**LAPORAN PROYEK DATA MINING**

***Fraud Detection with Binary Classification Using ANN Algorithm***



**Disusun oleh:**

12S19001 – Raynaldo Silalahi

12S19009 – Manuel Sigalingging

12S19040 – Abel M. Y. Tampubolon

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO INSTITUT TEKNOLOGI DEL**

**NOVEMBER 2022**

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc122727376)

[DAFTAR GAMBAR 3](#_Toc122727377)

[PENDAHULUAN 4](#_Toc122727378)

[BUSINESS UNDERSTANDING 6](#_Toc122727379)

[1.1 Menentukan Objektif Bisnis 6](#_Toc122727380)

[1.2 Menentukan Tujuan Bisnis 7](#_Toc122727381)

[1.3 Rencana Pengerjaan Proyek 7](#_Toc122727382)

[DATA UNDERSTANDING 9](#_Toc122727383)

[2.1 Mengumpulkan dan Menelaah Data 9](#_Toc122727385)

[2.2 Deskripsi Data 9](#_Toc122727386)

[2.3 Telaah Data 11](#_Toc122727387)

[2.4 Memvalidasi Data 16](#_Toc122727388)

[DATA PREPARATION 18](#_Toc122727389)

[3.1 Selecting Data 18](#_Toc122727391)

[3.2 Data Cleaning 19](#_Toc122727392)

[3.3 Construct Data 19](#_Toc122727393)

[3.4 Binning 20](#_Toc122727394)

[3.5 Standardization 21](#_Toc122727395)

[MODELLING 23](#_Toc122727396)

[4.1 Build Test Scenario 23](#_Toc122727398)

[4.2 Build Modeling 23](#_Toc122727399)

[MODEL EVALUATION 25](#_Toc122727400)

[5.1 Evaluation of Modelling Result 25](#_Toc122727402)

[5.2 Modelling Process Review 26](#_Toc122727403)

[DEPLOYMENT 27](#_Toc122727404)

[LAMPIRAN 28](#_Toc122727405)

# DAFTAR GAMBAR

[Gambar 1 4](file:///C:\Users\tampu\OneDrive\Documents\GitHub\Project-Data-Mining\Document\LaporanProyek_6.docx#_Toc122727668)

[Gambar 2 12](#_Toc122727669)

[Gambar 3 12](#_Toc122727670)

[Gambar 4 12](#_Toc122727671)

[Gambar 5 13](#_Toc122727672)

[Gambar 6 13](#_Toc122727673)

[Gambar 7 13](#_Toc122727674)

[Gambar 8 14](#_Toc122727675)

[Gambar 9 14](#_Toc122727676)

[Gambar 10 15](#_Toc122727677)

[Gambar 11 16](#_Toc122727678)

[Gambar 12 16](#_Toc122727679)

[Gambar 13 17](#_Toc122727680)

[Gambar 14 18](#_Toc122727681)

[Gambar 15 18](#_Toc122727682)

[Gambar 16 19](#_Toc122727683)

[Gambar 17 19](#_Toc122727684)

[Gambar 18 20](#_Toc122727685)

[Gambar 19 21](#_Toc122727686)

[Gambar 20 21](#_Toc122727687)

[Gambar 21 22](#_Toc122727688)

[Gambar 22 22](#_Toc122727689)

[Gambar 23 25](#_Toc122727690)

[Gambar 24 25](#_Toc122727691)

[Gambar 25 25](#_Toc122727692)

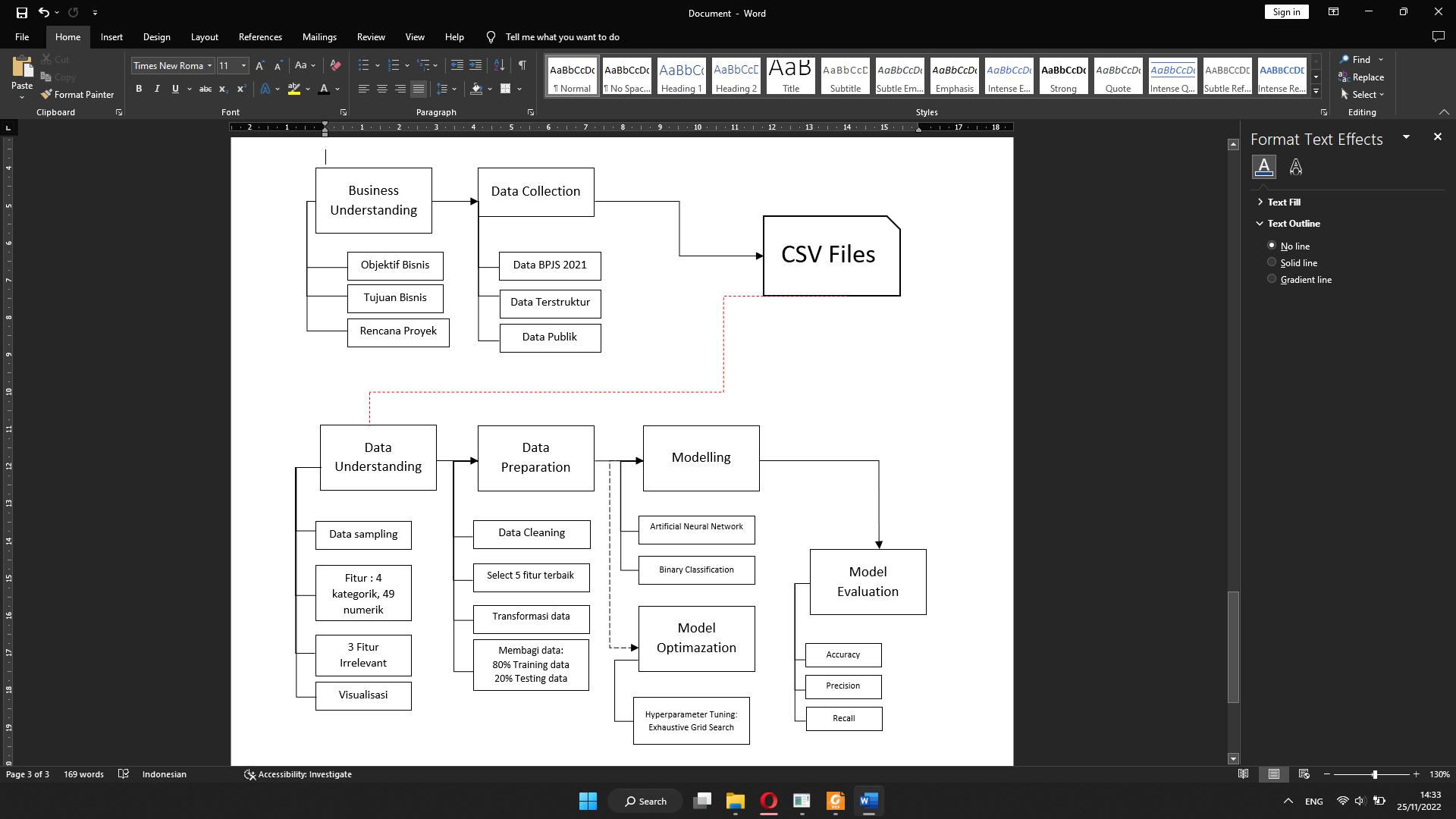
[Gambar 26 26](#_Toc122727693)

[Gambar 27 27](#_Toc122727694)

[Gambar 28 28](#_Toc122727695)

# PENDAHULUAN

Salah satu jenis cara dalam melakukan klasifikasi ialah Binary Classification. Binary Classification merupakan proses mengklasifikasikan dalam elemen yang membagi menjadi dua kelas antara normal dan abnormal (0 dan 1). Perencanaan proyek yang akan dibahas mengenai fraud detection terhadap data BPJS digunakan algoritma *Artificial Neural Network* atau ANN dalam melakukan klasifikasi. Metodologi pengerjaan proyek yang digunakan yakni metodologi CRISP-DM. Untuk tahapan terhadap pengerjaan proyek dengan metode CRISP-DM dapat dilihat pada gambar berikut:



Gambar 1

Tahapan dalam pengerjaan proyek terdiri atas  *business understanding* dimana menentukan objektif bisnis, tujuan bisnis, dan rencana pengerjaan proyek. Lalu *data collection*, pengumpulan dataset yang akan digunakan dalam membangun model. Selanjutnya *Data preparation*, untuk mempersiapkan data yang akan digunakan berupa membersihkan dan mentransformasi data. *Modelling* untuk membangun model klasifikasi dan *Model Evaluation* untuk mengevaluasi hasil akurasi dari model yang dibangun. Apabila model tercapai dibangun dan memenuhi target akurasi yang diinginkan, maka model akan dibangun pada tahap *Deployment Mode.*

# BUSINESS UNDERSTANDING

Proses awal dalam metodologi yang digunakan pada CRISP-DM diawali dengan business understanding, yang dimana pada proses tahapan ini memberikan penjelasan tentang objektif bisnis, tujuan bisnis, dan rencana pengerjaan proyek.

