**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**PROPOSAL PROYEK AKHIR DATA MINING**

**Prediksi Biaya Pasien pada Fasilitas Kesehatan Tingkat Pertama (FKTP) menggunakan Algoritma Regresi Multi Linear**

**`**

**OLEH:**

|  |  |
| --- | --- |
| 12S20013 | Lydia Tesalonika |
| 12S20020 | Wahyu Krisdangolyanti Simamora |
| 12S20053 | Andri Anjelia Hutapea |

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**LAGUBOTI**

**2023**

# DAFTAR ISI

[DAFTAR ISI ii](#_Toc155394718)

[DAFTAR TABEL iii](#_Toc155394719)

[DAFTAR GAMBAR iv](#_Toc155394720)

[BAB I BUSINESS UNDERSTANDING 1](#_Toc155394721)

[1.1 *Determine Business Objective* 1](#_Toc155394722)

[1.2 *Determine Project Goal* 2](#_Toc155394723)

[1.3 *Produce Project Plan* 2](#_Toc155394724)

[BAB II DATA UNDERSTANDING 5](#_Toc155394725)

[2.1 Data Collection 5](#_Toc155394726)

[2.2 Data Description 5](#_Toc155394727)

[2.3 Data Validation 7](#_Toc155394728)

[BAB III DATA PREPARATION 11](#_Toc155394729)

[3.1 Data selection 11](#_Toc155394730)

[3.2 Data Cleaning 13](#_Toc155394731)

[3.2.1 Missing and Duplicate Values 13](#_Toc155394732)

[BAB IV BUILD MODEL 21](#_Toc155394733)

[4.1 Select Modeling Technique 21](#_Toc155394734)

[4.2 Model Building 21](#_Toc155394735)

[BAB V MODEL EVALUATION 24](#_Toc155394736)

[5.1 Result Evaluation 24](#_Toc155394737)

[5.2 Result Review 24](#_Toc155394738)

[BAB VI DEPLOYMENT 25](#_Toc155394739)

[6.1 Deployment 25](#_Toc155394740)

[6.2 Web Application 26](#_Toc155394741)

[LAMPIRAN 27](#_Toc155394742)

# DAFTAR TABEL

[Tabel 1 Produce Project Plan 2](#_Toc155395718)

[Tabel 2 Data Description 5](#_Toc155395719)

[Tabel 3 Data Selection 11](#_Toc155395720)

# DAFTAR GAMBAR

[Gambar 1 Jenis Faskes 7](#_Toc155395739)

[Gambar 2 Tipe Faskes 8](#_Toc155395740)

[Gambar 3 Tingkat Layanan 8](#_Toc155395741)

[Gambar 4 Nama Tindakan 9](#_Toc155395742)

[Gambar 5 Biaya Tagih 9](#_Toc155395743)

[Gambar 6 Biaya Verifikasi 10](#_Toc155395744)

[Gambar 7 Hasil Run Data Selection 12](#_Toc155395745)

[Gambar 8 Hasil Run Data Selection 2 12](#_Toc155395746)

[Gambar 9 Hasil Run Data Selection 3 13](#_Toc155395747)

[Gambar 10 Missing dan Duplicate Values 14](#_Toc155395748)

[Gambar 11 Menampilkan Data dari Dataset 14](#_Toc155395749)

[Gambar 12 Hasil Run setelah Menampilkan Data 15](#_Toc155395750)

[Gambar 13 Hasil Run setelah Menampilkan Data 2 16](#_Toc155395751)

[Gambar 14 Menghilangkan Baris Duplikat 16](#_Toc155395752)

[Gambar 15 Mengganti Nilai yang Hilang (Missing Values) 16](#_Toc155395753)

[Gambar 16 Hasil Mengganti Nilai yang Hilang (Missing Values) 17](#_Toc155395754)

[Gambar 17 Menghapus Baris dengan Nilai yang Hilang 18](#_Toc155395755)

[Gambar 18 Hasil Run Menghapus Baris dengan Nilai yang Hilang 18](#_Toc155395756)

[Gambar 19 Hasil Run Menghapus Baris dengan Nilai yang Hilang 2 18](#_Toc155395757)

[Gambar 20 Menghapus Kolom 18](#_Toc155395758)

[Gambar 21 Menyaring data dan Menampilkan Isi 19](#_Toc155395759)

[Gambar 22 Hasil Menyaring data dan Menampilkan Isi 19](#_Toc155395760)

[Gambar 23 Hasil Menyaring data dan Menampilkan Isi 2 20](#_Toc155395761)

[Gambar 24 Menghapus Data 20](#_Toc155395762)

[Gambar 25 Mengimpor pustaka-pustaka 21](#_Toc155395763)

[Gambar 26 Read Data CSV 22](#_Toc155395764)

[Gambar 27 View Data Information 22](#_Toc155395765)

[Gambar 28 Describe the Data 22](#_Toc155395766)

[Gambar 29 Check Null Data 22](#_Toc155395767)

[Gambar 30 Variabel Independen 22](#_Toc155395768)

[Gambar 31 Subset Training dan Testing 23](#_Toc155395769)

[Gambar 32 Melatih Model dengan Testing 23](#_Toc155395770)

[Gambar 33 Menghitung Skor (R-Squared) 23](#_Toc155395771)

[Gambar 34 Menghitung Akurasi dan Regresi Ensembel Gradient Boosting 23](#_Toc155395772)

[Gambar 35 Result Evaluation 24](#_Toc155395773)

[Gambar 36 Model.pkl 25](#_Toc155395774)

[Gambar 37 FlashWebApp 25](#_Toc155395775)

[Gambar 38 Code untuk Memanggil Model 25](#_Toc155395776)

[Gambar 39 Run App 26](#_Toc155395777)

[Gambar 40 Web Application Sebelum Input Data 26](#_Toc155395778)

[Gambar 41 Web Application Setelah Input Data 26](#_Toc155395779)

# BAB I

# BUSINESS UNDERSTANDING

Business Understanding merupakan langkah pertama yang dilakukan dalam metodologi untuk memprediksi Biaya Pasien pada FKTP dengan Multi linear Regression Algorithm adalah untuk memehami bisnisnya. Bab ini akan menjelaskan pemahaman konten dan kebutuhan operasi data mining dari perspektif bisnis. Kegiatan ini meliputi Determine Business Objective, Determine Projeck Goal, dan Project Plan.

