BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Jantung merupakan organ penting dalam sistem tubuh manusia. Tugas jantung adalah memompa darah yang mengandung oksigen dan nutrisi ke seluruh tubuh. Penyakit jantung adalah salah satu bentuk penyakit kardiovaskular yang menjadi penyebab utama kematian di seluruh dunia. Penyakit jantung merupakan penyakit degeneratif yang terkait dengan gaya hidup masyarakat [48]. Beberapa faktor umum, seperti merokok, konsumsi alkohol dan kafein berlebihan, stres dan aktivitas fisik, serta faktor fisiologis lainnya seperti obesitas, hipertensi, dan kolesterol merupakan faktor penyakit jantung. Diagnosis medis penyakit jantung yang efektif dan akurat serta dini memainkan peran penting dalam penerapan langkah-langkah untuk mencegah kematian [49]. Mendeteksi dan mencegah faktor yang memiliki dampak terbesar pada penyakit jantung sangat penting dalam perawatan kesehatan. Perkembangan komputasi, pada gilirannya, memungkinkan penerapan metode pembelajaran mesin untuk mendeteksi "pola" dari data yang dapat memprediksi kondisi pasien [50].

3.2 Metode Penelitian

Metode penelitian yang dipilih merupakan metode penelitian kuantitatif dimana angka menjadi data utama atau acuan dalam pengukuran pada penelitian ini. Penelitian kuantitatif adalah penelitian, dimana penelitian sedang dalam proses, bagian menggunakan angka dari pengumpulan data, interpretasi, hingga hasil, atau penarikan kesimpulan. Disebut penelitian kuantitatif karena menghasilkan atau memerlukan data penelitian berupa angka (kuantitas) dan analisis statistik [51].

3.2.1 Metode *Framework* Penelitian

Metode penelitian kemudian didukung oleh teknik *data mining,* yaitu klasifikasi, dengan *framework data mining* yaitu KDD. Berikut merupakan perbandingan dari beberapa *framework data mining* yang sering digunakan dalam melakukan *data mining* selain KDD, yaitu CRISP-DM, dan SEMMA:

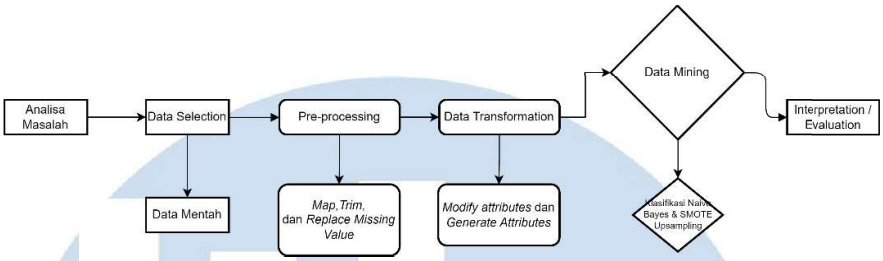
Tabel 3.1 Perbandingan KDD, CRISP-DM, dan SEMMA [32]

Indikator	KDD	CRISP-DM	SEMMA
Fase	- Data Selection - Preprocessing - Transformation - Data mining - Interpretation / Evaluation	- Business understanding - Data understanding - Data preparation - Modeling - Evaluation - Deployment	- Sample - Explore - Modify - Model - Assessment
Penggunaan	KDD membutuhkan pengetahuan sebelumnya yang relevan dan pemahaman singkat tentang domain dan tujuan aplikasi. proses KDD model bersifat iteratif dan interaktif.	Ini menyediakan seragam kerangka kerja dan pedoman untuk penambang data. CRISP-DM dirancang untuk digunakan dalam berbagai industri dan lingkungan bisnis.	membantu dalam menyediakan solusi untuk masalah bisnis dan tujuan. SEMMA terkait dengan penambang perusahaan SAS dan pada dasarnya merupakan organisasi logis dari alat fungsional untuk mereka.

Pemilihan *Knowledge Discovery in Database* didasari oleh penelitian Tabel 3.1, yang merupakan perbandingan pilihan *framework data mining* dan mengingat bahwa hasil proses KDD adalah untuk mengekstrak pengetahuan dari data dalam konteks basis data yang besar [38]. Ketiga *framework* dapat dikatakan merupakan adaptasi KDD, hanya saja CRISP-DM tidak seterus terang SEMMA dalam mengadopsi KDD [32].

