**DOKUMEN PROYEK**

# **12S4054 - PENAMBANGAN DATA**

***Fraud Detection (Binary Classification) BPJS Hackathon Using SVM Algorithm***

**Disusun Oleh:**

| **12S19051** | **Corri Hutahaean** |
| --- | --- |
| **12S19052** | **Mulyani Gabe Sayoni Simanjuntak** |
| **12S19053** | **Elysa Noelia Pangaribuan** |



**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO (FITE)**

**INSTITUT TEKNOLOGI DEL**

**TAHUN 2022/2023**

# **BAB 1**

**BUSINESS UNDERSTANDING**

Business Understanding adalah langkah pertama dalam CRISP-DM yang secara general digunakan untuk mendefinisikan proyek, tujuan dan kebutuhan dari sudut pandang bisnis, yang kemudian akan menerjemahkan pengetahuan yang sudah diperoleh ke dalam pendefinisian masalah pada data mining sehingga dapat dilakukan penyesuaian terhadap tujuan bisnis sehingga model terbaik dapat dibangun. Tahap business understanding juga merupakan tahap yang digunakan untuk mengetahui dan menentukan rencana dan strategi untuk mencapai tujuan yang sudah didefinisikan di awal. Pada tahap ini diperlukan pengetahuan dari objek bisnis tertentu, yaitu bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokkan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun [1].

Pada bab ini akan menjelaskan pemahaman mengenai aktivitas dalam menentukan sasaran bisnis, memahami situasi bisnis, menerjemahkan tujuan atau sasaran bisnis ke dalam data mining.Dalam sistem yang akan dikembangkan oleh penulis meliputi 4 tahap dalam business understanding, yaitu: determine business objectives, assess situation, determine data mining goals, dan produce project plan [2].

## **1. 1 Determine Business Objective**

Menganalisis data agar benar-benar memahami dari perspektif bisnis, mengenai apa yang ingin dicapai. Tujuan dilakukannya analisis adalah untuk mendapatkan faktor-faktor penting yang dapat mempengaruhi hasil proyek sehingga penelitian tidak akan menghasilkan jawaban yang benar atas pernyataan yang salah.

Tujuan bisnis dalam

## **1.2 DeterSituation Assessment**

Pada tahapan ini akan dilakukan pencarian fakta yang lebih terperinci mengenai semua sumber daya, kendala, asumsi, dan faktor lainnya yang harus dipertimbangkan dalam menentukan tujuan analisis data dan rencana penelitian. Tahapan ini bertujuan untuk memperluas detail dari analisis yang dihasilkan pada tahapan pertama. Sumber daya yang akan digunakan dalam penelitian adalah data dari BPJS Hackaton yang terdiri dari 200217 observasi dan 53 variable. Pada zaman sekarang sangat dibutuhkan sistem fraud detection yaitu untuk melakukan pendeteksian atau melakukan investigasi suatu kegiatan yang mencurigakan dan sekaligus mencegah terjadinya kecurangan. Tujuan pengerjaan proyek ini adalah membangun sebuah model dengan penggunaan teknik dalam *data mining* untuk mengetahui *fraud* dari data BPJS Hackaton dengan hasil apakah data *fraud* atau *non-fraud.*

## **1.3 Determine Data Mining Goal**

Pada tahap ini, akan ditentukan tujuan dalam terminologi bisnis. Tujuan data mining menyatakan tujuan proyek ini. Dan tujuan dari proyek ini adalah untuk mengembangkan sebuah model data mining untuk melakukan prediksi potensi terjadinya *fraud* pada klaim pelayanan Rumah Sakit berdasarkan dataset train yang terdiri dari 200.217 observasi dan 53 variabel.

## **1.4 Produce Project Plan**

Pada tahap ini, rencana yang akan dilaksanakan untuk mencapai tujuan data dijelaskan untuk pertambangan dan mencapai tujuan komersial. Rencana yang dibuat harus dapat ditentukan langkah-langkah yang harus diambil selama sisa proyek, termasuk pemilihan alat dan teknik awal. Algoritma yang akan digunakan dalam penelitian ini adalah Algoritma *SVM* yang akan membantu melakukan klasifikasi terhadap kecurangan pelayanan rumah sakit.