## Menentukan Objektif Bisnis

BPJS singkatan dari Badan Penyelenggara Jaminan Sosial merupakan sebuah badan hukum yang memiliki tujuan dalam memberikan jaminan mengenai perlindungan dan pelayanan kesehatan terhadap penggunanya. Untuk setiap jaminan kesehatan seseorang menjadi jaminan perlindungan bagi mereka yang telah membayarkan iuran wajib atau subsidi dari pemerintah kepada pihak BPJS. Iuran atau pembayaran yang dilakukan berdasarkan kelompok atau golongan kepesertaannya dalam BPJS.

Tetapi, munculnya masalah dalam pengelolaan dana pada BPJS yang kurang efektif mengenai jaminan sosial BPJS Kesehatan Indonesia yang diberikan. Pembayaran yang telah dilakukan tidak sesuai dengan daftar ketentuan yang telah ditetapkan oleh BPJS. Penyebab terjadinya disebabkan oleh oknum tertentu dengan melakukan manipulasi pengolahan data demi kepentingan sendiri dan peningkatan kerugian BPJS secara cepat.

Fraud merupakan suatu kegiatan yang sering terjadi di mana saja. Seluruh tindak *fraud/*penipuan tentu merugikan pihak yang menjadi korban. Pencegahan dalam *fraud/*penipuan harus diidentifikasi dengan serius agar mampu mengetahui pihak-pihak yang terjaring dalam penipuan. Salah satu tindak kejahatan *fraud/*penipuan yang terjadi pada BPJS menjadi adanya potensi yang menimbulkan kerugian kepada BPJS. Dalam menghindari timbulnya kerugian yang sangat besar tersebut, tentu dilakukannya upaya pencegahan. Salah satu upaya pencegahan yang mampu dilakukan yakni dengan menganalisis apa yang menjadi faktor yang mempengaruhi terjadinya *fraud* dan dilakukannya prediksi terhadap timbulnya terjadi penipuan dengan menggunakan *data mining*.

Dalam proyek ini mengubah klaim asuransi kesehatan pada pelayanan kesehatan di rumah sakit sebagai titik fokus untuk memprediksi *fraud,* sehingga objektif pada proyek ini yaitu:

* Melakukan prediksi potensi terjadinya *fraud* pada klaim pelayanan rumah sakit
* Mencari faktor-faktor yang mempengaruhi terjadinya penipuan

Proyek ini dapat dikatakan berhasil jika:

* Ditemukan prediksi terhadap potensi terjadinya *fraud* pada klaim pelayanan rumah sakit
* Ditemukan faktor-faktor yang mempengaruhi terjadinya *fraud*

## Menentukan Tujuan Bisnis

Penipuan yang terjadi menyebabkan kerugian yang dapat mengancam keberadaan bisnis menjadi berubah. Maka, tujuan bisnis dari pengerjaan proyek ini ialah melakukan prediksi terhadap munculnya atau terjadinya penipuan/*fraud* terhadap klaim dalam pelayanan BPJS Kesehatan dan melakukan evaluasi terhadap faktor yang mempengaruhi terhadap penipuan atau *fraud* dengan mengidentifikasi gambaran relasi antar data. Dengan mengetahui gambaran yang diperoleh, manfaat yang dapat diberikan dari penelitian ini dapat berguna untuk melakukan evaluasi proses terhadap klaim pelayanan BPJS rumah sakit ke depannya agar tujuan bisnisnya sesuai dengan yang diharapkan.

## Rencana Pengerjaan Proyek

Pada tahapan perencanaan yang dilakukan dalam proyek ini mengacu pada Standar Kompetensi Kerja Nasional: KepMen Ketenagakerjaan No 299 tahun 2020 dengan metodologi CRISP-DM. Berikut tahapan perencanaan dalam pengerjaan proyek.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aktivitas** | **Detail** | **Waktu (Minggu Ke-)** | | | | |
| **12** | **13** | **14** | **15** | **16** |
| Pemilihan Kasus dan Algoritma | Pemilihan Kasus |  |  |  |  |  |
| Penentuan Algoritma |  |  |  |  |  |
| Business Understanding | Menentukan Objektif Bisnis |  |  |  |  |  |
| Menentukan Tujuan Bisnis |  |  |  |  |  |
| Membuat Rencana Proyek |  |  |  |  |  |
| Data Understanding | Mengumpulkan Data |  |  |  |  |  |
| Mendeskripsikan dan Menelaah Data |  |  |  |  |  |
| Memvalidasi Data |  |  |  |  |  |
| Data Preparation | Memilah Data |  |  |  |  |  |
| Membersihkan Data |  |  |  |  |  |
| mengkonstruksi Data |  |  |  |  |  |
| Menentukan Label Data |  |  |  |  |  |
| Mengintegrasikan Data |  |  |  |  |  |
| Modelling | Membangun Skenario Pengujian |  |  |  |  |  |
| Membangun Model |  |  |  |  |  |
| Model Evaluation | Mengevaluasi Hasil Pemodelan |  |  |  |  |  |
| Melakukan Review Proses Pemodelan |  |  |  |  |  |

Dalam mendukung pada pengerjaan proyek untuk setiap tahapan, berikut spesifikasi dalam development environment yang digunakan untuk proyek ini.

* Tools : 1. Jupyter Notebook

2. Google Collab

* Bahasa Pemrograman : Python
* Algoritma : Artificial Neural Network

# DATA UNDERSTANDING



## Mengumpulkan dan Menelaah Data

Tahap kedua pada metodologi CRISP-DM adalah *data understanding,* dimana pada tahap ini akan mendapatkan informasi mengenai kekurangan dan kelebihan data, tingkat kesesuaian data dengan bisnis yang akan dipecahkan, hingga ketersediaan data. Tahapan ini diperlukan karena dataset belum tentu bisa langsung digunakan.

## Deskripsi Data

*Dataset* yang digunakan untuk *fraud detection* BPJS Kesehatan 2021 adalah fraud\_detection\_train *dataset*. Dataset ini berisi 53 atribut dan 200217 *rows*. Berikut tabel deskripsi setiap atributnya:

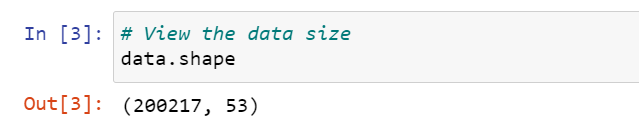
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Atribut | Keterangan | Tipe | Nilai |
| 1 | visit\_id | Id kunjungan | int64 | ID numerik |
| 2 | kdkc | Kode wilayah kantor cabang BPJS Kesehatan | int64 | Kode numerik |
| 3 | dati2 | Kode kabupaten/kota | int64 | Kode numerik |
| 4 | typeppk | Kode tipe rumah sakit | object | SC, C, B, SD, SB, A, ,D, I3, KM, KI, I2, I4, KJ, KL, I1, KB, KC, GD, SA, KP, KO, KG, HD, KT, KU |
| 5 | jkpst | Jenis kelamin peserta JKN-KIS | object | P (Perempuan), L (Laki-Laki) |
| 6 | umur | Umur peserta saat mendapatkan pelayanan di rumah sakit | int64 | 0-109 |
| 7 | jnspelsep | Tingkat layanan: 1 (rawat inap), 2 (rawat jalan) | int64 | 1: Rawat Inap  2: Rawat Jalan |
| 8 | los | Lama peserta dirawat di rumah sakit | int64 | 0-255 |
| 9 | cmg | Klasifikasi CMG (Case Mix Group) | object | 'F','E','Q','L','H','W','P','U','K','G','M','N','A','C', 'D', 'Z', 'J', 'O', 'S', 'I', 'V', 'T', 'B' |
| 10 | severitylevel | Tingkat urgensi | int64 | 0, 1, 2, 3 |
| 11 | diagprimer | Diagnosa primer | object | 'f00\_f99', 'e00\_e90', 'r00\_r99', 'j00\_j99', 's00\_t98', 'h00\_h59', 'm00\_m99', 'c00\_d48', 'z00\_z99', 'p00\_p96', 'h60\_h95', 'k00\_k93', 'g00\_g99', 'i00\_i99', 'l00\_l99',  'a00\_b99', 'n00\_n99', 'o00\_o99','d50\_d89',  'q00\_q99', 'u00\_u85' |
| 12 | dx2\_…\_… | Diagnosa Sekunder | int64 | terdapat 22 atribut dimana nilainya adalah 0 - 13 |
| 13 | proc…\_ | Kode Procedure | int64 | terdapat 19 atribut dimana nilainya adalah 0 - 23 |
| 14 | label | Flag fraud: 1 (Fraud), 0 (Tidak Fraud) | int64 | 1:fraud; 0:tidak fraud |

