

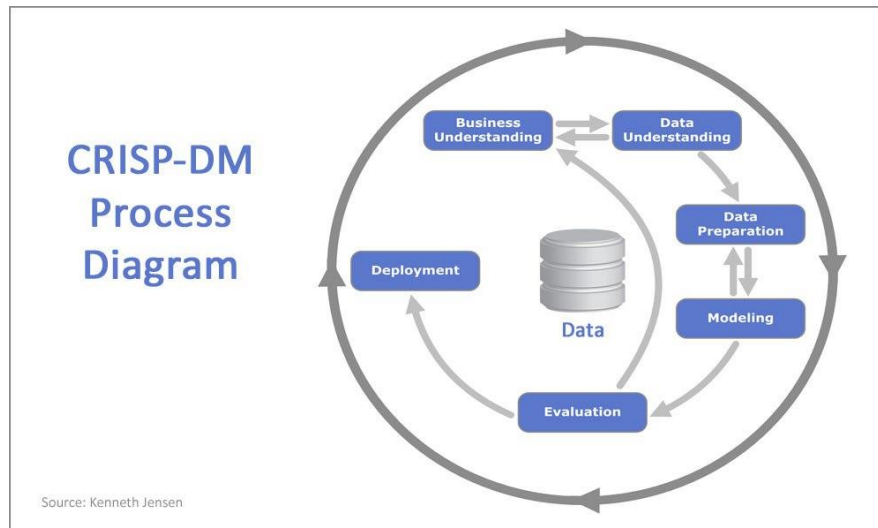
Metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM)
pada *Data Science* di Bidang Kesehatan

Data Science merupakan ilmu yang mempelajari dan menggunakan data dengan jumlah besar serta algoritma untuk mengolah data-data tersebut sehingga memberikan prediksi hasil. Dalam perkembangan teknologi serba digital saat ini *Data Science* banyak di aplikasikan untuk berbagai kepentingan, seperti di bidang kesehatan yang mampu mendiagnosa penyakit, mendeteksi anomali terhadap tubuh pasien, penerjemah bahasa isyarat, dll. Salah satu metodologi yang dapat digunakan dalam *Data Science* adalah CRISP-DM. Metodologi ini akan memberikan langkah-langkah serta alur dalam proyek *Data Science*, dengan langkah-langkah dan mengikuti prosedur yang ada maka proyek *Data Science* dapat menghasilkan hasil optimal yang dapat menjawab pertanyaan terhadap permasalahan yang ingin diselesaikan atau hasil prediksi yang dapat memberikan jawaban.

Cross Industry Standard Process for Data Mining (CRISP-DM) adalah metodologi dalam *Data Science* yang umum digunakan oleh sektor industri untuk mengolah data serta melakukan analisa untuk menentukan sebuah keputusan atau jawaban berdasarkan hasil prediksi. CRISP-DM merupakan salah satu metodologi atau model proses untuk penambangan dan menggali data informasi (*Data Mining Framework*). CRISP-DM memiliki 6 tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Dalam kasus dunia kesehatan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) untuk *Data Science* dapat diterapkan seperti sektor industri lainnya. Untuk contoh kasus *Data Science* dalam dunia kesehatan dengan metodologi CRISP-DM adalah diagnosa Diabetes Mellitus (DM). Diabetes Mellitus (DM) merupakan penyakit metabolik kronis yang ditandai oleh peningkatan kadar gula dalam darah, dimana seiring berjalannya waktu akan menyebabkan kerusakan pada organ tubuh lainnya. Dengan mengikuti langkah-langkah serta prosedur dalam CRISP-DM maka dapat dijelaskan bagaimana penerapannya dalam dunia kesehatan khususnya dalam diagnosa penyakit Diabetes Mellitus (DM).

Berdasarkan situs *World Health Organization* (WHO) sekitar 422 juta orang di seluruh dunia mengalami penyakit diabetes. Sedangkan menurut situs databox,

Indonesia menempati urutan ke-5 kasus diabetes terbanyak di dunia. Dari banyaknya kasus diabetes yang terjadi maka diperoleh data-data penderita diabetes yang dapat digunakan dan diolah. Dengan mengikuti metode CRISP-DM dapat digambarkan alur serta proses sebagai berikut:



1. *Business Understanding*

Dalam tahap pertama membutuhkan pengetahuan serta informasi yang relevan dari objek bisnis agar dalam membangun serta mendapatkan data-data yang sesuai terhadap objek bisnisnya, dalam hal ini juga bagaimana mencocokkan tujuan pemodelan untuk objek bisnis sehingga model yang efektif dapat dirancang. Dalam *Business Understanding* ini untuk contoh kasus bidang kesehatan diagnosa penyakit diabetes adalah bertumpuan pada masalah keterlambatan diagnosa penyakit menjadi penyebab utama lonjakan kematian, sehingga perlu dilakukan tindakan dini untuk memberikan prediksi atau hasil jawaban untuk penanganan terhadap pasien.