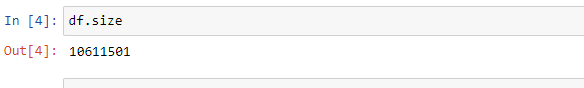
**BAB 2**

**DATA UNDERSTANDING**

*Data Understanding* adalah tahap pengumpulan data yang akan dilanjutkan dengan sebuah proses untuk memperoleh pemahaman mendalam mengenai data, mengidentifikasi kualitas data. serta memungkinkan untuk melakukan deteksi apabila terdapat sebuah bagian unik dari data yang baik digunakan sebagai hipotesis terhadap informasi yang tersembunyi. Tahap *data understanding* memberikan fondasi analitik dengan membuat ringkasan dan melakukan identifikasi potensi masalah dalam data yang harus dilakukan secara cermat.

## **2.1 Collect Initial Data**

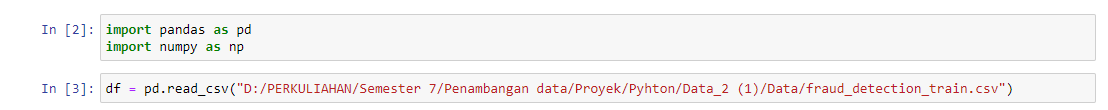
## Pada tahapan ini akan dilakukan pengumpulan data yang telah diberikan dosen pengampu yang diambil dari *case* salah satu lomba yaitu Hackaton. Dataset tersebut memiliki format file CSV (*Comma Separated Values*) sehingga datanya bersifat statis dan terstruktur. Ukuran data yang digunakan dalam kasus *fraud detection* menggunakan *Super Vector Machine adalah* 10611501.



Gambar 1. Informasi ukuran data

## **2.2 Describe Data**

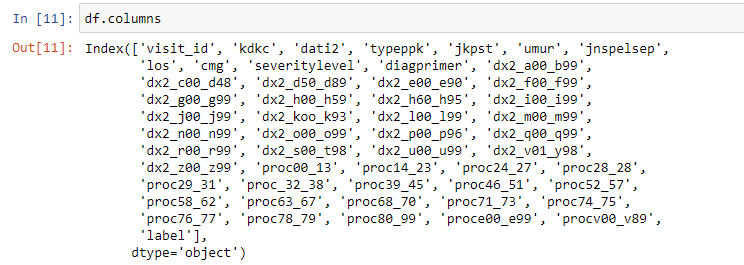
Pada tahapan ini akan dilakukan pendeskripsian terhadap data yang akan digunakan berupa keterangan mengenai format data, jumlah data, jumlah atribut dan fitur yang digunakan pada pengerjaan proyek. Dataset yang digunakan untuk melakukan prediksi terjadinya *fraud* pada klaim pelayanan Rumah Sakit. Untuk membaca dataset yang akan digunakan terlebih dahulu import library pandas untuk membaca data.



Gambar Kode untuk membaca dataset.

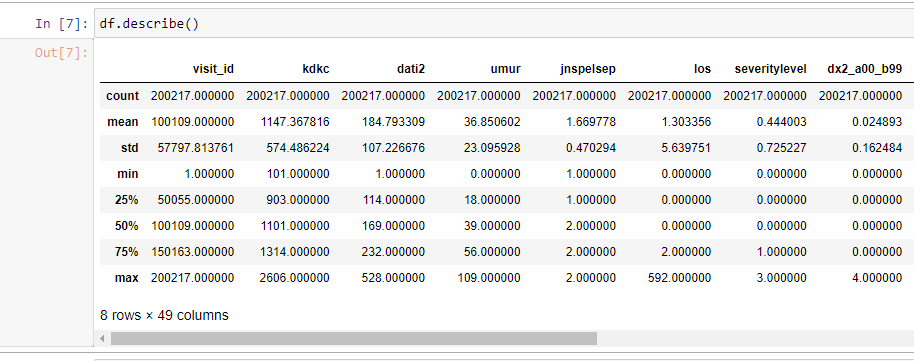
Data yang digunakan berjumlah 200.217 data dan terdapat 53 variabel, terdapat 4 atribut bertipe data kategorikal dan 49 atribut bertipe data numerik. Pada tahapan deskripsi data ini juga dapat memberikan informasi apa saja yang dapat digunakan untuk melakukan implementasi pada sistem yang dibangun. Untuk melihat dimensi dataset digunakan fungsi df.shape. Fungsi df.columns pada pandas digunakan untuk melihat kolom yang ada pada dataset.





