**LAPORAN PROYEK DATA MINING**

***Binary Classification using Random Forest***

****

**Disusun Oleh:**

12S18018 Yohana Polin Simatupang

12S18019 Maria Puspita Sari Nababan

12S18064 Letare Aiglien Saragih

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

# **BAB 1 BUSINESS UNDERSTANDING**

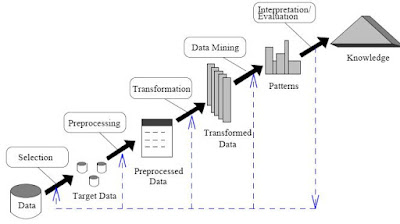
Pada pengerjaan proyek ini akan dilakukan sesuai dengan tahapan pada metodologi CRISP DM yang akan dimulai dengan tahapan *business understanding* yaitu memahami permasalahan bisnis untuk proses *data mining* yang akan dilakukan. Adapun yang termasuk bagian dari tahapan ini adalah menentukan tujuan bisnis, menentukan sasaran yang ingin dicapai dengan data mining, dan menghasilkan perencanaan proyek yang akan dilakukan.

## **1.1 Determine Business Objective**

Rumah sakit merupakan salah satu instansi yang bergerak sebagai pelayanan kesehatan bagi masyarakat. Dalam melaksanakan proses bisnisnya, peran BPJS cukup besar dalam mempengaruhi kualitas pelayanan bagi masyarakat. Namun dengan semakin banyak penggunaan BPJS Kesehatan, tidak jarang terjadi beberapa kecurangan (*fraud*) yang ditujukan untuk menguntungkan pihak tertentu. Pelaku yang terlibat bisa jadi adalah peserta BPJS Kesehatan, *fasilitator* kesehatan atau pembeli layanan kesehatan, penyedia obat dan alat kesehatan, dan pemangku kepentingan lainnya. Penanganan terkait masalah tersebut menjadi *concern* yang perlu untuk diatasi yang bertujuan untuk dapat mencegah dan mendeteksi berbagai indikasi potensi kecurangan sedini dan sesedikit mungkin. Sehingga dengan demikian biaya pelayanan kesehatan dapat dimanfaatkan semaksimal mungkin dalam memenuhi kepentingan dan pelayanan yang maksimal bagi masyarakat, serta untuk tetap menjaga *sustainability* BPJS Kesehatan,

## **1.2 Determine Data Mining Goal**

Tujuan bisnis pada penelitian ini adalah untuk melakukan prediksi potensi terjadinya penyimpangan (*fraud*) pada klaim pelayanan Rumah Sakit. Melihat jumlah data yang besar dan studi kasus yang akan diteliti untuk itu, dilakukan penerapan *data mining* untuk menemukan pola menarik dari data. *Data mining* dikelompokkan menjadi *description, estimation, prediction, classification, clustering*, dan *association* [ref: buku pang-ning tan]*.* Pada penelitian ini, penggunaan data mining bertujuan sebagai dasar dalam pengembangan sebuah model klasifikasi biner untuk menemukan fraud. Ketika melakukan proses data mining, harus dilakukan beberapa tahapan antara lain, pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola dan presentasi pengetahuan.



Tahapan Knowledge Discovery in Databases (KDD)

Dalam menemukan faktor apa saja yang menyebabkan terjadinya penyimpangan (*fraud*) pada layanan BPJS perlu digunakan data mining task dengan teknik asosiasi. *Association rule mining* adalah metode pembelajaran mesin berbasis aturan untuk menemukan hubungan yang menarik antara variabel dalam data yang berjumlah besar.

Algoritma yang akan untuk penelitian ini adalah Algoritma *Random Forest Classifier* (RFC). RFC merupakan metode klasifikasi yang *supervised* menggabungkan ratusan atau ribuan pohon keputusan, melatih masing-masing pohon pada serangkaian pengamatan yang sedikit berbeda, memisahkan simpul di setiap pohon dengan mempertimbangkan sejumlah fitur yang terbatas. Prediksi akhir dari random forest dibuat dengan merata-ratakan prediksi dari masing-masing pohon. Kelebihan dari algoritma ini adalah: menghasilkan eror yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, efektif untuk mengestimasi hilangnya data, memperkiraan variabel apa yang penting dalam klasifikasi dan menyediakan metode eksperimental untuk mendeteksi interaksi variabel.

