BAB III

METODOLOGI PENELITIAN

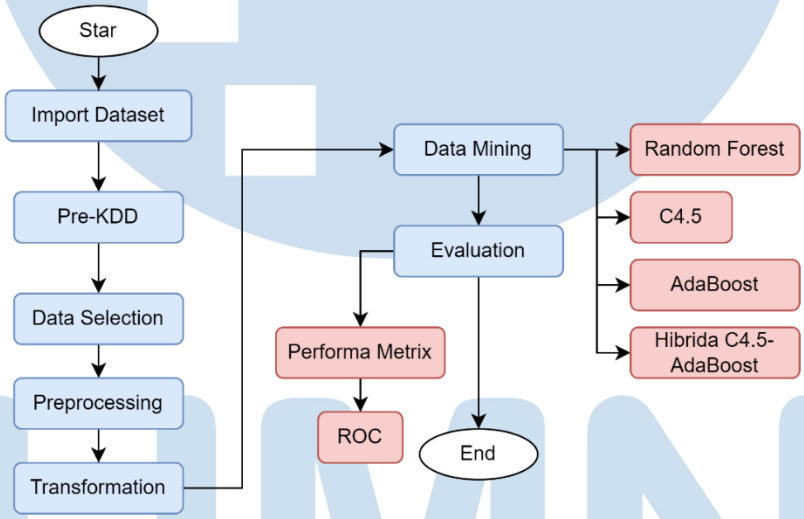
3.1 Gambaran Umum Objek Penelitian

Pada objek penelitian ini berfokus pada memprediksi penyakit hepatitis. Hepatitis adalah kondisi peradangan pada hati yang dapat disebabkan oleh berbagai faktor, termasuk infeksi virus, konsumsi alkohol berlebihan, dan penyakit autoimun. Ada beberapa jenis hepatitis yang umum, termasuk hepatitis A, B, C, D, dan E[1]. Hepatitis A disebabkan oleh virus hepatitis A (HAV) dan umumnya ditularkan melalui makanan atau minuman yang terkontaminasi. Sementara itu, hepatitis B disebabkan oleh virus hepatitis B (HBV) dan dapat ditularkan melalui kontak dengan darah, cairan tubuh, atau bahan terkontaminasi. Hepatitis C, disebabkan oleh virus hepatitis C (HCV), sering ditularkan melalui kontak dengan darah yang terinfeksi, seperti penggunaan jarum bersama. Hepatitis D hanya dapat menyebabkan infeksi pada individu yang sudah terinfeksi hepatitis B, sedangkan hepatitis E disebabkan oleh virus hepatitis E (HEV) dan biasanya ditularkan melalui air yang terkontaminasi. Gejala hepatitis meliputi kelelahan, demam, mual, muntah, nyeri perut, dan kuning pada kulit dan mata (ikterus), tetapi gejalanya bisa bervariasi tergantung pada jenis virusnya. Penanganan hepatitis tergantung pada jenis virus dan tingkat keparahannya, dengan pencegahan yang melibatkan vaksinasi, praktik kebersihan yang baik, dan menghindari perilaku berisiko[1]. Dengan pemahaman yang dampak yang ditimbulkan dari hepatitis perlu adanya kesadaran baik dari masyarakat ataupun instansi kesehatan dalam menyadari pentingnya bahaya hepatitis ini dengan berkunjung ke rumah sakit untuk dilakukannya madical chekup dan akan mencatat setiap data yang didapat sebagai pemahaman baru dalam menyadari penyakit hepatitis.

Data yang digunakan pada penelitian ini berasal dari Rumah Sakit Umumt Daerah Kota Tangerang yang diambil langsung oleh pihak rumah sakit pada bulan

Maret tahun 2024 dan direkap secara mandiri oleh peneliti selama dua minggu, dengan memenuhi persyaratan dan perizinan yang telah diselsaikan. Dataset yang akan digunakan dengan rentang waktu bulan Januari tahun 2019 hingga bulan Desember tahun 2023 dengan tujuan untuk mendukung penelitian dengan data yang nyata, sehingga memberikan hasil penelitian yang optimal. Dengan adanya analisa mengenai prediksi penyakit hepatitis diharapkan menjadi wawasan baru bagi para peneliti untuk melakukan penelitian dengan mengetahui tingkat prediksi penyakit hepatitis, sehingga dapat mencegahlebih awal penularan hepatitis dan tidak menjadi penyakit yang berbahaya sampai mengancam nyawa.

