**LAPORAN PROYEK DATA MINING**

***Case and Cost Prediction with Multiple Linear Regression***



**Disusun oleh:**

12S19003 Dian Filia Matondang

12S19005 Amelia Jane Audri Lumbanraja

12S19006 Cintya Yossy Silvana Nainggolan

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**JANUARI 2022**

**BAB 1**

**BUSINESS UNDERSTANDING**

Tahap pertama pada metodologi CRISP-DM untuk melakukan prediksi jumlah kasus dan unit cost adalah *business understanding.* Adapun aktivitas yang dilakukan antara lain menentukan sasaran bisnis, memahami situasi bisnis, menerjemahkan tujuan atau sasaran bisnis ke dalam tujuan data mining.

**1.1 *Determine Business Objective***

Banyaknya rumah sakit pada zaman sekarang ini sudah tidak perlu diherankan lagi, mengingat semakin banyaknya populasi manusia di seluruh dunia yang mana membutuhkan pelayanan kesehatan. Di indonesia sendiri, rumah sakit sudah disediakan di setiap daerah, agar masyarakat tidak kesusahan dalam melakukan pemeriksaan kesehatan. Hal tersebut tentu saja tidak lepas dari teknologi yang semakin canggih di zaman sekarang ini yang digunakan untuk mengumpulkan data dari pasien yang mengunjungi rumah sakit. Banyaknya data yang tersedia serta dengan berbagai kondisi yang harus diselesaikan, menjadi hal yang menarik untuk diteliti terutama melakukan prediksi pada jumlah kasus dan unit cost pada sebuah wilayah tertentu.

Dalam melakukan *case and cost prediction* data yang digunakan adalah data BPJS Hackathon. Yang mana data tersebut merupakan data yang kategorikal, dimana sebelum digunakan data harus dianalisis terlebih dahulu sebelum diimplementasikan ke dalam model yang akan digunakan. Pada proyek ini dalam melakukan *case and cost prediction* menggunakan pendekatan yaitu metode *Multiple Linear Regression*.

*Multiple Linear Regression* merupakan bentuk penyempurnaan dari *Single Linear Regression*. *Multiple Linear Regression* adalah teknik dari bidang statistika yang mana menggunakan *variable explanatory* yang digunakan untuk memprediksi dalam bentuk *response variable*. Metode ini digunakan untuk membuat perhitungan yang dapat menjelaskan dari dua atau tiga jenis variabel yang saling berkaitan dan juga *response variable* dengan memasukkan perhitungan linear pada data yang diteliti[1].

Berdasarkan penjelasan diatas, dapat diketahui bahwa hal yang dilakukan pada penelitian ini yaitu untuk melakukan prediksi pada *cost and case* pada data BPJS Hackathon, serta hal lain yang ingin dilakukan juga yaitu melakukan pengembangan model dari *Multiple Linear Regression*.

**1.2 *Determine Project Goal***

Adapun tujuan dari pengerjaan proyek ini adalah mengembangkan sebuah model untuk menaikkan akurasi dengan penggunaan teknik dalam Data Mining untuk mengetahui case and cost prediction dari data BPJS Hackathon.

**1.3 *Produce Project Plan***

Tahapan ini merupakan tahap perencanaan untuk mencapai tujuan pengerjaan proyek penelitian “*Case and Cost Prediction with Multiple Linear Regression”* adalah sebagai berikut:

| **Tahapan** | **Waktu Pengerjaan** | **Kegiatan** |
| --- | --- | --- |
| Business Understanding | 1 Minggu | Menentukan objektif bisnis, menentukan tujuan proyek serta membuat rencana proyek. |
| Data Understanding | 1 Minggu | Mengumpulkan data yang akan digunakan, menelaah data dan melakukan validasi pada data. |
| Data Preparation | 1 Minggu | Memilih data yang akan digunakan, membersihkan data, mengkonstruksi data, menentukan label data, dan mengintegrasikan data. |
| Modeling | 2 Minggu | Membangun skenario pengujian dan membangun model. |
| Evaluation | 1 Minggu | Melakukan evaluasi hasil pemodelan dan melakukan review terhadap proses pemodelan. |
| Deployment | 1 Minggu | Membuat rencana deployment model, Monitoring and Maintenance rencana deployment model dan meninjau proyek. |

