BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek yang diteliti adalah tahapan dalam penyakit ginjal kronis yang biasa kita sebut dengan PGK. PGK adalah penyakit dimana fungsi ginjal bertahap-tahap mulai menurun secara signifikan (Webster et al., 2017). Tanda-tanda kerusakan fungsi pada ginjal bisa berupa protein dalam urin ataupun kerusakan fisik pada ginjal (*Stages of Chronic Kidney Disease (CKD) - American Kidney Fund (AKF)*, 2020.). Dalam PGK sendiri terdapat 5 tahapan kerusakan pada ginjal. PGK sering disebut juga sebagai *"Silent Killer"* karena pasien yang mengidap PGK terlihat sama seperti orang biasa (*Kidney Disease | Lab Tests Online*, 2020.) sehingga dengan adanya penelitian prediksi pada PGK dapat mencegah PGK semakin menjadi parah (Rady & Anwar, 2019).

3.2 Metode Penelitian

3.2.1. *Data Collection*

Data yang dikumpulkan berasal dari *dataset* deteksi ginjal kronis dari *website The UCI Learning Repository* yang dikumpulkan oleh L.Jerlin Rubini di Alagappa University dengan judul *Early stage of Indians Chronic Kidney Disease(CKD)* (Dua & Graff, 2017). *Dataset* ini diambil dari *website The UCI Learning Repository* karena *website* tersebut sudah sering digunakan sebagai sumber *dataset* oleh pelajar, edukator dan peneliti

lainnya sehingga menjadikan *The UCI Learning Repository* menjadi salah satu dari 100 “makalah” yang paling sering dikutip dibidang ilmu komputer

Dataset PGK ini berisikan indikator-indikator atau variabel yang dipakai dalam mengidentifikasi PGK. *Dataset* ini sudah dipisahkan dan dipilih dari variabel-variabel pada *dataset* PGK awal dan dibuat menjadi *dataset* PGK baru. Berikut tabel 3.1 yang merupakan *dataset* PGK baru yang sudah dipilih terlebih dahulu oleh dr. Andre Dasta Chrisbeth Sinulingga, SpU yang merupakan dokter spesialis Urologi di RSUD Depok.

Tabel 3.1. Tabel *Dataset* PGK

No	Atribut	Representasi	keterangan	Tipe Data	deskripsi
1	*Age*	*Age*	umur	*numeric*	angka
2	*bp*	*blood pressure*	tekanan darah	*numeric*	Mm/Hg
3	*sg*	*specific gravity*	berat jenis urin hasil test *urinalysis*	nominal	1.005,1.010,1.015,1.020,1.025
4	*su*	*sugar*	gula darah dalam tubuh (glukosa)	nominal	0,1,2,3,4,5
5	*bu*	*blood urea*	kadar urea pada darah	*numeric*	Mgs/dl
6	*sc*	*serum creatinine*	kadar keratinin serum pada darah	*numeric*	Mgs/dl
7	*pot*	*potassium*	Kadar *potassium* pada darah	*numeric*	mEq/L
8	*hemo*	*hemoglobin*	kadar *hemoglobin* pada darah	*numeric*	gms

Tabel 3.2. Tabel *Dataset* PGK

9	*htn*	*hypertension*	mempunyai riwayat hipertensi	nominal	(ya/tidak)
10	*dm*	*diabetes mellitus*	mempunyai riwayat *diabetes mellitus*	nominal	(ya/tidak)
11	*cad*	*coronary artery disease*	mempunyai riwayat jantung koroner	nominal	(ya/tidak)
12	*ane*	*anemia*	mempunyai *anemia*	nominal	(ya/tidak)
13	*class*	*class*	Terkena PGK	nominal	(ya/tidak)