## Telaah Data

Exploratory Data Analysis merupakan salah satu pendekatan untuk memahami dan menganalisis data. Dengan melakukan EDA kita dapat mendapatkan informasi berupa konteks data, statistik dataset, keseimbangan dataset, tipe data pada tiap atribut, hingga kualitas data.

Hasil analisis EDA sebagai berikut:

* Melihat ukuran data dari dataset faud\_detection\_train.csv



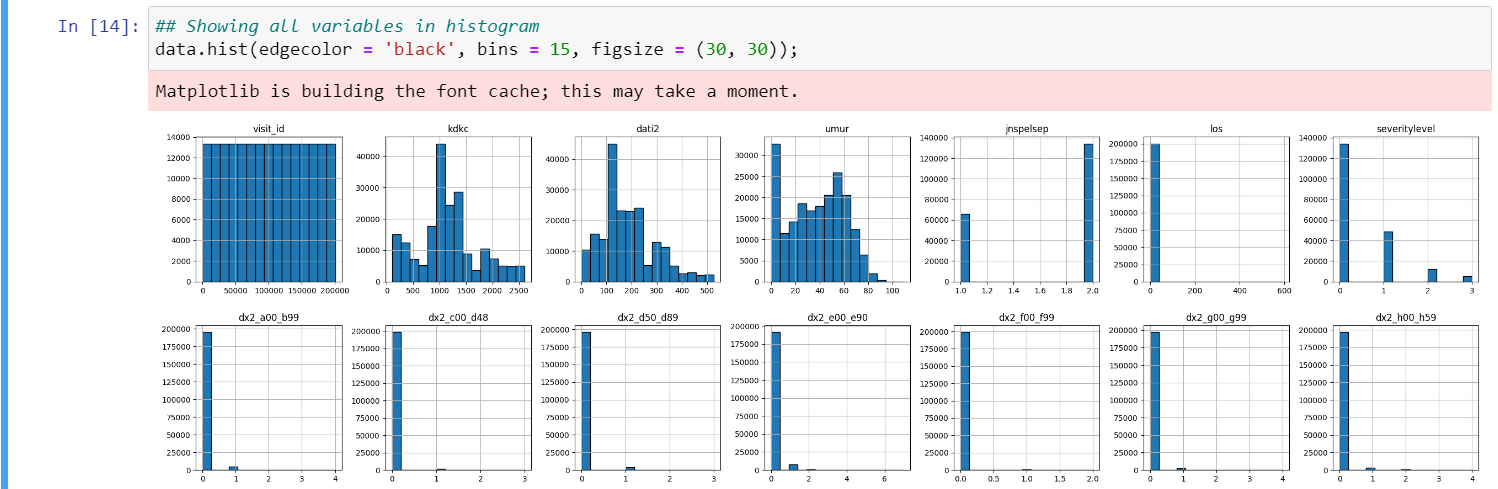
Gambar 2

* Melihat daftar fitur atau variabel yang ada pada dataset. pada dataset ini ditemukan 53 fitur.

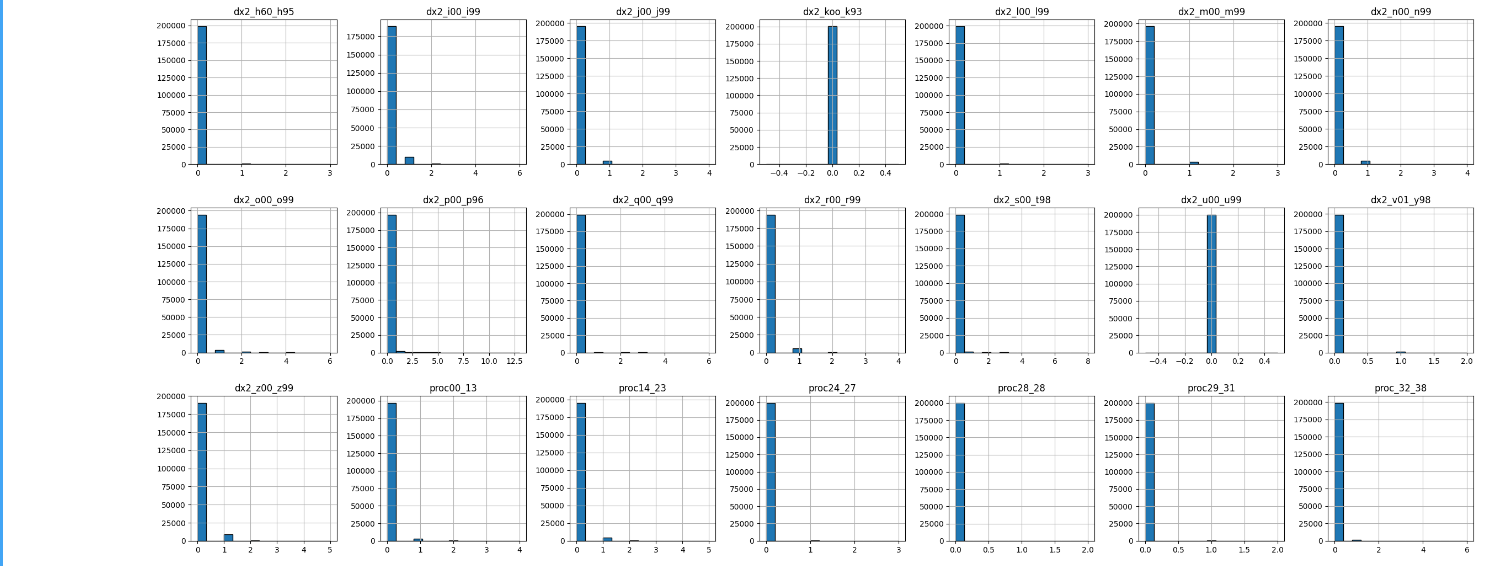


Gambar 3

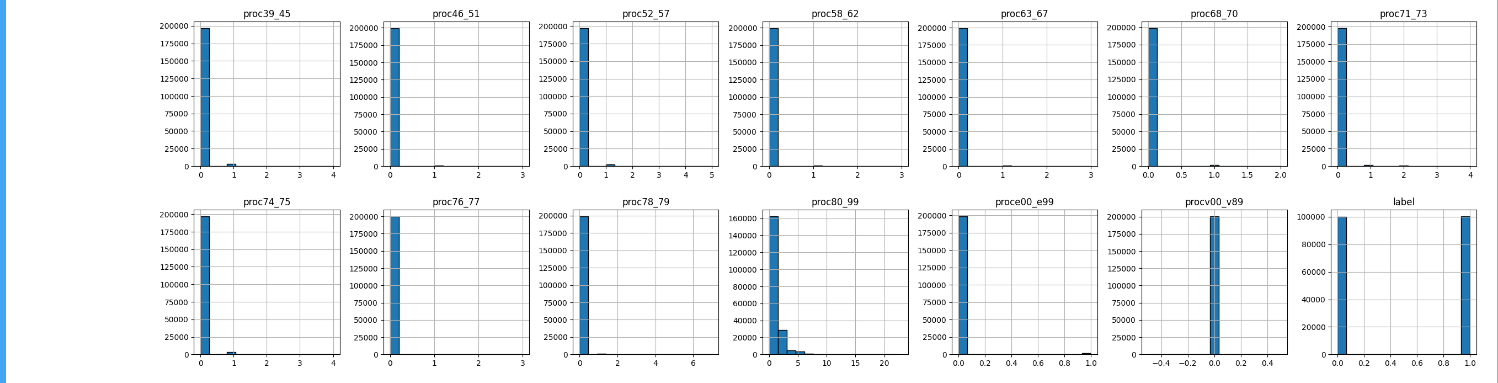
* Menampilkan semua variabel pada dataset dengan histogram



