BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian

Objek penelitian ini fokus kepada prediksi penyakit anemia dan klasifikasi jenis-jenis penyakit anemia. Anemia merupakan penyakit kekurangan sel darah merah pada tubuh manusia. Penyakit anemia dapat diidentifikasi melalui *Complete Blood Count* (CBC). CBC adalah sebuah tes yang dilakukan untuk mengevaluasi sel-sel yang bersirkulasi dalam darah, termasuk sel darah merah (*Red Blood Cells*), sel darah putih (*White Blood Cells*), dan trombosit [22]. Tes ini dapat mengevaluasi kesehatan secara keseluruhan dan mendeteksi berbagai penyakit dan kondisi. Data yang digunakan adalah data *Complete Blood Count* dari *National Health And Nutrition Examination Survey* pada tahun 2017-2018 yang terdiri dari 15 (lima belas) atribut. Penyakit anemia mempunyai 5 (lima) jenis, yaitu *anemia aplastic*, *anemia kronis*, *iron deficiency anemia*, t*halassemia*, dan *anemia of reneal disease*. Masing-masing jenis ini mempunyai gejala, faktor penyebab, dan pengobatan yang berbeda-beda.

3.2. Metode Penelitian

Metode penelitian ini menggunakan teknik *data mining* dengan teknik *classification*. Pemilihan metode ini berdasarkan perbandingan tiga *framework data mining*. Di bawah ini merupakan perbandingan tahapan antara *framework* CRISP-DM, SEMMA, dan KDD *Process* :

Tabel 3.1. Perbandingan Teknik *data mining*

Data mining Framework	**CRISP-DM**	**SEMMA**	**KDD *Process***
No. of Steps	**6**	**5**	**7**
Name of Steps	*Business Understanding*	-	*Pre KDD*
	Data Understanding	*Sample*	*Selection*
		Explore	*Pre Processing*
	Data Preparation	*Modify*	*Transformation*
	Modeling	*Model*	*Data mining*
	Evaluation	*Assesment*	*Interpretation/Evaluation*
	Deployment	-	*Post KDD*

Berdasarkan tabel 3.1 perbandingan teknik *data mining*, penelitian ini menggunakan *framework* CRISP-DM dikarenakan merupakan tahapan pengembangan dari KDD [23]. Pemilihan CRISP-DM juga didukung berdasarkan penelitian sebelumnya yang menyatakan bahwa CRISP-DM ada pada tingkatan nomor satu pada *KDNuggets Poll on Data mining Methodology* pada tahun 2007 dan 2014 [23]. Pada tahun 2020, CRISP-DM masih menempati peringkat nomor satu berdasarkan *polling* yang dilakukan oleh organisasi Data Science Projet Management USA [24]. Penerapan *framewok* CRISP-DM pada penelitian ini memanfaatkan *tools data mining* yang dipilih berdasarkan perbandingan. Di bawah ini merupakan perbandingan antara *tools RapidMiner*, WEKA, dan R Studio :

Tabel 3.2. Perbandingan *Tools Data mining* [25]**,** [26]

Performance	**RapidMiner**	**WEKA**	**R Studio**
Access	*Free Community Edition, Commercial Enterprise Edition*	*Open Source*	*Open Source*
Programming Language	JAVA	JAVA	R *interpreted language*
Launch Date	2001	2002	1997
Development Team	*Rapid-I Foundation*	University Of Waikato	R *Foundation*
Kelebihan	-Mempunyai statistikal dan prediktif analisis yang mudah di implementasi pada sistem -Mempunyai kapabilitas algoritma terbanyak -Mempunyai *user interace* yang menarik dan lebih grafis (GUI / *graphic user interface*). -Mampu melakukan operasi parameter *machine learning* / metode statistic -Mampu validasi model dengan *cross-validation* dengan *independent validation set*	-*Software* WEKA berlisensi *open source*, sehingga tidak berbayar -Dapat berjalan diberbagai macam *platform*, sehingga bersifat *portable* -Memiliki GUI yang mudah dipahami oleh pengguna awam -Mampu diimplementasikan menggunakan bahasa pemrogram an Java -Tingkatan pengguna WEKA bisa untuk pemula maupun *expert* dikarenakan banyaknya fitur built-in *-Less programming / coding knowledge*.	-*Software* R Studio berlisensi *open source*, sehingga tidak berbayar -*Commands* pada R lebih fleksibel karena dapat melakukan *edited, rerun, commented, shared*, dsb. -Lebih mudah terkoneksi dengan berbagai macam format *database*. -R lebih *up-to-date* untuk metode analisis yang baru.
Kekurangan	-Diharuskan mempunyai *license RapidMiner* community untuk mengakses aplikasi. -*More*	-Tidak dapat memasukan data yang banyak untuk diproses (*overloading*) -Tidak dapat	-Tidak otomatis dalam melakukan operasi parameter *machine learning* / metode statistic -Memiliki