3.2.2 Alur Penelitian

Alur penelitian mengacu pada *framework* penelitian terpilih yaitu *knowledge discovery in database*. Berikut ini merupakan alur penelitian tahapan dari KDD yang akan diimplementasikan dalam penelitian ini.



Gambar 3.1 Alur Penelitian

Gambar 3.1 menunjukan alur penelitian, dimana alur penelitian yang mengacu pada *framework knowledge discovery in database*. Memulai dari *data selection, data pre-processing, data transformation, data mining,* dan *interpretation / evaluation*. Penjelasan lengkap terkait alur penelitian, diantaranya:

A. *Data selection*

Pemilihan data merupakan tahapan pertama dalam *framework* ini, dimana, pada penelitian ini data diambil dari *website* Kaggle.com, data yang diambil merupakan data dengan 319,795 jumlah data, dengan 18 atribut.

B. *Data Preprocessing*

Pada tahap ini data disiapkan sebelum masuk ke transformasi data, dimana, proses yang dilakukan adalah *formatting data* mulai dari penggunaan operator *Trim* yang digunakan untuk menghapus spasi berlebihan pada nilai di sebuah kolom, dan *Replace Missing Value* yang digunakan baik untuk menggantikan atau menghapus *missing value*. Pemformatan data dilakukan untuk memeriksa setiap baris data tentang kumpulan data yang digunakan, misalnya memperhatikan format, isi baris data yang kosong hapus baris data dan aktivitas lain yang terkait dengan membangun kumpulan data yang lebih baik dan teratur.

C. *Data Transformation*

Transformation merupakan proses mengubah data yang dipilih menjadi sesuatu yang cocok untuk proses penambangan data. Kemudian setelah itu, menggunakan operator *generates attributes* yang digunakan untuk

memperkaya dan memperluas data yang ada, sehingga memungkinkan analisis yang lebih kaya dan komprehensif. Data baru yang akan di *generate* nantinya memuat kategori atau pengelompokan daripada nilai BMI.

D. ***Data Mining***

Pada tahap ini dilakukan pemilihan dan pembentukan model algoritma yang dibutuhkan dan diinginkan sesuai dengan kebutuhan penelitian, dimana, pada penelitian ini menggunakan algoritma *naïve bayes untuk* klasifikasi terhadap pasien penderita penyakit jantung. Dalam penggunaannya operator *Naïve Bayes Kernel* sendiri membutuhkan label untuk dijadikan kelas target dalam klasifikasi.

E. ***Interpretation / Evaluation***

Pada tahap ini hasil data diinterpretasikan melalui visualisasi data yang menarik yang menyajikan pengetahuan atas setiap proses penelitian yang dilakukan. Lalu kemudian evaluasi adalah bentuk dari kesimpulan atas hasil yang didapatkan, dimana evaluasi akan berdasar pada pilihan *underfitting* kondisi dimana, data hanya bekerja dengan baik pada data pelatihan tetapi berkinerja buruk pada data pengujian [52], o*verfitting* kondisi dimana, saat pengujian dengan data uji menghasilkan varian tinggi. Kemudian model tidak mengkategorikan data dengan benar, karena terlalu banyak detail dan noise [52], dan *good fit* yaitu kondisi dimana, model dikatakan memiliki nilai yang baik pada *data training* sebagus nilai yang *data testing* [52].

3.3 Teknik Pengumpulan Data

Pengumpulan data dilakukan dengan menggunakan data sekunder atau data yang sudah terlebih dahulu dikumpulkan. Pada penelitian ini data berasal dari *website* Kaggle.com [50]. Adanya ketersediaan dataset publik yang sudah memenuhi kebutuhan penelitian ini membantu peneliti, sehingga penggunaanya menghemat waktu dan upaya yang diperlukan. Data terdiri atas 319,795 jumlah data, dengan 18 atribut, dengan 9 tipe data *boolean*, 5 *string* dan 4 *decimal* [50].