## *Determine Business Objective*

Dalam proyek ini, Kami akan menyelidiki dan menganalisis sampel data yang berkaitan dengan Badan Penyelenggara Jaminan Sosial Kesehatan (BPJS Kesehatan) dalam rangka proyek ini. Tujuan utama BPJS Kesehatan adalah memberikan jaminan kesehatan kepada semua orang di Indonesia. Data yang akan kami analisis mencakup berbagai elemen terkait, seperti informasi tentang kepersertaan, FKTP, FKTRTL, FKTP non kapitasi, dan diagnosissekunder. Dataset yang dipilih pada proyek ini adalah dataset FKTP-non kapitasi untuk tahun 2021.

BPJS (Badan Penyelenggara Jaminan Sosial), merupakan Badan Hukum Publik yang bertanggung jawab atas jaminan perlindungan dan pelayanan kesehatan bagi pengguna atau masyarakat.BPJS Kesehatan menghimpun dan menyimpan banyak data, termasuk keanggotaan peserta, klaim pengobatan, biaya layanan kesehatan, dan informasi lainnya. Analisis dataset ini dapat memberi tahu kita tentang tren kesehatan masyarakat, meningkatkan efisiensi sistem, dan membuat kebijakan yang lebih baik.

Metode Regresi adalah metode statistik yang digunakan untuk menganalisis hubungan antara dua atau lebih variabel. Hal yang dipelajari disebut variabel respon, sedangkan hal yang mempengaruhi variabel respon disebut variabel prediktor. Variabel respon adalah variabel ingin diketahui pengaruhnya terhadap variabel lain. Variabel respon diketahui sebagai variabel terikat, variabel dependent atau variabel keluaran. Sedangkan, variabel Prediktor adalah variabel yang memiliki pengaruh terhadap variabel respon. Variabel prediktor disebut sebagai variabel bebas, variabel independen atau variabel masukan.

Regresi Multi Linear adalah cara untuk memahami hubungan antara satu variabel yang ingin diprediksi dengan dua atau lebih variabel yang memungkinannya. Misalnya, jika kita ingin memprediksi harga rumah, kita bisa melibatkan variabel seperti luas tanah, jumlah kamar tidur, dan jarak ke pusat kota.

Algoritma ini mencari persamaan matematis yang paling baik mencocokkan pola hubungan di antara variabel-variabel tersebut. Tujuannya adalah menemukan seberapa besar setiap variabel berkontribusi terhadap hasil akhir. Dengan koefisien yang ditemukan, kita dapat membuat prediksi berdasarkan nilai-nilai variabel independen. Regresi Multi Linear sangat berguna untuk memahami dan memprediksi hubungan kompleks dalam berbagai bidang, termasuk ekonomi dan ilmu sosial.

## *Determine Project Goal*

Melalui analisis data BPJS Kesehatan, proyek ini bertujuan memberikan beberapa keuntungan yang signifikan. Pertama, analisis tersebut diharapkan dapat memberikan wawasan yang diperlukan untuk meningkatkan efisiensi dan efektivitas layanan kesehatan kepada peserta BPJS. Dengan begitu, upaya ini akan membantu optimalisasi layanan kesehatan sesuai dengan kebutuhan peserta. Selain itu, tujuan proyek ini adalah memberikan kontribusi penting dalam proses perencanaan kebijakan kesehatan.

Diharapkan analisis data mendalam ini akan menyediakan informasi yang relevan dan akurat untuk perumusan kebijakan yang lebih baik. Dengan pemahaman yang lebih baik tentang dinamika partisipasi dan kebutuhan layanan kesehatan, diharapkan kebijakan yang diimplementasikan dapat lebih sesuai dengan kebutuhan masyarakat yang dilayani oleh BPJS Kesehatan. Hal ini dapat memberikan gambaran mengenai faktor-faktor apa saja yang berkontribusi terhadap biaya perawatan pasien di FKTP.

Tujuan akhirnya adalah menghasilkan model prediksi yang dapat membantu pihak BPJS Kesehatan dalam merencanakan dan mengelola alokasi sumber daya dengan lebih efektif. Dengan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi biaya perawatan, BPJS Kesehatan dapat mengoptimalkan pendanaan dan meningkatkan efisiensi layanan kesehatan di FKTP.

## *Produce Project Plan*

Produce plant pada proyek ini dapat dilihat pada tabel dibawah ini yang terdiri dari aktivitas, detail dan durasi pengerjaan proyek, dapat dilihat pada tabel dibawah ini:

Tabel 1 Produce Project Plan

|  |  |  |
| --- | --- | --- |
| **Aktivitas** | **Detail** | **Durasi** |
| Pemilihan Kasus | Pemilihan kasus | 1 Hari |
| Pemilihan Algoritma | 1 Hari |
| Business Understanding | Menentukan Objektif Bisnis | 2 Hari |
| Menentukan Tujuan Bisnis | 2 Hari |
| Membuat Rencana proyek | 2 Hari |
| Data Understanding | Mengumpulkan Data | 1 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 | 3 Hari |
| Mengintegrasikan Data | 5 hari |
| Modeling | Membangun Skenario Pengujian | 5 Hari |
| Membangun Model | 3 Hari |
| Model Evaluation | Mengevaluasi Hasil Pemodelan | 2 Hari |
| Melakukan Review Proses Pemodelan | 4 Hari |
| Deployment | Melakukan Deployment Model | 1 Hari |
| Membuat laporan akhir Proyek | - |

Dalam pelaksanaan proyek penelitian, Python digunakan sebagai bahasa pemrograman utama. Python adalah bahasa pemrograman yang sangat populer dalam konteks Data Science, Machine Learning, dan Internet of Things (IoT). Python juga dikenal sebagai bahasa pemrograman yang dapat menjalankan sejumlah instruksi secara langsung dengan metode orientasi objek (Object Oriented Programming) dan menggunakan semantik dinamis untuk meningkatkan tingkat keterbacaan syntax.

Proyek ini juga mengadopsi metode algoritma Multi Linear untuk mengevaluasi hubungan positif dan negatif antara variabel independen dan variabel dependen. Algoritma ini memungkinkan prediksi nilai variabel dependen berdasarkan perubahan nilai variabel independen. Cara kerja Multi Regresi melibatkan penyusunan grafik garis antara dua variabel data, x dan y. Variabel independen, x, terletak di sepanjang sumbu horizontal, dan tujuan utama analisis regresi ini adalah untuk memahami hubungan sebab-akibat antara variabel satu dengan variabel lainnya.