## **1.3 Produce Project Plan**

Tahap perencanaan yang dilakukan untuk mencapai tujuan data mining dan mencapai tujuan bisnis pada penelitian ‘*Binary Classification using Random Forest*’ ini adalah sebagai berikut:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Aktivitas** | **Sub Aktivitas** | **Durasi** | **Sumber daya yang dibutuhkan** | **Ketergantungan** |
| Pemilihan Kasus dan Algoritma | Pemilihan Kasus | 1 | Semua analisis | - |
| Penentuan Algoritma | 6 |
| Business Understanding | Menentukan Objektif Bisnis | 1 | Semua analisis | Pemilihan kasus dan algoritma |
| Menentukan Tujuan Bisnis | 1 |
| Membuat Rencana Proyek | 1 |
| Data Understanding | Mengumpulkan Data | 1 | Semua analisis | Data dan teknologi |
| Menelaah Data | 1 |
| Memvalidasi Data | 1 |
| Data Preparation | Memilah Data | 1 | *Data mining consultant,* beberapa *database analyst time* | Data dan teknologi |
| mengkonstruksi Data | 4 |
| Menentukan Label Data | 1 |
| Membersihkan Data | 4 |
| Modeling | Membangun Skenario Pengujian | 3 | *Data mining consultant,* beberapa *database analyst time* | Algoritma |
| Membangun Model | 7 |
| Model Evaluation | Mengevaluasi Hasil Pemodelan | 5 | Semua analisis | Model yang telah dibuat |
| Melakukan Review Proses Pemodelan | 4 |
| Deployment | melakukan Deployment Model | 2 | *Data mining consultant,* beberapa *database analyst time* | Penerapan model berdasarkan data dan algoritma yang dipilih |
| Membuat laporan akhir Proyek | 4 |

Dalam pelaksanaan proyek dalam penelitian ini, diperlukan tools data mining yang mendukung metode untuk berbagai tahapan proses. Tools dan teknik yang digunakan dapat mempengaruhi keseluruhan proyek. Tools yang digunakan dalam mengerjakan proyek ini adalah python. Python adalah bahasa pemrograman berorientasi objek yang digunakan dalam pengembangan perangkat lunak maupun dalam analisis dan data science. Python memiliki berbagai library yang menyediakan fungsi untuk melakukan analisis data, memproses data, memvisualisasikan data, dll.

# **BAB 2 DATA UNDERSTANDING**

Tahap kedua pada metodologi CRISP-DM setelah *business understanding* dalam melakukan metodologi *data science* adalah *data understanding*. Pada bab ini akan dijelaskan mengenai pengumpulan *initial data*, analisis untuk dapat memahami data yang akan digunakan dalam penelitian serta verifikasi pada kualitas data.

## **2.1 Collect Initial Data**

Langkah *data understanding* diawali dengan pengumpulan data yang akan digunakan pada proses *data science.* Data yang akan digunakan dalam kasus *binary classification*  menggunakan *Random Forest Classification* (RFC) adalah data BPJS Kesehatan yang berasal dari dataset yang digunakan dalam kompetisi Hackathon.

## **2.2 Analysis Data**

Dataset train yang digunakan untuk memprediksi penyimpangan (fraud) pada layanan BPJS terdiri dari 200217 observasi dan 53 variable. Kemudian perlu dilakukan Exploratory Data Analysis (EDA). EDA digunakan untuk memahami data, mendapatkan konteks data, memahami variabel dan hubungan di antara variabel, dan merumuskan hipotesis yang berguna dalam membangun model prediksi. Atribut atau fitur pada dataset tidak semua diperlukan dalam menganalisis. Fitur atau atribut yang digunakan merupakan atribut yang relevan dan sesuai dengan tujuan proyek. Adapun fitur atau variabel yang dibutuhkan dalam penelitian ini antara lain, yaitu:

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Variabel** | **Tipe Variabel** | **Deskripsi** |
| 1 | visit\_id |  | id kunjungan |
| 2 | kdkc |  | kode wilayah kantor cabang BPJS Kesehatan |
| 3 | dati2 |  | kode kabupaten/kota |
| 4 | typeppk |  | kode tipe Rumah Sakit |
| 5 | jkpst |  | jenis kelamin peserta JKN-KIS |
| 6 | umur |  | umur peserta saat mendapatkan pelayanan rumah sakit |
| 7 | jnspelsep |  | tingkat pelayanan; 1:rawat inap; 2. rawat jalan |
| 8 | los |  | lama peserta dirawat di rumah sakit |
| 9 | cmg |  | klasifikasi CMG (Case Mix Group) |
| 10 | severitylevel |  | tingkat urgensi |
| 11 | diagprimer |  | diagnosa primer |
| 12 | dx2\_...\_... |  | diagnosa sekunder |
| 13 | proc..\_... |  | kode kelompok procedure |
| 14 | label |  | flag fraud; 1:fraud; 0:tidak fraud |

Untuk mendapatkan hasil analisa dataset yang lebih baik, makan perlu dilakukan pengidentifikasian kembali subset data yang relevan untuk kemudian digunakan pada tahapan selanjutnya yang sesuai dengan tujuan data mining pada penelitian ini.

## **2.3 Verify Data Quality**

Tahapan selanjutnya adalah melakukan verifikasi terhadap kualitas data yang digunakan. Untuk mendapatkan data yang berkualitas baik, perlu dilakukan pembersihan data (*dat cleaning*). *Data cleaning* pada proses data mining dapat mengurangi jumlah dan kompleksitas data. Salah satu aspek yang menyebabkan kualitas data menjadi kurang baik adalah terjadinya *missing value* atau terdapat data yang hilang pada dataset yang digunakan. Untuk mengantisipasi hal tersebut terlebih dahulu dilakukan pemeriksaan apakah terdapat data yang hilang (*missing*) atau bernilai kosong. Pemeriksaan dilakukan menggunakan fungsi pada python yaitu *df.isna().* Adapun hasil yang didapatkan dari pemeriksaan tersebut adalah bahwa pada dataset yang digunakan tidak terdapat *missing value*.

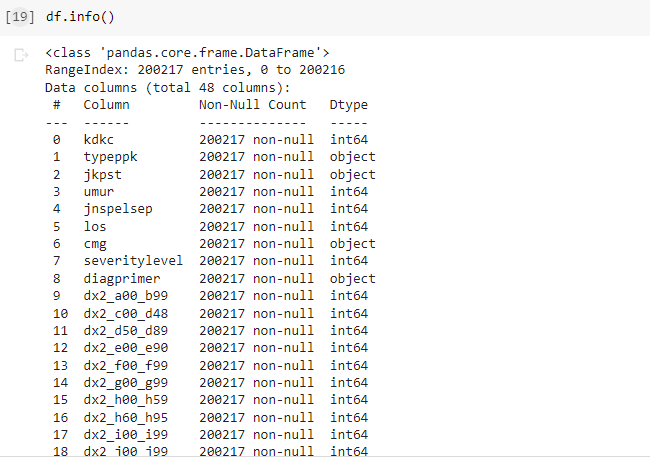
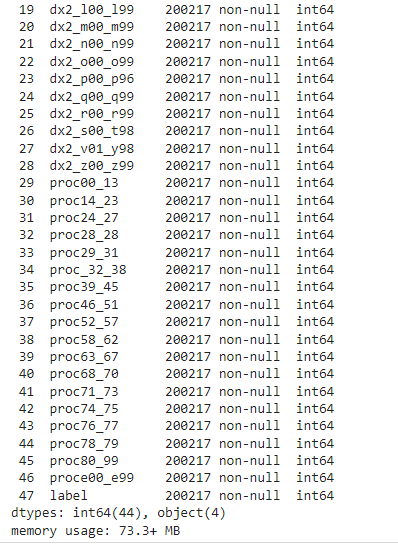
# **BAB 3 DATA PREPARATION**

## **3.1 Sorting Data**

Data yang akan digunakan dalam proses *data mining* terlebih dahulu perlu dipersiapkan dengan baik. Fase *sorting* merupakan tahapan untuk melakukan pemilihan pada atribut yang akan digunakan. Atribut yang tidak digunakan akan *di drop*.



