BAB III
METODOLOGI PENELITIAN

3.1 Subjek dan Objek Penelitian

Subjek dalam penelitian ini adalah data *Body signal of smoking* (Kaggle). Objek dalam penelitian ini adalah pengklasifikasian untuk mendapatkan akurasi yang terbaik diantara algoritma C4.5, *Naïve Bayes* dan *K-Nearest Neighbor*.

3.2 Alat dan Bahan Penelitian

Dalam melakukan penelitian diperlukan alat dan bahan dalam mencapai tujuan dari penelitian.

1. Alat

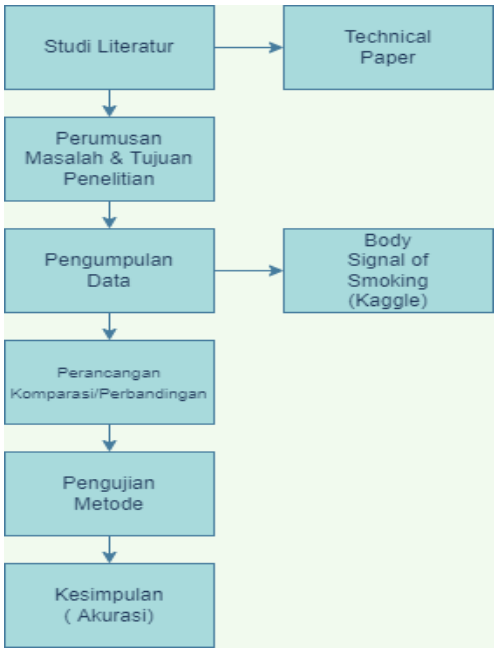
   Alat yang dipakai untuk penelitian ini terdiri dari perangkat keras dan perangkat lunak. Perangkat keras yang dipakai untuk penelitian ini yaitu laptop Asus 64-bit dengan RAM 8GB. Sedangkan, perangkat lunak yang dipakai dalam penelitian adalah *software* RapidMiner dan Sistem Operasi Windows 11.

2. Bahan

   Bahan yang dipakai dalam penelitian ini yaitu data *Body signal of smoking*. Data yang diapakai dalam penelitian ini adalah sejumlah 55692 data.

3.3 Diagram Alir Penelitian

Diagram alir penelitian menggambarkan tahapan-tahapan yang dijalankan agar mempermudah penelitian sehingga penelitian bisa berjalan baik serta sistematis dan sesuai dengan tujuan yang ingin dicapai.



Gambar 3.1 Flowchart Metode Penelitian

3.3.1 Studi Literatur

Studi literatur dalam penelitian ini berupa referensi yang didapatkan dari jurnal, buku, dan website (*online research*) yang sesuai dengan perokok aktif dan pasif serta yang berkaitan dengan algoritma yang digunakan. Studi literatur ini dilakukan sebagai landasan penelitian ini.

3.3.2 Perumusan Masalah dan Tujuan Penelitian

Tahapan sesudah penulis menjalankan analisa terhadap data yang telah didapatkan yaitu tahapan perumusan masalah dan tujuan penelitian yang ingin dicapai dalam penelitian ini. Perumusan masalah serta tujuan penelitian yang diinginkan berdasarkan pada hasil analisa data pada tahap sebelumnya.

Merokok adalah kebiasaan yang sering dilakukan di berbagai negara salah satunya adalah Indonesia. Merokok mempunyai dampak yang buruk bagi kesehatan. Dampak buruknya tidak hanya bagi perokok itu sendiri, tetapi juga bagi orang-orang di sekitar perokok tersebut.

3.3.3 Pengumpulan Data

Pengumpulan data dilakukan melalui website Kaggle, data yang diambil adalah dataset *body signal of smoking* tahun 2022. Sumber data yang dipakai dalam penelitian adalah data yang berasal dari: *https://www.kaggle.com /datasets/kukuroo3/body-signal-of-smoking/* yaitu data *Body Signal Of Smoking* dengan jumlah data sebanyak 55.692 data.