Performance	RapidMiner	WEKA	R Studio
	programming / coding knowledge.	menyimpan parameter untuk di *apply* pada *future* dataset. -Tidak otomatis dalam melakukan operasi parameter *machine learning* / metode statistik -Tidak bisa menyimpan proses saat validasi model dengan *cross-validation,* maka diharuskan *rebuild model.*	keterbatasan error measurement saat validasi model dengan *cross-validation.* *-More programming / coding knowledge.*

Berdasarkan tabel 3.2 perbandingan *tools data mining*, penelitian ini menggunakan *tools RapidMiner* dikarenakan terdapat beberapa unggulan dibandingkan dengan *tools* lain nya. Pemilihan *RapidMiner* didukung dengan penelitian yang pernah dilakukan sebelumnya pada tahun 2015 yang menyimpulkan bahwa *RapidMiner* memiliki kecepatan yang lebih unggul dari pada WEKA [27]. *RapidMiner* mempunyai persentase pengguna sebesar 51.2%, sementara R Studio sebesar 46.6% pada survey *software data science* yang dilakukan oleh KDnuggets [26]. *RapidMiner* juga dikenal dengan *drag-and-drop* yang memudahkan pengguna dalam mengaksesnya [21].

Tabel 3.3. Tabel Perbandingan Penggunaan Algoritma

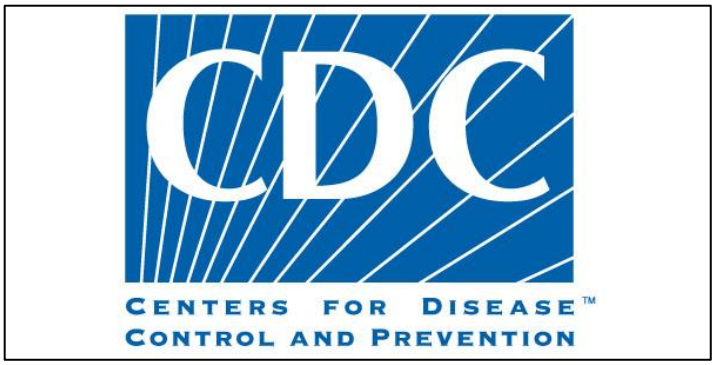
Perbedaan	**Algoritma**		
	Naïve Bayes	***J48 Decision Tree***	***Random Forest***
Output	Grafik simple distribusi atribut	1 pohon keputusan	Beberapa pohon keputusan
Kapasitas Memori	-	-	*Memory-intensive*
Persentase pengguna pada *RapidMiner Community*	12.6%	19.3%	2.5%
Kelebihan	- *Updatable* - Mampu handle data paling banyak - Dapat *handle missing value*	- Dapat *handle missing value* - Paling hemat waktu dalam pembuatan model	- Mampu mengatasi data yang memiliki atribut tidak lengkap - Dapat digunakan pada regresi dan klasifikasi

Adapun perbedaan perbandingan penggunaan algoritma seperti pada tabel 3.3, *J48 Decision Tree* dan *Random Forest* menghasilkan pohon keputusan, sedangkan *Naïve Bayes* menghasilkan sebuah grafik. *Random forest* mempunyai *alert* dengan algoritma *memory-intensive* atau memakan cukup banyak memori untuk memproses pembuatan model, maka dari itu harus dipastikan mesin yang dibuat sudah cukup baik. *Decision Tree* mempunyai peringkat tertinggi pengguna model pada *RapidMiner Community* jika dibandingkan dengan *Naïve Bayes* dan *Random Forest*. Ketiga algoritma ini juga mempunyai kelebihan dan kekurangannya masing-masing, maka dari itu penelitian ini akan membandingkan performa dari *Naïve Bayes*, *J48 Decision Tree*, dan *Random Forest*.