Dalam pelaksanaan proyek ini, tools yang digunakan yaitu Python, yang merupakan salah satu bahasa pemrograman yang dapat melakukan eksekusi sejumlah instruksi multiguna secara langsung (interpretatif) dengan metode orientasi objek (*Object Oriented Programming*) serta menggunakan semantik dinamis untuk memberikan tingkat keterbacaan syntax. Python juga merupakan bahasa populer yang berkaitan dengan *Data Science, Machine Learning, dan Internet of Things* (IoT).

Adapun metode atau algoritma yang akan digunakan dalam proyek ini adalah *Multiple Linear Regression*. *Multiple Linear Regression* merupakan generalisasi regresi linier sederhana, dalam arti bahwa pendekatan ini memungkinkan untuk mengevaluasi hubungan linier antara variabel respon (kuantitatif) dan beberapa variabel penjelas (kuantitatif atau kualitatif). *Multiple Linear Regression* digunakan untuk mengestimasikan hubungan antara dua atau lebih *independent variables* dan satu *dependent variable*.

**BAB 2**

**DATA UNDERSTANDING**

Tahap data understanding merupakan tahapan pemahaman terhadap data yang akan digunakan. Tahapan ini dimulai dari mengumpulkan data, mendeskripsikan data, dan memahami data yang akan digunakan dalam proyek.

**2.1 Collecting Data**

Tahap pengumpulan data merupakan tahap awal untuk menemukan data yang akan digunakan dalam penelitian. Maka dari itu dataset yang akan digunakan untuk memprediksi jumlah kasus dan unit cost pada sebuah daerah akibat penambahan Rumah Sakit kerja sama berdasarkan dataset train yaitu data case\_cost\_prediction\_train.csv.

**2.2 Describe Data**

Tahap mendeskripsikan data merupakan tahap kedua untuk mendeskripsikan dataset g digunakan untuk memprediksi jumlah kasus dan unit cost pada sebuah daerah akibat penambahan Rumah Sakit kerjasama adalah case\_cost\_prediction\_train.csv. Dataset tersebut terdiri atas 57971 observasi dan 36 variable. Berikut tabel yang membahas terkait atribut pada dataset.

| **No.** | **Nama Atribut** | **Deskripsi** |
| --- | --- | --- |
| 1 | *row\_id* | ID dari setiap data |
| 2 | *tglpelayanan* | periode bulan pelayanan di rumah sakit |
| 3 | *kddati2* | kode dari setiap kabupaten/kota |
| 4 | *tkp* | tingkat pelayanan; 30:rawat jalan; 40:rawat inap |
| 5 | peserta | jumlah peserta aktif 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 |

Dari rincian dataset tersebut, akan dilakukan EDA (Exploratory Data Analysis) terhadap dataset untuk menganalisis karakteristik utama dari dataset. Pada pengerjaan proyek ini, hanya beberapa atribut yang relevan dengan tujuan proyek penelitian, tidak semua atribut dalam dataset digunakan. Atribut yang paling sesuai untuk melakukan prediksi pada jumlah kasus dan unit cost adalah atribut *case*, *unit\_cost*, dan beberapa atribut relevan lainnya. Oleh karena itu, adapun atribut dataset yang akan digunakan untuk hipotesis adalah sebagai berikut:

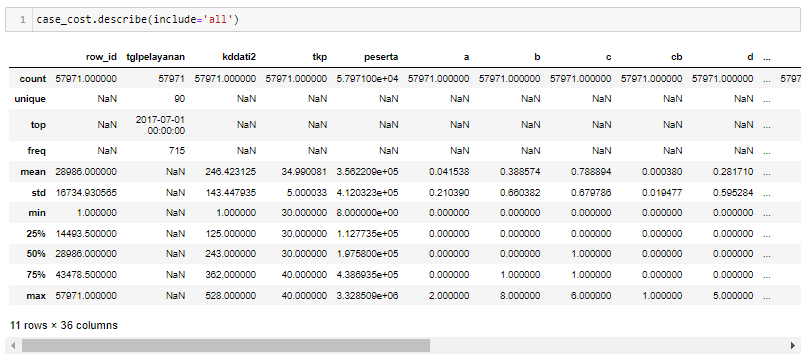
* atribut *kddati2* berguna untuk mengetahui kasus per kabupaten/kota berdasarkan kode dari setiap kabupaten/kota.
* atribut *peserta* berguna untuk mengetahui jumlah dari peserta aktif dari setiap kabupaten/kota.
* atribut *case* berguna untuk mengetahui kasus kunjungan ke rumah sakit.
* atribut *unit\_cost* berguna untuk mengetahui jumlah biaya pelayanan rumah sakit.

Dari beberapa hipotesis di atas, terdapat atribut dataset *kddati2*, *peserta*, *case*, dan *unit\_cost* yang berpengaruh pada jumlah kasus dan unit cost dan sesuai dengan tujuan proyek yaitu mengembangkan model data mining untuk melakukan prediksi pada jumlah kasus dan unit cost di sebuah daerah akibat penambahan Rumah Sakit kerja sama.

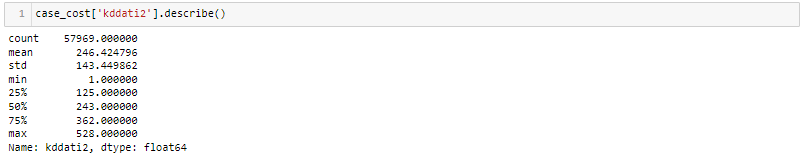
**2.3 Validation Data**

Pada tahap ini dilakukan validasi terhadap data yang akan digunakan dengan cara memeriksa kelengkapan data yang bertujuan untuk menghindari terjadinya *error* ataupun *missing value* yang terjadi akibat masalah *input data*. Oleh karena itu, dilakukan pemeriksaan terhadap atribut utama yang akan digunakan pada dataset.