Gambar 2. Kode untuk melihat kolom pada dataset

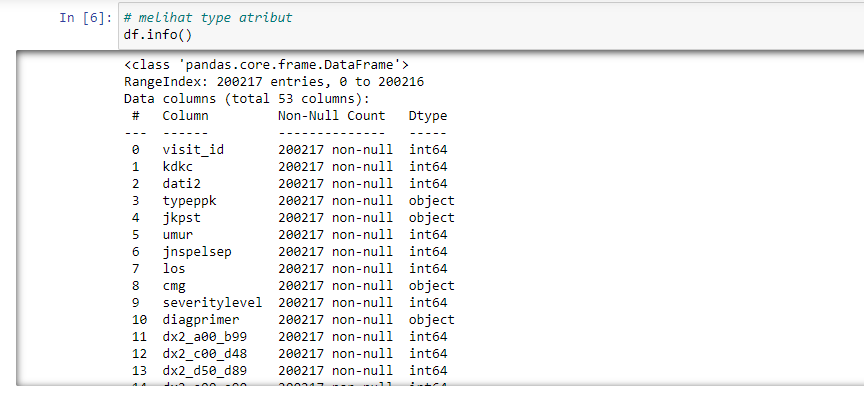
Pada hasil yang telah diberikan nama dari 53 atribut yang ada pada dataset. Selanjutnya, untuk melihat detail statistik seperti persentil, rata-rata, standar deviasi dan lainnya dari atribut dalam dataset digunakan fungsi df.describe().



Gambar Kode untuk detail statistik dataset

**2.3. Explore Data**

Pada bagian ini, dataset ditelaah untuk mendapatkan informasi terkait kondisi dari dataset. Penelaahan terhadap data dimulai dengan memperhatikan informasi terkait kolom yang terdapat di dalam dataset beserta data deskripsinya. Eksplorasi data dimulai dengan memperhatikan informasi terkait tipe dari setiap fitur/ variabel dataset dengan menggunakan fungsi info()

****

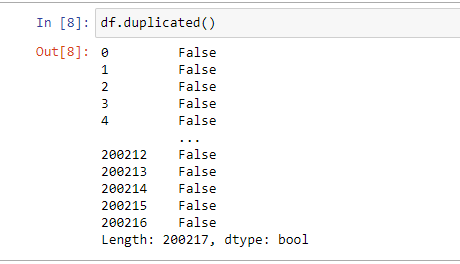
Gambar. Kode untuk melihat type atribut

Informasi Fitur/variabel yang menunjukkan tipe data dan keterangan dari setiap variabel pada dataset

Table Keterangan Atribut

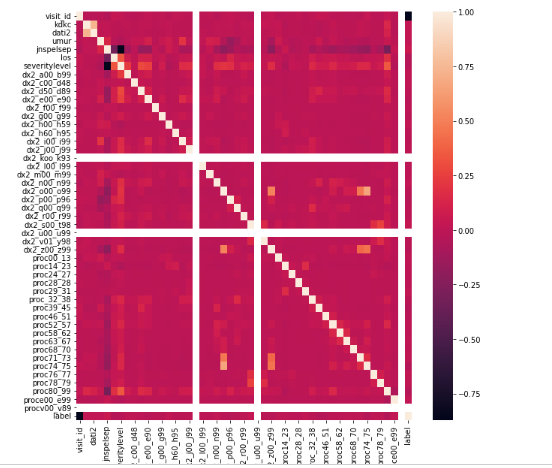
| **No** | **Variabel** | **Tipe Variabel** | **Deskripsi** |
| --- | --- | --- | --- |
| 1 | visit\_id | int64 | id kunjungan |
| 2 | Kdkc | int64 | kode wilayah kantor cabang BPJS Kesehatan |
| 3 | dati2 | int64 | kode kabupaten/kota |
| 4 | typeppk | object | kode tipe Rumah Sakit |
| 5 | jkpst | object | jenis kelamin peserta JKN-KIS |
| 6 | umur | int64 | umur peserta saat mendapatkan pelayanan rumah sakit |
| 7 | jnspelsep | int64 | tingkat pelayanan; 1:rawat inap; 2: rawat jalan |
| 8 | los | int64 | lama peserta dirawat di rumah sakit |
| 9 | cmg | object | klasifikasi CMG (Case Mix Group) |
| 10 | severitylevel | int64 | tingkat urgensi |
| 11 | diagprimer | object | diagnosa primer |
| 12 | dx2\_...\_... | int64 | diagnosa sekunder |
| 13 | proc..\_... | int64 | kode kelompok procedure |
| 14 | label | int64 | flag fraud; 1:fraud; 0:tidak fraud |

Setelah itu, dataset dipastikan kembali apakah masih ada duplikasi data, yaitu dengan menggunakan fungsi df.duplicate(),. Berikut dapat dilihat dari data tersebut bahwa tidak ada data yang duplikasi.