Gambar 4

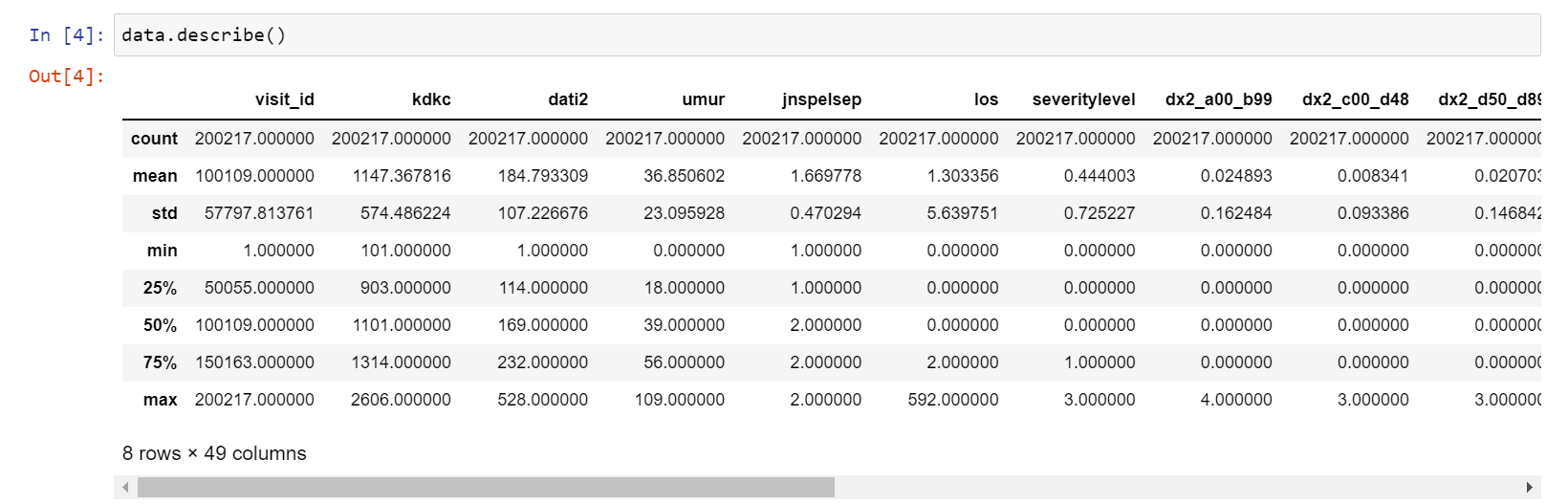


Gambar 5



Gambar 6

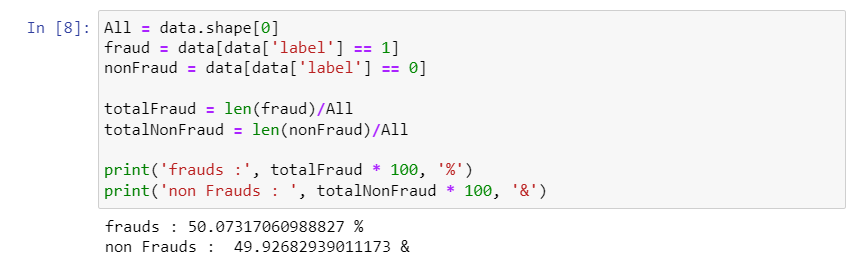
* Memeriksa statistik dataset yang digunakan dengan menggunakan fungsi *data.describe()*. Adapun kode program beserta output sebagai berikut:



Gambar 7

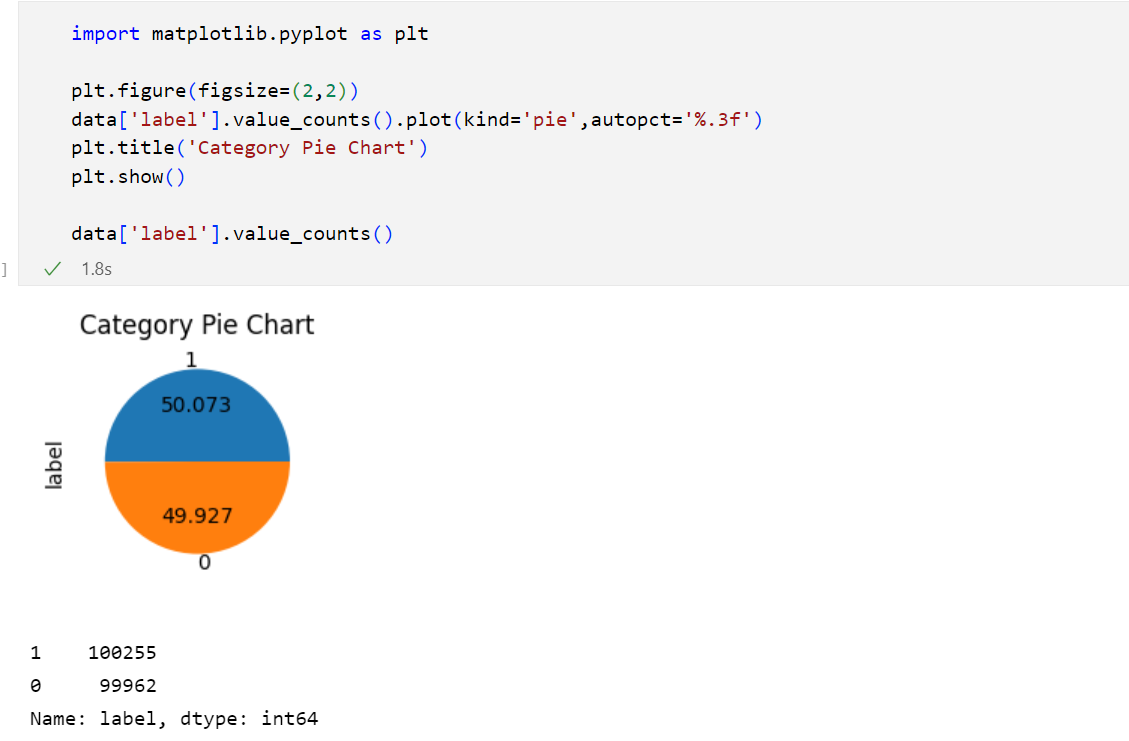
Output menampilkan rangkuman data set yang meliputi *count*, *mean*, *standard of deviation*, *minimum value*, *quantile* 1 (25%), *quantile* 2 (50%), *quantile* 3 (75%), dan *maximum value*.

* Memeriksa keseimbangan dataset yang digunakan. Adapun kode program yang digunakan sebagai berikut:



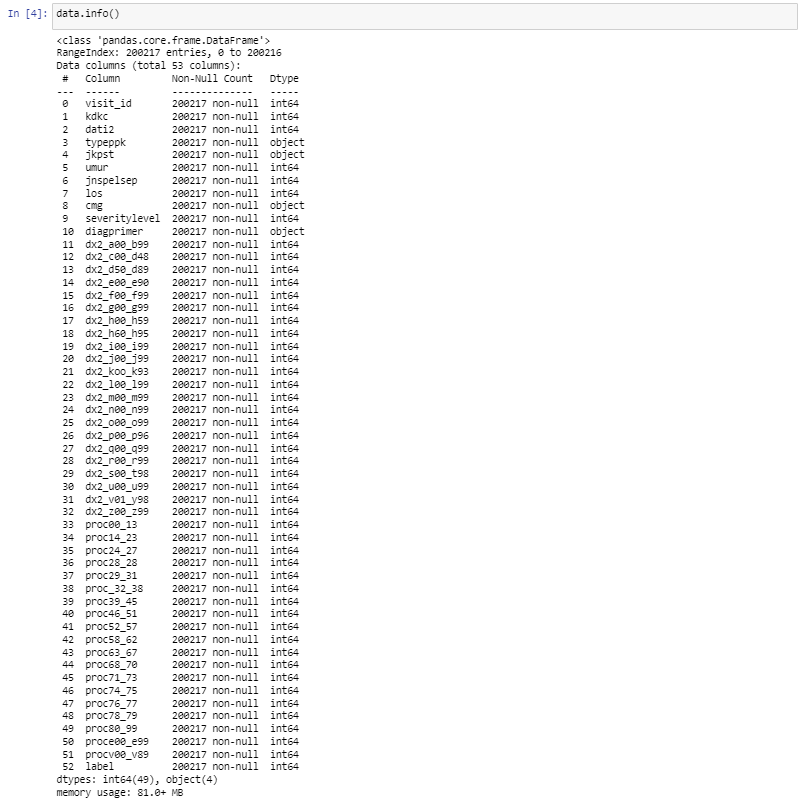
Gambar 8

* Memvisualisasikan keseimbangan data



Gambar 9

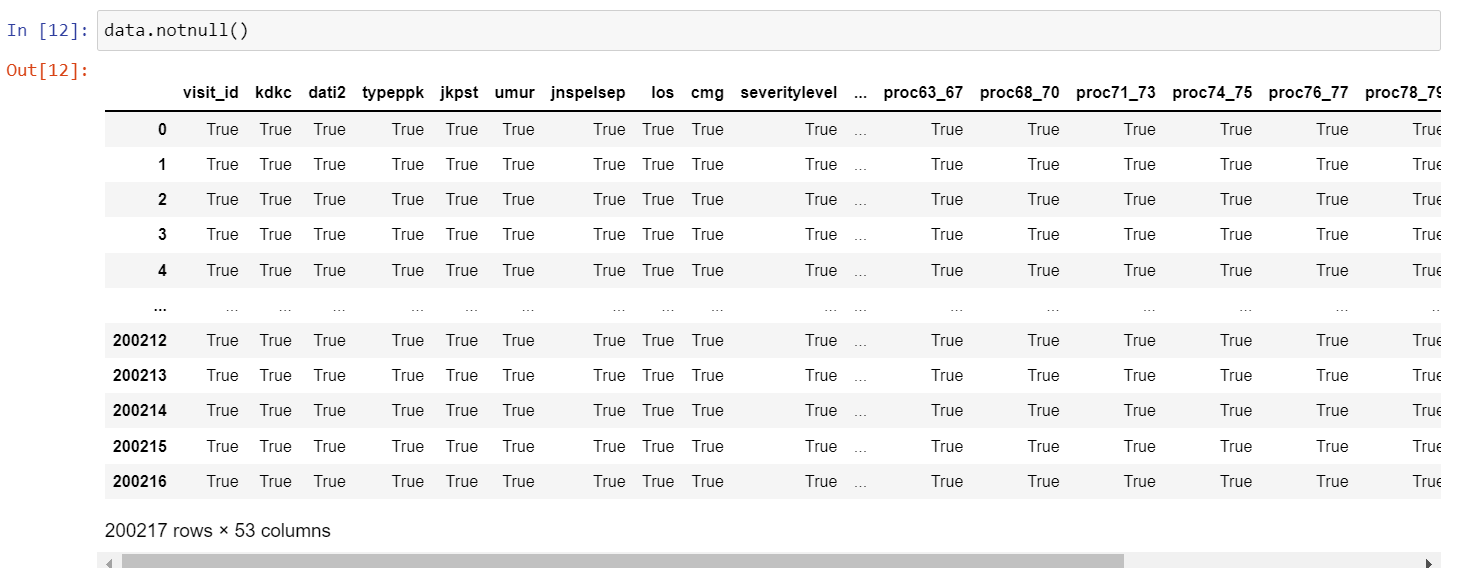
* Memeriksa data tipe dari atribut pada dataset yang digunakan.



Gambar 10

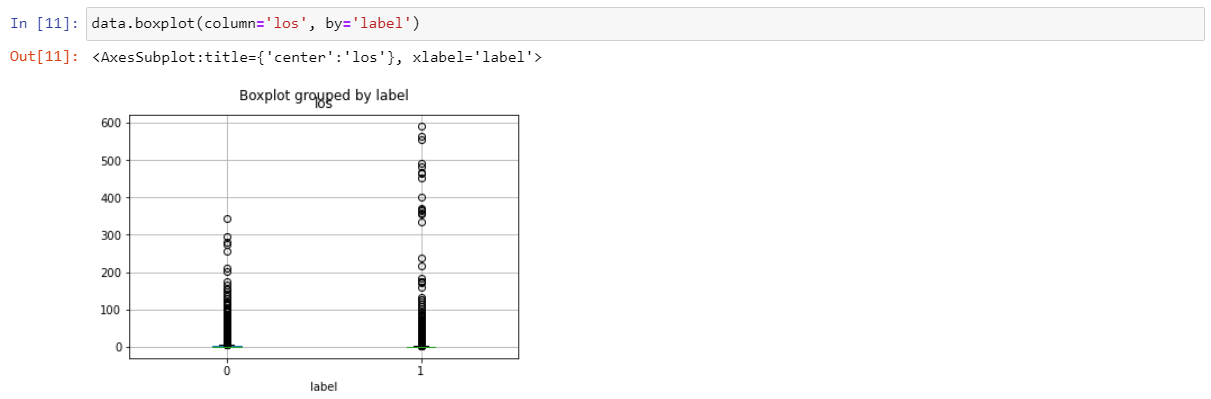
Fitur pada dataset yang digunakan terbagi menjadi 2 kategori yaitu 4 fitur non-numerik dan 49 fitur numerik.

* Memeriksa nilai null pada dataset dengan menggunakan fungsi notnull().Jika tidak ada data yang hilang, maka pada tabel akan menghasilkan nilai true. Kode program dan output yang digunakan adalah sebagai berikut:



Gambar 11

* Memeriksa *outlier* pada dataset. *Outlier* yang ditemukan berdasarkan atribut los dan label dengan kode program dan output yang digunakan sebagai berikut:



Gambar 12

* Memeriksa korelasi yang ada pada setiap atribut pada dataset dengan heatmap dan output yang dihasilkan adalah sebagai berikut:



Gambar 13

## Memvalidasi Data

Pada tahapan validasi data dilakukan evaluasi untuk memastikan kualitas dan kelengkapan *dataset* yang digunakan pada proyek. Pada tahap ini dilakukan pengecekan kembali terhadap atribut yang tidak lengkap, pembersihan data (*data cleaning*) untuk memastikan data konsisten dan relevan, serta mengurangi jumlah dan kompleksitas data. Memperkirakan apakah semua value dan ejaan nilai-nilai rasional serta apakah fitur dengan value yang berbeda memiliki pengertian yang sama. Berikut adalah hasil yang ditemukan saat dilakukan penelusuran:

1. Dataset sudah seimbang dengan 100255 label *fraud* dan 99962 label tidak *fraud*;
2. Dataset terdiri atas 49 fitur numerik dan 4 fitur non-numerik;
3. Ditemukan 3 atribut yang hanya memiliki 1 nilai unik;
4. Tidak ditemukan data bernilai *null*; serta
5. Terdapat outlier pada atribut los.

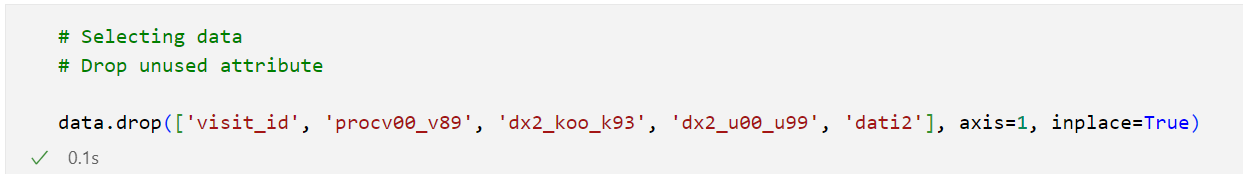
# DATA PREPARATION

Data preparation merupakan tahapan ketiga pada metode ANN. Sebelum dilakukan pemodelan, data terlebih dahulu harus diperbaiki terlebih dahulu. Ada beberapa sub proses pada data preparation antara lain:



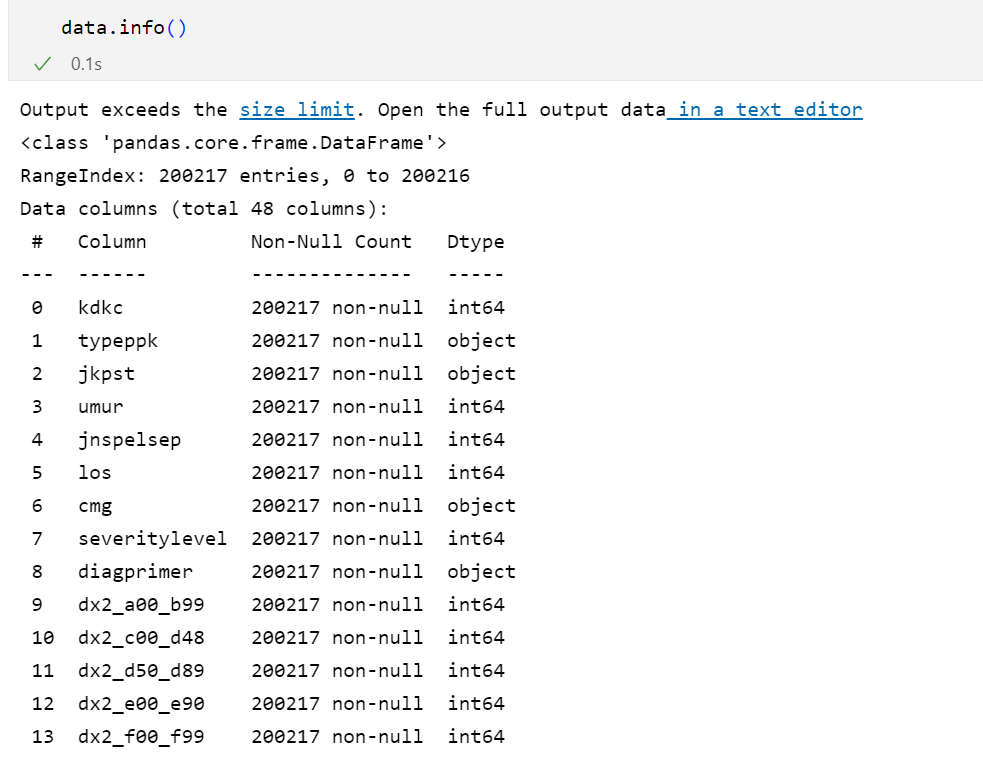
## Selecting Data

Sebelum data digunakan dalam pemodelan, kita terlebih dahulu menyiapkan data dengan baik. Ada beberapa atribut yang tidak diperlukan dalam pemodelan nanti. Atribut tersebut akan di drop agar data menjadi lebih efisien.

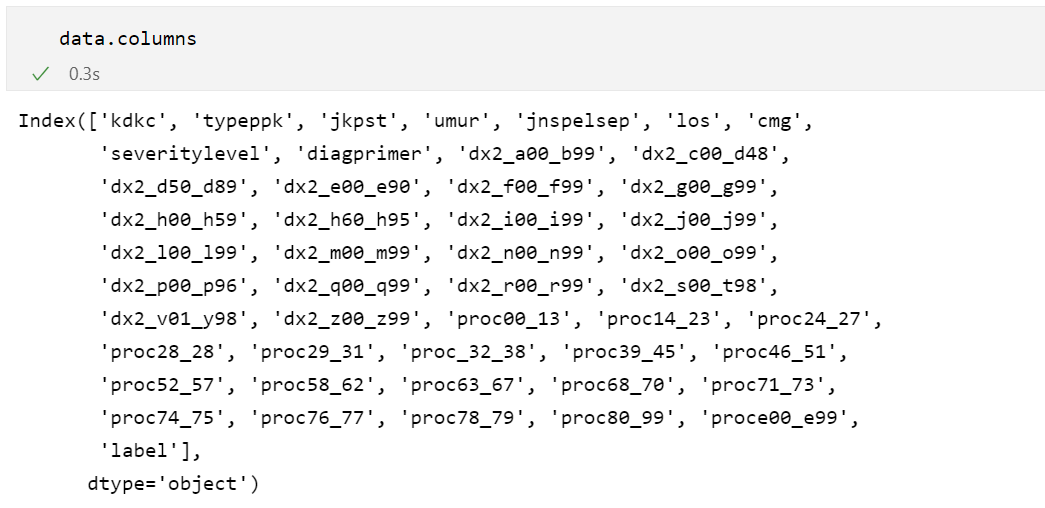


Gambar 14

Berikut adalah potongan tampilan data setelah beberapa atribut didrop.



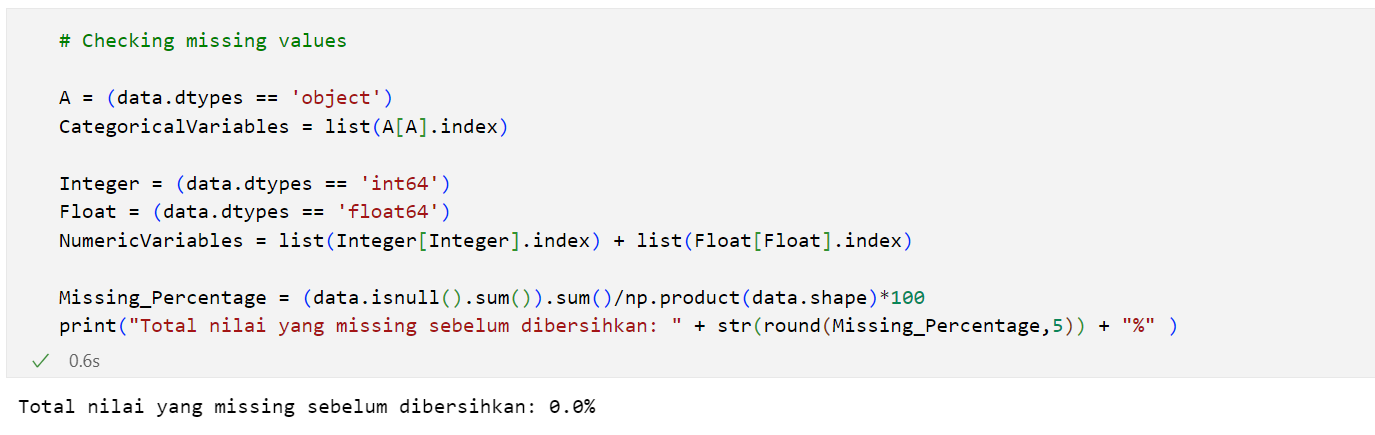
Gambar 15



Gambar 16

## Data Cleaning

Data yang dipilih kemungkinan belum bersih sehingga dibutuhkan proses pembersihan data. Proses yang dilakukan pada data *cleaning* adalah menghapus objek data yang tidak memiliki atau mengandung nilai (*missing value)* dan menghapus atau mengeliminasi atribut yang tidak relevan.

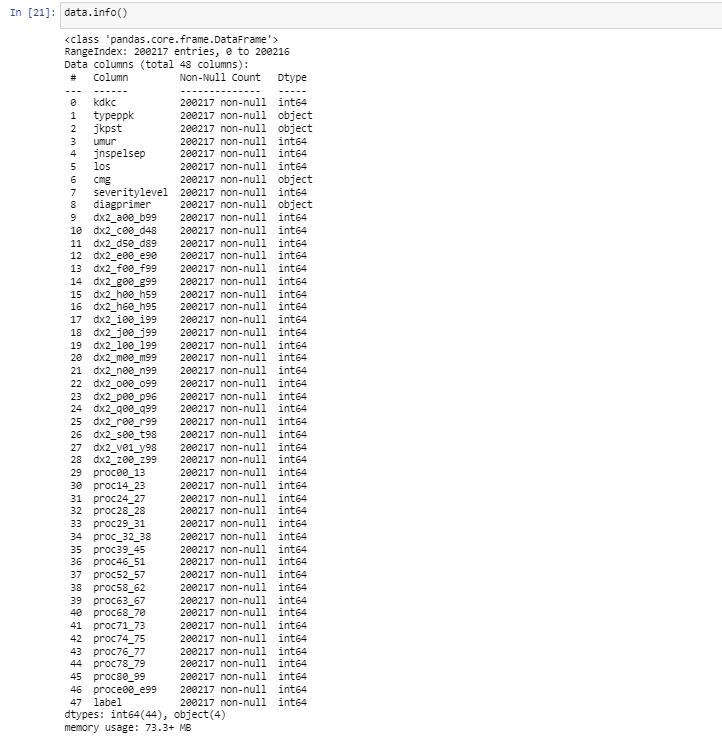


Gambar 17

Berdasarkan output pengecekan *missing value*, dapat ditarik kesimpulan bahwa tidak ada atribut yang memiliki nilai *null*. Proses selanjutnya setelah ini adalah mengeliminasi objek data yang tidak relevan.