3.2 Alur Penelitian



Gambar 3. 1 Alur Penelitian

Pada gambar 3.1 merupakan alur penelitian yang akan dijalankan untuk mencapai tujuan dan menyelesaikan masalah pada rumusan masalah. Penelitian ini dimulai dengan import dataset yang telah diambil secara langsung di RSUD Kota Tangerang pada Bulan Maret 2023. Selanjutnya, yaitu Pre-KDD dan langkah-langkah pada *framework Knowledge Discovery in Databases* (KDD).

1.2.1 Pre-KDD

Pada tahapan ini merupakan tahapan pengenalan dataset sebelum masuk dalam tahapan pada *Knowledge Discovery in Database*. Tahapan ini

menjelaskan tentang dataset yang akan digunakan, seperti penjelasan mengenai librari yang akan digunakan, penjelasan singkat mengenai atribut pada dataset, dan menampilkan grafik untuk memberikan pemahaman singkat pada dataset. Setelah pengenalan mengenai dataset, barulah melakukan penelitian sesuai dengan alur penelitian menggunakan model data mining yaitu *Knowledge Discovery in Database* atau yang biasa disingkat KDD. Pada Gambar 2.1 terdapat lima tahapan dalam proses KDD yang akan di gunakan dalam penelitian dengan penjelasan sebagai berikut:

1.2.2 Data Selection

Tahapan pertama KDD adalah *data selection.* Ini adalah pemahaman tahapan pertama pada *Knowledge Discovery in Database* yang bertujuan untuk memilih variabel-variabel yang akan digunakan pada penelitian. Sumber data yang akan digunakan berasal dari Rumah Sakit Daerah Umum Kota Tangerang, yang didapat pada bulan Maret 2024. Pada tabel 3.1 merupakan variabel dari dataset yang didapat berdasarkan hasil cek lab darah dan cek lab hati yang akan digunakan peneliti untuk pemodelan.

Tabel 3. 1 Variabel Dataset

Variable	Keterangan
Hemaglobin	Protein dalam sel darah merah yang mengangkut oksigen dari paru-paru ke jaringan tubuh.
Hematokrit	Persentase volume sel darah merah dalam darah total
Leukosit	Sel darah putih yang berfungsi dalam sistem kekebalan tubuh untuk melawan infeksi dan penyakit.
Tromobosit	Sel kecil dalam darah yang berperan dalam pembekuan darah untuk menghentikan perdarahan.
Eritrosit	Sel darah merah yang mengangkut oksigen dari paru-paru ke seluruh tubuh.
Eosinofil	Jenis sel darah putih yang berperan dalam respons alergi dan melawan infeksi parasit.
Neutrofil Segmen	Salah satu jenis sel darah putih yang menjadi pertahanan utama tubuh terhadap infeksi bakteri.
Limfosit	Sel darah putih yang berperan dalam respons

	kekebalan tubuh terhadap infeksi dan penyakit
Monosit	Sel darah putih yang berfungsi sebagai fagosit, menelan dan mencerna materi asing dalam tubuh.
SGOT (AST)	Enzim yang terdapat di dalam sel hati dan otot; kenaikan levelnya bisa menandakan kerusakan hati atau jaringan otot.
SGPT (ALT)	Enzim yang terutama terdapat di dalam sel hati; peningkatan levelnya dapat mengindikasikan kerusakan hati.
Bilirubin Direk	Salah satu bentuk bilirubin yang dihasilkan dari pemecahan sel darah merah, dan harus dikeluarkan dari tubuh melalui hati.
Bilirubin Indirek	Bentuk bilirubin yang belum diubah oleh hati untuk diekskresikan dari tubuh.
Bilirubin Total	Jumlah keseluruhan bilirubin dalam darah, yang mencakup baik bentuk langsung maupun tidak langsung.