Gambar Kode untuk melihat duplikat

Korelasi antar atribut perlu untuk ditinjau. Maka dari itu, korelasi antar masing-masing atribut divisualisasikan menggunakan visualisasi heatmap. Ukuran dari visualisasi diatur menggunakan library matplotlib, sementara untuk menampilkan heatmap menggunakan library seaborn. Menggunakan korelasi ini dapat memperlihatkan bahwa setiap atribut tersebut memiliki korelasi sebesar 0. Semakin terang hasil dari sebuah pemetaan dari heatmap, maka semakin rendah korelasi antara atributnya.



Gambar Visualisasi korelasi antar atribut

Setelah berhasil mendapatkan korelasi antar atribut, maka langkah selanjutnya melakukan penelaahan terhadap value dari masing-masing atribut. Dari tampilan visualisasi histogram dibawah. Tampilan value atribut dalam dataset, maka didapatkan hal-hal berikut ini :

1. Atribut yang memiliki variasi value terbanyak adalah atribut kdkc, dati2, dan umur.
2. Atribut kdkc dengan value 1000 memiliki frekuensi tertinggi dan value 2250 memiliki frekuensi terendah.
3. Atribut kdkc menunjukkan kode wilayah kantor cabang BPJS Kesehatan, yang

menunjukkan bahwa kode kdkc dengan value sekitar 1000 memiliki jumlah pasien

terbanyak.

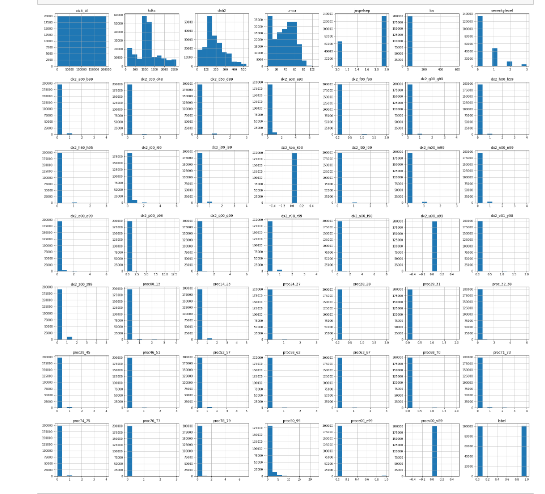
1. Atribut dati2 dengan value 100 memiliki frekuensi tertinggi dan value 500 terendah

yang menunjukkan bahwa kabupaten dengan kode 100 memiliki jumlah pasien

tertinggi.

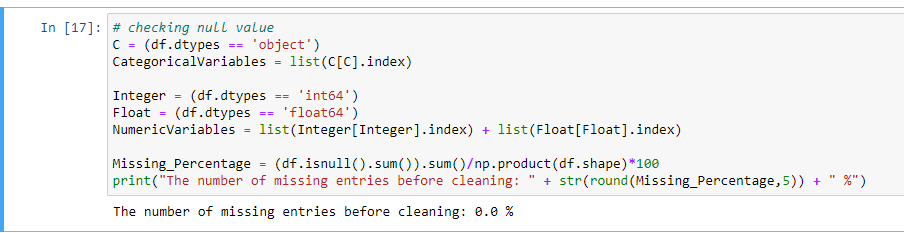
1. Untuk umur pasien dengan jumlah terbanyak adalah pasien dengan umur sekitar 0

bulan dan umum pasien dengan jumlah terkecil adalah umur 80.



Gambar Value setiap Atribut

Gambar diatas menunjukkan visualisasi dari keseimbangan data yang ada di masing-masing fitur. Dapat dilihat dari data diatas dianggap keseimbangan pada fitur label. Label merupakan fitur acuan dalam melatih dan menguji data apakah data tersebut fraud atau tidak fraud.



