**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**LAPORAN AKHIR PROYEK AKHIR MATA KULIAH**

**12S4054 - DATA MINING**



***Fraud Detection Using SVM Algorithm***

**OLEH:**

|  |  |
| --- | --- |
| **12S17009** | **Prince Ephraim Prabowo Silaban** |
| **12S17043** | **Enjelin Ida Hutahaean** |
| **12S18004** | **Rosalia Pane** |
| **12S18017** | **Putri Yohana Panjaitan** |

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

**DAFTAR ISI**

[**BAB 1  
BUSINESS UNDERSTANDING**](#_heading=h.gjdgxs) **6**

[**1.1 Determine Business Objectives**](#_heading=h.30j0zll) **6**

[**1.2 Access the Situation**](#_heading=h.1fob9te) **6**

[**1.3 Determine Data Mining Goals**](#_heading=h.3znysh7) **7**

[**1.4 Produce Project Plan**](#_heading=h.2et92p0) **7**

[**BAB 2  
DATA UNDERSTANDING**](#_heading=h.3dy6vkm) **9**

[**2.1 Collect the Initial Data**](#_heading=h.1t3h5sf) **9**

[**2.2 Describe the Data**](#_heading=h.4d34og8) **9**

[**2.3 Explore the Data**](#_heading=h.17dp8vu) **10**

[**2.4 Verify Data Quality**](#_heading=h.3rdcrjn) **11**

[**BAB 3  
DATA PREPARATION**](#_heading=h.8qp47qfm3flg) **12**

[**3.1 Package**](#_heading=h.bhny42cijryj) **12**

[**3.2 Dataset Description**](#_heading=h.g6m7epmazcg5) **12**

[**3.3 Clean Data**](#_heading=h.o25ygpcsqp31) **14**

[**BAB 4  
MODELLING**](#_heading=h.jutv1y360bmx) **16**

**DAFTAR GAMBAR**

**DAFTAR TABEL**

[**Tabel 1 Perencanaan Proyek** 5](#_heading=h.tyjcwt)

[**Tabel 2 Deskripsi dataset fraud\_detection\_train.csv** 7](#_heading=h.2s8eyo1)

# **BAB 1 BUSINESS UNDERSTANDING**

*Business Understanding* merupakan tahap awal atau tahap pemahaman dalam penelitian. Pada tahap ini dibutuhkan pemahaman mengenai substansi dari kegiatan *data mining* yang akan dilakukan serta kebutuhan dari sebuah perspektif bisnis. Pada tahap *Business Understanding* juga diperlukan pemahaman tentang latar belakang dan tujuan pada proses bisnis yang berhubungan dengan *Fraud Detection.*

## **1.1 Determine Business Objectives**

Pada tahap *Determine Business Objectives*, dijelaskan tujuan bisnis untuk menentukan faktor-faktor penting dalam penelitian yang direncanakan dan memastikan bahwa hasil akhir dari penelitian sesuai dengan yang diharapkan. Semakin berkembangnya teknologi maka semakin banyak informasi yang tersedia. Informasi dapat diakses dengan mudah melalui penggunaan teknologi yang dikaji agar lebih efisien dan optimal melalui internet.

*Business Objectives* dari penelitian ini adalah melakukan *Fraud Detection* dengan menggunakan data dari BPJS Hackathon*.* Setiap atribut yang terdapat pada data BPJS Hackathon akan dianalisis dan dilakukan pemodelan dengan menggunakan algoritma Support Vector Machine (SVM) menggunakan bahasa pemrograman Python.

## **1.2 Access the Situation**

Proyek ini akan melibatkan pencarian fakta yang lebih rinci untuk semua sumber daya (*sources)* seperti sumber daya perangkat keras, sumber daya data (*data sources)* dan sumber daya personal.