## Construct Data

Pada tahap ini, dilakukan transformasi dengan tipe kategorikmenjadi fitur numerik. Sehingga pada tahap data *construction* dilakukan agar data kemudian dapat di normalisasi. Tahap pertama yang dilakukan adalah pengecekan tipe data pada dataset dengan menggunakan fungsi *data.info()*.



Gambar 18

Dikarenakan ada 4 atribut yang memiliki tipe data kategorik, perlu dilakukan transformasi data pada tipe data atribut dengan menjalankan potongan kode:

## Binning

Pada tahap ini adalah proses transformasi data dengan menggunakan metode *binning*. Metode *binning* adalah metode yang digunakan untuk mengelompokkan data beratribut numerik menjadi beberapa bin yang akan memudahkan dalam memahami persebaran data. Berdasarkan hasil analisis, ditemukan dua atribut yang bertipe data numerik dengan persebaran data yang cukup luas, yaitu los dan umur. Oleh karena itu proses binning akan dilakukan.

Untuk atribut umur akan dibagi menjadi 5 kategori, sesuai dengan standar usia WHO

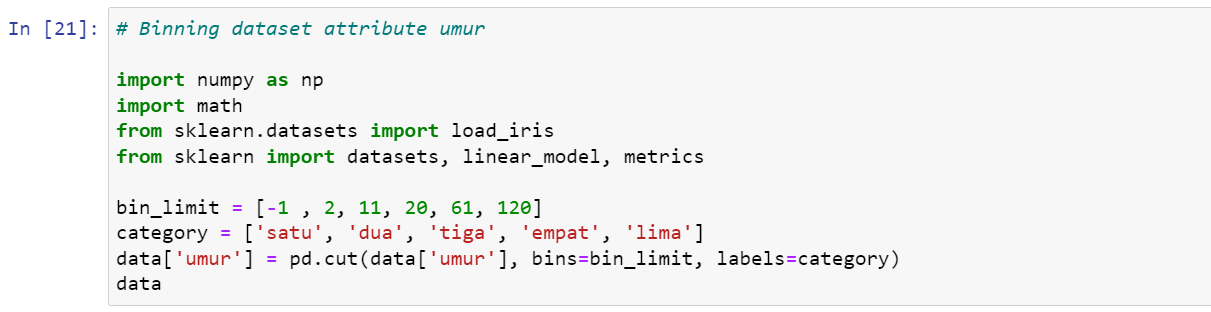
Bin 1: umur <=1,

Bin 2: 2<= umur >=10,

Bin 3: 11<= umur >=19,

Bin 4: 20<= umur >=60,

Bin 5: umur > 60



Gambar 19

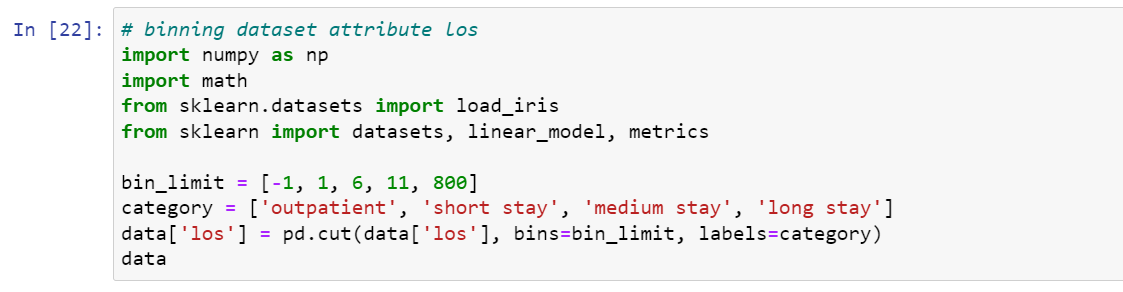
Untuk fitur los yang bekaitan dengan jnpspsleep yaitu rawat inap atau rawat jalan. Kategori ini akan dibagi menjadi 4 bagian yaitu

0 : rawat jalan,

1-5 : short stay,

6- 10 : medium stay,

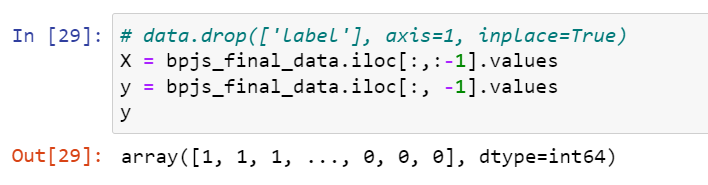
> 10 : long stay



Gambar 20

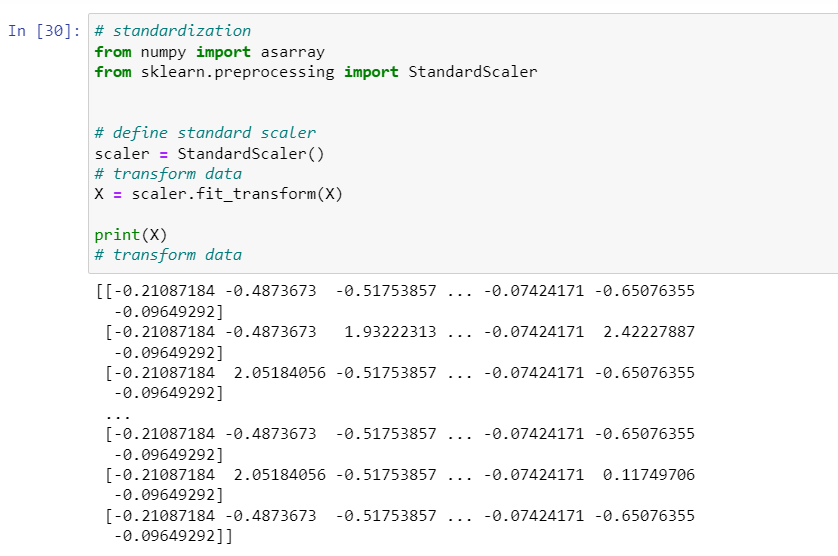
## Standardization

Pada tahapan standarisasi ini diperlukan agar dataset mendapatkan hasil yang lebih baik. Terlebih dahulu data harus disimpan ke dalam variabel X dan y seperti dibawah ini.



Gambar 21

Penerapan standarisasi akan mengubah data mentah menjadi informasi yang dapat digunakan sebelum dilakukannya pemodelan. Teknik yang digunakan dengan menskalakan data sehingga memiliki mean = 0 dan standar deviasi =1. Standarisasi dilakukan dengan menggunakan fungsi standardscaler dan diperoleh hasil standarisasi sebagai berikut.



Gambar 22

# MODELLING

Tahapan selanjutnya pada metodologi CRISP-DM untuk melakukan binary classification dalam mendeteksi fraud pada BPJS adalah modeling. Pada bab ini akan dijelaskan mengenai pemilihan teknik modelling, dan menghasilkan *test scenario* serta teknik membangun model yang akan dibangun.



## Build Test Scenario

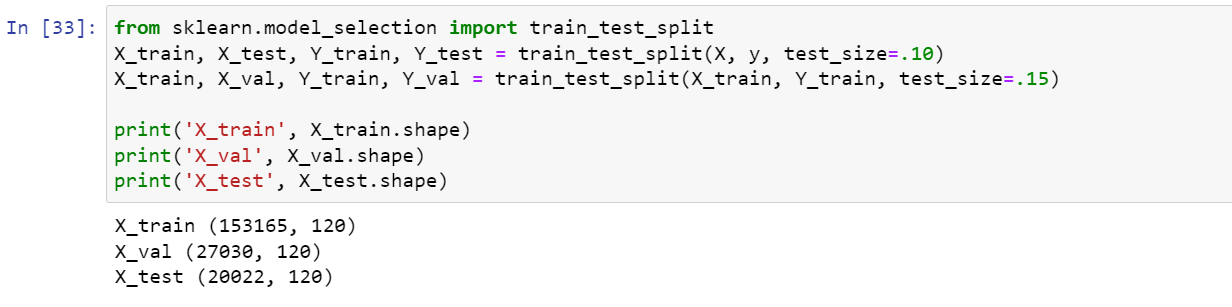
Dalam melakukan proses *data mining,* sebelum melakukan model diperlukan adanya perancangan bagaimana model yang akan dibangun. Analisis pengujian model adalah sebagai berikut.