Gambar Kode untuk melihat *Missing value*

**2.4 Verify Data Quality**

*Data validation* dilakukan dengan maksud memastikan bahwa pemodelan terjadi pada data yang benar. Data yang salah yang digunakan sebagai data pelatihan untuk model akan menghasilkan pengetahuan yang salah. Validasi data dilakukan segera setelah persiapan data,

dan sebelum pemodelan data. Itu karena selama persiapan data ada kemungkinan besar terjadi kesalahan terutama dalam skenario yang kompleks. Validasi data harus dilakukan dengan melibatkan minimal satu orang eksternal yang memiliki pemahaman yang tepat tentang data dan bisnis. Dataset yang digunakan terdiri atas data kuantitatif, yaitu data yang dapat diukur (measurable) atau dapat dihitung sebagai angka atau bilangan. Data tersebut dapat berupa bilangan diskrit atau bilangan kontinu. Data kuantitatif memiliki kecenderungan dapat dianalisis dengan teknik statistik. Data yang termasuk kuantitatif pada dataset adalah Quantity (QTY) dan Value.

**BAB 3**

**DATA PREPARATION**

Data preparation merupakan langkah setelah dilakukannya pengumpulan data awal yang telah dilakukan pada fase crisp-dm sebelumnya, yaitu business understanding. Pada tahap data preparation ini, dilakukan proses menyiapkan data awal, memilih variabel yang akan dianalisis dan membersihkan data. Dalam pengerjaan proyek, bahasa pemrograman yang digunakan adalah pemrograman python dengan software pengolah data Jupyter Notebook. Pada bab ini akan dijelaskan mengenai proses apa saja yang akan dilakukan untuk mempersiapkan data seperti sorting, cleaning, construction, binning dan normalization.

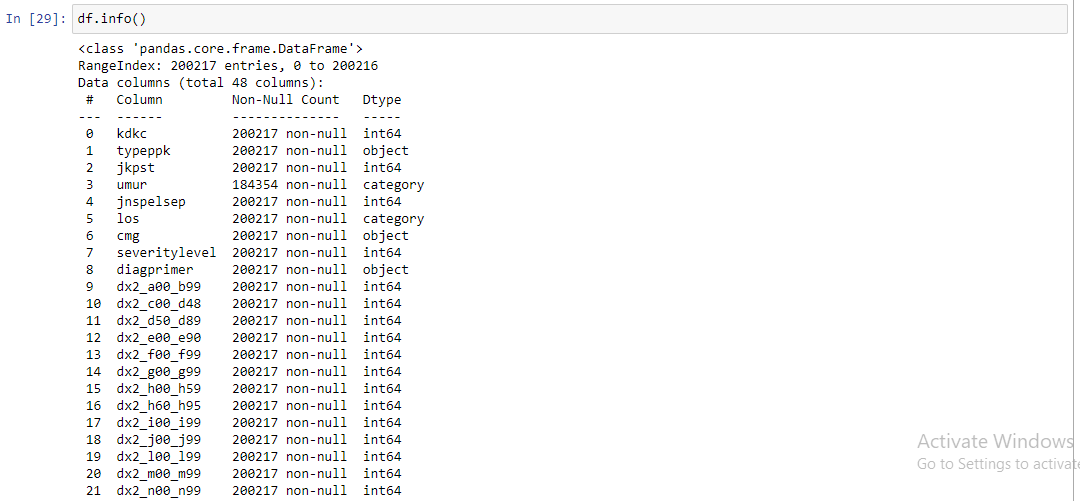
***3.1. Sorting Data***

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



**Gambar.** Kode untuk menghapus fitur tertentu.

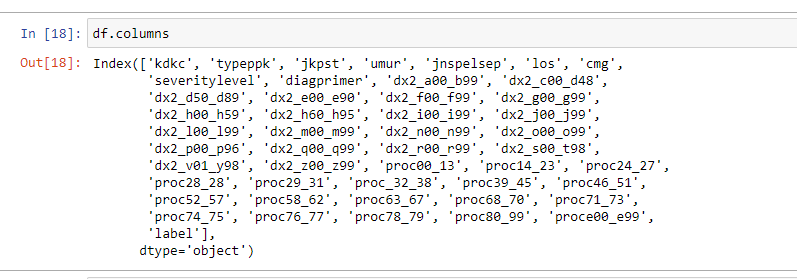
Atribut-atribut tersebut di *drop* bertujuan agar data dapat digunakan lebih efisien dan efektif dalam pengolahan data termasuk dalam penggunaan memory. Berikut ini merupakan tampilan setelah atribut yang tidak digunakan telah di *drop.*

**

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



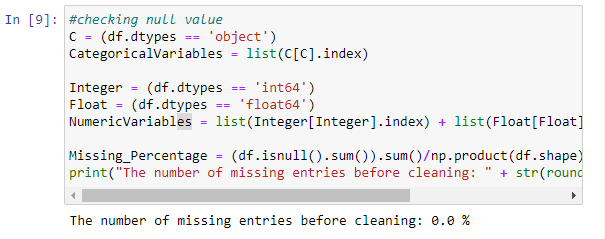




Gambar. informasi mengenai 53 fitur pada dataset.