# BAB II

# DATA UNDERSTANDING

Pada tahapan *data understanding/* pemahaman data, akan dijelaskan bahwa tahapan ini dimulai dengan pengumpulan data awal untuk mengetahui karakteristik dan variabel-variabel utama data. Ini memungkinkan tim untuk membuat hipotesis awal tentang informasi yang tersembunyi dan mendapatkan pemahaman awal tentang data.

## Data Collection

Pengumpulan data merupakan tahap awal yang dilakukan guna memperoleh data yang akan digunakan. Oleh karena itu dataset yang akan digunakan untuk memprediksi biaya Pasien ke Fasilitas Kesehatan Tahap Pertama (FKTP) dengan Algoritma Regresi Multi Linear adalah dataset fktp-non kapitas tahun 2021 yaitu fktpnonkapitasi.dta.

## Data Description

Dataset yang digunakan dalam melakukan prediksi biaya pasien ke FKTP adalah terdiri beberapa data dan 21 variabel. Berikut akan dibahas mengenai atribut yang tersedia pada dataset:

Tabel 2 Data Description

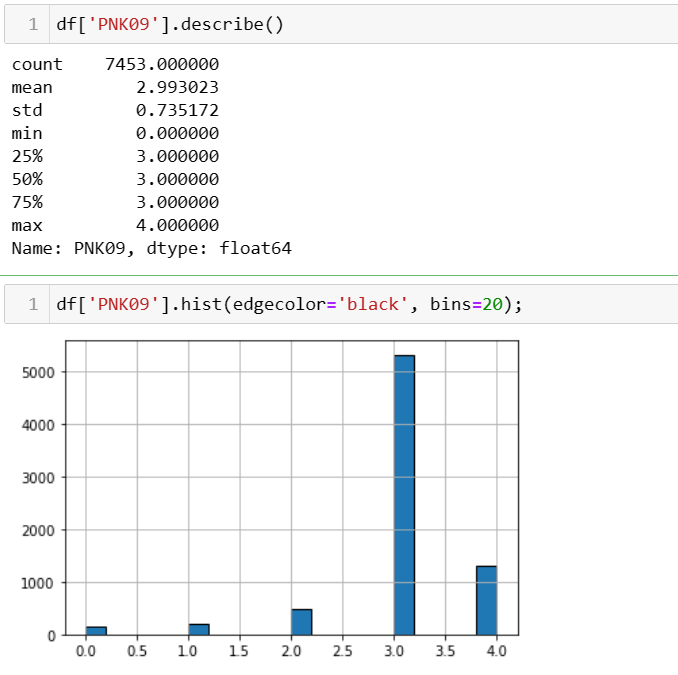
|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Variabel** | **Label variabel** | **Deskripsi** |
|  | PSTV01 | Nomor Peserta | Nomor identifikasi peserta yang bersifat unik dan telah dideidentifikasi untuk melindungi identitas peserta sebenarnya |
|  | PSTV02 | Nomor keluarga | Nomor yang mengidentifikasi kepala keluarga dalam sampel dan berfungsi sebagai penanda keluarga (peserta BPJS Kesehatan dalam satu keluarga memiliki nomor kepala keluarga yang sama) |
|  | PSTV15 | Bobot | Faktor pengali yang menggambarkan jumlah individu di dalam populasi diwakili oleh individu di dalam sampel |
|  | PNK02 | ID Kunjungan | Nomor identifikasi unik untuk menandakan setiap kunjungan FKTP oleh peserta |
|  | PNK03 | Tanggal kunjungan | Tanggal melakukan kunjungan |
|  | PNK04 | Tanggal tindakan | Tanggal melakukan tindakan |
|  | PNK05 | Tanggal pulang | Tanggal menyelesaikan kunjungan |
|  | PNK06 | Provinsi faskes | Provinsi tempat peserta mengakses fasilitas kesehatan tingkat pratama (FKTP) |
|  | PNK07 | Kode Kab/Kota faskes | Kabupaten/kota tempat peserta mengakses fasilitas kesehatan tingkat pratama (FKTP) |
|  | PNK08 | Kepemilikan faskes | Kepemilikan dari fasilitas kesehatan tingkat pratama (FKTP) tempat peserta berkunjung |
|  | PNK09 | Jenis faskes | Jenis dari fasilitas kesehatan tingkat pertama (FKTP) tempat peserta berkunjung |
|  | PNK10 | Tipe faskes | Tipe dari fasilitas kesehatan tingkat pertama (FKTP) tempat peserta berkunjung |
|  | PNK11 | Tingkat layanan | Tingkat layanan yang diterima peserta di FKTP |
|  | PNK12 | Segmen peserta | Segmen peserta saat mengakses FKTP |
|  | PNK13 | Kode dan Nama diagnosis berdasarkan ICD-10 (3 digit) | Kode dan nama diagnosis berdasarkan 3 digit pertama kode ICD 10 yang diperoleh dari hasil input sistem informasi BPJS Kesehata |
|  | PNK13A | Kode diagnosis berdasarkan ICD-10 (3 digit) | Kode dan nama diagnosis berdasarkan 3 digit pertama kode ICD 10 yang diperoleh dari hasil input sistem informasi BPJS Kesehatan |
|  | PNK14 | Kode diagnosis (3-5 digit) | Kode diagnosis menurut ICD 10 (jumlah digit tidak sama pada semua observasi dengan rentang 3-5 digit kode ICD 10) |
|  | PNK15 | Nama Diagnosis | Nama diagnosis yang terbaca oleh sistem informasi BPJS Kesehatan berdasarkan kode diagnosis yang ter-input dalam sistem |
|  | PNK16 | Nama Tindakan | Nama jenis tindakan yang dilakukan kepada pasien |
|  | PNK17 | Biaya tagih | Biaya yang ditagihkan fasilitas kesehatan untuk setiap ID kunjunga |
|  | PNK18 | Biaya verifikasi | Biaya yang diverifikasi BPJS Kesehatan untuk setiap nomor ID Kunjungan |

Dataset fktp-non kapitasi akan melalui sebuah proses untuk melakukan peninjauan dan analisis data guna memperoleh pemahaman dasar mengenai data tersebut yang disebut sebagai *Exploratory Data Analysis (EDA).* Adapun tujuan dilakukannya EDA ini adalah untuk memahami struktur data dan menentukan variabel yang relevan untuk digunkan pada proyek.