1. *Data sources* yang digunakan pada proyek ini adalah dataset *Fraud Detection train* pada studi kasus BPJS Hackathon.
2. Sumber daya perangkat keras yang digunakan pada proyek ini adalah laptop IdeaPad Lenovo 4GB RAM, Processor Intel Core i5-7200U Dual Core 2.5 GHZ Turbo Boost 3.1 GHZ, CD/DVD ROM Drive.
3. Sumber daya personal pada proyek ini terdiri dari 4 orang mahasiswa yang berperan pada pengerjaan proyek mulai dari tahap *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment.*

## **1.3 Determine Data Mining Goals**

*Determine Data Mining Goals* adalah tahapan untuk mengubah pengetahuan pada domain bisnis menjadi sebuah *definisi problem data mining* serta untuk menetapkan tujuan *data mining.* Tujuan data mining dalam pengerjaan proyek ini adalah untuk menggali *Discovering Knowledge* mengenai pola (*pattern*) item mengenai *Fraud Detection* menggunakan dataset BPJS Hackathon.

## **1.4 Produce Project Plan**

**Tabel 1 Perencanaan Proyek**

Tahapan yang dilakukan disini adalah memaparkan rancangan kerja yang ditujukan untuk mencapai tujuan dari data mining sehingga mampu untuk mencapai tujuan bisnis, kemudian menentukan teknik dan t*ools* yang selanjutnya akan dipergunakan.

*Project plan* proyek yaitu untuk dapat menyelesaikan tujuan data mining serta mencapai tujuan bisnis adalah sebagai berikut :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tahapan** | **Waktu** | **Sumber daya yang dibutuhkan** | **Kegiatan** | **Ketergantungan** |
| *Business*  *Understanding* | 3 hari | Semua *analysts* | Menentukan tujuan utama bisnis, melakukan penilaian terhadap situasi, menentukan tujuan *data mining,* dan membuat *project plan*. | Perkembangan penerapan konsep *data mining* |
| *Data understanding* | 4 hari | Semua *analysts* | Mengumpulkan data yang akan digunakan, mendeskripsikan data, melakukan eksplorasi data, dan memverifikasi kualitas data. | Masalah data dan teknologi |
| *Data preparation* | 2 minggu | *Data mining*  *consultant,* beberapa *database analyst time* | Memilih data yang akan digunakan, membersihkan data dari *noise* atau *outlier*, membangun data, menggabungkan data, dan membuat format data. | Masalah data dan teknologi |
| *Modelling* | 1 minggu | *Data mining*  *consultant,* beberapa *database analyst time* | Memilih teknik pemodelan, membuat *Test Design,* membangun model, dan  menilai model | Ketidakmampuan  menemukan model  yang tepat |
| *Evaluation* | 3 hari | Semua *analysts* | Mengevaluasi hasil, meninjau proses, dan menentukan  tahapan selanjutnya | Ketidakmampuan untuk menerapkan hasil, terjadi kesalahan pada proses pengerjaan  proyek, perkembangan penerapan konsep *data mining* |
| *Deployment* | 1 minggu | *Data mining consultant,* beberapa *database analyst time* | Membuat *plan deployment, Monitoring and Maintenance*  *Plan* dan meninjau proyek | Ketidakmampuan  untuk menerapkan hasil,  perkembangan penerapan konsep *data mining* |

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.

*Python* menyediakan *library* seperti *scikit-learn,* Keras, TensorFlow untuk membantu dalam pembuatan model *data mining* dengan cepat. Selain itu, terdapat juga *library* yang dapat digunakan untuk membagi *dataset* menjadi data *training* dan data *test*, misalnya menggunakan *cross-validation*. Metode atau algoritma yang akan digunakan dalam proyek ini adalah algoritma Support Vector Machine (SVM) yang termasuk dalam *Supervised Learning* pada penambangan data (*Data Mining*).

# **BAB 2 DATA UNDERSTANDING**

*Data Understanding* atau pemahaman data merupakan tahap pengumpulan data awal dan meneliti data yang bertujuan untuk mengidentifikasi dan mempelajari data untuk bisa mengenal data yang akan dipakai. Tahap ini mencoba mengidentifikasikan masalah yang berkaitan dengan kualitas data, mendeteksi subset yang menarik dari data untuk membuat hipotesa awal.

## **2.1 Collect the Initial Data**

*Collect Initial Data* adalah proses pengumpulan data untuk dapat digunakan, data dapat diperoleh dengan melakukan kuesioner, wawancara, mengambil langsung sampel data dari lapangan, maupun dari internet. Penulis akan menggunakan data *Fraud Detection train* pada studi kasus BPJS Hackathon.