3.3. Teknik Pengumpulan Data

3.3.1. *Data Collection*

Penelitian ini menggunakan dataset yang berasal dari organisasi bernama Centers For Disease Control And Prevention (CDC). CDC mempunyai program rutin yang dilakukan yaitu *National Health and Nutrition Examination Survey* (NHANES). NHANES adalah program yang dirancang untuk menilai kesehatan dan status gizi orang dewasa dan anak-anak di Amerika Serikat. Program NHANES dimulai pada awal tahun 1960 dan telah dilakukan sebagai serangkaian survei yang berfokus pada berbagai kelompok populasi atau topik kesehatan [28]. Dataset yang digunakan merupakan data *complete blood count* (CBC) pada tahun 2017-2018. Berikut adalah gambar 3.1 yang merupakan logo CDC.



Gambar 3.1. Logo CDC (Centers For Disease Control anad Prevention) [28]

Berikut ini merupakan tabel 3.3 atribut data *Complete Blood Count*, isi dari dataset tersebut yang berisikan indikator-indikator atau variabel yang dipakai dalam mengidentifikasi anemia :

Tabel 3.4. Atribut data *Complete Blood Count* NHANES

No	*Attribute*	Representasi	*Attribute Value*	*Attribute Category*
1	SEQN	Respondent sequence number	numeric	numeric
2	RIDAGEYR	Age in years of the participant	0 – 150	Age
3	RIAGENDR	Gender of the participant.	1 2	Male Female
4	RIDETH1	Race/Hispanic origin	1 2 3 4 5	Mexican America Other Hispanic Non-Hispanic White Non-Hispanic Black Other Race
5	DMDCITZN	Citizen Status	1 2 7 9	Citizen by birth or naturalization Not a citizen of the US Refused Don't Know
6	LBXMCVSI	Mean cell volume (fL)	<80 80 – 100 >100	Microcytic Normocytic Macrocytic
7	LBXHCT	Hematocrit (%)	<37 37.0 – 50.0	Low Normal Severe
8	LBXHGB	Hemoglobin (g/dL)	<10 10 – 12	Severe Modarate
9	LBXMC	Mean Cell Hemoglobin Concentration (g/dL)	<32 32 - 36	Hyphochronic Normochronic
10	LBXRDW	Red cell distribution width (%)	>14.6 11.6 – 14.6	High Normal
11	LBXWBCSI	White blood cell count	1.9 – 74.2	-

No	*Attribute*	Representasi	*Attribute Value*	*Attribute Category*
12	LBXRBCSI	Red blood cell count	2.32 – 7.84	-
13	BMXWT	Weight (kg)	3.2 - 242.6	-
14	BMXHT	Standing Height (cm)	78.3 - 197.7	-
15	Anemia_prediction	Anemia Prediction	1 2	Yes No

3.3.2. Variabel Penelitian

Pengaruh suatu hasil prediksi penyakit anemia dilandasi oleh variable-variabel pada dataset. Variable ini terbagi menjadi 2 (dua) kategori, yaitu variabel dependen dan variabel independen.

A. Variabel Dependen

Variabel dependen adalah variabel yang bergantung pada faktor-faktor lain yang diukur [29]. Berdasarkan dataset NHANES mengenai *Complete Blood Count* variabel yang dapat dikategorikan sebagai variabel dependen adalah *Anemia Prediction*.

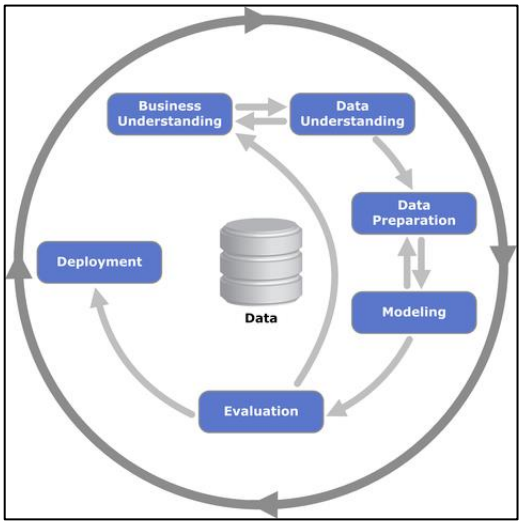
B. Variabel Independen

Variabel yang stabil dan tidak terpengaruh oleh variabel yang lain. Variabel ini mempengaruhi variabel dependen [29]. Berdasarkan dataset NHANES mengenai *Complete Blood Count* variabel yang dapat dikategorikan sebagai variabel independen adalah usia, jenis kelamin,

berat badan, tinggi badan, *Race/Hispanic origin*, *citizen status, mean cell volume*, *hematocrit*, *hemoglobin*, *Mean Cell Hemoglobin Concentration*, *Red cell distribution width*, *White blood cell count*, dan *Red blood cell count*.