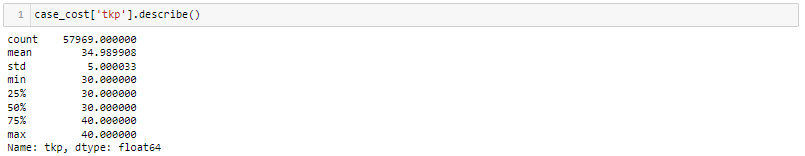
* All Atribut



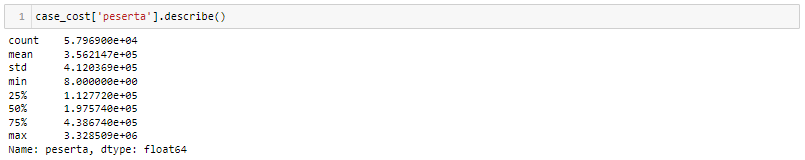
* Atribut *kddati2*

**

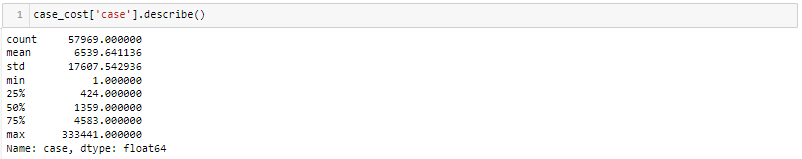
* Atribut *tkp*

**

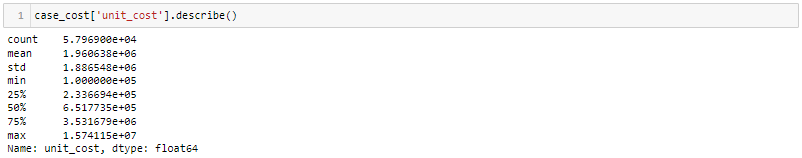
* Atribut *peserta*

**

* Atribut *case*

**

* Atribut *unit\_cost*



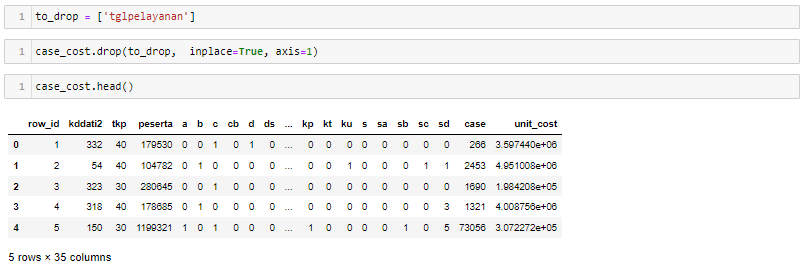
**BAB 3**

**DATA PREPARATION**

*Data preparation* dilakukan agar data yang digunakan memiliki kualitas yang baik.pada *data preparation* telah dilakukan *data cleaning, data integration, data transformation,* dan *data reduction*. *Data preparation* adalah tahap dimana data yang dikumpulkan telah dilakukan tahap CRISP-DM lalu dilanjut ke tahap persiapan data dan pemilihan variabel yang akan digunakan. Untuk kode program yang akan dibuat menggunakan software *Jupyter Notebook*.

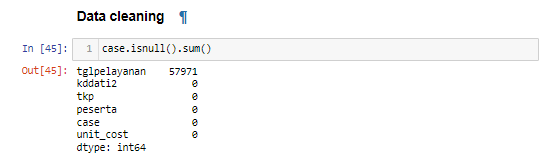
**3.1 Data Selection**

Pada tahapan ini akan dilakukan pemilihan data yang akan digunakan dalam proyek. Dalam melakukan pemilihan data, dapat memasukkan kode ***to\_drop()*** dan memasukkan data yang tidak dibutuhkan untuk pengerjaan proyek. Data yang dihapus yaitu 'tglpelayanan’.



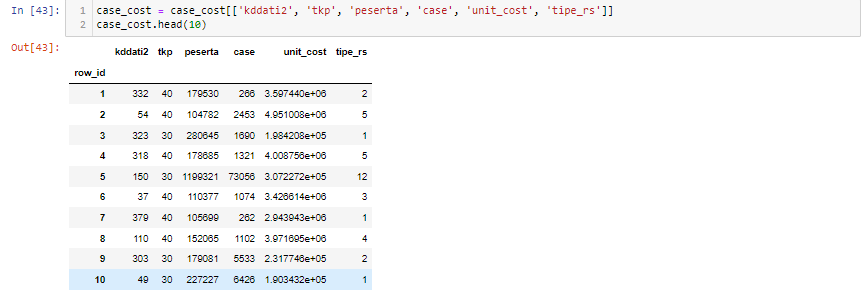
**3.2 Data Cleaning**

Data cleaning dilakukan untuk menghasilkan data yang berkualitas dengan menghapus object data yang tidak bernilai (*missing value*).

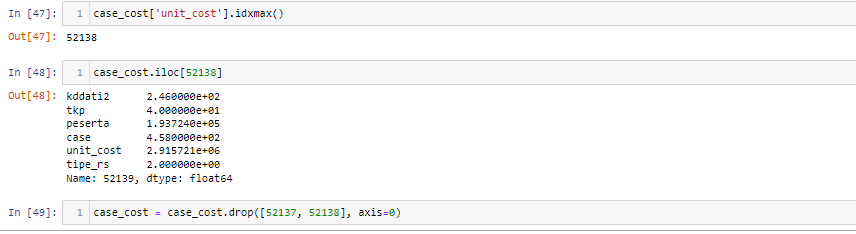


**3.3 Data Construct**

Data construct merupakan bagian dari data transformasi yang digunakan untuk merepresentasikan fitur, menentukan korelasi, dan mengintegrasi data. Representasi fitur dilakukan agar kompleksitas data berkurang, untuk memilih fitur yang lebih optimal dan meningkatkan akurasi.