**3.2 *Cleaning Data***

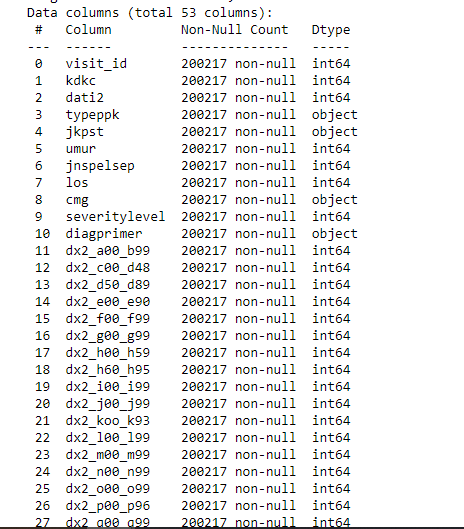
Dalam tahap *data cleaning* akan dilakukan pembersihan data. Pembersihan data akan menangani objek data yang kosong (*missing value*). Maka dari itu, terlebih dahulu dilakukan pemeriksaan data untuk memeriksa apakah terdapat nilai yang hilang pada dataset yang digunakan.

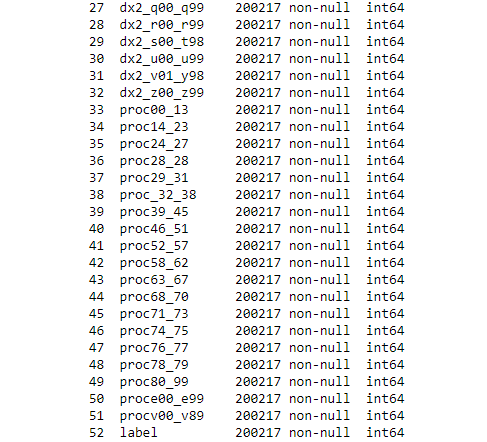


Gambar. Kode untuk melihat *missing value*

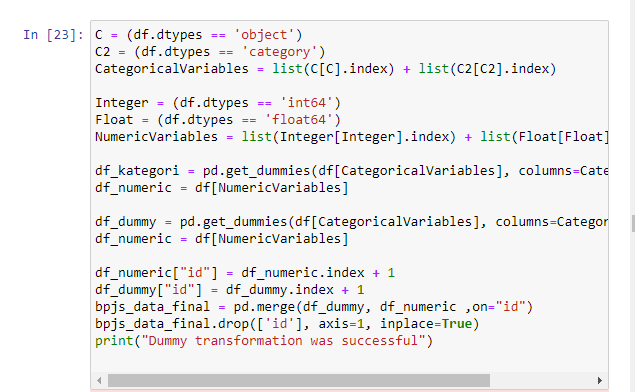
**3.3 *Construct Data***

Pada fase ini akan 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:

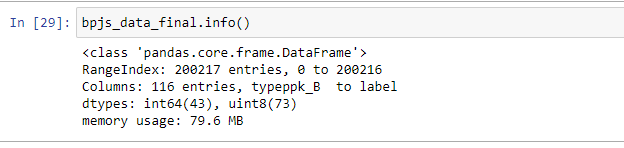




Dapat dilihat dari gambar diatas bahwa terdapat 4 atribut yang bertipe data kategorik (object64), maka perlu dilakukan transformasi data. Untuk itu perlu dilakukan transformasi data tipe pada atribut dengan menjalankan potongan kode berikut:



Setelah dilakukan transformasi data, maka dilakukan pengecekan pada tipe atribut menggunakan fungsi df.info ( )



Gambar Pengecekan atribut fitur setelah Transformasi

**3.4 *Binning***

Di langkah ini dilakukan proses transformasi data dengan menggunakan metode binning. Metode ini akan digunakan untuk mengelompokkan data numerik menjadi beberapa bin dengan tujuan memudahkan pemahaman pada persebaran data yang digunakan. Berdasarkan analisis yang didapatkan, diketahui bahwa fitur umur dan LoS merupakan nilai bertipe numerik dan memiliki persebaran data yang tidak merata. Oleh karena itu akan dilakukan proses binning pada kedua fitur tersebut. Untuk fitur umur akan dibagi menjadi 5 kategori dengan bin yang ditentukan adalah sesuai dengan kategori usia berdasarkan WHO yaitu sebagai berikut.