2. *Data Understanding*

Pada tahap selanjutnya adalah *Data Understanding*, yaitu untuk memeriksa data-data yang dibutuhkan untuk proses prediksi sehingga pemahaman terhadap data dapat mengidentifikasi masalah dalam data-data yang tersedia. Pada proses ini akan memberikan fondasi dalam analisis untuk membuat ringkasaan informasi data-data dan mengidentifikasi permasalahan dalam data, dengan menggunakan Analisis Data Eksplorasi/ *Exploratory Data Analysis* (EDA) sehingga memahami informasi data secara mendalam serta bagaimana gambaran awal informasi. Dalam kasus

diagnosa penyakit diabetes akan mengumpulkan semua data yang akan digunakan untuk memahami informasi data dalam penelitian yaitu dengan *Early Stage Diabetes Risk Prediction Dataset*. Lalu set data akan menampilkan beberapa sampel data dari dataset tersebut untuk mengetahui isi serta informasi dari dataset. Langkah selanjutnya adalah melakukan penjabaran data dan *check missing value*. *Check missing value* juga memeriksa kualitas dari set data diabetes yang akan digunakan meliputi data yang hilang/ *missing value*.

3. *Data Preparation*

Selanjutnya adalah *Data Preparation* adalah untuk memperbaiki masalah dalam data serta kemudian memilih permasalahan dan variabel yang akan dianalisis yang sesuai untuk penelitian. Tahap ini sering dilakukan tinjauan kembali saat menemukan masalah-masalah lain pada saat pembangunan sebuah model. Sehingga perlu dilakukan iterasi sampai menemukan model yang cocok dengan data yang akan analisis. Sehingga pada tahap ini akan melakukan *data cleaning* dan *transform data*. Dari tahapan *Data Understanding* dapat mengetahui bagaimana kualitas dari data dengan *missing value check* yang mengetahui jumlah dari data yang hilang. Untuk mengatasi permasalahan tersebut dapat melakukan proses *data cleaning*. *Data cleaning* berguna untuk mengatasi *missing value*, hal tersebut dapat diatasi dengan menggunakan Library Pandas pada bahasa pemrograman Python dengan cara menghapus baris data yang memiliki *missing value* sehingga *missing value* dapat diatasi.



4. *Modeling*

Selanjutnya *Modeling* adalah untuk membuat model prediktif atau deskriptif. Pada proses ini menerapkan metode statistika dan juga *Machine Learning* untuk menentukan teknik penambangan data, alat bantu penambangan data, dan algoritma dalam penambangan data. Pada proses ini adalah penerapan teknik dan algoritma penambangan data kepada data yang sudah tersedia. Jika diperlukan penyesuaian kembali terhadap data-data untuk teknik penambangan data tertentu, dapat kembali pada tahap *Data Preparation* untuk melakukan penyesuaian. Beberapa *Modeling* yang umum digunakan adalah *classification*, *scoring*, *ranking*, *clustering*, *finding relation*, dan *characterization*.

5. *Evaluation*

Pada tahap *Evaluation* adalah melakukan interpretasi terhadap hasil dari penambangan data/ *data mining* dari proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai sesuai permasalahan pada tahapan pertama. Evaluasi yang dilakukan dapat dilihat dari akurasi dan nilai AUC yang dihasilkan oleh masing-masing model. Setelah dilakukan evaluasi kinerja pada model yang dihasilkan, maka akan dilakukan uji *paired t-test* untuk mengetahui apakah terdapat perbedaan akurasi dari sebelum dan sesudah dilakukan optimasi.

6. *Deployment*

Tahapan terakhir adalah *Deployment* aktivitas penyebaran. Dari hasil akurasi dan nilai AUC yang telah dihasilkan dari model yang ditentukan dan sudah melakukan tahap evaluasi, lalu dibuat laporan sederhana dalam bentuk sebuah informasi pengetahuan dan dipresentasikan sehingga dapat digunakan oleh pengguna/ digunakan untuk proses selanjutnya dengan data-data baru yang memiliki kasus serupa.

Jadi *Cross Industry Standard Process for Data Mining* (CRISP-DM) adalah metodologi dalam *Data Science* yang umum digunakan oleh sektor industri untuk mengolah data serta melakukan analisa untuk menentukan sebuah keputusan atau jawaban berdasarkan hasil prediksi. Dalam kasus dunia kesehatan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) untuk *Data Science* dapat

diterapkan seperti halnya dengan sektor industri lainnya. Pada contoh kasus penerapan *Data Science* dengan metodologi CRISP-DM pada bidang kesehatan adalah melakukan diagnosa penyakit Diabetes Mellitus (DM). Dari banyaknya kasus penyakit diabetes yang terjadi maka diperoleh data-data penderita diabetes yang dapat digunakan dan diolah untuk melakukan prediksi serta diagnosa penyakit. Dengan mengikuti metodologi CRISP-DM melalui 6 tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*, maka akan memberikan alur kerja dan hasil model untuk melakukan prediksi serta jawaban atas permasalahan yang ingin dipecahkan, sehingga untuk kasus dalam diagnosa penyakit Diabetes Mellitus (DM) mampu memberikan jawaban atau hasil prediksi penyakit sebagai peringatan dini terhadap penderita diabetes. Maka dari itu, penerapan *Data Science* dengan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) tidak hanya digunakan untuk beberapa bentuk industri tetapi di bidang kesehatan juga bisa diaplikasikan untuk memberikan prediksi hasil dan jawaban atas permasalahan yang ingin dipecahkan, seperti halnya mendiagnosa suatu penyakit atau mendeteksi anomali terhadap tubuh pasien.

Referensi

Wiratama, Muhammad Abid, dan Windha Mega Pradnya. 2022. *Optimasi Algoritma Data Mining Menggunakan Backward Elimination Untuk Klasifikasi Penyakit Diabetes*. Amikom Yogyakarta University