Setelah itu hapus *outlier* untuk meningkatkan hasil secara signifikan



**3.4 Labeling Data**

Sebelum membangun model, data perlu untuk dilabelkan terlebih dahulu. Data BPJS Hackathon dibagi menjadi 2 kategori yaitu data training dan validation, untuk SVM problem:



**BAB 4**

**MODELLING**

Persiapan data yang akan digunakan dalam membangun model telah dilakukan pada bab sebelumnya. Pada bab ini akan dibangun model Prediction Regression dengan implementasi algoritma *Multiple Linear Regression*. Dalam kasus BPJS Hackathon ini, adapun tujuan dari pembangunan model adalah untuk melakukan prediksi jumlah kasus dan unit cost pada sebuah daerah akibat penambahan Rumah Sakit di suatu wilayah. Regression yang memiliki kata dasar regresi dengan arti sebagai proses memprediksi nilai yang kontinu. Banyak jenis algoritma yang dapat digunakan untuk memprediksi kasus ini, salah satunya dengan metode *Multiple Linear Regression*. Dengan harapan algoritma ini akan lebih mudah dipahami cara kerja nya dan diimplementasikan dalam penyelesaian proyek Data Mining. Adapun tahapan yang akan dilakukan dalam pembangunan model adalah sebagai berikut.

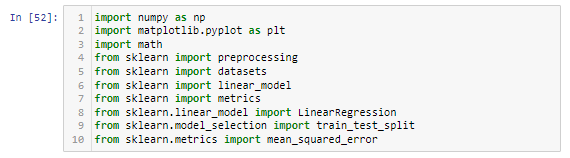
**4.1 Testing Scenario**

Sub bab ini menjelaskantentang bagaimana proses testing dilakukan. Langkah pertama, untuk case akan dibuat test berbasis di x sebagai independen dan y sebagai dependen variabelnya. Dimana x berisi atribut kddati2, tkp, peserta, unit\_cost, dan tipe\_rs. Sedangkan untuk y berisi atribut case. Langkah kedua, untuk cost akan dibuat test berbasis di x dan y. Untuk x berisi atribut kddati2, tkp, peserta, case, dan tipe\_rs sebagai variabel independennya, sedangkan untuk y berisi atribut unit\_cost sebagai variabel dependennya. Kemudian, setelah variabel ditentukan, kami menggunakan algoritma *multiple linear regression* dalam membangun model dengan *gradient boosting* untuk meningkatkan *score* akurasinya.

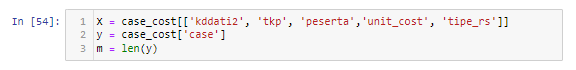
**4.2 Model Building**

Pada subbab ini akan menjelaskan tahapan-tahapan membangun model dalam mencari prediksi *cash and cost*. Berikut tahapannya:

1. *Import library* yang akan digunakan

****

1. Berikut atribut yang akan digunakan dalam menghitung nilai *case*

****

1. Bagi data menjadi *train* dan *test*

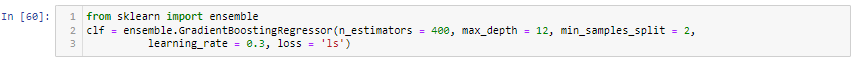
****

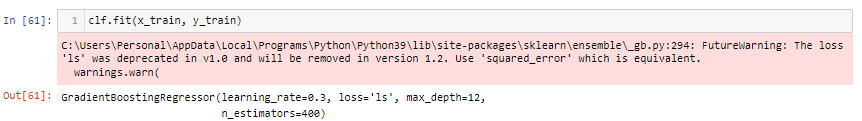
1. Dalam menghitung nilai *case*, proyek ini menggunakan algoritma *Linear Regression* dan mendapatkan score sebagai berikut.

