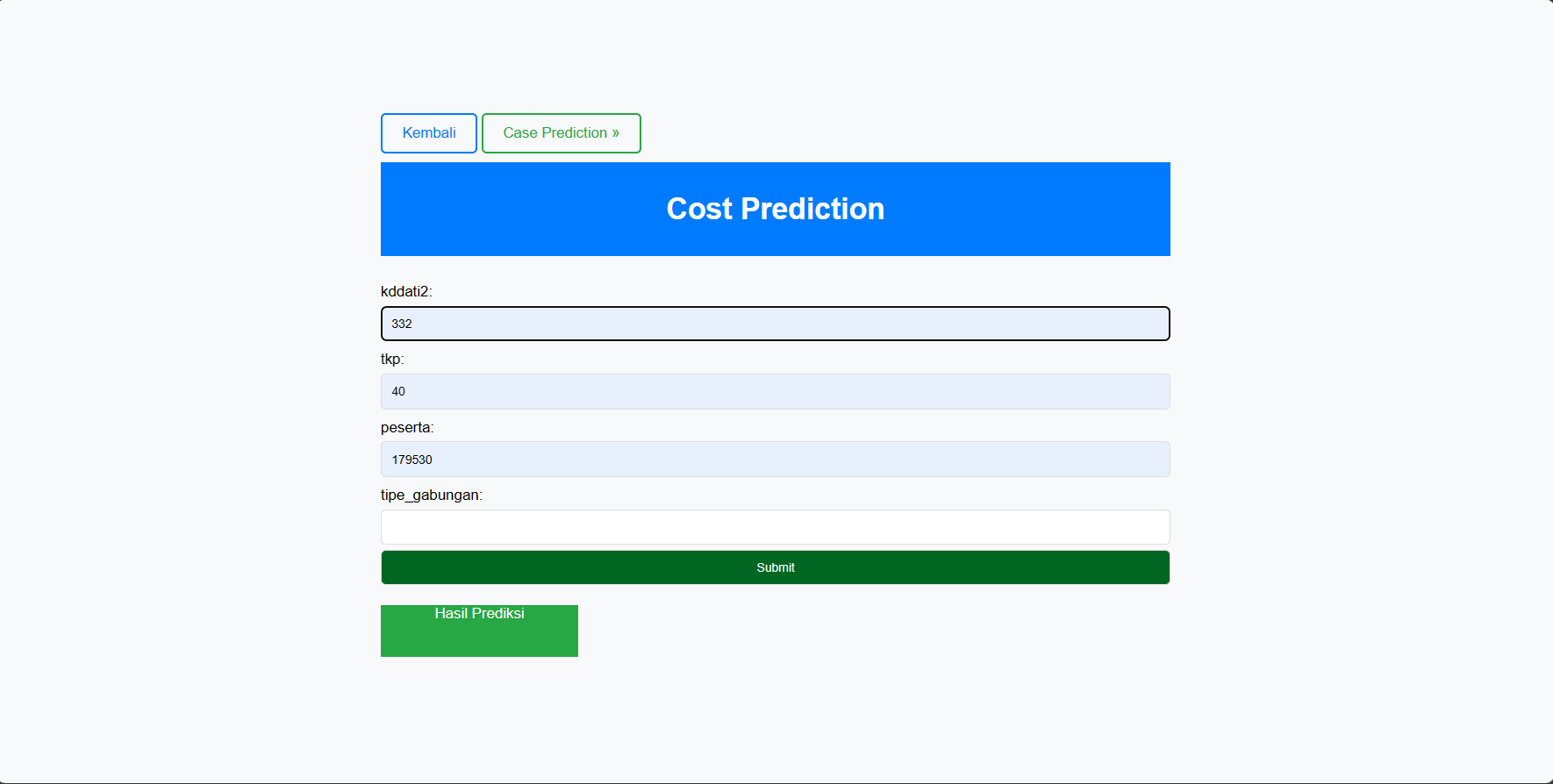
Pada bagian web application dibawah, akan ditampikan tampilan user interface dari case and cost prediction untuk data BPJS pada proyek ini

