**LAPORAN PROYEK DATA MINING**

******

**Disusun Oleh :**

**12S20019 Kristina Margaret Sitorus**

**12S20042 Mastawila Febryanti Simanjuntak**

**12S20043 Sevia Rajagukguk**

**12S20052 Eka Rohani Gultom**

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

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

**2023**

**DAFTAR ISI**

**DAFTAR TABEL**

**DAFTAR GAMBAR**

# **BAB I**

# **Business Understanding**

Pada bab ini akan dijelaskan mengenai pemahaman substansi dari aktivitas data mining yang akan dilaksanakan serta kebutuhan dari perspektif bisnis. Aktivitasnya antara lain menentukan objektif bisnis, menentukan tujuan bisnis, dan membuat rencana proyek

# **1.1 Menentukan Objektif Bisnis**

Pada tahapan business understanding dilakukan pemahaman terhadap data yaitu data sampel BPJS Kesehatan tahun 2021. Menentukan objektif bisnis merupakan tahapan awal pemahaman dengan melihat objektif bisnis pada data BPJS tersebut. Objektif bisnis dilakukan untuk dapat menganalisa kebutuhan bisnis tersebut yaitu BPJS. BPJS (Badan Penyelenggara Jaminan Sosial) adalah lembaga publik yang bertanggung jawab pada jaminan terhadap perlindungan dan pelayanan kesehatan bagi para pengguna atau masyarakat. BPJS dibentuk untuk menyediakan perlindungan bagi masyarakat terutama dalam hal kesehatan, ketenagakerjaan, dan jaminan sosial. Lembaga ini berdiri dengan program jaminan sosial sesuai dengan undang - undang yang berlaku.

Pada data sampel BPJS kesehatan berisi berbagai informasi dalam memanfaatkan layanan kesehatan . Data tersebut memiliki 135 variabel yang berisikan informasi demografis, pemanfaatan pelayanan kesehatan dan biaya kesehatan. Variabel pada data tersebut memiliki potensi yang dapat dimanfaatkan pada berbagai bidang yaitu analisa kesehatan populasi, evaluasi kinerja penyedia layanan kesehatan, manajemen biaya dan efisiensi, prediksi kebutuhan layanan kesehatan, peningkatan kualitas kesehatan, evaluasi program pencegahan dan pengobatan, keamanan data dan kepatuhan regulasi, penyelidikan penyakit spesifik, dan pengembangan program kesehatan.

Objektif bisnis tersebut ditetapkan dengan baik untuk memastikan bahwa analisis terhadap data tersebut telah sesuai dengan proses bisnis untuk mencapai keberhasilan bisnis terhadap arah dan pengambilan keputusan terhadap proyek tersebut.

# **1.2 Menentukan Tujuan Bisnis**

Tujuan pengerjaan proyek ini adalah membangun sebuah model dengan penggunaan D*ata Mining* untuk

# **1.3 Membuat Rencana Proyek**

Berikut ini adalah rencana proyek yang akan dikerjakan. Rencana proyek berisi tahapan, waktu pengerjaan dan kegiatan yang dilakukan setiap tahapan untuk mencapai tujuan pada proyek *data mining* dan tujuan bisnis dalam

| Tahapan | Waktu Pengerjaan | Kegiatan |
| --- | --- | --- |
| *Business Understanding* | 4 hari | Menentukan objektif bisnis, menentukan tujuan proyek serta membuat rencana proyek |
| *Data Understanding* | 3 hari | Menelaah data yang akan digunakan dan melakukan validasi pada data tersebut. |
| *Data Preparation* | 4 hari | Memilih data yang akan digunakan, membersihkan data, konstruksi data, menentukan label data dan mengintegrasikan data yang akan digunakan. |
| *Modeling* | 3 hari | Membangun skenario pengujian dan membangun model. |
| *Evaluation* | 3 hari | Melakukan evaluasi hasil pemodelan dan melakukan review terhadap proses pemodelan. |
| *Deployment* | 5 hari | Membuat rencana deployment model, monitoring and maintenance terhadap deployment model proyek. |