1.2.3 Pre-Processing / Cleaning

Tahapan selanjutnya adalah *pre-processing / cleaning* yang merupakan proses menggabungkan, membersihkan, memeriksa data, dan menganalisa semua antribut pada data yang ingin digunakan. Pada tahapan *pre-processing / cleaning* akan dilakukannya pembersihan data yang hilang. Bagian data *cleansing* akan dilakukannya penghapus atau pembersihan pada data yang tidak terisi atau missing value pada setiap variabel yang tidak relevant ataupun eror dengan mengganti dengan semua nilai dengan bantuan library KNNImputer, sehingga data yang hilang dapat teisi berdasarkan data pada kedekatan pada data lengkap[24]. Ketika data yang hilang sudah terisi, selanjutnya akan dilakukan. Pengabungan data lengkap dengan data yang baru terisi pada tahapan *transformation.*

1.2.4 Transformation

Tahapan ketiga adalah *data preparation*. *Data prepatration* merupakan sebuah proses untuk mentranformasi atau mengubah data yang belum siap ke dalam bentuk data yang siap untuk proses tahapan selanjutnya yaitu *data mining*. Data yang sudah tidak memiliki missing value, akan dilakukan

pembuatan target class label baru, dimana target class ini akan berfokus pada dua atribut yaitu SGOT (AST), dan SGPT (ALT) dengan kriteria pada standar enzim rumah sakit yaitu di bawah 40 dan 41 U/L berdasarkan dataset yang peneliti dapat[32].

1.2.5 Data Mining

Tahapan keempat yaitu *data mining* yang merupakan langkah untuk menentukan dan menganalisis untuk mendukung proses yang diinginkan. Penelitian kali ini akan membandingkan algoritma *Random Forest*, *C4.5*, *AdaBoost* dan hibrida *C4.5-AdaBoost* sebelum dan sesudah penerapan optimalisasi *hyperparameter*. Pemilihan algoritma ini berdasarkan penelitian terdahulu mengenai klasifikasi hepatitis. Pembuatan model ini akan menggunakan bantuan *python*.

1.2.6 Interpratation / Evaluation

Tahapan terakhir atau kelima yaitu *interpretaion / evaluation* merupakan tahapan peninjauan sebuah hasil dari proses yang telah dilakukan. Tahapan ini berfokus pada apakah hasil sudah sesuai dengan tujuan pada penelitian. Akhir dari penelitian ini menggunakan hasil dari tingkat akurasi, presisi, f1-score dan presisi sebagai pengukuran sebuah perfoma dari dua algoritma yang dibuat. Serta, didukung dengan tingkat ROC pada pemodelan algoritma.

3.3 Metode Penelitian

Komponen penting dari setiap studi ilmiah adalah metode penelitian. Ada sejumlah teknik penelitian yang sering digunakan, yaitu teknik kualitatif dan kuantitatif. Metode penelitian kualitatif bersifat deskriptif dan lebih cenderung menggunakan wawancara yang bertujuan untuk memahami pemahaman yang berbeda-beda dari setiap individunya[33], sedangkan metode kuantitatif didasarkan pada faktor penjelas seperti variabel atau parameter yang ditentukan terlebih dahulu untuk mengetahui hubungan antara variabel[33],

Penelitian ini menggunakan teknik penelitian kuantitatif. Pada penelitian ini

dengan metode kuantitatif akan menggunakan *framework* KDD untuk menyelesaikan penelitian ini serta menggunakan *tools Google Colab* sebagai *software* dan juga menggunakan bahasa *pemograman python* untuk melakukan perbandingan antara empat model yaitu algoritma *Random Forest, C4.5, AdaBoost* dan hibrida *C4.5-AdaBoost* untukmenghasilakan akurasi, presisi, recall, dan f1-score. Terdapat perbandingan algoritma yang dipakai pada penelitian ini pada tabel berikut:

Tabel 3. 2 Perbandingan Algoritma

Algoritma	Kelebihan	Kekurangan
Random Forest	-. Random Forest mampu mengatasi masalah overfitting yang sering terjadi dalam model machine learning. -. Random Forest cenderung lebih stabil daripada pohon keputusan tunggal. -. Random Forest dapat memberikan perkiraan kepentingan fitur yang digunakan dalam prediksi. -. Random Forest dapat bekerja dengan baik pada dataset yang tidak seimbang, di mana jumlah contoh dari kelas target yang berbeda tidak proporsional	-. Random Forest menghasilkan prediksi akurat, namun menghadapi kendala dalam interpretabilitas model karena terdiri dari banyak pohon keputusan yang sulit diinterpretasikan secara langsung. -. Proses tuning parameter seperti jumlah pohon, jumlah fitur yang diambil secara acak, dan kriteria pemisahan pada simpul pohon dapat menjadi tugas yang memerlukan eksperimen dan tuning cermat. -. Kesulitan juga muncul saat menangani data berdimensi tinggi, di mana pengambilan sampel acak dan pemilihan fitur acak mungkin tidak efektif dalam menghasilkan prediksi yang akurat.
C4.5	-. Metode yang mudah dipahami karena hasilnya dapat divisualisasikan dalam bentuk pohon keputusan yang mudah dimengerti. -. Cocok untuk menangani data yang memiliki pola non-linear atau hubungan antara variabel	-. Cenderung overfiting atau kelebihan fitting pada data training yang dapat mengurangi performa pada data testing atau validasi. -. Cenderung tidak stabil terhadap perubahan data sehingga model dapat berubah

	yang kompleks. -. Tidak diperlukannya normalisasi data seperti pada beberapa metode machine learning.	secara signifikan jika data berubah. -. Tidak mampu menangani data kontinu atau numerik yang berkelanjutan dengan baik.
AdaBoost	-. AdaBoost biasanya menghasilkan model yang memiliki akurasi tinggi karena fokus pada mengidentifikasi pola yang sulit diakui oleh model lemah. -. AdaBoost cenderung tidak rentan terhadap overfitting karena memberikan penalti pada data yang salah diklasifikasikan. -. Algoritma ini relatif mudah diimplementasikan dan tidak memerlukan penyesuaian parameter yang rumit.	-. AdaBoost dapat menjadi sensitif terhadap noise dan outlier, yang dapat mempengaruhi performa model. -. Waktu pelatihan AdaBoost mungkin lebih lama dibandingkan dengan beberapa algoritma lainnya, terutama jika kompleksitas model lemah tinggi. -. Meskipun relatif mudah diimplementasikan, AdaBoost memerlukan penyesuaian parameter, seperti jumlah iterasi (jumlah model lemah) dan tingkat pembelajaran.

Sumber: [*Dq.Lab*]

Berdasarkan tabel 3.2 diatas dapat dilihat kelebihan dan kekurangan algoritma yang akan digunakan. Algoritma *Random Forest* memiliki kelebihan dalam mengatasi *overfitting*, stabilitas model, memberikan perkiraan kepentingan fitur, dan kinerja baik pada dataset tidak seimbang. Namun, kelemahannya mencakup kompleksitas interpretasi model dan penyesuaian parameter yang memerlukan eksperimen, sedangkan algoritma *C4.5* unggul dalam kemudahan interpretasi dan penanganan pola *non-linear* tanpa perlu normalisasi data, tetapi rentan terhadap *overfitting* dan tidak stabil terhadap perubahan data. *AdaBoost* menghasilkan model akurat dan tidak rentan terhadap overfitting, tetapi dapat sensitif terhadap *noise* dan memerlukan waktu pelatihan lebih lama. Meskipun mudah diimplementasikan, *AdaBoost* memerlukan penyesuaian parameter.