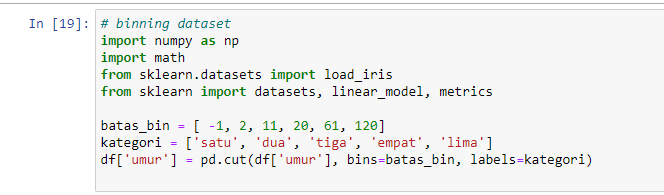
Bin 1: umur <=1,

Bin 2: 2<= umur >=10,

Bin 3: 11<= umur >=19,

Bin 4: 20<= umur >=60,

Bin 5: umur > 60



Gambar. Binning untuk fungsi fitur umur

Untuk fitur LoS yang memiliki hubungan terhadap jnpsplesep yang terkait pada tipe rawat inap atau rawat jalan selanjutnya akan dikelompokkan menjadi 4 kategori yaitu ‘rawat jalan’, ‘short stay’, ‘medium stay’, ‘long stay’. Penentuan bin adalah sebagai berikut. los = 0

0 : rawat jalan,

1-5 : short stay,

6- 10 : medium stay,

> 10 : long stay

gambar

**3.5 *Standardization***

Pada tahap ini dilakukan standarisasi pada data yang telah diolah sebelumnya agar mendapatkan hasil yang lebih baik lagi. Maka untuk menjalankan perintah ini, terlebih dahulu dilakukan pembagian data ke dalam variabel x dan y seperti berikut.

gambar

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.

gambar

*Output* yang dihasilkan adalah sebagai berikut:

gambar

**BAB 4**

**MODELLING**

Pada bab ini akan menjelaskan mengenai pemilihan teknik modelling, menghasilkan test design, membangun model, dan menilai model yang telah dibangun. Pada modelling ini, data preparation akan dioperasikan yang kemudian akan menjelaskan masalah bisnis yang ditimbulkan selama proses business understanding.

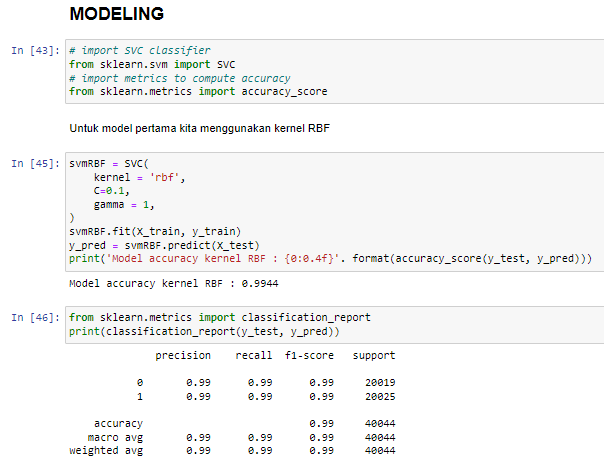
***4.1 Selection Modelling Technique***

Teknik pemodelan yang digunakan pada proyek ini didorong oleh tujuan penambangan data yang ingin dicapai dalam proyek. Penerapan algoritma support vector machine cocok digunakan dalam teknik pemodelan dalam pengerjaan proyek ini dikarenakan SVM adalah algoritma pembelajaran terawasi yang sangat efektif digunakan untuk classification. Dalam algoritma SVM, pada data pelatihan, algoritma mencoba menemukan hyperplane optimal terbaik yang dapat digunakan untuk melakukan klasifikasi data. Biasanya dalam SVM akan bekerja dengan menemukan contoh yang paling mirip antar kelas sehingga akan dijadikan sebagai vektor pendukung. Untuk menentukan model yang sesuai biasanya akan didasarkan pada pertimbangan berikut:

1. Tipe data yang tersedia untuk mining
2. Tujuan data mining
3. Persyaratan pemodelan khusus

**4.1.1 *Modelling Techniques***

Teknik pemodelan yang digunakan pada proyek ini adalah algoritma support vector machine (SVM) sesuai dengan tujuan data mining yaitu menggali Discovering Knowledge mengenai pola (pattern) item mengenai Fraud Detection menggunakan dataset BPJS Hackathon. Algoritma support vector machine (SVM) adalah sebuah algoritma klasifikasi berdasarkan prinsip linear classifier yang mampu menyelesaikan permasalahan dengan waktu komputasi lebih cepat daripada SVM standar untuk data yang berukuran besar.