****

****

1. Meningkatkan *score* dengan *Gradient Boosting*.

****

****

1. Untuk menghitung nilai *cost*, berikut atribut yang akan digunakan

****

1. Bagi data menjadi *train* dan *test*

****

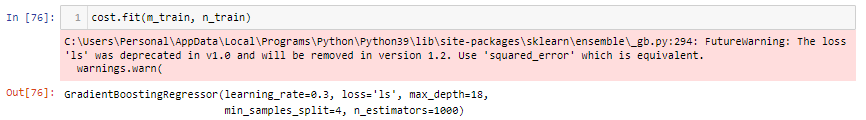
1. Dalam menghitung nilai *cost*, proyek ini menggunakan algoritma *Linear Regression* dan mendapatkan score sebagai berikut.

****

****

1. Tingkatkan *score* dengan *gradient boosting*

****

****

**BAB 5**

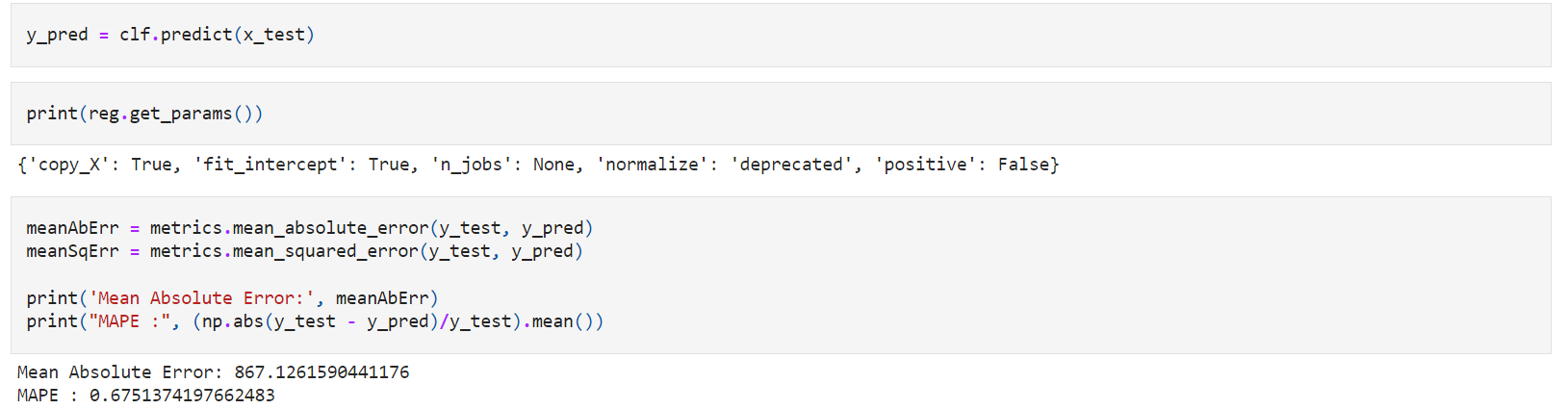
**Evaluation**

Pada tahap ini akan dilakukan evaluasi terhadap model yang digunakan yaitu *Multiple Linear Regression* dalam melakukan *cost and case prediction*. Evaluasi dilakukan untuk menilai model apakah memenuhi tujuan ada tahap *business understanding* yaitu menaikkan akurasi dalam memprediksi *case and cost*.