3.4 Teknik Pengumpulan Data

3.4.1 Data Collection

Pada penelitian ini menggunakan data yang berjenis Studi Kasus. yang didapat langsung dari pihak RSUD Kota Tangerang pada bulan Maret tahun 2024 dengan memenuhi persyaratan yang diberikan dari pihak rumah sakit seperti melampirkan Proposal BAB 1 sampai 3 dan Perijinan Pengambilan Data dari Universitas. Pengambilan dataset dilakukan diruang Instalasi Rekam Medis selama dua minggu. Dataset yang akan digunakan peneliti merupakan dataset penyakit Hepatitis dengan rentang waktu bulan Januari 2019 hingga bulan Desember 2023. Peneliti melakukan pengecekan secara satu per satu pada data yang didapat, dan kemudian di catat dan disimpan ke dalam file excel. Pada Gambar 3.2 menujukan tampilan isi dari dataset yang diperoleh secara langsung setelah pengecekan pada data rekam medis dan direkap secara mandiri.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Tahun	Jenis Rawat	No. RM	enis Kelami	Umur	Hemaglobin	Hematokrit	Leukosit	Trombosit	Eritrosit	Eosinofil	utrofil Segm	Limfosit	Monosit	SGOT (AST	SGPT (ALT	ilirubin Dire	lirubin Indir	ilirubin Tot	Albumin
2	2020	Inap	226255	L	8	11,4	32	4,95	346	4,95	2	84	12	2						
3	2020	Inap	227675	L	22	15,2	45	4,5	325	5,25	1	59	28	12	735	1405				
4	2020	Inap	220427	P	37	12	35	12,6	213	4,14	0	92	5	3	145	72	6,41	1,79	8,2	
5	2020	Inap	221217	L	74	11	29	7	119	4,97	1	76	8	15	231	181	9,48	3,36	12,84	
6	2020	Inap	229771	L	51	11,4	34	11,3	149	3,79	1	83	7	9	136	57				2,1
7	2020	Inap	218411	L	64	13,2	37	10,9	126	3,66	0	78	12	10	72	84				
8	2020	Inap	224711	L	56	11,1	32	16,3	274	3,7	0	86	10	4	44	16	2,01	0,85	2,86	2,3
9	2021	Inap	258431	L	39	11,2	32	13,3	50	4,09	0	46	44	4	52	18	0,28	0,25	0,53	2,8
10	2021	Inap	246217	L	34	12	35	8,8	212	3,99	1	61	61	11	96	46	1,58	0,55	2,13	1,8
11	2021	Inap	176293	P	50	13,2	42	13,4	449	4,99	0	80	12	8	171	359	10,08	0,65	10,73	
12	2021	Inap	245117	L	1	10,4	31	6,6	343	4,31	0	41	49	10	239	118	4,54	0,08	4,62	
13	2021	Inap	185166	L	28	14	40	8,3	249	4,42	5	57	28	9	247	395	6,16	1,25	7,41	
14	2021	Inap	245217	L	55	11,2	34	16	170	4,09	3	75	15	7	83	264	6,97	1,38	8,35	
15	2021	Inap	241599	P	1	8,4	28	3,92	8	268	5	23	64	8	44	200				

Gambar 3. 2 Raw Dataset

3.5 Teknik Analisis Data

Pada penelitian ini menggunakan teknik *data mining* disertai dengan *teknik klasifikasi* yang didasari dengan perbandingan framework data mining. Pada *Tabel 3.3* menunjukkan perbanding antara kerangka kerja dan tujuan setiap framework CRISP-DM dan KDD:

Tabel 3. 3 Perbandingan Framework

Faktor Pembeda	CRISP-DM	KDD
Langkah – Langkah	*Business Understanding, Data Understanding,*	*Data Selection, Pre-Processing,*

Penerapan	*Data Preparation, Modeling, Evaluation*	*Tranformation, Data mining, Interpretation / Evaluation*
Tujuan	Memberikan sebuah solusi terhadap masalah sekelompok bisnis tertentu	Memberikan informasi tersembunyi pada pola atau trend yang terdapat pada dataset

Sumber: [*sis.binus.ac.id]*

Berdasarkan pada tabel 3.3 metode KDD memiliki langkah yang tepat dalam hal melakukan analisis data untuk mengetahui informasi pada suatu data dan berfokus pada penemuan pengetahuan. Dapat disimpulkan, disimpulkan bahwa metode KDD merupakan metode yang paling sesuai untuk diterapkan pada analisis data yang dilakukan pada penelitian ini. Dapat disimpulkan bahwa dalam penelitian ini, metode KDD diadopsi sebagaimetode penelitian. Penerapan *framework* KDD pada penelitian inimenggunakan bahasa pemrograman yang dipilih berdasarkan perbandingan. Pada tabel 3. 4 merupakan perbandingan antara bahasa pemrograman *Python* dan bahasa pemrograman R.