Berikut data yang menunjukkan sebagian dari data set

ID	gender	age	height	weight	waist	eyesight_left	eyesight_right	hearing_left	hearing_right	systolic	relaxation	fasting blood...	Cholesterol
0	F	40	155	60	81.300	1.200	1	1	1	114	73	94	215
1	F	40	160	60	81	0.800	0.600	1	1	119	70	130	192
2	M	55	170	60	80	0.800	0.800	1	1	138	86	89	242
3	M	40	165	70	88	1.500	1.500	1	1	100	60	96	322
4	F	40	155	60	86	1	1	1	1	120	74	80	184
5	M	30	180	75	85	1.200	1.200	1	1	128	76	95	217
6	M	40	160	60	85.500	1	1	1	1	116	82	94	226
7	M	45	165	90	96	1.200	1	1	1	153	96	158	222
9	F	50	150	60	85	0.700	0.800	1	1	115	74	86	210
10	M	45	175	75	89	1	1	1	1	113	64	94	198
11	M	30	175	65	89	1.500	1.500	1	1	130	77	100	184
12	M	30	170	75	87	1.200	1.200	1	1	124	78	101	184

Gambar 3.2 Dataset *Body Signal Of Smoking*

3.3.3.1 Pengolahan Data Awal

Pengolahan data awal adalah pengolahan data menjadi bentuk yang diinginkan, lalu dikelompokkan dan ditentukan atribut datanya. Pengolahan data awal terdiri dari :

1. *Select Data*

   Pada *select data* dilakukan pemilihan atribut data yang akan dianalisis, dari total 55.629 data dengan 25 atribut dan 1 Class. Atribut yang digunakan adalah *ID : index, gender, age : 5-years gap, height, weight, waist : Waist circumference length, eyesight(left), eyesight(right), hearing(left), hearing(right), systolic : Blood pressure, relaxation : Blood pressure, fasting blood sugar, Cholesterol : total,*

triglyceride, HDL : cholesterol type, LDL : cholesterol type, hemoglobin, Urine protein, serum creatinine, Gtp : γ-GTP, oral : Oral Examination status, dental caries, tartar : tartar status, smoking merupakan atribut dalam dataset. Atribut serta deskripsi tentang dataset yang digunakan akan diterangkan pada Tabel 3.1 berikut ini:

Tabel 3.1 Atribut serta Deskripsinya

Atribut	Type	Deskripsi
gender	Text	jenis kelamin
age	Integer	umur
height	Integer	Tinggi diukur dalam cm
weight	Integer	berat (kg)
waist	Integer	pinggang(cm): Panjang lingkar pinggang diukur dalam cm
eyesight_kiri	Integer	penglihatan (kiri)
eyesight_right	Integer	penglihatan (kanan)
Hearing_left	Integer	pendengaran(kiri)
Hearing_right	Integer	pendengaran(kanan)
systolic: Blood pressure	Integer	Tekanan darah sistolik
relaxation: Blood pressure	Integer	Tekanan darah relaksasi
fasting blood sugar	Integer	gula darah puasa
cholesterol: total	Integer	Kolesterol
triglyceride	Integer	Trigliserida
HDL: cholesterol type	Integer	HDL: jenis kolesterol
LDL: cholesterol type	Integer	LDL: jenis kolesterol
hemoglobin	Integer	hemoglobin
Urine protein	Integer	Protein urin

Atribut	Type	Deskripsi
serum creatinine	Integer	kreatinin serum
AST		Aspartate transminase
ALT		Alanine transminase
Gtp: γ-GTP	Integer	Gtp: γ-GTP
oral: Oral Examination status	Integer	oral: Status Ujian Lisan
dental caries	Integer	karies gigi
tartar: tartar status	Integer	karang gigi: status karang gigi
smoking	Integer	Perokok (*Class Atribut*)

2. *Preprocessing Data*

   Data *Body Signal Of Smoking* terdiri dari 25 atribut serta 1 class. Kelas Atribut memiliki 2 nilai yaitu perokok aktif dan perokok pasif yang direpresentasikan dengan angka 1 sebagai perokok aktif dan angka 0 sebagai perokok pasif. Data preprocessing digunakan untuk membersihkan data dari nilai yang inkonsisten dan missing value tersebut. berikut tahapan yang dilakukan pada data *preprocessing*, yaitu :

   a. *Replace Missing Value*

      Untuk mengetahui data yang mengandung *missing value* pada aplikasi Rapidminer menggunakan *tools replace missing values*.

   b. *Data Integration*

      Dalam tahap ini untuk menggabungkan data dari berbagai sumber menjadi satu kesatuan data yang lebih besar.