A. Case Prediction



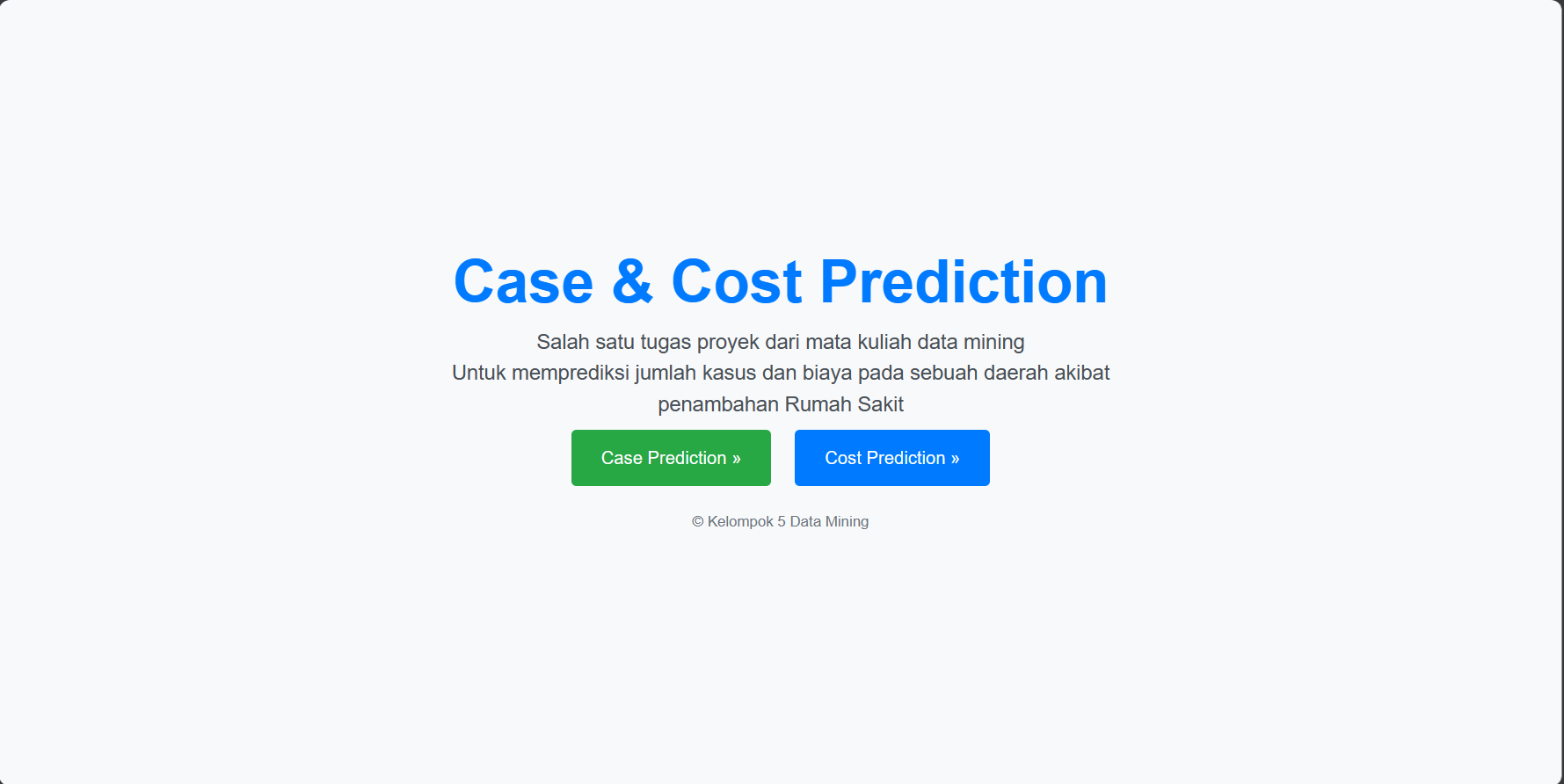
**Gambar 2. Halaman Case Prediction**

B. Cost Prediction



**Gambar 3. Cost Prediction**

C. Homepage



**Gambar 4. Homepage**