****

**4.1.2 Modelling Assumptions**

Pada teknik ini akan dilakukan teknik pemodelan dengan *Support Vector Machine (SVM)* yang memerlukan asumsi spesifik terhadap data, yaitu semua atribut yang sama, tidak ada missing value. Pembuatan bin terlebih dahulu sebelum dilakukan penerapan algoritma SVM pada data tersebut untuk atribut yang tidak kategorikal (nominal).

***4.2. Generate Test Design***

Sebelum dilakukannya pembangunan model, perlu dilakukan perancangan terhadap bagaimana model akan diuji. Untuk mendapatkan hasil test design yang komprehensif yaitu dengan cara menentukan data yang akan menguji kriteria. Dimana kriteria model ini akan dinilai bergantung pada data mining goals pada model yang akan dibangun.

**4.2.1 Test Design**

Desain pengujian (test design) merupakan gambaran langkah-langkah yang akan dilakukan untuk menguji model yang dihasilkan. Pada proyek ini, langkah-langkah untuk menguji model adalah sebagai berikut :

1. Mengekstrak test data yaitu record yang tidak digunakan dalam training set.
2. Menghitung instance yang benar di mana premisnya mengarah ke kesimpulan.
3. Menghitung confidence setiap aturan dari jumlah yang benar.
4. Mencetak aturan asosiasi terbaik dengan judul.

**4.3 Build Model**

Pada proses pembuatan model, terdapat tiga informasi yang akan digunakan dalam keputusan data mining, diantaranya:

1. Parameter settings Parameter settings adalah pengaturan parameter yang mencakup catatan mengenai parameter yang memberikan hasil yang terbaik.
2. Models Models dimana model aktual yang diproduksi

**4.3.1 Parameter Settings**

Pada sebagian besar teknik modeling mempunyai beberapa parameter yang dapat disesuaikan untuk mengamati dan mengendalikan proses modeling. Pada proyek ini menggunakan parameter C, kernel, dan gamma untuk menentukan nilai parameter-parameter model.

**4.3.2 Models**

Pada bagian ini, setelah menentukan parameter yang akan dipakai dan dibutuhkan pada proyek, langkah selanjutnya adalah mengeksekusi model untuk menghasilkan result atau output yang terlihat.

**4.4 Assess Model**

Assess model merupakan tahapan yang dilakukan untuk menilai kesesuaian model yang telah dibangun dengan kriteria sukses yang telah didefinisikan. Secara umum, hasil yang diperoleh dari pembangunan model dengan menggunakan algoritma SVM telah menghasilkan rule yang baik.

**BAB 5**

**EVALUATION**

Pada tahap Evaluation (Evaluasi), akan dijelaskan mengenai evaluasi terhadap model untuk memprediksi Fraud Detection yang dihasilkan dengan menggunakan algoritma SVM. Evaluasi adalah fase interpretasi terhadap hasil data mining. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap modelling sesuai dengan sasaran yang ingin dicapai dalam tahap business understanding.

**5.1 Evaluate Result**

Tahap ini dilakukan untuk mengetahui performa SVM dengan menggunakan dataset yang diperoleh. Dari pemodelan yang dilakukan pada tahap sebelumnya, dilakukan implementasi menggunakan bahasa pemrograman python.

**5.2 *Determine Next Steps***

Tahapan ini menentukan langkah apa yang akan diambil selanjutnya. Berdasarkan hasil evaluasi terhadap model yang digunakan dengan algoritma SVM, jika telah menghasilkan yang terbaik maka diputuskan pengerjaan proyek akan dilanjutkan ke tahap akhir yakni deployment.

**DAFTAR PUSTAKA**

[1] A. A. P. d. A. Purwarianti, "Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM (Studi Kasus: X Music Indonesia)," Jurnal Sarjana Institut Teknologi Bandung bidang Teknik Elektro dan Informatika, 2011.

[2] T. P. a. Y. C. I. Budiman, "Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan

Pola Proporsi Pelaksanaan Tridharma," J. Sist. Inf. BISNIS, 2014.