BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian

Pada penelitian ini objek yang akan diteliti adalah mengenai penyakit jantung koroner. Penyakit ini merupakan bagian dari penyakit kardiovaskular yang memang dikenal sebagai penyakit yang menyerang organ jantung dan arteri yang menyebabkan gangguan aliran darah pada tubuh[8]. Penyakit jantung koroner dianggap merupakan jenis penyakit kardiovaskular paling berbahaya dan telah menjadi alasan kematian terbesar didunia[9]. Dimana tidak sedikit dari mereka yang terlihat sehat dan dan bugar tiba – tiba mengalami kematian mendadak yang disebabkan oleh serangan jantung yang diakibatkan oleh penyakit jantung koroner yang tidak disadari oleh pasien[46]. Sehingga dengan adanya anlisa mengenai risiko penyakit jantung, akan lebih banyak pasien jantung koroner yang mendapatkan perawatan oleh tim medis sejak dini dan dapat mengurangi angka kematian yang diakibatkan oleh penyakit jantung koroner.

3.2 Metode Penelitian

3.2.1 Metode Penerapan Data Mining

Terdapat dua metode data *analytics* yang akan menjadi pertimbangan untuk diterapkan pada penelitian ini diantaranya CRISP – DM dan SEMMA yang akan dibahas secara lebih detail pada **tabel 3.1**.

Tabel 3.1 Perbandingan metode penelitian

Faktor Pembeda	CRISP - DM	SEMMA
Langkah – Langkah penerapan [30][47]	*Business Understanding,* Data *Understanding,* Data *Preparation, Modeling, Evaluation*	*Sample, Explore, Modify, Model, and Assess*
Keakuratan [48]	CRISP-DM merupakan metode yang	Kerugian dari penggunaan SEMMA

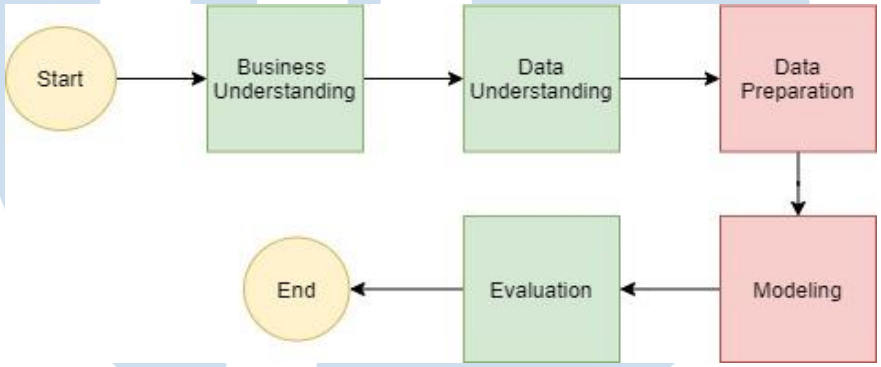
Faktor Pembeda	CRISP - DM	SEMMA
	direkomendasikan untuk melakukan proyek data *mining*, karena metode ini memiliki semua dokumentasi yang tersedia, *detailed phases, tasks* dan *activities*, serta tahap pengembangan tahap pertama, yaitu: memfasilitasi pemahaman masalah dan mengubahnya menjadi solusi atas masalah data *mining*.	meningkat dan menunjukkan bahwa metode SEMMA tidak cocok untuk diterapkan dalam studi data *mining* seperti yang dilakukan dalam ini penelitian ini.
Penggunaan Tools[48]	Metode CRISP – DM lebih popular dilakukan dengan menggunakan *tools* yang berbasis *Python-built scrip* seperti Rstudio, RapidMiner dan lain sebagainya	Sementara metode SEMMA dirancang untuk bekerja dengan *tools* khusus serperti *SAS® Enterprise Miner™*

Seperti yang dapat dilihat pada tabel perbandingan diatas, bahwa metode CRISP – DM memiliki langkah yang lebih berkaitan dengan dalam melakukan analisis data dibandingkan dengan SEMMA. Sehingga dapat dikatakan bahwa metode CRISP – DM merupakan metode yang paling sesuai untuk diterapkan pada studi terkait data analisis seperti yang akan dilakukan pada penelitian ini. Dan dapat disimpulkan bahwa penelitian ini

akan mengadaptasi metode CRISP – DM sebagai metode penelitian yang digunakan.

3.2.2 CRISP – DM

Penelitian ini dilakukan *menggunakan Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai model alur kerja dalam menganalisis risiko penyakit jantung koroner. Dimana seperti yang telah dijabarkan secara singkat pada **gambar 3.1** dibawah ini.



Gambar 3.1 CRISP – DM flowchart

Berikut merupakan langkah kerja dari CRISP – DM yang akan diimplementasikan pada penelitian ini.