Atribut tersebut di *drop* dengan tujuan agar data yang digunakan lebih efisien dan efektif dalam pengolahan data termasuk dalam penggunaan memory. Berikut adalah tampilan setelah atribut yang tidak digunakan telah di *drop*

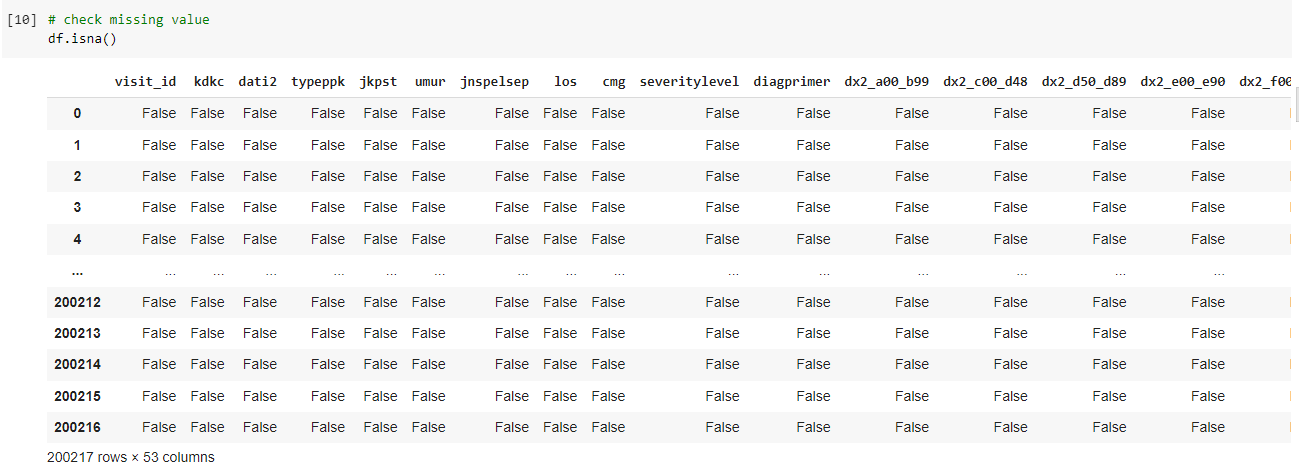
 

Nilai penggunaan *memory* menjadi berkurang setelah dilakukan pemilihan atribut yang diperlukan yaitu sebagai berikut

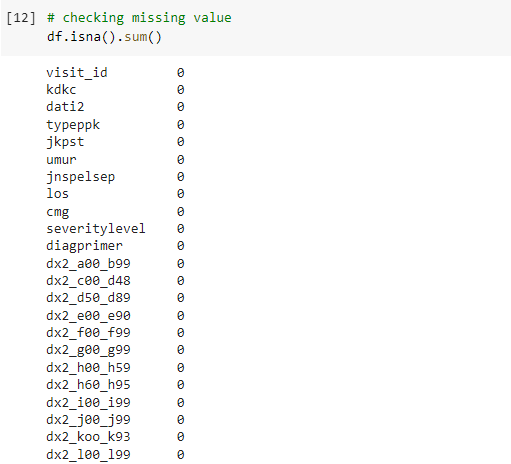
 

## **3.2 Cleaning Data**

Fase ini merupakan tahapan untuk melakukan pembersihan data. Pembersihan data yang dilakukan adalah menangani objek data yang kosong (*missing value*). Untuk itu, terlebih dahulu dilakukan pemeriksaan data dengan menggunakan fungsi df.isna()