Dalam proses pengerjaan tiap tahapan pada proyek ini, berikut ini adalah beberapa spesifikasi *development environment* yang digunakan adalah sebagai berikut

| Tools | * Jupyter Notebook * Google Collab * Visual Studio Code |
| --- | --- |
| Bahasa Pemrograman | Python 3 |
| Algoritma |  |
| Web Interface |  |
| Deployment Cloud |  |

# **BAB II**

# **DATA UNDERSTANDING**

Dalam tahapan data understanding yang merupakan tahapan pemahaman terhadap data yang akan digunakan, tahapan ini dimulai dari mengumpulkan data, mendeskripsikan data dan memahami data yang akan digunakan dalam penelitian.

# **2.1 Mengumpulkan Data**

Pengumpulan data merupakan tahap awal untuk menemukan data yang akan digunakan dalam penelitian. Data yang digunakan adalah data sampel BPJS Kesehatan 2015-2021. Data Sampel BPJS Kesehatan 2015-2021 adalah kumpulan data yang berisi informasi mengenai peserta BPJS Kesehatan, termasuk data demografi, data pelayanan kesehatan, dan data pembiayaan kesehatan. Data ini dirilis oleh BPJS Kesehatan setiap tahun untuk mendukung penelitian dan pengembangan di bidang kesehatan masyarakat.

Data sampel ini terdiri dari 2.305.435 peserta BPJS Kesehatan, yang diambil dari populasi database BPJS Kesehatan dengan menggunakan metode sampel acak. Data ini mencakup periode waktu 2015-2021, sehingga dapat digunakan untuk melihat tren kesehatan masyarakat dalam jangka waktu tersebut.

# **2.2 Menelaah Data**

Menelaah data pada data understanding adalah proses untuk mendapatkan pemahaman awal mengenai data dengan melakukan analisis dan visualisasi data. Data sampel kepesertaan kontekstual Diabetes Mellitus (DM) adalah seluruh peserta tersampel yang ada dalam sistem informasi BPJS Kesehatan sebagai representative peserta JKN-KIS pernah didiagnosis DM. Data kepesertaan terkait karakteristik selain jenis kelamin bersifat dinamis, dan data kepesertaan berikut menunjukkan status peserta pada tanggal 31 Desember 2021. Hasil analisis pada tabel adalah jumlah dan persentase tertimbang terhadap populasi menggunakan variabel bobot.

Berikut tabel yang membahas terkait atribut pada dataset kepesertaan pada file DM 2019 kepesertaan.dta

| **No.** | ***Variable*** | **Nama *variable*** | **Tipe *variable*** | |
| --- | --- | --- | --- | --- |
| 1. | PSTV01 | Nomor peserta | Integer | Kategorik (Nominal) |
| 2. | PSTV02 | Nomor keluarga | Integer | Kategorik (Nominal) |
| 3. | PSTV03 | Tanggal lahir peserta | String | Numerik (Interval-scaled) |
| 4. | PSTV04 | Hubungan keluarga | String | Kategorik (Nominal) |
| 5. | PSTV05 | Jenis kelamin | String | Kategorik (Nominal) |
| 6. | PSTV06 | Status perkawinan | String | Kategorik (Nominal) |
| 7. | PSTV07 | Kelas rawat | String | Kategorik (Nominal) |
| 8. | PSTV08 | Segmentasi peserta | String | Kategorik (Nominal) |
| 9. | PSTV09 | Provinsi tempat tinggal peserta | String | Kategorik (Nominal) |
| 10. | PSTV10 | Kabupaten/Kota Tempat Tinggal Peserta | String | Kategorik (Nominal) |
| 11. | PSTV11 | Kepemilikan faskes | String | Kategorik (Nominal) |
| 12. | PSTV12 | Jenis faskes | String | Kategorik (Nominal) |
| 13. | PSTV13 | "Provinsi Fasilitas Kesehatan Peserta Terdaftar | String | Kategorik (Nominal) |
| 14. | PSTV14 | Kabupaten/Kota Fasilitas Kesehatan Peserta Terdaftar | String | Kategorik (Nominal) |
| 15. | PSTV15 | Bobot | Float | Numerik (Ratio) |
| 16. | PSTV16 | Tahun Sample | Integer | Numerik (Ratio) |
| 17. | PSTV17 | Status Kepesertaan | String | Numerik (Ordinal) |
| 18 | PSTV18 | Tahun Meninggal | Integer | Numerik (Ratio) |