3.2.2.1 Business Understading

Tahapan *Business Understanding*, merupakan tahapan pertama yang akan dilakukan pada penelitian ini. Tahapan pertama ini akan berfokus untuk memahami tujuan dan persyaratan yang dibutuhkan dari proyek / penelitian yang dilakukan. Yang kemudian dapat digunakan sebagai definisi masalah utama penelitian serta menjadi panduan untuk membuat rencana proyek awal yang berfungsi untuk mencapai tujuan penelitian[49]. *Business Perspective* dari penelitian ini sendiri adalah untuk melakukan analisa tingkat risiko penyakit jantung koroner mengingat tingginya angka kematian yang diakibatkan oleh jantung koroner. Sehingga dengan adanya analisa ini, diharapkan

penyakit jantung koroner dapat diketahui sejak dini dan dapat ditangani dengan perawatan yang tepat dari tenaga kerja medis.

3.2.2.2 Data Understading

Tahapan kedua adalah *Data Understanding* yang merupakan tahapan yang melibatkan pengumpulan *dataset* yang akan digunakan dan melakukan ekplorasi *variable – variable* yang ada pada *dataset* serta melibatkan pengidentifikasian potensi kualitas data, menemukan *insight* yang ada pada data yang dapat mengungkap informasi tersembunyi pada data yang digunakan[49]. Beberapa *dataset* yang ditemukan dalam penelitian terdahulu yang juga membahas mengenai analisa risiko penyakit jantung koroner yang akan dijelaskan pada **tabel 3.2** dibawah ini

Tabel 3.2 Penggunaan Dataset Heart Disease

Nama Jurnal	***Cleveland Heart Disease* UCI Machine Learning Respitory**	***Z-Alizadeh Sani* Data Set**	***Medica Norte* Hospital Dataset**
"A Survey on Heart Disease Early Prediction Methodologies" (2021)[50]			
"Heart Disease Prediction Using Machine Learning Algorithms" (2021)[51]			
"Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico"(2020) [52]			

Nama Jurnal	***Cleveland Heart Disease* UCI Machine Learning Respitory**	***Z-Alizadeh Sani* Data Set**	***Medica Norte* Hospital Dataset**
"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" (2019)[53]			
"Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung" (2020)[54]			
"Prediction system for heart disease using Naive Bayes and particle swarm Optimization" (2018)[55]			
"Early Prediction of Heart Disease Using Decision Tree Algorithm" (2017)[56]			
Total Penggunaan Dataset	**5**	**1**	**1**

Dapat disimpulkan bahwa dataset *Cleveland Heart Disease* merupakan salah satu *dataset* yang paling banyak digunakan dalam penelitian yang menganalisis penyakit jantung koroner. Sehingga

pada penelitian ini akan menggunakan *dataset Cleveland Heart Disease* dimana *dataset* ini memiliki 14 kolom yang akan dijelaskan secara lebih lengkap pada **tabel 3.3** :

Tabel 3.3 Definisi Variable Dataset

Variable	**Definisi**	***Range Validation***
Age	Umur dalam satuan tahun	-
Sex	Jenis kelamin (1 = *male*; 0 = *female*)	-
Cp	Tipe *chest pain* (1: *typical angina* ; 2: *atypical angina* ; 3: *non-anginal pain* ; 4: *asymptomatic*)	>=4: *asymptomatic*[57]
Trestbps	*Resting blood pressure*	>140 mmHg[58]
Chol	Kadar kolesterol dalam satuan mg/dl	>240 mg/dl[59]
Fbs	Kadar gula > 120 mg/dl (1 = *true*; 0 = *false*)	>200 mg/dl[60]
Restecg	Hasil *electrocardiographic* (0: normal; 1: abnormal; 2: adanya kemungkinan *ventricular hypertrophy*)	>=2: *ventricular hypertrophy*[61]
Thalach	Catatan *heart rate* maksimal	>149 bpm[62]
Exang	*exercise induced angina* (1 = *yes*; 0 = *no*)	1: *yes*[63]
Oldpeak	ST *depression*	-
Slope	ST *segment* (1: *upsloping*; 2: *flat*; 3: *downsloping*	-
Ca	*Major vessels* (0-3) (ditandai dengan *flourosopy*)	-
Thal	3 = normal; 6 = fixed defect;	>6: fixed

Variable	Definisi	*Range Validation*
	7 = *reversable defect*	*defect*[64]
Target	1 = *yes*, 0 = *no*	-

3.2.2.3 Data Preparation

Tahapan selanjutnya merupakan tahapan yang akan mengubah data mentah yang belum siap digunakan menjadi informasi yang lebih berguna dan dapat diproses secara lebih lanjut. Dalam melakukan tahapan data *preparation* pada penelitian ini, terdiri dari beberapa langkah lainnya yaitu seperti[49] :

