# DOKUMEN PROYEK

## 12S4054 – PENAMBANGAN DATA

## *Regression Problem from Case and Cost Prediction using Random Forest*


**Disusun oleh:**

12S18020 Dita L. Sastri Sihombing

12S18029 Estomihi Rascana Sirait

12S18061 Angela Friscilia Simamora

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

# DAFTAR ISI

# DAFTAR TABEL

# BAB 1 BUSINESS UNDERSTANDING

Langkah pertama dalam metodologi CRISP-DM dalam melakukan prediksi pada jumlah kasus dan unit cost pada sebuah wilayah diakibatkan adanya penambahan rumah sakit adalah business understanding. Sehingga dalam pembahasan bab ini akan dijelaskan terkait aktivitas data mining untuk meningkatkan pemahaman diantaranya dengan menentukan Tujuan bisnis, Tujuan proyek serta menghasilkan Rencana proyek.

## 1.1 Determine Business Objective

Dalam lingkungan kita sehari hari tentunya sudah banyak pemanfaatan teknologi yang telah memberikan kita kemudahan dalam penyelesaian beberapa masalah. Perkembangan yang sangat pesat teknologi menguasai berbagai sektor dengan cepat. Begitu juga dimana semakin banyak penambahan rumah sakit di wilayah tertentu, semakin banyak data dan berbagai kondisi harus diselesaikan, melakukan prediksi pada jumlah kasus dan unit cost pada sebuah wilayah tertentu yang menjadi topik pembahasan dalam proyek ini dengan menggunakan data BPJS Hackathon (Case and Cost prediction). Untuk melakukan prediksi dengan data yang telah kita miliki pendekatan yang digunakan yaitu metode *Regression* Problem dengan algoritma *Random Forest* yang merupakan bagian dari teknik teknik dalam Data Mining. *Regression* merupakan proses memprediksi nilai kontinu yang terdiri dari variabel bebas(x) dan variabel tak bebas (y), *Regression* dan *Classification* adalah *prediction problem* pada s*uvervised learning.* Sedangkan algoritma *Random Forest* adalah algoritma yang diinterpretasikan dalam satu pohon model/*decision tree* yang menggunakan kualitas pohon keputusan dan random atau acak dimana algoritma ini adalah keputusan yang dibuat secara acak/*random forest.* Dengan beberapa kelebihan random forest juga memiliki kelemahan dimana cenderung menyebabkan data *Underfitting* . Underfitting adalah keadaan dimana sebuah model memiliki kinerja buruk untuk data *training* maupun *test* data dan dapat diatasi dengan *regression random forest*[citation].

Data Mining merupakan sebuah konsep yang diperuntukkan untuk menemukan pengetahuan atau informasi berharga dari sekumpulan data [Citation]. Data Mining juga merupakan proses semi otomatis yang menerapkan ilmu matematika, teknik statistik dan *machine learning* , ada banyak

teknik yang terkandung salah satunya ialah teknik *prediktif* yaitu melakukan prediksi terhadap data dengan menggunakan hasil hasil yang telah diperoleh dari data berbeda.

Data BPJS Hackathon ini merupakan sekumpulan data yang kategorial, dimana kita perlu melakukan analisis data sebelum dijadikan sebagai pembentuk Model yang akan dibangun dan mengimplementasikan metode dengan algoritma yang telah dipilih dan dijelaskan sebelumnya, sehingga setiap personel dalam kelompok ini haruslah mengupayakan pembagian waktu dalam menganalisis, mengetahui dan mencapai *Case* and *Cost prediction* pada data apakah menyebabkan *fraud* atau *non fraud.* Sehingga dapat dijelaskan objektif yang akan dicapai dalam pengerjaan proyek ini ialah:

1. Men-check data apakah ada data yang *missing* atau *duplicate.*Jika ditemukan data sesuai maka dilakukan analisis *Exploratory Data Analysis* (EDA), karena data yang besar memungkin terjadinya *redudancy* perlu dilakukan berulang-ulang
2. Identifikasi faktor apa yang menyebabkan data berpotensi menghasilkan *fraud.*
3. Meningkatkan performa Data BPJS Hackathon itu sendiri dengan model yang akan menjadi output proyek ini

Pekerjaan didalam membangun model ini akan dikatakan sukses apabila:

- Dihasilkan faktor penyebab data fraud
- Dikerjakan dengan tepat waktu dengan hasil sebaik mungkin

## 1.2 Determine Project Goal

Tujuan pengerjaan proyek ini adalah Membangun sebuah Model dengan penggunaan Teknik dalam Data Mining untuk mengetahui *case* and *cost prediction* dari data BPJS Hackathon dengan hasil apakah data dikatakan *fraud* atau *non fraud.*

## 1.3 Produce Project Plan

Tahap perencanaan yang dilakukan untuk mencapai tujuan pengerjaan proyek penelitian "*Regression Problem from Case and Cost Prediction using Random Forest* " adalah sebagai berikut:

**Tabel 1. Jadwal Pelaksanaan Proyek**

| Tahapan | Waktu Pengerjaan | Kegiatan |
|---|---|---|
| *Business Understanding* | 3 hari | Menentukan objektif bisnis, menentukan tujuan proyek serta membuat rencana proyek. |
| *Data Understanding* | 3 hari | Mengumpulkan data yang akan digunakan, menelaah data dan melakukan validasi pada data. |
| *Data Preparation* | 4 hari | Memilih data yang akan digunakan, 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. |

Dalam pelaksanaan proyek penelitian, adapun *tools* yang digunakan yaitu *Python* adalah salah satu bahasa pemrograman yang dapat melakukan eksekusi sejumlah instruksi multi guna secara langsung (interpretatif) dengan metode orientasi objek (*Object Oriented Programming*) serta menggunakan semantik dinamis untuk memberikan tingkat keterbacaan syntax. *Python*

juga merupakan salah satu bahasa populer yang berkaitan dengan *Data Science, Machine Learning*, dan *Internet of Things* (IoT). Sebagian lain mengartikan *Python* sebagai bahasa yang kemampuan, menggabungkan kapabilitas, dan sintaksis kode yang sangat jelas, dan juga dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif.

