

DOKUMEN PROYEK

12S4054 – PENAMBANGAN DATA

***Regression Problem from Case and Cost Prediction using
Random Forest***



Disusun oleh:

12S18020 Dita L. Sastri Sihombing

12S18029 Estomihi Rascana Sirait

12S18061 Angela Friscilia Simamora

**PROGRAM STUDI SARJANA SISTEM INFORMASI
FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO
INSTITUT TEKNOLOGI DEL
2021**

DAFTAR ISI

DAFTAR TABEL

BAB 1 BUSINESS UNDERSTANDING

Langkah pertama dalam metodologi CRISP-DM dalam melakukan prediksi pada jumlah kasus dan unit cost pada sebuah wilayah diakibatkan adanya penambahan rumah sakit adalah business understanding. Sehingga dalam pembahasan bab ini akan dijelaskan terkait aktivitas data mining untuk meningkatkan pemahaman diantaranya dengan menentukan Tujuan bisnis, Tujuan proyek serta menghasilkan Rencana proyek.

1.1 Determine Business Objective

Dalam lingkungan kita sehari-hari tentunya sudah banyak pemanfaatan teknologi yang telah memberikan kita kemudahan dalam penyelesaian beberapa masalah. Perkembangan yang sangat pesat teknologi menguasai berbagai sektor dengan cepat. Begitu juga dimana semakin banyak penambahan rumah sakit di wilayah tertentu, semakin banyak data dan berbagai kondisi harus diselesaikan, melakukan prediksi pada jumlah kasus dan unit cost pada sebuah wilayah tertentu yang menjadi topik pembahasan dalam proyek ini dengan menggunakan data BPJS Hackathon (Case and Cost prediction). Untuk melakukan prediksi dengan data yang telah kita miliki pendekatan yang digunakan yaitu metode *Regression Problem* dengan algoritma *Random Forest* yang merupakan bagian dari teknik teknik dalam Data Mining. *Regression* merupakan proses memprediksi nilai kontinu yang terdiri dari variabel bebas(x) dan variabel tak bebas (y), *Regression* dan *Classification* adalah *prediction problem* pada *supervised learning*. Sedangkan algoritma *Random Forest* adalah algoritma yang diinterpretasikan dalam satu pohon model/*decision tree* yang menggunakan kualitas pohon keputusan dan random atau acak dimana algoritma ini adalah keputusan yang dibuat secara acak/*random forest*. Dengan beberapa kelebihan random forest juga memiliki kelemahan dimana cenderung menyebabkan data

Underfitting . Underfitting adalah keadaan dimana sebuah model memiliki kinerja buruk untuk data *training* maupun *test* data dan dapat diatasi dengan *regression random forest*[citation].

Data Mining merupakan sebuah konsep yang diperuntukkan untuk menemukan pengetahuan atau informasi berharga dari sekumpulan data [Citation]. Data Mining juga merupakan proses semi otomatis yang menerapkan ilmu matematika, teknik statistik dan *machine learning* , ada banyak teknik yang terkandung salah satunya ialah teknik *prediktif* yaitu melakukan prediksi terhadap data dengan menggunakan hasil yang telah diperoleh dari data berbeda.

Data BPJS Hackathon ini merupakan sekumpulan data yang kategorial, dimana kita perlu melakukan analisis data sebelum dijadikan sebagai pembentuk Model yang akan dibangun dan mengimplementasikan metode dengan algoritma yang telah dipilih dan dijelaskan sebelumnya, sehingga setiap personel dalam kelompok ini haruslah mengupayakan pembagian waktu dalam menganalisis, mengetahui dan mencapai *Case and Cost prediction* pada data apakah menyebabkan *fraud* atau *non fraud*. Sehingga dapat dijelaskan objektif yang akan dicapai dalam pengerjaan proyek ini ialah:

1. Men-check data apakah ada data yang *missing* atau *duplicate*. Jika ditemukan data sesuai maka dilakukan analisis *Exploratory Data Analysis* (EDA), karena data yang besar mungkin terjadinya *redundancy* perlu dilakukan berulang-ulang
2. Identifikasi faktor apa yang menyebabkan data berpotensi menghasilkan *fraud*.
3. Meningkatkan performa Data BPJS Hackathon itu sendiri dengan model yang akan menjadi output proyek ini

pekerjaan didalam membangun model ini akan dikatakan sukses apabila:

- Dihasilkan faktor penyebab data fraud

- Dikerjakan dengan tepat waktu dengan hasil sebaik mungkin

1.2 Determine Project Goal

Tujuan pengerjaan proyek ini adalah Membangun sebuah Model dengan penggunaan Teknik dalam Data Mining untuk mengetahui *case and cost prediction* dari data BPJS Hackathon dengan hasil apakah data dikatakan *fraud* atau *non fraud*.

1.3 Produce Project Plan

Tahap perencanaan yang dilakukan untuk mencapai tujuan pengerjaan proyek penelitian “*Regression Problem from Case and Cost Prediction using Random Forest*” adalah sebagai berikut:

Tabel 1. Jadwal Pelaksanaan Proyek

Tahapan	Waktu Pengerjaan	Kegiatan
<i>Business Understanding</i>	2 hari	Menentukan objektif bisnis, menentukan tujuan proyek serta membuat rencana proyek.
<i>Data Understanding</i>	3 hari	Mengumpulkan data yang akan digunakan, menelaah data dan melakukan validasi pada data.
<i>Data Preparation</i>	4 hari	Memilih data yang akan digunakan, membersihkan data, mengkonstruksi data, menentukan label data, dan mengintegrasikan data.
<i>Modeling</i>	3 hari	Membangun skenario pengujian dan membangun model.
<i>Evaluation</i>	2 hari	Melakukan evaluasi hasil pemodelan dan melakukan review terhadap proses pemodelan.
<i>Deployment</i>	4 hari	Membuat rencana deployment model, Monitoring and Maintenance rencana deployment model dan meninjau proyek.

Dalam pelaksanaan proyek penelitian, adapun *tools* yang digunakan yaitu *Python* adalah salah satu bahasa pemrograman yang dapat melakukan eksekusi sejumlah instruksi multi guna secara langsung (interpretatif) dengan metode orientasi objek (*Object Oriented Programming*) serta menggunakan semantik dinamis untuk memberikan tingkat keterbacaan syntax. *Python* juga merupakan salah satu bahasa populer yang berkaitan dengan *Data Science*, *Machine Learning*, dan *Internet of Things* (IoT). Sebagian lain mengartikan *Python* sebagai bahasa yang kemampuan, menggabungkan kapabilitas, dan sintaksis kode yang sangat jelas, dan juga dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif.

Selain itu untuk metode atau algoritma yang akan digunakan dalam proyek ini adalah *Random Forest*. *Random forest* adalah kombinasi dari masing – masing *tree* yang baik kemudian dikombinasikan ke dalam satu model. *Random Forest* bergantung pada sebuah nilai vector random dengan distribusi yang sama pada semua *tree* yang masing masing *decision tree* memiliki kedalaman yang maksimal. Penggunaan *tree* yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil voting dari *tree* yang terbentuk.

BAB 2 DATA UNDERSTANDING

BAB 3 DATA PREPARATION

BAB 4 MODELLING

BAB 5 EVALUATION

BAB 6 DEPLOYMENT

DAFTAR PUSTAKA