- ***Data Cleansing***

  Untuk menjaga kualitas data yang digunakan pada analisa, data *cleansing* diperlukan untuk menemukan dan menghilangkan *variable* dan *value* baik yang tidak digunakan, tidak *relevant* dengan penelitian yang dilakukan ataupun eror (*missing values*). Dengan ini, hasil analisa akan lebih terpercaya dan efektif serta efisien[65]. Dan dimulai dari tahapan ini, akan dibutuhkan *tools* yang berfungsi untuk mengolah data mulai dari *exploration* sampai pada pembangunan model solusi. Pada penelitian ini, terdapat dua bahasa pemograman yang menjadi pertimbangan yang akan dibahas secara lebih lengkap pada **tabel 3.4** berikut ini.

Tabel 3.4 Perbandingan Tools

Faktor pembeda	Python	RStudio
Fleksibilitas Software	Dapat diakses menggunakan akun *google*	Terpaku pada aplikasi desktop yang harus di *install* ulang setiap berpindah *device*.

Faktor pembeda	Python	RStudio
	dan dikerjakan dari manapun	
Kompleksitas UI (Visual)	*Notebook Python* terlihat lebih *simple* dan *modern*	Pada aplikasi RStudio, terdiri dari beberapa worksheet yang terkesan kaku dan membingungkan
Kompleksitas penggunaan (package)	Hanya perlu menginstall *package* tertentu	Seluruh *package* yang ingin digunakan harus diinstall secara manual terlebih dahulu
Memori yang dibutuhkan	Versi *Web-based* dari Python tidak membutuhkan memori internal yang besar karena tidak perlu melakukan instalisasi program	Program dan *package / library* yang digunakan harus di install terlebih dahulu sehingga akan memakan memori internal yang cukup besar

Pertimbangan yang dilakukan dalam memilih *software* yang akan digunakan pada penelitian ini didasari oleh kebutuhan dan tujuan dilakukannya penelitian ini. Sehingga dapat disimpulkan bahwa bahasa pemograman yang akan

digunakan pada penelitian ini adalah bahasa pemograman *Python* menggunakan *software google Colab* yang memiliki *environment jupyter* serta mendukung hampir semua *library* yang diperlukan pada proses pengembangan AI (*Artificial Intelegence*)[66]. Dan dengan menggunakan *google colab*, akan lebih meminimalisir waktu yang digunakan karena pada saat menggunakan *google colab*, tidak perlu dilakukan instalisasi pada *device* yang digunakan sehingga penggunaan *tools* lebih sederhana dan efisien[67].

- ***Data splitting***

  Tahapan ini dilakukan untuk membagi dataset menjadi dua bagian yang terdiri dari data *training* dan data *test* dengan *range* data yang disesuaikan dengan kebutuhan dan tujuan penelitian. Selain itu, pada tahapan ini pula, data *train* akan dibagi kembali menjadi dua bagian yaitu data *train* dan juga data *validation*[68]. Data *splitting* dilakukan untuk mendapatkan hasil evaluasi yang lebih baik bandingkan dengan model yang ada pada penelitian sebelumnya[69] dengan menggunakan *dataset heart disease UCI Machine Learning Respitory* yang akan dibahas secara lebih detail pada sub - bab **3.5 Teknik Pengambilan Sample**.

3.2.2.4 Modeling

Pada tahapan ini, akan melibatkan pemilihan dan pengembangan teknik analisa dan model / algoritma yang akan digunakan untuk melaksanakan suatu penelitian[49]. Berikut ini merupakan akurasi dari beberapa algoritma yang akan dibandingkan untuk mendapatkan pengetahuan mengenai algoritma mana yang memiliki akurasi tertinggi dimana penggunaan akurasi merupakan tolak ukur seseorang dalam mempercayai suatu model / hasil. Algoritma yang akan menjadi

perbandingan untuk digunakan pada penelitian ini diantaranya adalah LSTM (*Long short term memory*), KNN (*K-Nearest Neighbor*), SVM (*Support Vector Machine*), NB (*Naïve bayes*). Perbandingan dapat dilihat pada **tabel 3.5** dibawah ini.

Tabel 3.5 Perbandingan algoritma

Nama artikel	**SINGLE LSTM**	**KNN**	**SVM**	**STACKED LSTM**	**RF**
"Heart disease prediction based on random forest and LSTM" (2020)[15]	84.5%	-	-	-	-
"Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method" (2020)[16]	81.3%	82.8%	81.3%	-	83.6%
"A stacked LSTM for atrial fbrillation	-	-	-	92%	-

Nama artikel	SINGLE LSTM	KNN	SVM	STACKED LSTM	RF
prediction based on multivariate ECGs"(2020)[43]					
"Comparison of Machine Learning Algorithms in Data classification" (2018)[70]	-	-	82%	-	83%
Maksimal Akurasi	**84.5%**	**82.8%**	**82%**	**92%**	**83.6%**

Dari tabel diatas dapat disimpulkan bahwa algoritma *stacked long short term memory* dan *single long short term memory* merupakan dua algoritma dengan akurasi tertinggi. Sehingga algoritma *stacked long short term memory* dan *single long short term memory* akan menjadi algoritma yang digunakan pada tahapan *modeling* untuk membangun solusi atas permasalahan dari penelitian ini.