# **2.3 Memvalidasi Data**

Pada tahap ini dilakukan validasi terhadap data yang akan digunakan dengan memeriksa kelengkapan data untuk menghindari terjadinya *error* ataupun masalah *input data* yang terjadi *missing value*. *Missing data* adalah row atau baris pada data yang bernilai *null/Nan/NaT,* beberapa *feature* pada *row* yang kosong(*missing*)sudah ditangani dengan memberikan nilai “*missing*”. *Duplicate value* merupakan row atau baris pada data yang memiliki nilai (*value*) yang sama. Informasi mengenai *missing* *value* dan *duplicate value* untuk setiap data dapat dilihat pada tabel berikut.

| **No.** | **Data** | ***Missing Value*** | ***Duplicate Value*** |
| --- | --- | --- | --- |
| 1. | Kepesertaan | - | - |
| 2. | FKTP-Kapitasi | - | 2 rows |
| 3. | FKTP Non-Kapitasi | 42.503 rows | 38 rows |
| 4. | FKRTL | - | - |
| 5. | FKRL Diagnosis Sekunder | - | - |

# **BAB III**

# **DATA PREPARATION**

Data Preparation akan dilakukan untuk menghasilkan data yang memiliki kualitas baik. Berdasarkan penjelasan data pada bab 2, data preparation dilakukan dengan beberapa tahapan meliputi *data cleaning, data integration*, *data transformation* dan *data reduction.* Data preparation merupakan tahap setelah dilakukannya pengumpulan data awal yang telah dilakukan pada fase crisp-dm sebelumnya yaitu business understanding kemudian dilakukan proses menyiapkan data, memilih variabel yang akan dianalisis dan membersihkan data. Pemrograman dilakukan menggunakan software pengolah data **Jupyter Notebook**

# **3.1 Data Selection**

Dalam hal ini akan dipilih data data yang diperlukan. Data yang diperlukan untuk pengerjaan proyek ini adalah data Kepesertaan, FKTP dan FKTRL. Ketiga data tersebut nantinya akan diintegrasikan. Data sampel BPJS Kesehatan 2015-2021 terdiri dari tiga file, yaitu file kepesertaan, file FKTP, dan file FKTRL. File kepesertaan berisi informasi tentang peserta BPJS Kesehatan, seperti nomor kartu peserta, tanggal lahir, jenis kelamin, alamat, jenis kepesertaan, dan status kepesertaan. File FKTP berisi informasi tentang fasilitas kesehatan tingkat pertama (FKTP), seperti kode FKTP, nama FKTP, jenis FKTP, alamat FKTP, dan nomor SIP tenaga kesehatan. File FKTRL berisi informasi tentang fasilitas kesehatan tingkat lanjutan (FKTRL), seperti kode FKTRL, nama FKTRL, jenis FKTRL, alamat FKTRL, dan nomor SIP tenaga kesehatan.

Untuk melakukan data selection, BPJS Kesehatan terlebih dahulu membagi basis data peserta menjadi beberapa strata. Strata ini dibuat berdasarkan berbagai karakteristik, seperti provinsi, kabupaten/kota, jenis kepesertaan, dan usia. Pemilihan data sampel dilakukan dengan mempertimbangkan aspek kepesertaan dan interaksi dengan fasilitas kesehatan tingkat pertama dan tingkat lanjutan.