   c. *Data Transformation*

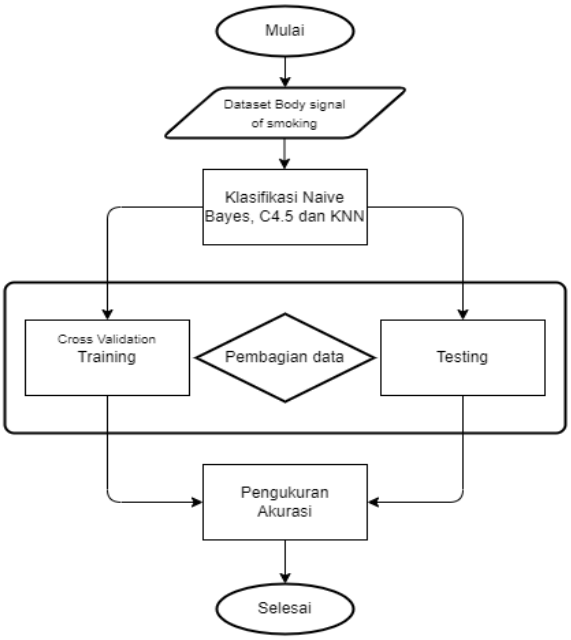
      Pada tahap ini, data akan diubah struktur data, format data, atau nilai data sedemikian rupa sehingga menghasilkan dataset yang sesuai untuk proses mining.

d. *Data Reduction*

Pada tahap ini adalah tahap pengurangan data yaitu untuk mengurangi jumlah sampel data yang diambil.

3.3.4 Perancangan komparasi *Naïve Bayes*, C4.5 dan K-NN

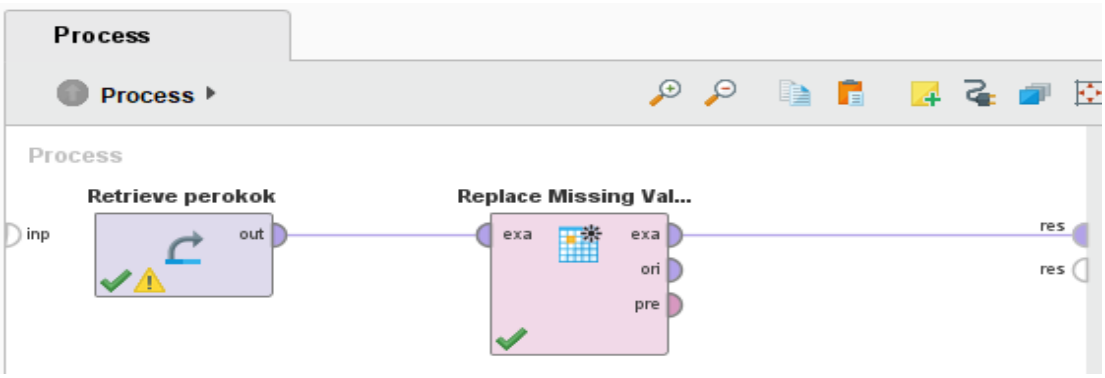
Dalam melakukan perancangan memakai algoritma *Naïve Bayes* dan C4.5 yang menggunakan aplikasi RapidMiner. Parameter yang dipakai berupa atribut dari dataset *body signal of smokings*. Gambar 3.2 menggambarkan bentuk perancangan model algoritma *Naïve Bayes*, C4.5 serta K-NN.



Gambar 3.3 Diagram Alur Perancangan Model

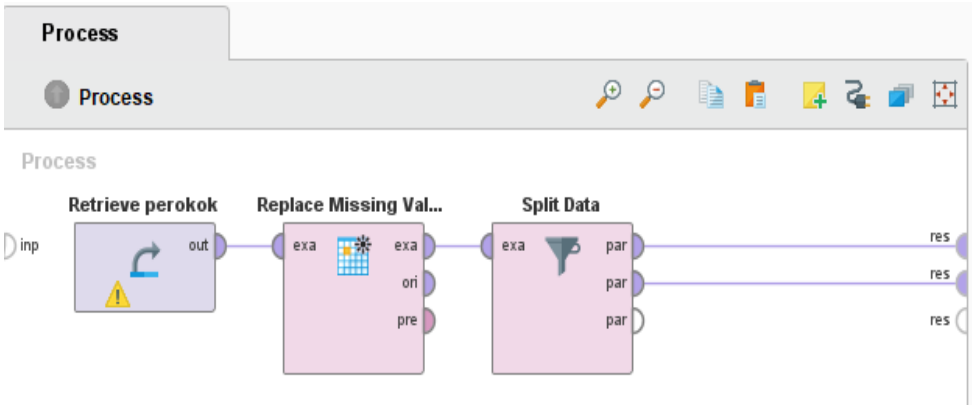
Langkah-langkah dalam perancangan komparasi Naïve Bayes, C4.5 dan K-NN adalah:

1. Dataset yang dipakai menggunakan dataset dari *body signal of smoking* pada tahun 2022 dan sebanyak 55692 data lewat kaggle di mana data tersebut sudah tidak memiliki *missing value* serta siap untuk diuji.