3.4. Teknik Analisis Data

Dalam penelitian ini, alur penelitian yang digunakan disesuaikan dengan model *data mining* yaitu *Cross Industry Standard Process for Data mining* atau yang biasa disingkat menjadi CRISP-DM. Berdasarkan gambar 3.2 terdapat enam tahapan dalam proses CRISP-DM yang akan di implementasi dengan penjelasan sebagai berikut :



Gambar 3.2 Alur Penelitian CRISP-DM [19]

3.4.1. *Business Undertanding*

Tahapan pertama pada CRISP-DM adalah *business understanding* yang merupakan tahapan awal untuk memahami dan menentukan target atau tujuan pembentukan penelitian meliputi seluruh aspek kebutuhan bisnis [19]. Pada

penelitian kali ini tujuan utama dari pembentukan model adalah untuk prediksi klasifikasi jenis penyakit anemia dengan membandingan 3 (tiga) algoritma *supervised learning* dan mencari hasil yang optimal. Berdasarkan hasil pemodelan yang paling optimal, nantinya data prediksi penderita anemia akan dianalisa dengan data demografi. Data yang dipakai merupakan data *Complete Blood Count* dari program NHANES yang dilakukan oleh salah satu organisasi kesehatan di Amerika Serikat yaitu CDC. *Tools data mining* yang akan dipakai adalah *RapidMiner*.

3.4.2. *Data Understanding*

Tahapan kedua adalah *data understanding* yang merupakan tahapan mengumpulkan data, memahami data, serta membuat analisis terkait semua atribut pada data yang digunakan [19]. Penelitian ini akan menggunakan 3 (tiga) dataset NHANES yaitu *data laboratory*, data demografi, dan data *examination*. Pemilihan atribut disesuaikan baik dari segi penamaan dan kebutuhan penelitian. Ketiga data tersebut nantinya akan disatukan pada tahapan selanjutnya yaitu *data preparation* dengan atribut ID SEQN.

3.4.3. *Data Preparation*

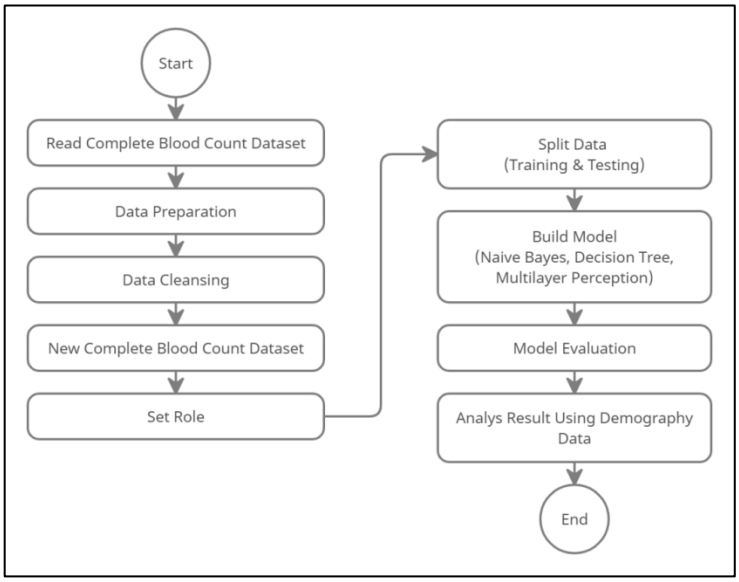
Tahapan ketiga adalah *data preparation* yang merupakan pembersihan dan pemrosesan data. Proses ini termasuk menghilangkan kebisingan jika sesuai, mengumpulkan informasi yang diperlukan untuk memodelkan atau memperhitungkan kebisingan, memutuskan strategi untuk menangani bidang data yang hilang, dan menghitung informasi urutan waktu dan perubahan yang

diketahui [19]. Penelitian ini membagi data preparation menjadi tiga bagian, yaitu *data cleansing*, *set parameter*, dan *split data*.