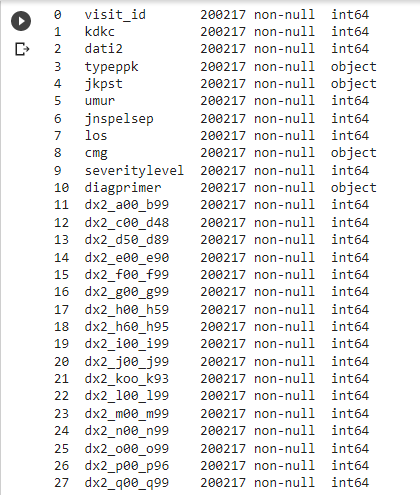


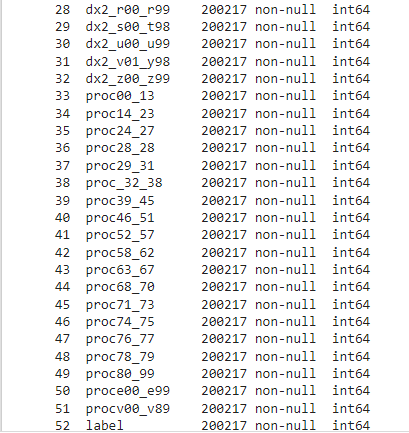
Python Pandas memungkinkan kita dapat menemukan *missing value* secara cepat dengan fungsi isna().Fungsi isna() akan mengembalikan nilai boolean dari dataset yang diperiksa. Hasil keluaran berupa **False** menunjukkan bahwa pada cell tersebut tidak terdapat nilai yang kosong (*missing*). Agregasi data dengan fungsi sum() ditujukan agar dapat memahami data dengan lebih baik. Agregasi sum()akan menjumlahkan semua cell yang kosong apabila terdapat nilai yang kosong pada atribut tertentu.



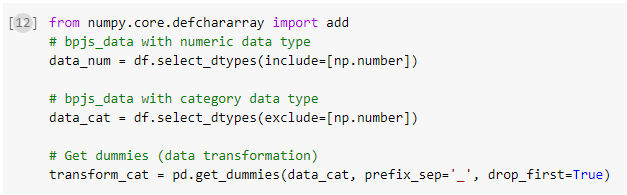
## **3.3 Construct Data**

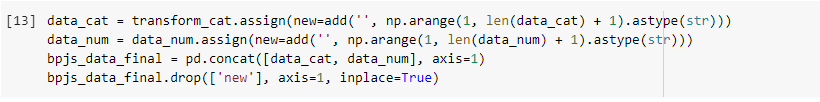
Fase ini merupakan tahapan untuk melakukan konstruksi pada data. Adapun konstruksi yang dilakukan adalah transformasi atribut dengan tipe kategorik menjadi numerik. Hal ini bertujuan agar data kemudian dapat di normalisasi. Untuk tahap pada konstruksi data dilakukan pengecekan tipe data pada dataset menggunakan fungsi df.info(), dan output yang dihasilkan adalah sebagai berikut:



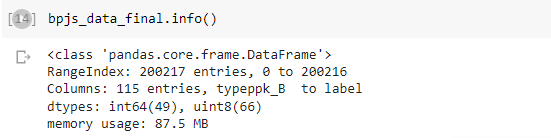


Dapat dilihat pada gambar di atas, terdapat 4 atribut yang bertipe data kategorikal (object64), untuk itu perlu dilakukan transformasi data. Untuk itu perlu dilakukan transformasi data tipe pada atribut dengan menjalankan potongan kode berikut:



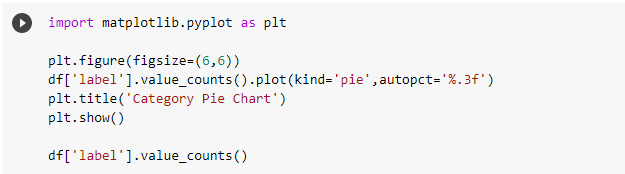


Setelah transformasi berhasil, dilakukan pengecekan kembali pada type atribut menggunakan fungsi df.info()

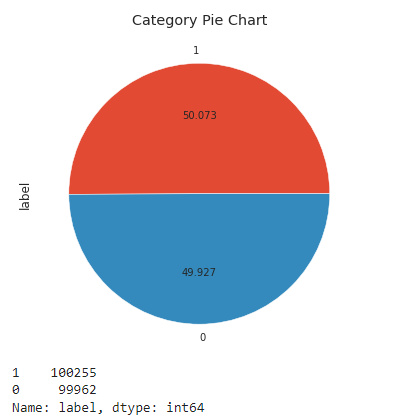


## **3.4 Define Data Labels**

Pada tahap ini akan dilakukan pendefinisian label data yang akan digunakan. Tahap ini bertujuan untuk memastikan bahwa data yang digunakan merupakan *balance dataset*. Berikut merupakan potongan kode yang menunjukkan bagaimana feature label divisualisasikan.