Tabel 3. 4 Perbandingan Bahasa Pemrograman

Bahasa Pemrograman	Kelebihan	Kekurangan
Python	-. Memiliki Sintaks sederhana dan mudah dipelajari. -. Python lebih cocok untuk machine learning dan deep learning. -. Tools yang dapat digunakansangat banyak -. Lebih Populer di kalangan pengguna	-. Library yang banyak dan berkemungkinan bisa rumit untuk memahami semuanya. -. Statistik python kurang kuat.
R	-. R lebih cocok utnuk statisticallearning. -. Library yang sedikit dan mudah diketahui	-. Sintaks yang relatif kompleks dan pembelajaran yangtidak langsung. -. Tools yang dapat digunakan hanya ada

	-. Statistik R sangat kuat.	beberapa saja. -. Kurang Populer di kalangan pengguna.

Sumber: [*Dqlab.id*]

Berdasarkan tabel 3.4 menujukan kelebihan yang diberikan oleh bahasa pemograman *python* dapat membantu peneliti dalam melakukan langkah *data mining*, ini dikarenakan *python* memiliki sintaks yang sederhana dan juga lebih cocok untuk machine learning dan deep learning, maka dari itu peneliti menggunakan memilih bahasa pemograman python sebagai bahasa pemograman untuk membuat sebuah model algoritma *Random Forest, C4.5, AdaBoost* dan hibrida *C4.5-AdaBoost*. Dalam pengkodean menggunakan bahasa pemrograman python, peneliti menggunakan *tools* dari *google colab* dalam membantu pemrograman python peneliti. Berikut tabel perbandingan fitur pada *google colab* dan *jupyter notebook*:

Tabel 3. 5 Perbandingan Tools

Feature	***Google Colab***	***Jupyter Notebook***
Cloud-based	-. Memiliki sistem penyimpanan secara otomatis, dan dicadangkanke cloud tanpa harus melakukannya.	-. Jupyter Notebook tidak memiliki sistem penyimpanan cloud, ini dikarekan jupyter notebook dijalankan pada local michine dan disimpan ke dalam hard disk laptop atau komputer.
File Syncing	-. Google Colab dapat diakses melalui hardware apa saja seperti hp, laptop, komputer, dan lain-lain. ini dikarekan akses untuk membukanya hanya perlu melalui browser.	-. Hanya laptop atau komputer yang tentunya harus memiliki file yang sama yang bisa membuka jupyter notebook.
File Sharing	-. Google Colab memiliki fitur untuk berbagi seperti halnya Google Docs, dengan adanya pengguna dapat berkolaborasi dengan orang lain hanya dengan menghubungkannya ke Googel Colab.	-. Jupyter Notebook tidak memiliki fitur untuk berkolaborasi.

Library Install	-. Google Colab sudah menginstall hampir semua library yang ingin digunakan oleh pengguna, ini memungkinkan pengguna tidak perlu menggunakan ruang hard disk dan waktu untuk mengunduk library yang ingin digunakan.	-. Dengan Jupyter Notebook, pengguna harus menginstal setiap library yang ingin gunakan ke perangkat pengguna menggunakan pip atau pengelola paket lainnya. Pengguna juga akan dibatasi oleh RAM, ruang disk, GPU, dan CPU yang tersedia di kompute.
File View Without Install	-. Googel Colab berbasis cloud, pengguna dapat membuka file tanpa menginstall apa pun, pengguna dapat membuka browser google colab dimana pun dan kapanpun.	-. Jupiter Notebook berbasis local, jadi setiap adalibrary atau file baru yang digunakan harus lah menginstall terlebih dahulu.

Sumber: [*Dqlab.id*]

Berdasarkan Tabel 3.5 perbandingan *tools* untuk penelitian, penelitian ini menggunakan *tools Google Colab* dikarenakan terdapat beberapa keunggulan dibandingkan *tools* lainnya. Pemilihan *Google Colab* sebagai *tools* yang digunakan, dikarenakan fitur *File Syncing* dan *Library Install* yang memudahkan peneliti menggunakan dan mengaksesnya dimanapun dan kapanpun.