Selain itu untuk metode atau algoritma yang akan digunakan dalam proyek ini adalah *Random Forest*. *Random forest* adalah kombinasi dari masing – masing *tree* yang baik kemudian dikombinasikan ke dalam satu model. *Random Forest* bergantung pada sebuah nilai vector random dengan distribusi yang sama pada semua *tree* yang masing masing *decision tree* memiliki kedalaman yang maksimal. Penggunaan *tree* yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil voting dari *tree* yang terbentuk.

# 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

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

Dataset yang digunakan untuk memprediksi jumlah kasus dan unit cost pada sebuah daerah akibat penambahan Rumah Sakit kerja sama adalah *case_cost_prediction_train.csv*. Dataset tersebut terdiri atas 57971 observasi dan 36 variable. Berikut tabel yang membahas terkait atribut pada dataset.

**Tabel 2. Tabel Atribut Dataset**

| No. | Atribut | Deskripsi |
|---|---|---|
| 1 | *row_id* | ID dari setiap data |
| 2 | *tglpelayanan* | 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 |

Terkait dataset yang tersebut, maka selanjutnya dilakukan EDA (*Exploratory Data Analysis*) terhadap dataset yang digunakan untuk menganalisis karakteristik utama dataset. Dalam pengerjaan proyek, tidak semua atribut dalam dataset digunakan karena hanya terdapat beberapa atribut yang relevan dengan tujuan proyek penelitian. Untuk itu atribut yang paling sesuai dalam melakukan prediksi pada jumlah kasus dan unit cost adalah atribut *case*, atribut *unit_cost* dan beberapa atribut yang relevan. Maka dari itu berikut beberapa hipotesis terkait atribut 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.

Dari beberapa hipotesis tersebut, terdapat atribut *kddati2, peserta, case,* dan *unit_cost* yang berpengaruh pada jumlah kasus dan unit cost tersebut dan dapat digunakan sesuai tujuan proyek yaitu mengembangkan model data mining untuk melakukan prediksi pada jumlah kasus dan unit cost pada sebuah daerah akibat penambahan Rumah Sakit kerja sama.

### 2.3 Validation Data

Pada tahap ini dilakukan validasi terhadap data yang akan digunakan dengan memeriksa kelengkapan data untuk menghindari terjadinya *error* ataupun masalah *input data* yang terjadi *missing value*. Maka dari itu berikut pemeriksaan terhadap atribut utama yang akan digunakan pada dataset.

- Atribut *kddati2*

```
casecostprediction['kddati2'].describe()
```

```
count    57971.000000
mean       246.423125
std        143.447935
min          1.000000
25%        125.000000
50%        243.000000
75%        362.000000
max        528.000000
Name: kddati2, dtype: float64
```

- Atribut *peserta*

```
casecostprediction['peserta'].describe()
```

```
count    5.797100e+04
mean     3.562209e+05
std      4.120323e+05
min      8.000000e+00
25%      1.127735e+05
50%      1.975800e+05
75%      4.386935e+05
max      3.328509e+06
Name: peserta, dtype: float64
```

- Atribut *case*

```
casecostprediction['case'].describe()
```

```
count     57971.000000
mean       6539.418451
std       17607.280021
min           1.000000
25%         424.000000
50%        1359.000000
75%        4583.000000
max      333441.000000
Name: case, dtype: float64
```

- Atribut *unit_cost*

```
casecostprediction['unit_cost'].describe()
```

```
count    5.797100e+04
mean     1.961092e+06
std      1.889367e+06
min      1.000000e+05
25%      2.336742e+05
50%      6.547994e+05
75%      3.531702e+06
max      2.690550e+07
Name: unit_cost, dtype: float64
```

# BAB 3 DATA PREPARATION

Data Preparation akan dilakukan untuk menghasilkan data yang memiliki kualitas baik. Berdasarkan penjelasan data pada bab 2, data preparation dilakukan dengan beberapa tahapan meliputi *data cleaning, data integration*, *data transformation* dan *data reduction.*

## 3.1 Memilah data

Dalam hal ini akan dipilih data data yang diperlukan. Untuk data yang missing value, tidak relevan akan dihapus. Dimana kita dapat mengetahui sebuah data tidak memiliki missing value dengan menggunakan kode isnull pada program kita.

## 3.2 Cleaning data

Untuk menghasilkan data berkualitas, maka perlu dilakukan cleaning data. kita harus mengetahui terlebih dahulu terkait data unik yang kita miliki dari data, sehingga dapat dilakukan drop pada atribut yang memiliki 1 nilai.

# BAB 4 MODELLING

# BAB 5 EVALUATION

# BAB 6 DEPLOYMENT

# DAFTAR PUSTAKA