Gambar 3.4 Proses *Missing value*

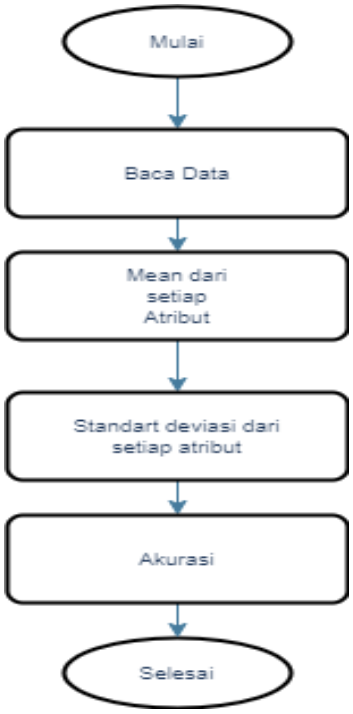
2. Langkah berikutnya adalah *split-data*, dataset terbagi dua yaitu data training dan data testing. Pembagian dilaksanakan dengan random berdasarkan nilai iterasi memakai RapidMiner. Untuk nilai iterasinya adalah 0.8 untuk data testing dan 0.2 untuk data latih (training).



Gambar 3.5 Proses *Split Data*

3. Sesudah mempunyai data uji kemudian dilaksanakan klasifikasi memakai algoritma C4.5, *Naive Bayes* dan KNN. Pengklasifikasian data dilaksanakan secara terpisah untuk masing-masing algoritma.
4. Perhitungan akurasi dilaksanakan menggunakan klasifikasi yang terdiri dari dua kelas, perokok aktif dan perokok pasif. Penilaian klasifikasi tiap metode diuji tingkat akurasinya memakai persamaan akurasi. Perbandingan kedua metode

dilaksanakan sesudah hasil akurasi keduanya keluar. RapidMiner digunakan dalam pengukuran akurasi.



Gambar 3.6 Flowchart Klasifikasi Naïve Bayes

Gambar 3.3 memperlihatkan *flowchart* algoritma *Naive Bayes* yang dipakai dalam penelitian ini. Diawali dengan pembacaan data training *body signal of smoking* yang telah dibagi lewat *cross validation* setelah itu diteruskan proses pemodelan untuk menghitung nilai rata--rata dan standar deviasi dari masing-masing atribut pada dataset *body signal*. Setelah itu memasuki proses penyelesaian memakai data uji/testing.

Gambar 3.7 *Flowchart* Klasifikasi C4.5 (*Decision Tree*)

Diketahui pada gambar di atas memperlihatkan alur tahapan algoritma C4.5 dengan melaksanakan klasifikasi. Dimulai dengan pembuatan atribut kemudian menghitung nilai entropi pada atribut yang sudah dibuat, setelah menghitung entropi dilanjutkan dengan menghitung nilai gain pada masing- masing atribut maka akan keluar hasil gain tertinggi. Maka didapatkanlah rules pohon keputusan dan data akan diuji untuk mendapatkan akurasi.

Gambar 3.8 *Flowchart* Klasifikasi K-*Nearest Neighbor*

Diketahui pada *flowchart* di atas memperlihatkan alur tahapan algoritma *K-Nearest Neighbor* dengan melaksanakan klasifikasi. Dimulai dengan menentukan jumlah *Nearest Neighbor* K setelah menentukan tetangga terdekat terhadap K maka kita dapat melakukan perhitungan jarak *Test* Data ke *Training* Data, setelahh dapat hasil perhitungannya maka kita dapat mengurutkan data berdasarkan data yang mempunyai jarak terkecil lalu mengelompokkan data *test* berdasarkan label mayoritas pada K.

3.4 Pengujian Metode

Langkah berikutnya dilaksanakan pengujian akurasi dari ketiga algoritma dengan menganalisis hasil algoritma terbaik dalam melaksanakan klasifikasi prediksi perokok aktif dan perokok pasif. Hasil penentuan akurasi menentukan ketepatan dari klasifikasi yang dilaksanakan oleh algoritma *Naïve Bayes*, C4.5 dan KNN.

3.5 Kesimpulan

Tahap kesimpulan menerangkan cara menarik kesimpulan berdasarkan hasil dari tiap algoritma *Naive Bayes*, C4.5 dan K-NN, dan dilakukan pembahasan yang berguna dalam menghasilkan fakta yang sesuai tentang hasil penelitian yang sudah dilaksanakan.