## **2.2 Describe the Data**

Pada tahap *describe the data,* penulis akan memahami karakteristik dari data dengan menganalisa setiap atribut yang berada di dalam data, termasuk juga melakukan analisis apakah ada objek data yang bersifat *noisy*. Hal ini dilakukan untuk memperoleh informasi terkait data yang akan digunakan. Pada tahap ini, penulis akan melakukan *exploratory data analysis* (EDA) untuk memahami karakteristik dari data.

Adapun deskripsi dari data yang akan digunakan dapat dilihat pada tabel berikut.

**Tabel 2 Deskripsi dataset fraud\_detection\_train.csv**

|  |  |  |
| --- | --- | --- |
| ***Variable name*** | **Tipe Atribut** | ***Variable description*** |
| visit\_id |  | id kunjungan |
| kdkc |  | Kode wilayah kantor cabang BPJS Kesehatan |
| dati2 |  | Kode kabupaten/kota |
| typeppk |  | Kode tipe dari rumah sakit |
| jkpst |  | Jenis kelamin peserta JKN-KIS |
| umur |  | Umur peserta saat mendapatkan pelayanan rumah sakit |
| jnspelsep |  | Tingkat pelayanan: 1.rawat inap; 2.rawat jalan; |
| los |  | Lama peserta dirawat di rumah sakit |
| cmg |  | Klasifikasi CMG (Case Mix Group) |
| severitylevel |  | Tingkat urgensi |
| diagprimer |  | Diagnosa primer |
| dx2\_a00\_b99 - dx2\_z00\_z99 |  | Diagnosa sekunder |
| proc00\_13 – procv00\_v89 |  | Kode kelompok procedure |
| label |  | Flag fraud: 1:fraud; 0:tidak fraud |

## **2.3 Explore the Data**

Pada tahap *Exploratory Data Analysis* (EDA) diperlukan sebagai sebuah pendekatan dalam menganalisis dataset untuk meringkas karakteristik utama *dataset*. Biasanya dilakukan dengan menggunakan metode visual. 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. Eksplorasi data juga memperhatikan ekstensi dari data yang akan digunakan. Oleh karena itu eksplorasi data pada penelitian ini akan dilakukan dengan melakukan analisis terhadap dimensi dari data yang digunakan, termasuk mengelompokkan data berdasarkan variabel target.

## **2.4 Verify Data Quality**

Pada tahap *verify data quality,* dilakukan verifikasi terhadap pengerjaan eksplorasi data untuk memastikan tidak ada data yang bersifat *noisy*. Hal ini dilakukan untuk menghindari kesalahan pada tahap pemodelan. Tahap mengevaluasi kualitas data dan kelengkapan data atau nilai-nilai yang hilang sering terjadi, terutama jika data yang dikumpulkan di jangka waktu yang lama. Memeriksa atribut yang hilang atau kosong. Menilai apakah semua nilai masuk akal, ejaan nilai-nilai, dan apakah atribut dengan nilai yang berbeda memiliki arti yang sama.

# **BAB 3 DATA PREPARATION**

*Data preparation* merupakan tahap setelah dilakukan pengumpulan data awal yang telah dilakukan pada fase 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.

## ***3.1 Package***

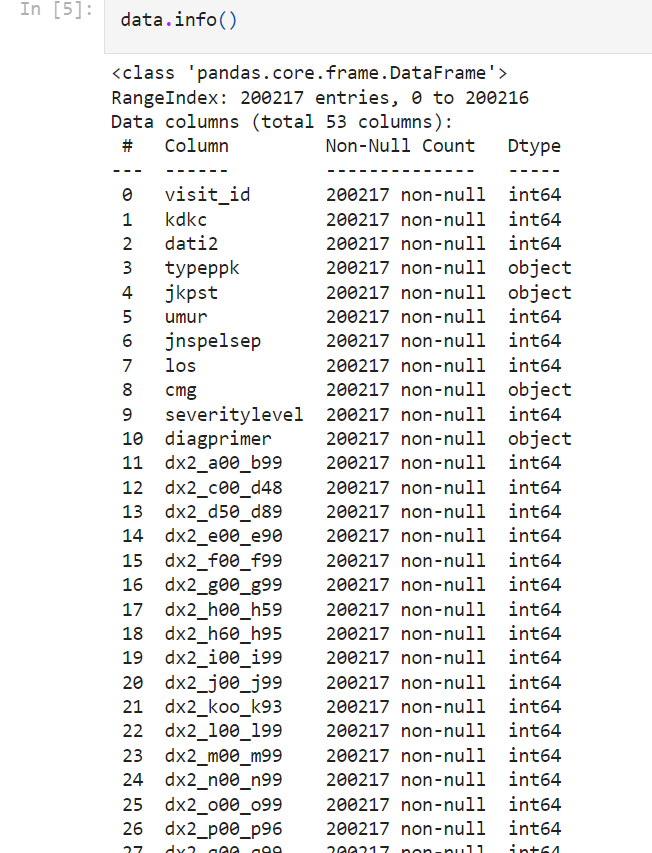
Untuk dapat menjalankan beberapa kode program yang akan dijalankan, dibutuhkan beberapa *package* yang harus diinstal, yaitu:

1. **Pandas**, untuk memuat sebuah file ke dalam tabel virtual seperti *spreadsheet*, mengumpulkan data, dan mengolahnya.
2. **Numpy**, untuk operasi vektor dan matriks serta analisis data.
3. **Matplotlib**, untuk menyajikan data ke dalam visual yang lebih menarik dan rapi.

## ***3.2 Dataset Description***

Pada fase ini, *dataset* akan dideskripsikan dengan memanfaatkan bahasa pemrograman python. Berikut beberapa fungsi yang dijalankan dalam mendeskripsikan *dataset* tersebut:

1. **.info()**, untuk menampilkan gambaran mengenai *dataset*.



**Gambar Fungsi .data.info()**

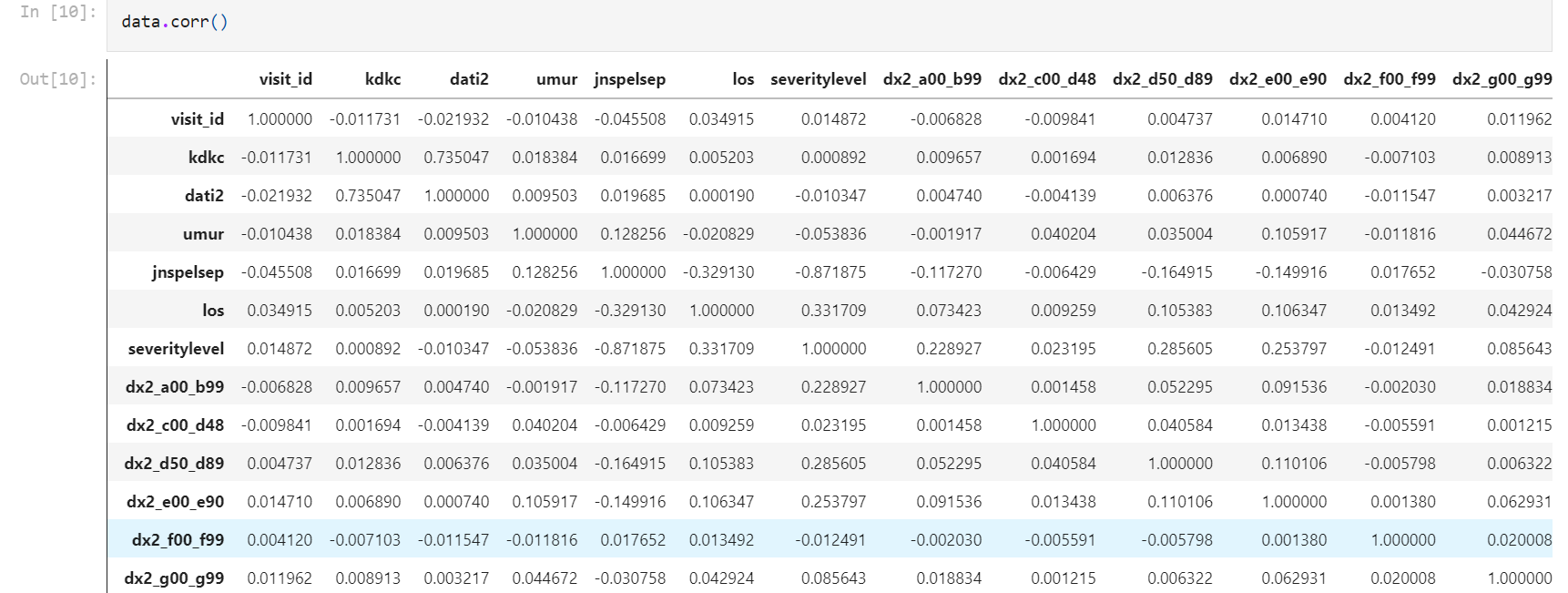
**Gambar** Error! Use the Home tab to apply 0 to the text that you want to appear here.**‑2 Fungsi .info()**