3.2.2.5 Evaluation

Tahapan *evaluation* akan melibatkan proses peninjauan dan interpretasi hasil analisa yang disesuaikan dengan tujuan dan kriteria keberhasilan yang telah dijabarkan pada tahapan *business understanding*[49]. Dimana pada penelitian ini, evaluasi yang digunakan adalah dengan akurasi atau tingkat keberhasilan analisa

dari model – model analisa yang telah dibangun pada tahapan *modeling* yang didapatkan dari hasil penggunaan *confusion matrix* dan akan dibahas secara lebih lengkap pada sub – bab **3.6 Teknik Analisis Data**.

3.2.2.6 Deployment

Pada penelitian ini hanya membahas sampai pada tahapan evaluasi dan seperti yang telah disebutkan sebelumnya dalam sub-bab **1.2 batasan masalah**.

3.3 Variabel Penelitian

3.3.1 Variable Independent

Penelitian ini akan menggunakan seluruh *variable* yang terdapat pada *dataset* yaitu diantaranya adalah *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca,* dan *thal* sebagai *variable Independent*. Namun diantara ketiga belas *variable* tersebut, *chol* dan *trestbps* merupakan faktor yang paling mempengaruhi seseorang terjangkit penjakit jantung koroner[71][72].

3.3.2 Variable Dependent

Variable Dependent dari penelitian ini adalah *target.* Dimana *target* merupakan hasil diagnosa penyakit jantung yang dihasilkan dari pengaruh penggunaan *variable – variable* lainnya pada *dataset* yang digunakan.

3.4 Teknik Pengumpulan Data

Penelitian ini akan menggunakan data tersier yang merupakan *dataset* yang diekstrasi secara langsung melalui *UCI Machine Learning Respitory* dengan nama *heart disease dataset* atau yang juga biasa disebut dengan *cleveland heart disease dataset UCI*[73]. Dan sesuai dengan namanya dataset ini diperoleh dari pengumpulan data pasien pada *database* yang dimiliki *Cleveland Clinic Foundation.* **Gambar 3.2** dibawah ini menunjukan tampilan *raw dataset* dalam bentuk *excel* setelah pengunduhan dari *UCI Machine Learning Respitory*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.03	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.05	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.04	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.08	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.06	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.04	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.03	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0.00	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.05	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.06	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.02	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.02	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.06	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.08	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1.00	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.06	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0.00	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.06	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.05	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.08	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.05	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.04	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0.00	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1.00	1	0	2	1
26	40	1	3	140	199	0	1	178	1	1.04	2	0	3	1
27	71	0	1	160	302	0	1	162	0	0.04	2	2	2	1
28	59	1	2	150	212	1	1	157	0	1.06	2	0	2	1
29	51	1	2	110	175	0	1	123	0	0.06	2	0	2	1
30	65	0	2	140	417	1	0	157	0	0.08	2	1	2	1
31	53	1	2	130	197	1	0	152	0	1.02	0	0	2	1

Gambar 3.2 Raw Dataset

3.5 Teknik Pengambilan Sampel

Salah satu cara yang dapat digunakan untuk pengambilan *sample* data adalah dengan menggunakan skema validasi *train-test*. Hal ini digunakan untuk mendapatkan dan juga meningkatkan hasil evaluasi dari model yang akan dibangun pada penelitian ini. Dimana data *train* (80% dari total *dataset*) digunakan untuk melatih model yang dibangun sementara data *test* (20% dari total *dataset*) akan digunakan untuk percobaan pengimplementasian model[74].

3.6 Teknik Analisis Data

Tujuan dari dilakukannya penelitian ini adalah untuk menemukan dan membandingkan tingkat keberhasilan / keakurasian dari algoritma yaitu *Long Short Term Memory* menggunakan dua pendekatan. Yaitu dengan *Single Long Short Term Memory* dan *Stacked Long Short Term Memory* dalam memprediski risiko penyakit jantung koroner. Analisa dilakukan menggunakan Bahasa pemograman *python* dan ditulis dengan menggunakan *tools google colab*[75]. Dimana hasil analisa dapat dikatakan baik jika angka akurasi yang dihasilkan atas model yang dibangun mencapai angka 80 – 110%[76].