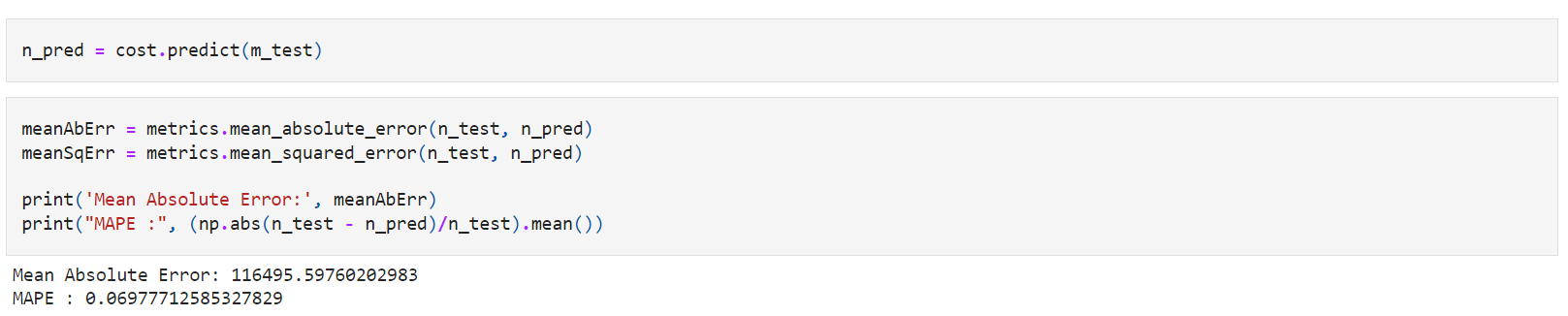
**5.1 Evaluate Results**

Pada sub bab ini dijelaskan tentang hasil dan evaluasi dari model *case* dan *cost* yang telah dibangun dalam proyek ini. Adapun hasil yang didapat adalah sebagai berikut:

1. *Case*



1. *Cost*



**5.2 Evaluate Process**

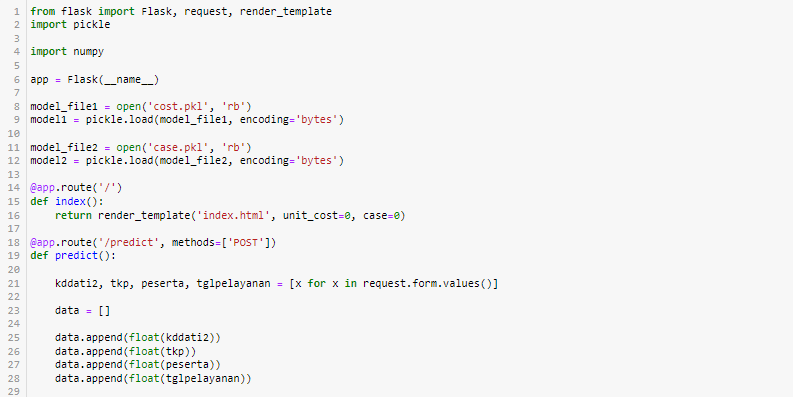
Pada sub bab ini akan dilakukan peninjauan nilai *case* dan *cost* yang didapat dari model yang telah dibangun. Untuk nilai *case* yang ditetapkan yaitu MAE < 900 dan MAPE < 90% dan hasil yang didapat sudah sesuai dengan yang ditetapkan yaitu nilai MAE = 867,1 dan MAPE = 67,5%. Sedangkan nilai *cost* yang ditetapkan yaitu MAE < 97.000 dan MAPE < 70%. Pada kasus *cost*, hanya nilai MAPE yang sesuai dengan yang ditetapkan yaitu MAPE = 6,97%, sementara MAE yang didapatkan belum sesuai dengan nilai yang ditetapkan yaitu MAE = 116.495,5.