# **3.2 Data Cleaning**

Pada dasarnya data yang diperoleh dari suatu instansi maupun eksperimen memiliki kualitas yang kurang baik disebabkan oleh adanya data yang hilang (*missing data*), *noisy, inconsistency,* dan *discrepancy* antara data yang duplikat yang kemungkinan disebabkan oleh salah ketik(*human error)*  ataupun kesalahan akibat *equipment malfunction*. Berdasarkan hasil *Exploratory Data Analysis (EDA),* ditemukan adanya nilai *duplicate* pada FKTP Kapitasi dan Non Kapitasi . Nilai *duplicate*  akan dihapus karena dapat menyebabkan *overlapping* dimana data saling tumpang tindih dan menyebabkan *inconsistency* data. Pada FKTP Non Kapitasi terdapat nilai *null* pada kolom PNK04 (tanggal melakukan tindakan) sebanyak 42.503 baris atau 40,689%. Pada row yang bernilai *null*, untuk *feature* bertipe kategorik akan diisi dengan nilai “MISSING” kemudian untuk *feature* bertipe numerik akan diisi dengan nilai 0.

# **3.3 Data Construct** Data Construct mengacu pada struktur dan elemen-elemen data yang membentuk suatu dataset. Dataset BPJS Kesehatan mencakup berbagai variabel dan indikator terkait dengan pelayanan kesehatan, klaim, peserta, dan aspek lain dari sistem kesehatan.

# Dataset BPJS Kesehatan tahun 2015-2021 mencakup beragam informasi terkait dengan program jaminan kesehatan nasional di Indonesia. Data ini mencakup berbagai variabel, termasuk informasi peserta, klaim pelayanan kesehatan, pembayaran, dan statistik lainnya. Setiap baris dataset mungkin mewakili satu kejadian atau transaksi, dengan kolom-kolom yang mewakili atribut-atribut spesifik seperti nomor peserta, jenis pelayanan kesehatan, biaya klaim, tanggal layanan, dan lain sebagainya. Analisis data dari dataset ini dapat memberikan wawasan yang berharga terkait dengan pengelolaan klaim, pemantauan partisipasi peserta, tren biaya pelayanan kesehatan, dan evaluasi kinerja program jaminan kesehatan nasional selama periode 2015-2021.

# **3.4 Labeling Data**

Labelling data pada sampel BPJS Kesehatan 2015-2021 merujuk pada proses memberikan kategori atau label pada setiap instance data untuk memudahkan analisis dan pengolahan lebih lanjut. Labelling dapat mencakup atribut seperti status kepesertaan, jenis pelayanan kesehatan, atau kategori tertentu yang relevan dengan tujuan analisis.

# **3.5 Data Integration**

Pada tahap *data preprocessing* ini terdapat proses penggabungan data sampel yang tersedia seperti yang sudah dijabarkan sebelumnya. Sebelum melakukan penggabungan data, perlu dilihat keterkaitan antar *feature* sesuai dengan tujuan penelitian. Saat menggabungkan data penting untuk mengetahui jenis hubungan antara data, luaran yang diinginkan dan dilakukan sesuai dengan hirarki data sampel. Jenis hubungan yang sesuai dengan data sampel ini adalah  *one to many* atau 1:M, yaitu hubungan antara Kepesertaan dengan Pelayanan FKTP dan FKTRL. Berikut merupakan langkah-langkah yang perlu diperhatikan sebelum melakukan penggabungan data sampel BPJS Kesehatan:

1. Menentukan variabel penentu identitas peserta yang berlaku umum pada seluruh subset data sampel (dalam hal ini adalah nomor peserta)
2. Mengurutkan kedua data file berdasarkan variabel penentu identitas
3. Menentukan basis *file* (primer) yaitu file yang membentuk unit analisis

Pada jenis kardinalitas *one to many* , *file* dasar atau *master*  adalah file dengan banyak entitas. Misaknya, jika menggabungkan data dari kepesertaan dengan pelayanan FKTP, file dasar seharusnya pelayanan FKTP.