Pada bagian *data cleansing*, seluruh atribut yang sudah dipilih akan dihapus dalam satu *row* apabila salah satu atribut mempunyai *missing value*. Pada bagian set parameter, atribut *anemia_prediction* akan dijadikan jenis label. Paremeter yang akan digunakan mengacu pada klasifikasi dari masing-masing jenis anemia [11]. Pada bagian *split* data, data akan dibagi menjadi 2 (dua) bagian yaitu data *training* dan data *testing* dengan presentase *training* sebesar 70% dan *testing* sebesar 30%. Pembagian persentase ini merujuk pada penelitian [30] yang menyatakan bahwa pada proses *cross-validation* terdapat pilihan metode *LOOCV*, *v-fold*, serta *x-fold* dan metode yang memberikan hasil yang paling optimal adalah *v-fold*. Penarapan pembagian *data testing* sebesar 30% merupakan versi *v-fold* yang disederhanakan. Jika tahapan *data preparation* sudah dipastikan selesai, maka data dapat diolah ke tahapan *data modeling*.

3.4.4. *Data Modeling*

Tahapan keempat yaitu *data modeling* yang merupakan tahapan menentukan dan menganalisis untuk mendukung proses yang diinginkan. Tahapan ini adalah proses mencocokan model dengan dataset yang dipakai agar sesuai dengan hasil yang diharapkan [19]. Penelitian kali ini akan membandingkan algoritma *Naive Bayes*, *Decision Tree*, dan *Random Forest*. Pemilihan algoritma ini berdasarkan penelitian terdahulu mengenai prediksi penyakit anemia menggunakan *data mining*. Pembuatan model ini akan menggunakan bantuan *tools RapidMiner*.



Gambar 3.3. *Flowchart Modeling Phase*

Berdasarkan *flowchart* yang ditampilkan pada gambar 3.3, terdapat 9 (sembilan) proses yang dilakukan dalam pembuatan model klasifikasi *data mining*, berikut merupakan penjelasan dari setiap tahapan pada gambar tersebut :

a. Mulai dengan membaca data *Complete Blood Count* yang didapatkan dari NHANES

b. Memilih dan memisahkan atribut yang akan dipakai

c. Melakukan data *cleansing* dan menangani *missing value*

d. Dataset CBC yang baru siap dipakai untuk membuat model

e. Memiih atribut yang dipakai sebagai label untuk klasifikasi

f. Membagi data menjadi data *training* sebesar 70% dan data *testing* sebesar 30%

g. Membentuk model menggunakan algoritma *Naïve Bayes*, *J48 Decision Tree*, dan *Random Forest*

h. Mengevaluasi model dengan melihat *performance* dari tiga model yang dibentuk

i. Menganalisa hasil klasifikasi dengan data demografi

3.4.5. *Evaluation*

Tahapan kelima yaitu *evaluation* yang merupakan tahapan penilaian atas model yang sudah dibentuk dari tahapan sebelumnya. Konsentrasi dari tahapan ini adalah apakah hasil sudah sesuai dengan tujuan bisnis pada tahapan pertama CRISP-DM dan bagaimana kecenderungannya diantara positif atau negatif [19]. Penelitian ini menggunakan hasil akurasi, sensitivitas, dan presisi sebagai pengukuran performa dari masing-masing model yang sudah dibentuk. Jika sudah ada hasil performa model yang paling optimal, data prediksi penderita anemia akan dianalisa dengan data demografi pada tahapan selanjutnya yaitu *deployment*.

3.4.6. *Deployment*

Tahapan terakhir dari CRISP-DM yaitu deployment yang merupakan tahapan yang melibatkan organisasi dan presentasi pengetahuan berdasarkan hasil model yang sudah dibentuk. Tujuannya untuk memudahkan *end-user* dalam memahami hasil dan memperjelas masalah bisnis terselesaikan [19]. Penelitian ini memanfaatkan teknik visualisasi data menyesuaikan dengan kebutuhan organisasi [13]. Visualisasi yang dihasilkan berupa *dashboard* dengan menganalisa data predikisi penderita anemia dari model paling optimal dengan data demografi seperti jenis kelamin, rata-rata usia, berat badan, tinggi badan, kategori *race/origin*, dan kategori *citizen*.