1. **.describe()**, untuk menampilkan berbagai ringkasan atau deskripsi statistik data, seperti jumlah data di setiap kolom (count), rata-rata nilai per kolom (mean), standar deviasi (std), nilai minimum (min), nilai maksimum (max), serta batas nilai dari masing-masing kuartil (25%, 50%, 75%). Berikut beberapa ringkasan atau deskripsi statistik data pada atribut yang bertipe data numerik.



Gambar 2. Fungsi .describe()

1. **.data.corr(),** untuk melakukan korelasi berpasangan



Gambar 3. Fungsi .data.corr()

1. **.head(),** untuk melihat 5 sampel data teratas.



Gambar 4. Fungsi .head()

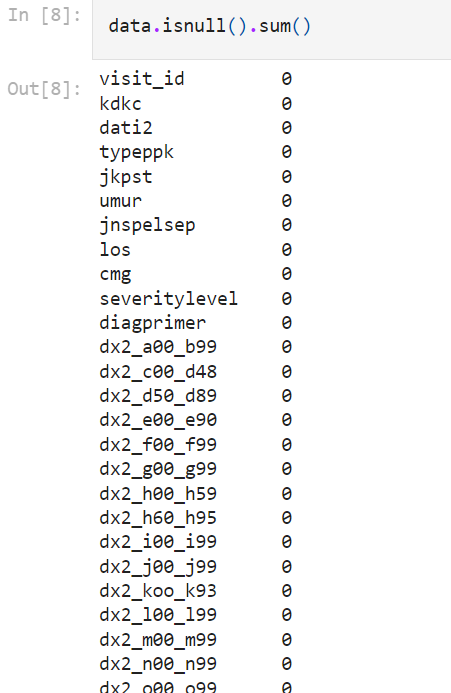
## ***3.3 Clean Data***

Pada fase ini dilakukan pembersihan data. Data cleaning yang dilakukan adalah dengan cara menghapus objek data yang tidak mengandung nilai (missing value).

1. **Fungsi data.isnull().sum()**

Fungsi Isnull () berguna untuk memeriksa suatu kolom ada datanya atau tidak. dan jika tidak ada datanya atau NULL, maka diberi data pengganti.

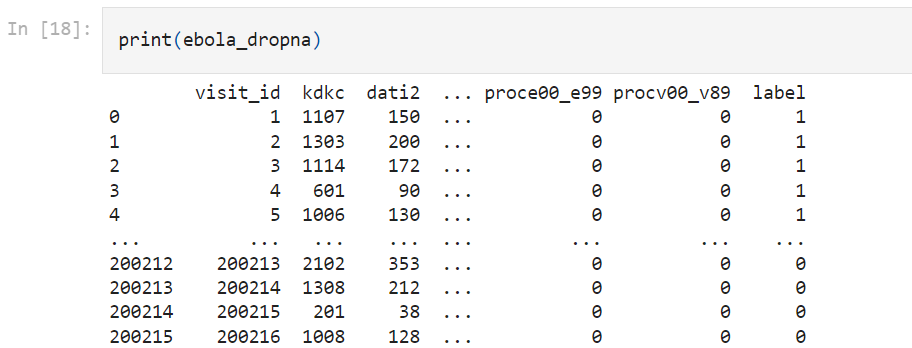
Fungsi sum() berguna untuk memudahkan dalam memahami data, maka perlu dilakukan agregasi data dengan . Dengan menggunakan fungsi sum(), maka akan diketahui berapa jumlah data yang missing value dan berasal dari atribut apa.



Gambar 5. Fungsi data.isnull().sum()

1. **Fungsi .dropna()**

fungsi .dropna() untuk menghilangkan data yang hilang. Setelah fungsi .dropna() dijalankan, maka data yang mengandung missing value terhapus.



Gambar 6. Fungsi .dropna()

# **BAB 4 MODELLING**