**LAPORAN PROYEK DATA MINING**

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



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

**PROGRAM STUDI SARJANA SISTEM INFORMASI FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO INSTITUT TEKNOLOGI DEL**

**2022**

# **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 Situation 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. Sistem Fraud Detection bertujuan untuk mencari kecurangan-kecurangan baru dan menemukan kecurangan tradisional yang pernah terjadi. Hal ini memiliki keuntungan bagi organisasi, antara lain:

1. Mengurangi tingkat tindak kecurangan
2. Mengurangi biaya yang berkaitan dengan kecurangan
3. Mencari karyawan yang berpotensi melakukan tindak kecurangan
4. Melakukan kontrol organisasi
5. Meningkatkan kinerja organisasi
6. Mendapatkan kepercayaan dan kepercayaan diri shareholders

## **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.

## **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. Data yang digunakan berjumlah 200.217 data dan terdapat 53 variabel, terdapat 4 atribut bertipe data string 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.

## **2.3 Explore Data**

Pada tahap ini, akan dilakukan eksplorasi terhadap atribut yang akan digunakan. Fungsi eksplorasi atribut ini dilakukan adalah untuk melihat kesesuaian atribut dengan tujuan proyek yang akan dilakukan. Adapun tujuan dilaksanakannya proyek ini adalah untuk melakukan prediksi terjadinya *fraud* pada klaim pelayanan Rumah Sakit berdasarkan dataset train yang diberikan. Atribut yang relevan untuk tujuan proyek yang dilakukan adalah sebagai berikut :

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

## **2.4 Verify Data Quality**

Pada tahap ini dilakukan evaluasi dan kelengkapan data yang dapat berupa melakukan pemeriksaan terhadap data yang hilang, *noise* atau atribut yang kosong serta seluruh isi dari data apakah dapat dimengerti dan sesuai dengan atributnya. Yang kemudian jika terdapat hal tersebut, maka akan dilakukan pembersihan data dengan tujuan menghasilkan data yang konsisten, akurat dan lengkap yang diperlukan dalam proyek ini.

**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.

***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.
4. **Seaborn**, untuk menyajikan data ke dalam visualisasi data secara statistik (dibangun di atas matplotlib.
5. **Sklearn,** untuk membantu melakukan processing data ataupun melakukan training data untuk kebutuhan machine learning atau data science

**3.2 *Dataset Description***

**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.