## Data Validation

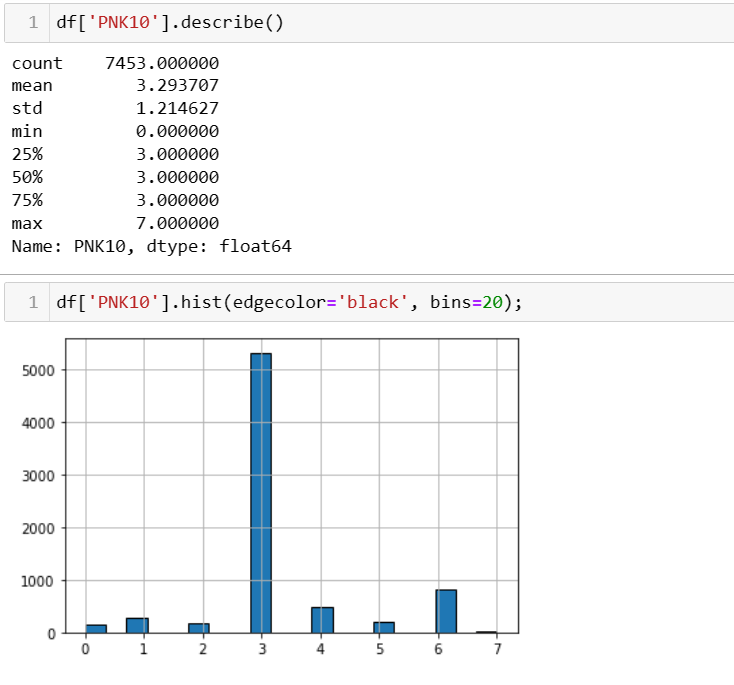
Pada tahap ini, data yang akan digunakan divalidasi untuk menghindari kesalahan atau masalah input yang tidak memiliki nilai. Validasi data merupakan proses memastikan bahwa data yang digunakan dalam analisis atau model memiliki kualitas yang baik dan dapat diandalkan. Terdapat 6 variabel yang digunakan dalam program ini yaitu 5 variabel bebas dan 1 variabel terikat