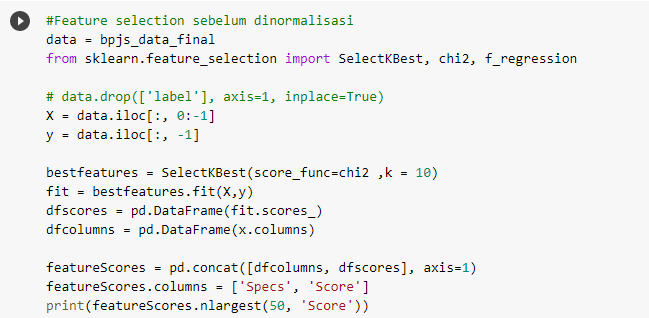
Dan dihasilkan output sebagai berikut:



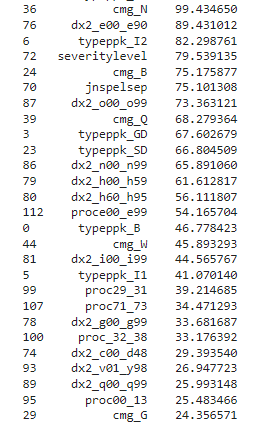
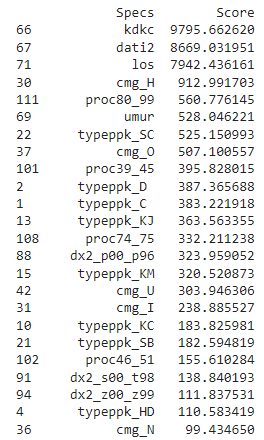
Berdasarkan pie chart diatas, dapat dilihat bahwa jumlah data yang menunjukkan *fraud* dan tidak *fraud* pada feature Label adalah seimbang.

## **3.5 Best Feature Selection**

Pada tahap ini akan dilakukan *Feature Selection*, yang bertujuan untuk melakukan analisa pada data dengan memilih *feature* yang memiliki dampak atau berpengaruh terhadap proses data dan memisahkan *feature* yang tidak memiliki dampak terhadap proses data. Pada tahap ini, setiap *feature* akan diurutkan sesuai dengan feature yang paling berpengaruh. *Feature selection* dapat dijalankan menggunakan potongan kode berikut:



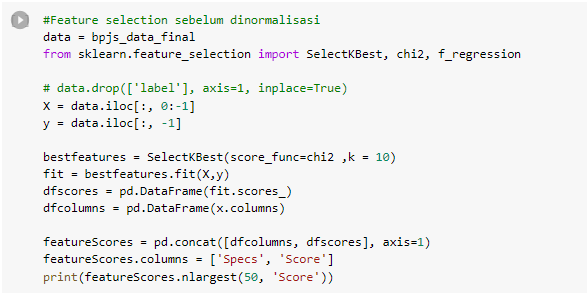
Dan output yang diperoleh adalah sebagai berikut:



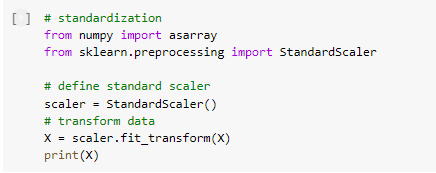
Output diatas menunjukkan urutan dari 50 *feature* pada data menggunakan *feature selection*. Berdasarkan hasil tersebut dapat dilihat bahwa *feature* paling berpengaruh adalah kdkc dan *feature* paling tidak berpengaruh adalah cmg\_G.

## **3.6 Normalization**

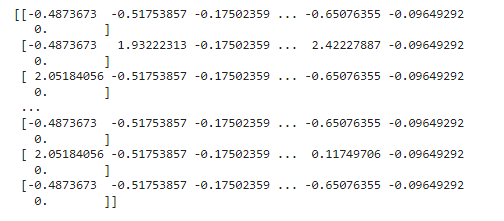
Untuk tahap ini akan dilakukan normalisasi dan standarisasi pada data yang telah diolah sebelumnya untuk mendapatkan hasil yang lebih baik. Sebelumnya normalisasi dan standarisasi dijalankan, terlebih dahulu data dibagi dan disimpan dalam variabel X dan y seperti dibawah ini.