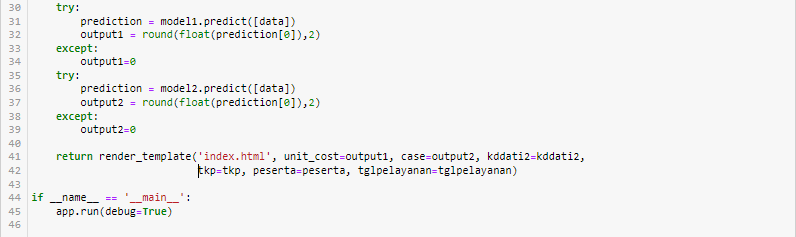
**BAB 6**

**Deployment**

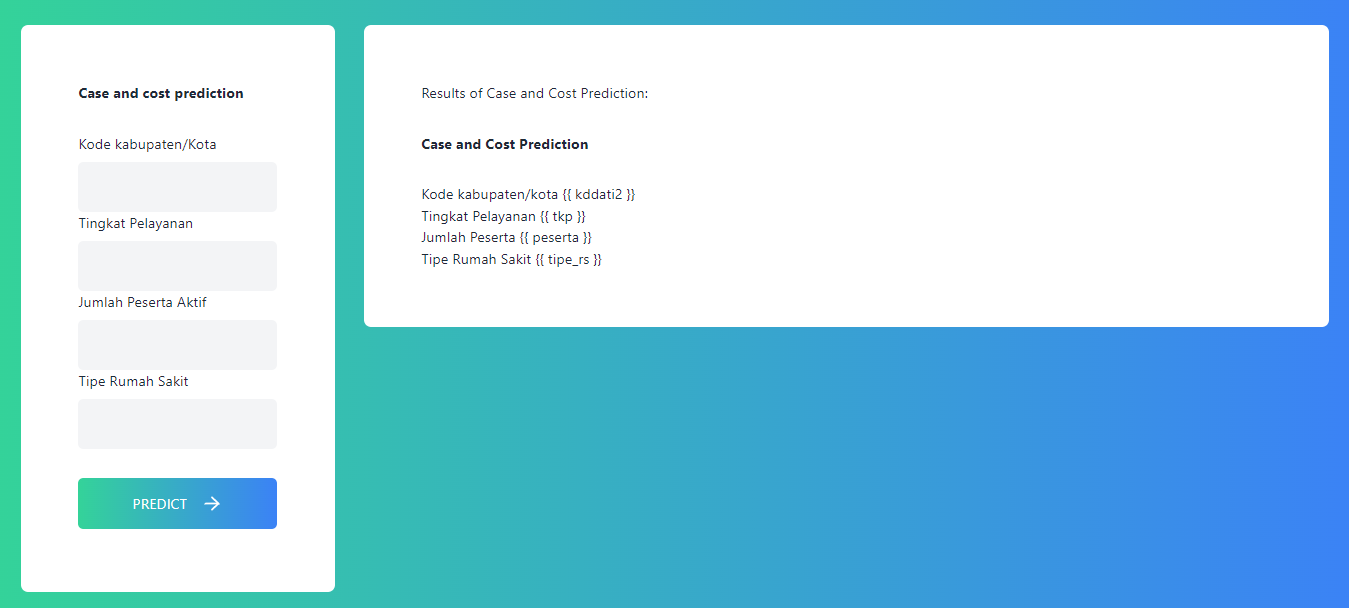
Pada bab ini, akan dijelaskan tahapan-tahapan dari deployment pada proyek yang dikerjakan. Deployment pada proyek ini dilakukan dengan menggunakan cloud platform dari heroku. Heroku merupakan sebuah media yang digunakan untuk mendeploy website proyek yang dikerjakan.

Berikut ini merupakan code untuk melakukan deployment.





Berikut ini merupakan tampilan aplikasi website yang telah dikembangkan. Untuk melakukan prediksi terhadap kasus dan biaya penanganan dapat mengisi form yang disediakan yaitu mengisi kode kabupaten atau kota, tingkat pelayanan, jumlah peserta aktif, dan tipe rumah sakit lalu menekan *button* “Predict”.



Referensi

[1] Ana Mariyam Puspitasari, Dian Eka Ratnawati, Agus Wahyu Widodo, “Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine, Vol. 2, No. 2, Februari 2018.