1. Variabel PNK09 (Jenis Faskes)



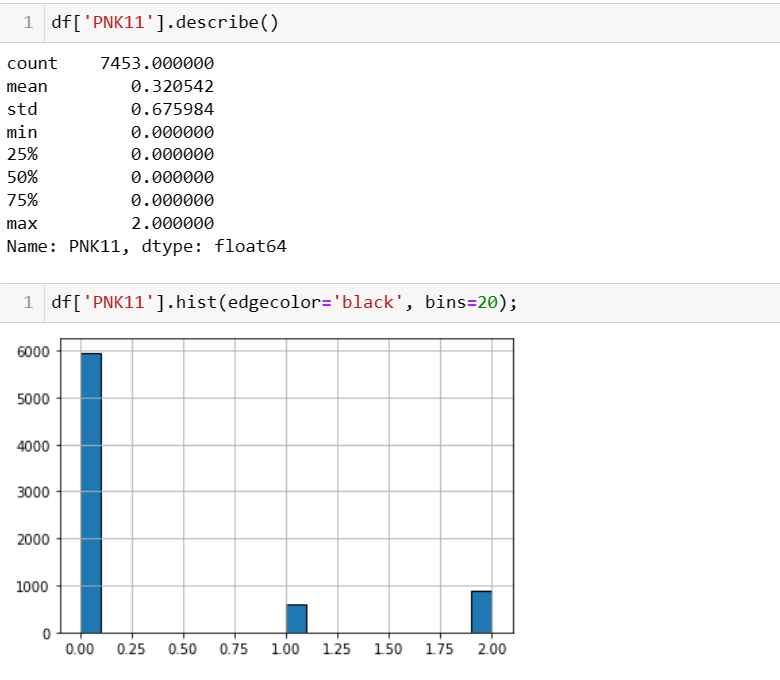
Gambar 1 Jenis Faskes

1. Variabel PNK10 (Tipe Faskes)



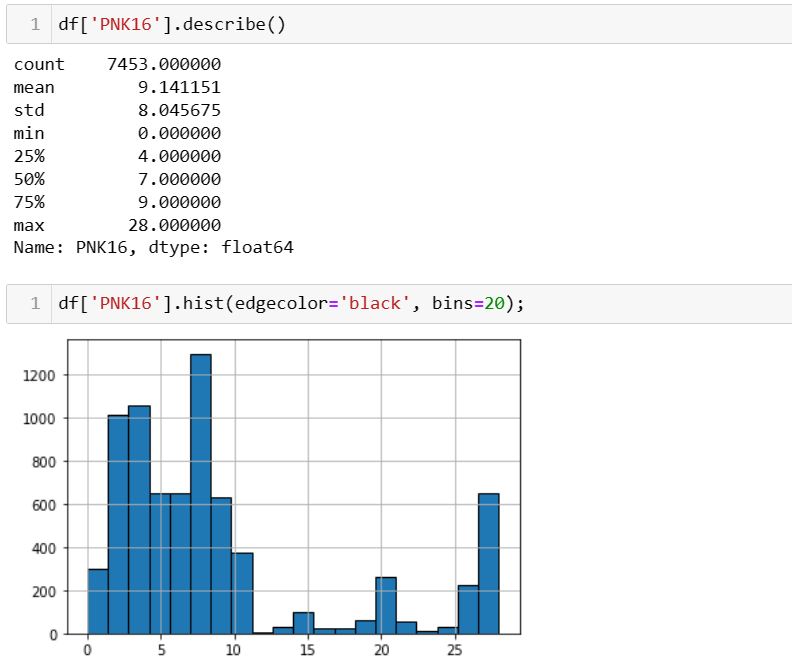
Gambar 2 Tipe Faskes

1. Variabel PNK11 (Tingkat Layanan)



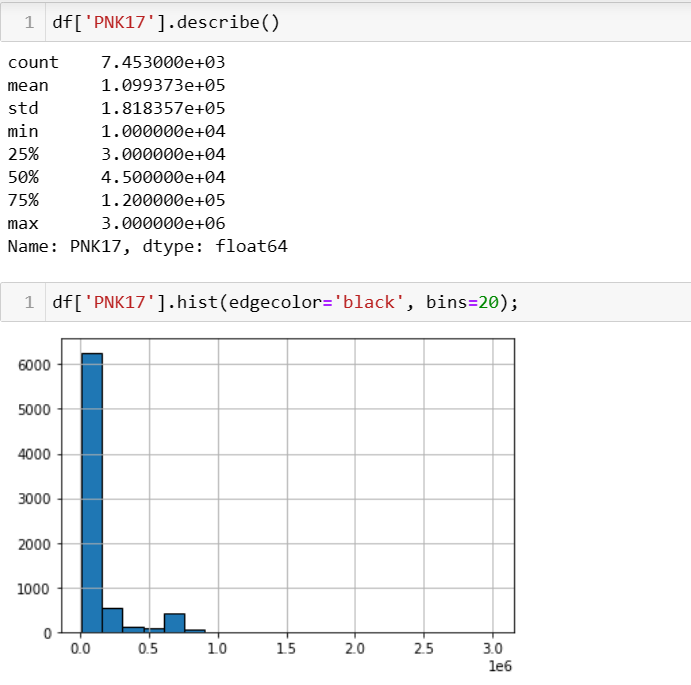
Gambar 3 Tingkat Layanan

1. Variabel PNK16 (Nama Tindakan)



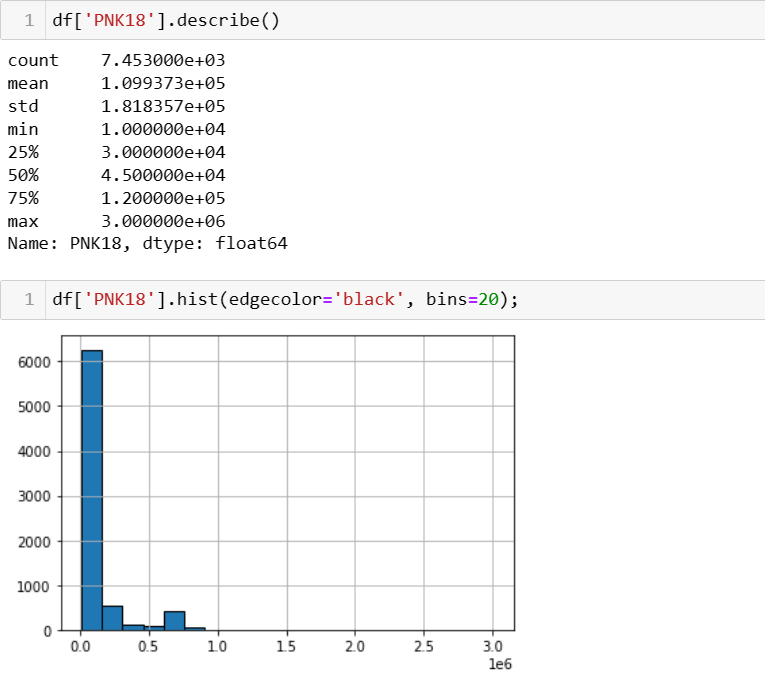
Gambar 4 Nama Tindakan

1. Variabel PNK17 (Biaya Tagih)



Gambar 5 Biaya Tagih

1. Variabel PNK18 (Biaya Verifikasi)



Gambar 6 Biaya Verifikasi

# BAB III

# DATA PREPARATION

Pada tahapan *data preparation* tim akan melakukan *data preparation*  dengan melakukan langkah-langkah untuk memastikan bahwa kualitas baik pada data. Proses dalam *data preparation* ini terdiri dari data selection, *data cleaning, data transformation,* dan *data reduction.*

## Data selection

Dalam tahapan ini akan dipilih data yang dibutuhkan. Maka sebelumnya kita harus membaca dan menganalisis data yang telah tersedia yaitu dataset TB20152021\_fktpnonkapitasi.dta. Data yang digunakan adalah data FKTP non kapitasi berdasarkan karakteristik fasilitas kesehatan tahun 2021. Dan untuk memperoleh data khusus tahun 2021 akan dilakukan data cleaning pada tahapan proses selanjutnya.

Tabel 3 Data Selection

|  |
| --- |
| df = pd.read\_stata("TB20152021\_fktpnonkapitasi.dta")  #melihat dataset  print(df) |

|  |
| --- |
|  |
| Gambar 7 Hasil Run Data Selection    Gambar 8 Hasil Run Data Selection 2    Gambar 9 Hasil Run Data Selection 3 |

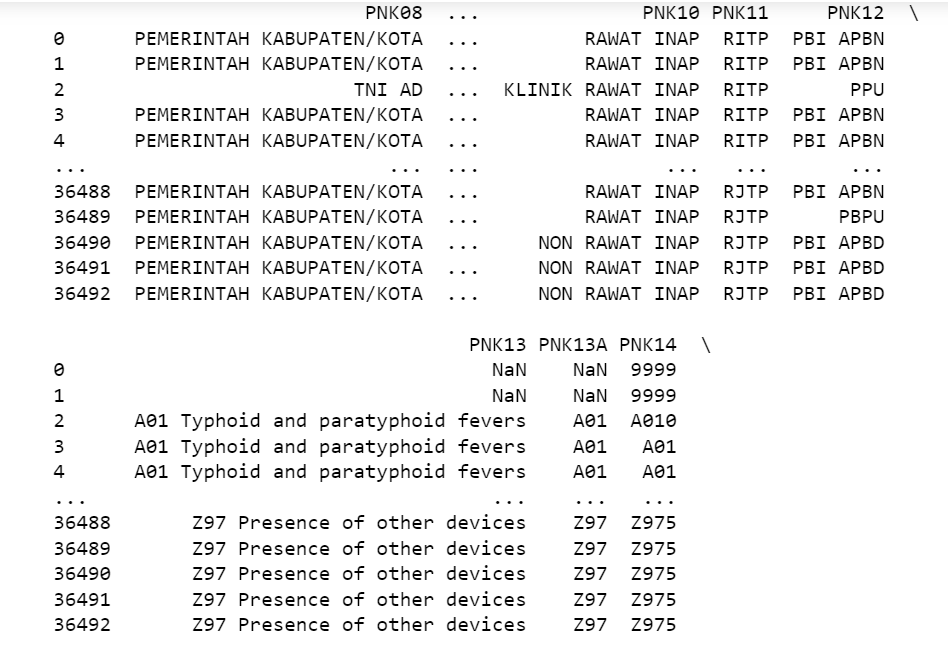
## Data Cleaning

Untuk menghasilkan data yang berkualitas, dibutuhkan *cleaning data* untuk membersihkan data. Pada tahapan ini, 3 proses utama yang dilakukan adalah

* Mengisi nilai-nilai yang hilang
* Memeriksa dan menghapus data duplicate
* Menghapus data tahun 2015-2019 karena data yang digunakan adalah data tahun 2021

### Missing and Duplicate Values

Pada dataset terdapat missing values yaitu nilai ‘NaN’ dan akan dilakukan cleaning data duplicate values pada dataset.