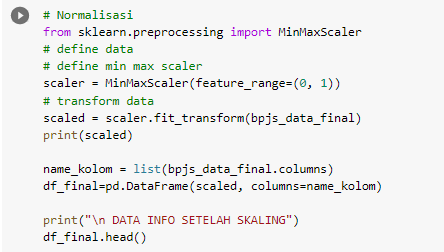
Penerapan standarisasi berfokus pada mengubah data mentah menjadi informasi yang dapat digunakan sebelum dianalisis. Merupakan teknik yang menskalakan data sehingga memiliki mean = 0 dan standar deviasi =1.



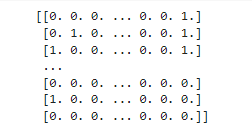
Dan output yang diperoleh adalah sebagai berikut:



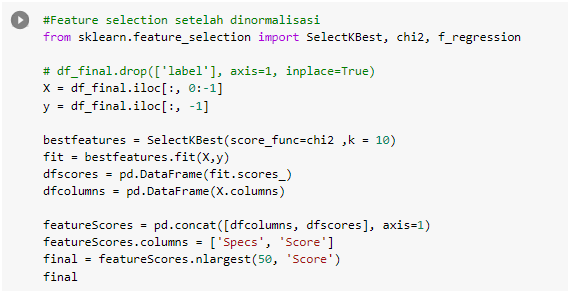
Penerapan normalisasi bertujuan untuk membuat data train menjadi kurang peka terhadap skala fitur sehingga koefisien menghasilkan pecahan yang lebih baik ketika dibagi. Teknik penskalaan dengan nilai-nilai digeser dan diubah skalanya sehingga nilainya berkisar antara 0 dan 1 atau rentang [0,1].



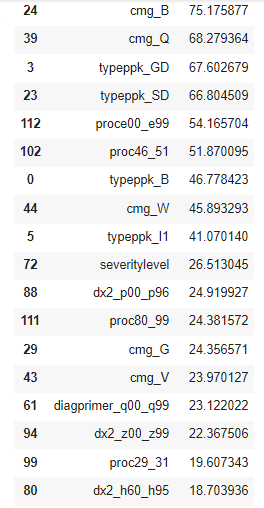
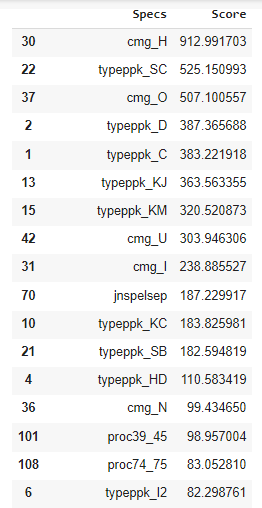
Dan output yang diperoleh adalah sebagai berikut:

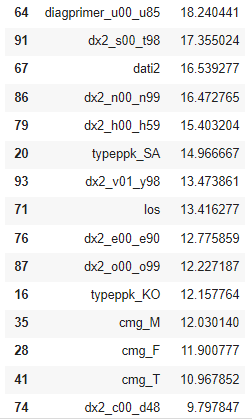


Setelah proses normalisasi dan standarisasi dilakukan, ditemukan fakta bahwa penerapan normalisasi menghasilkan data *scaling* yang lebih baik. Untuk itu, data hasil normalisasi yang kemudian akan digunakan untuk penerapan *feature selection*.



Dan output yang diperoleh adalh sebagai berikut:





Dapat dilihat terjadi perubah output saat fungsi feature selection dijalankan sebelum dan sesudah proses normalisasi dilakukan. Sebelum dilakukan feature selection, *feature* paling berpengaruh adalah kdkc dan *feature* paling tidak berpengaruh adalah cmg\_G. Sementara setelah dilakukan normalisasi, *feature* paling berpengaruh adalah cmg\_H dan *feature* paling tidak berpengaruh adalah dx2\_c00\_d48.

# **BAB 4 MODELLING**

## **4.1 Build Test Scenario**

## **4.2 Model Building**

# **BAB 5 MODEL EVALUATION**

## **5.1 Evaluation of Modeling Result**

## **5.2 Modeling Process Review**

# **BAB 6 DEPLOYMENT**

## **6.1 Model Development**

## **6.2 Final Report**