Tabel 3.2. Penjelasan isi data

Nama atribut	Tipe data	Deskripsi
Heart Disease	Boolean	Responden yang menderita penyakit jantung.
BMI	Decimal	Body Mass Index
Smoking	Boolean	Keterangan terkait merokok atau tidak.
AlcoholDrink	Boolean	Keterangan terkait meminum alcohol yang berlebih atau tidak per minggunya.
Stroke	Boolean	Responden yang pernah mengalami stroke.
PhysicalHealth	Decimal	Kesehatan fisik responden selama 30 hari kebelakang.
MentalHealth	Decimal	Kesehatan mental responden selama 30 hari kebelakang.
DiffWalking	Boolean	Mengalami kesulitan yang serius dalam berjalan atau menaiki sebuah tangga.
Sex	String	Keterangan gender atau jenis kelamin.
AgeCategory	String	Rentang umur
Race	String	Entitas respoden
Diabetic	String	Responden yang pernah mengalami diabetes.
PhysicalActivity	Boolean	Melakukan aktivitas yang melibatkan fisik atau olahraga selama 30 hari kebelakang, namun bukan sebuah pekerjaan harian.
GenHealth	String	Penilaian secara general atau secara keseluruhan mengenai tingkat kesehatan dihari itu.
SleepTime	Decimal	Rata rata berapa jam waktu tidur dalam periode 24 jam.
Asthma	Boolean	Responden yang mengalami penyakit asma.
KidneyDisease	Boolean	Tidak termasuk batu ginjal, infeksi kandung kemih atau inkontinensia, apakah responden pernah diberitahu bahwa menderita penyakit ginjal.
SkinCancer	Boolean	Responden yang mengalami penyakit kulit.

Berdasarkan tabel 3.3 isi data digunakan sebagai indikator klasifikasi untuk penyakit jantung, selain dari pada itu juga isi data dapat digunakan sebagai bentuk

visualisasi yang akan membantu untuk memperjelas informasi yang ada pada data tersebut.

3.3.1 Populasi dan Sampel

Populasi pada data ini adalah setiap atribut yang ada. Sampel pada data ini adalah setiap responden yang teridentifikasi memiliki penyakit jantung pada dataset yang diperoleh.

3.3.2 Periode Pengambilan Data

Dataset ini pada *website* Kaggle.com merupakan dataset dataset berasal dari CDC dan merupakan bagian utama dari *Behavioral Risk Factor Surveillance System* (BRFSS). Data yang diambil merupakan data BRFSS tahun 2020. Dataset ini diakses pada tanggal 28 Juni 2023.

3.4 Variabel Penelitian

Variabel penelitian adalah entitas yang terkait dengan topik (khusus). Variabel penelitian dapat dikelompokkan menjadi dua jenis, yaitu variabel independen dan variabel dependen [53].

3.4.1 Variabel Independen

Variabel independen, sering disebut variabel bebas atau variabel yang memengaruhi. Variabel independen juga dapat diartikan sebagai kondisi atau sebagai nilai tampaknya memaksakan (memodifikasi) kondisi atau nilai lain [53]. Pada penelitian ini yang menjadi variabel independen adalah BMI, Smoking, AlcoholDrink, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer.

3.4.2 Variabel Dependen

Variabel dependen, biasa disebut variabel terikat merupakan variabel yang disebabkan oleh perubahan variabel lain [53]. Pada penelitian ini yang menjadi variabel dependen adalah HeartDisease dan BMI Category.

3.5 Teknik Analisis Data

Secara umum, teknik analisis data dibagi menjadi dua bagian, yaitu analisis kuantitatif dan kualitatif. Itulah satu-satunya perbedaan antara kedua teknik analisis ini dapat dilihat pada jenis datanya. Penelitian ini menggunakan teknik analisis data kuantitatif [54]. Analisis dilakukan dengan menggunakan rapidminer *software data mining* yang mudah penggunaannya, dimana analisis yang dilakukan pada penelitian ini adalah perhitungan *confusion matrix* untuk melihat kinerja klasifikasi lewat akurasi, presisi, dan *recall*. Selain itu, perhitungan dilakukan atas kurva ROC untuk melihat nilai AUC, perhitungan ini dilakukan dengan metode optimasi SMOTE untuk algoritma *naïve bayes*.