Gambar 10 Missing dan Duplicate Values

Pada tahapan ini, proses cleaning akan menampilkan data dari dataset dan selanjutnya akan menghapus data duplikat,



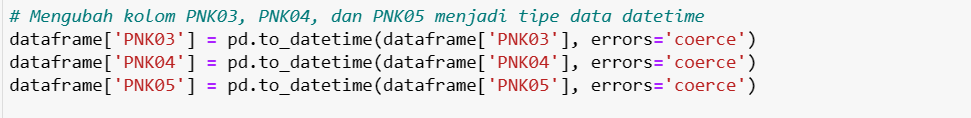
Gambar 11 Menampilkan Data dari Dataset

|  |
| --- |
| Inserting image...    Gambar 12 Hasil Run setelah Menampilkan Data    Gambar 13 Hasil Run setelah Menampilkan Data 2    Gambar 14 Menghilangkan Baris Duplikat    Gambar 15 Mengganti Nilai yang Hilang (Missing Values)    Inserting image...    Gambar 16 Hasil Mengganti Nilai yang Hilang (Missing Values) |

Selanjutnya adalah dengan menghapus data dengan nilai yang hilang:

|  |
| --- |
| Gambar 17 Menghapus Baris dengan Nilai yang Hilang    Gambar 18 Hasil Run Menghapus Baris dengan Nilai yang Hilang    Gambar 19 Hasil Run Menghapus Baris dengan Nilai yang Hilang 2 |

Berikut code untuk mengubah type data yang berisi `tanggal-bulan-tahun` menjadi type data *datetime*



Gambar 20 Menghapus Kolom

Code untuk menampilkan data yang hanya memiliki tahun 2021, karena sample yang digunakan adalah data dengan tahun 2021.

|  |
| --- |
| Gambar 21 Menyaring data dan Menampilkan Isi |
| Gambar 22 Hasil Menyaring data dan Menampilkan Isi |
| Gambar 23 Hasil Menyaring data dan Menampilkan Isi 2 |

Menghapus data dengan tahun 2015, 2016, 2017, 2018, 2019 dan 2020.

|  |
| --- |
| Gambar 24 Menghapus Data |

# BAB IV

# BUILD MODEL

Pada bab sebelumnya, tim telah mempersiapkan data untuk dapat digunakan dalam membangun model.

## Select Modeling Technique

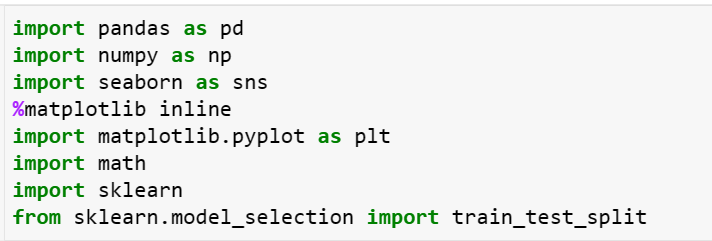
Pada proses ini, tim akan membangun model Regresi, yaitu Regresi Multi Linear. Alasan tim menerapkan algoritma ini adalah untuk membantu tim dalam membuat sebuah keputusan. Banyak algoritma yang dapat melakukan Prediksi, namun tidak semua algoritma memiliki performansi yang sama karena dipengaruhi oleh kualitas data yang dimiliki. Untuk itu tim akan memilih model Regresi yang kinerjanya lebih baik dari pada regresi lainnya.

Metode Regresi adalah metode statistik yang digunakan untuk menganalisis hubungan antara dua atau lebih variabel. Hal yang dipelajari disebut variabel respon, sedangkan hal yang mempengaruhi variabel respon disebut variabel prediktor. Variabel respon adalah variabel ingin diketahui pengaruhnya terhadap variabel lain. Variabel respon diketahui sebagai variabel terikat, variabel dependent atau variabel keluaran. Sedangkan, variabel Prediktor adalah variabel yang memiliki pengaruh terhadap variabel respon. Variabel prediktor disebut sebagai variabel bebas, variabel independen atau variabel masukan.

## Model Building

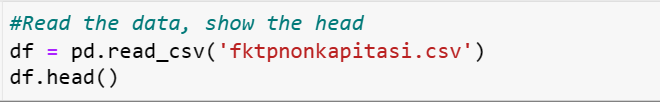
Pada sub bab ini akan menggambarkan bagaimana tahapan pembangunan model, yang mana ini digunakan untuk mencari hasil dari prediksi biaya pasien pada FKTP. Berikut merupakan proses membangun model:

1. Mengimpor pustaka-pustaka yang dibutuhkan dalam analisis data dan pembangunan model



Gambar 25 Mengimpor pustaka-pustaka

1. Membaca data dari file CSV yang digunakan yaitu fktpnonkapitasi.csv ke dalam sebuah DataFrame menggunakan pustaka Pandas (‘pd’).



Gambar 26 Read Data CSV

1. Menampilkan informasi umum tentang dataset: jumlah baris dan kolom, tipe data setiap kolom.



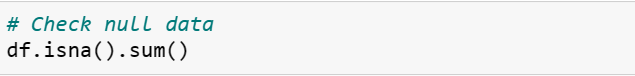
Gambar 27 View Data Information

1. Menyajikan ringkasan statistik deskriptif dari dataset



Gambar 28 Describe the Data

1. Memeriksa nilai null(nilai kosong) dalam setiap kolom.



Gambar 29 Check Null Data

1. Memisahkan fitur (variabel independen) dan variabel target(variabel dependen) dari dataset. Variabel x berisi atribut yang digunakan untuk melakukan prediksi atau analisis, sedangkan variabel y berisi nilai yang ingin diprediksi atau dianalisis. Variabel m akan menyimpan jumlah baris data yang dapat digunakan.