1. Model dengan menggunakan seluruh features

Pada model ini, akan dibangun menggunakan seluruh features pada dataset. Sebelumya diketahui terdapat 53 features sebelum dilakukan *data preprocessing*. Pada model ini akan dilakukan klasifikasi menggunakan *Artificial Neural Network*. Hasil klasisifikasi yang dilakukan menghasilkan akurasi untuk data train dan data test masing-masing sebesar 0.64 dan 0.64

## Build Modeling

*Binary Classification* dengan algoritma ANN akan dirancang menggunakan bahasa pemrograman *python* dengan menggunakan pustaka *python* yaitu *scikit-learn*. *Scikit-learn* adalah salah satu pustakayang disediakan *python* untuk membangun model machine learning seperti klasifikasi ini. Pada tahap pemodelan ini, dataset yang digunakan merupakan dataset yang telah diproses sebelumnya seperti yang sudah dijelaskan pada bab 2 dan 3. Untuk pengimplementasian model ANN, tahap pertama yang dilakukan adalah membagi 3, yaitu: data latih, validasi, dan data uji dengan persentase 75% untuk data latih 15% untuk validasi, dan 10% untuk data uji. Data latih akan digunakan untuk membangun model dan data uji akan digunakan untuk menguji model yang telah dibangun.



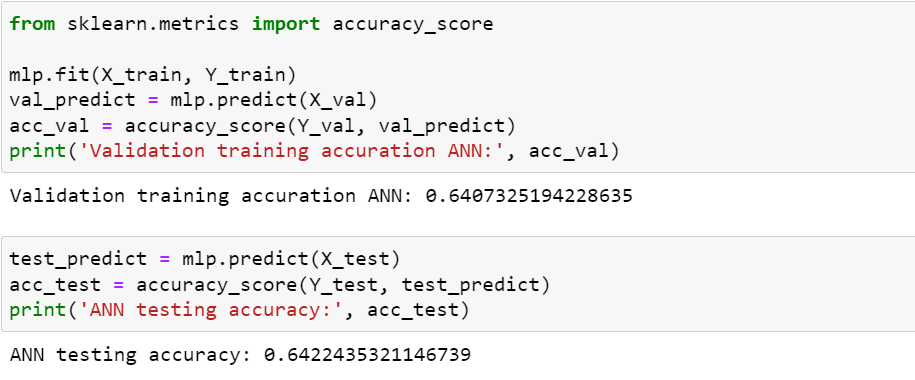
Gambar 23

Kemudian dilakukan pendefinisian informasi yang dibutuhkan dalam melakukan klasifikasi yaitu membuat model menggunakn algoritma yang sudah ditentukan.



Gambar 24

Setelah dilakukan pemanggilan algoritma, maka tahapan selanjutnya adalah menampilkan akurasi data latih dan data uji.



Gambar 25

Selanjutnya kita menyimpan model yang sudah dihasilkan dengan menggunakan library pickle.



Gambar 26

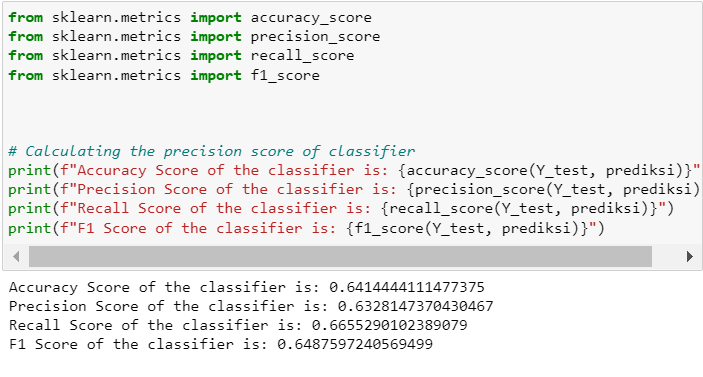
# MODEL EVALUATION

Pada bab ini akan dijelaskan mengenai evaluasi terhadap model pendeteksi potensi kecurangan pada layanan BPJS yang dihasilkan menggunakan algoritma *Artificial Neural Network.* Evaluasi adalah fase interpretasi terhadap hasil penambangan data. Evaluasi akan dilakukan secara mendalam agar hasil pada tahap *modelling* sesuai dengan sasaran yang akan dicapai.



## Evaluation of Modelling Result

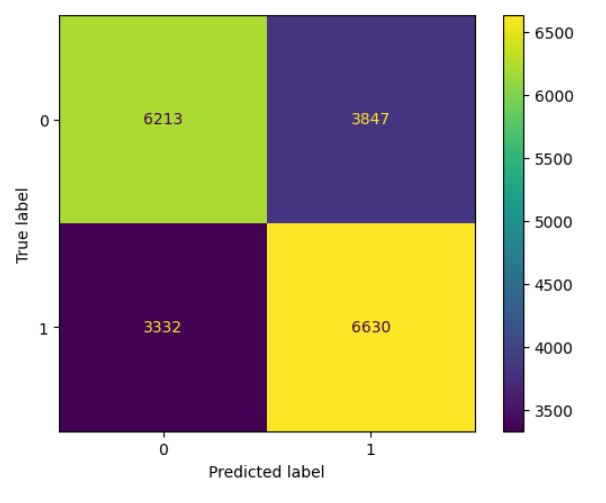
Tahap ini dilakukan untuk mengetahui performa *binary classification* untuk mendeteksi *fraud* menggunakan *confusion matrix* berdasarkandataset yang sudah digunakan yaitu data BPJS yang berasal dari kompetisi Hackathon. Pada tahap pembangunan model, telah dilakukan penilaian akurasi terhadap data latih dan data uji. Dan pada tahap ini dilakukan evaluasi pemodelan dengan melihat *precision, recall* dan *accuracy* yang dilakukan adalah sebagai berikut:



Gambar 27

Berdasarkan hasil yang diperoleh dari pembangunan model dengan menggunakan algoritma ANN telah menghasilkan model dengan akurasi cukup baik dengan score > 0.60 dimana nilai ini sudah memenuhi standar dan ketentuan pembangunan proyek. Model yang dibangun telah cukup baik dalam menerapkan algoritma ANN untuk mendeteksi kecurangan pada layanan BPJS.

Selanjutnya evaluasi dilanjutkan dengan melakukan pemetaan kesesuaian output dari model menggunakan visualisasi heatmap, dan diperoleh hasil sebagai berikut:



Gambar 28

Karena penelitian ini merupakan *binary classification*, maka output akhir dari pemodelan ini adalah binary, dimana angka 0 memiliki arti terdapat tidak *fraud* dan 1 berarti terdapat *fraud*. Merujuk pada heatmap yang diperoleh dapat dilihat hubungan *predicted lable* dengan *true lable* dalam menentukan validasi data. Data valid merupakan data yang diprediksi tidak *fraud* berjumlah 6556 dan data yang diprediksi *fraud* 6303. Sementara untuk data yang diprediksi tidak *fraud* tetapi kebenarannya adalah *fraud* berjumlah 3493 dan data yang diprediksi *fraud* tetapi kebenarannya adalah tidak *fraud* berjumlah 3670.

## Modelling Process Review

Tahap ini memeriksa Kembali semua tahapan yang dilakukan di awal dan berguna untuk memastikan bahwa tidak ada hal yang terabaikan atau terlewati. Dengan menggunakan metodologi CRISP-DM, maka dapat ditelaah bahwa:

* Tahapan EDA akan sangat membantu dalam pemilihan atribut yang berkaitan dalam mendeteksi terjadinya deteksi fraud pada layanan BPJS.
* Data preparation, pada proses pembersihan data dan transformasi data sehingga akan didapatkan data yang baik digunakan untuk modelling.

# DEPLOYMENT

Tahap terakhir pada pelaksanaan data mining menggunakan metodologi CRISP\_DM adalah deployment. Dalam bab ini akan dijelaskan mengenai deployment yang akan dihasilkan.

# LAMPIRAN