Gambar 30 Variabel Independen

1. Membagi data menjadi subset pelatihan (training) dan pengujian (testing)



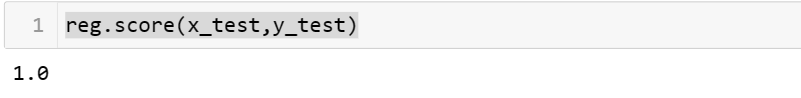
Gambar 31 Subset Training dan Testing

1. Melatih model dengan data pelatihan, setelah proses pelatihan selesai maka model regresi linear akan digunakan untuk melakukan prediksi pada data baru atau dievaluasi kinerjanya pada data pengujian.



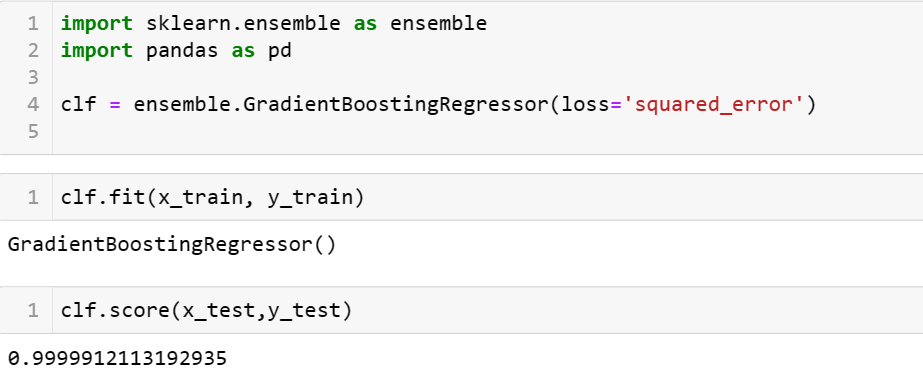
Gambar 32 Melatih Model dengan Testing

1. Menghitung skor atau koefisien determinasi (R-Squared) pada data pengujian (x\_test dan y\_test)



Gambar 33 Menghitung Skor (R-Squared)

1. Menggunakan model regresi ensemble Gradient Boosting dari scikit-learn untuk meningkatkan kinerja dan akurasi prediksi.



Gambar 34 Menghitung Akurasi dan Regresi Ensembel Gradient Boosting

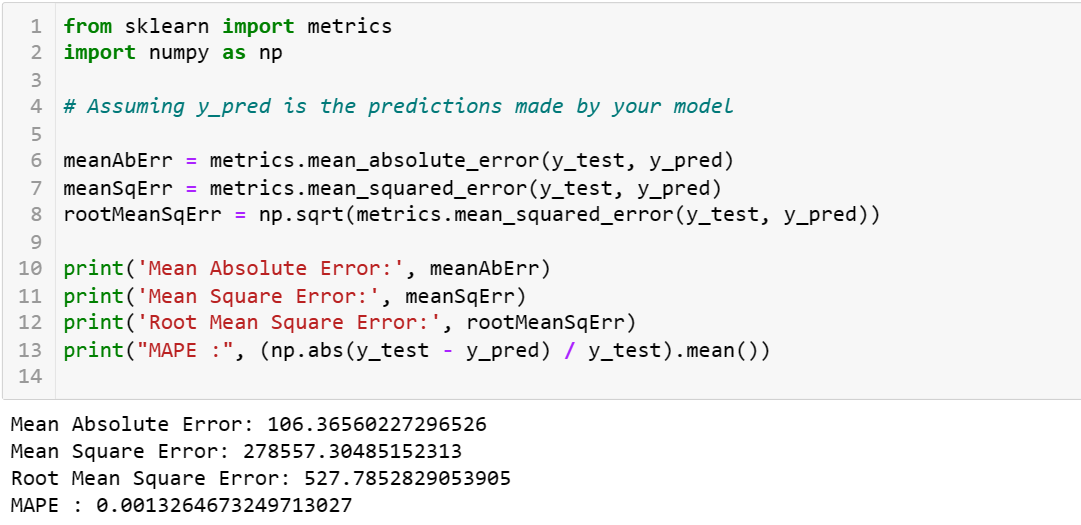
# BAB V

# MODEL EVALUATION

Pada bab ini akan menjelaskan model evaluation yang meliputi Result Evaluation, dan Result Review.

## Result Evaluation

Pada sub-bab ini akan diterangkan result/ hasil dan evaluation dari model yang sudah dibuat.



Gambar 35 Result Evaluation

## Result Review

Hasil yang didapatkan dari hasil evaluasi, yaitu dimana, Evaluasi yang harus dipenuhi untuk Regression Problem adalah MAE < 900 dan MAPE < 70%.

1. MAE (Mean Absolute Error) dari program ini adalah 106.36560227296526 yang berarti bahwa telah **memenuhi** evaluasi atau lebih kecil dari 900.
2. MAPE (Mean Absolute Percentage Error) dari program ini adalah 0.0013264673249713027 atau setara dengan 14% artinya untuk hasil MAPE ini telah **memenuhi** evaluasi atau lebih kecil 70%

# BAB VI

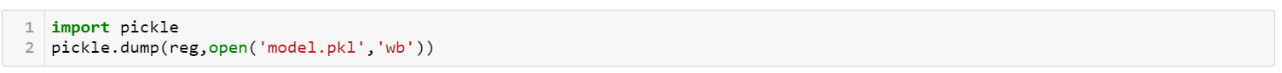
# DEPLOYMENT

Tahap keenam ini akan dilakukan *cost prediction* deployment. Pada bab ini kelompok kami akan melakukan *cost prediction* deployment dan akan di jelaskan pengerjaan dari deployment pada lokal komputer.

## Deployment

Pada fase plan deployment ini, model yang telah terbentuk pada fase modeling akan digunakan sesuai dengan tujuan data mining yang dibutuhkan. Adapun tahapan untuk melakukan deployment sebagai berikut:

1. Membuat model dalam bentuk file pickle (model.pkl) dan menyimpannya di local komputer dengan format model.pkl dengan code:

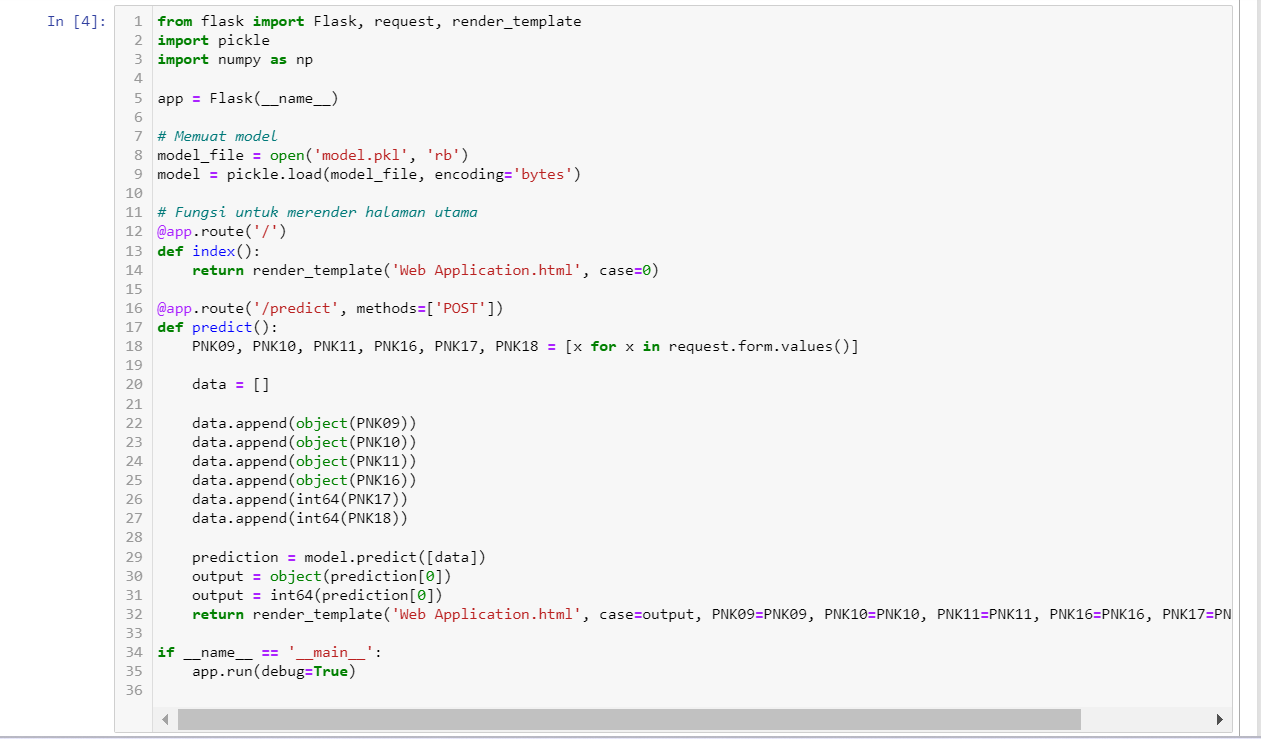


Gambar 36 Model.pkl



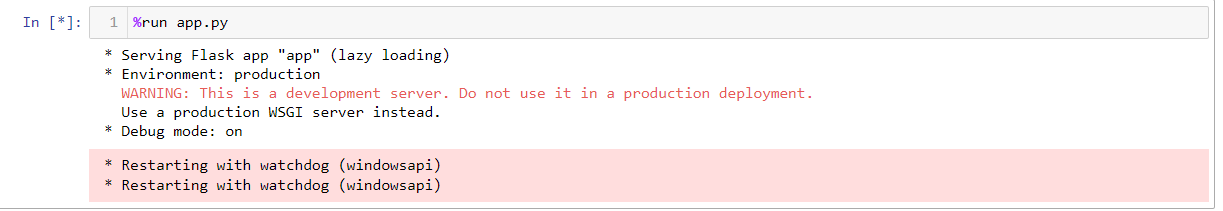
Gambar 37 FlashWebApp

1. Membuat code untuk memanggil model. Code dibawah akan disimpan dalam format .py dan berfungsi untuk menghubungkan dengan template htmlnya. Berikut codenya:



Gambar 38 Code untuk Memanggil Model

1. Menjalankan code berikut untuk mendapatkan URL dari deployment.



Gambar 39 Run App

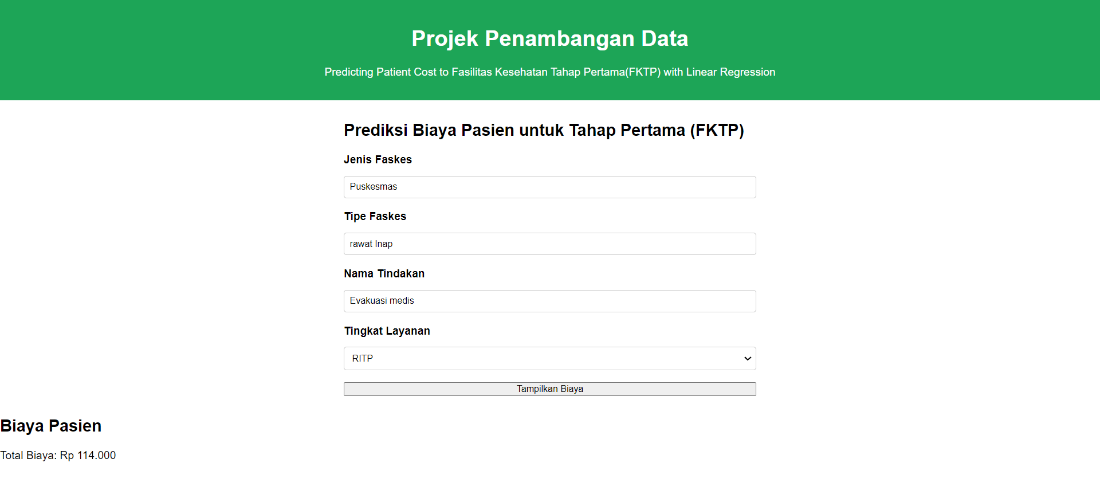
Pada tahap terakhir saat melakukan *deployment* di local komputer tidak berhasil, ditandai dengan ketidakmunculan hasil yaitu http/127.0.0.1.5000/ (tidak dapat running).

## Web Application

Pada bagian web application dibawah, akan ditampilkan tampilan front end dari Prediksi Biaya Pasien pada FKTP dengan Algoritma Regresi Multi linear, web application berhasil di deploy dan dapat dilihat pada gambar di bawah ini.



Gambar 40 Web Application Sebelum Input Data



Gambar 41 Web Application Setelah Input Data

# LAMPIRAN

1. Link Github: [lydiaharianja/Group-8\_Project-Data-Mining\_Case-3-Using-Regression (github.com)](https://github.com/lydiaharianja/Group-8_Project-Data-Mining_Case-3-Using-Regression)
2. Link Youtube : <https://youtu.be/6zQeGRHIZig>