3.2.2. Variabel Independen

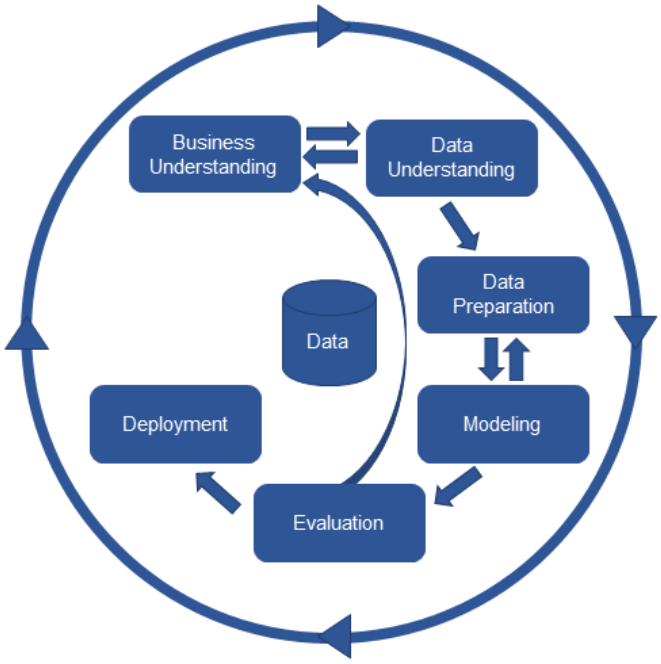
Variabel Independen adalah variabel yang mempengaruhi variabel lain. Dalam *dataset*, yang merupakan variabel independen adalah *age, blood preassure, specific gravity, sugar, blood urea, serum creatinine, potassium, hemoglobin, hypertension, diabetes mellitus, coronary artery disease,* dan *anemia.*

3.2.3. Variabel Dependen

Variabel Dependen adalah variabel yang dipengaruhi oleh variabel independen. Dalam *dataset* ini, yang merupakan variabel dependen adalah *class,* yaitu apakah pasien terkena PGK dengan deskripsi *yes* untuk terkena PGK dan *no* untuk tidak terkena PGK

3.3. Alur Penelitian

Dalam penelitian ini, model *data mining* yang digunakan adalah *Cross Industry Standard Process for Data mining* atau yang biasa disingkat menjadi *CRISP-DM*. Terpilih nya metode *CRISP-DM* dalam penelitian ini bukan tanpa sebab, Menurut (Sharma et al., 2012) *CRISP-DM* merupakan salah satu metode *data mining* yang paling sering dipakai, serta menurut (Azvedo & Santos, 2008) metode *CRISP-DM* dianggap teknologi yang neutral, industri independen dan merupakan standar de-facto untuk *data mining*. Berdasarkan *polling online* di KDNuggets pada tahun 2014, 45% responden memilih *CRISP-DM* sebagai metode utama dalam data analisis, *data mining* ataupun projek *data science* lainnya (Piatetsky, 2014) . Berikut merupakan proses *CRISP-DM* yang berbentuk *flowchart* (Sharma et al., 2017).



Gambar 3.1. *flowchart* CRISP-DM

Sumber: (Sharma et al., 2017)

Dari gambar 3.1 dapat dilihat bahwa *flowchart* dari *CRISP-DM* yang akan diimplementasikan dalam penelitian ini dan yang akan diterangkan oleh penjelasan berikut :

3.3.1. *Business Understanding*

Pada tahap *business understanding,* akan dicarinya arah dan tujuan serta strategi awal pada penelitian ini. Akan ditentukan masalah bisnis dalam penyakit PGK yang diduga sebagai "*silent killer*" menurut *U.S Centers for Disease Control and Prevention* (CDC). CDC adalah lembaga kesehatan masyarakat nasional di Amerika Serikat yang memiliki tujuan utama untuk melindungi kesehatan dan keselamatan publik melalui pengendalian dan pencegahan penyakit, cedera, dan disabilitas di Amerika Serikat dan internasional.

Selain itu, jika PGK terlambat terdeteksi maka peluang untuk sembuh juga akan semakin mengecil sehingga dibutuhkan tenaga medis yang lebih ahli dalam penanganannya serta memerlukan biaya yang lebih besar untuk pengobatannya dan peluang penyembuhan akan semakin kecil. Selain itu, atribut yang diperlukan untuk melakukan tes kesehatan untuk PGK juga cukup banyak sehingga membutuhkan biaya yang cukup mahal.

3.3.2. Data Understanding

Tahap *data understanding* ini bertujuan untuk menganalisasi data yang telah di kumpulkan. Pengumpulan data pada penelitian ini di ambil dari *website The UCI Learning Repository*. *Dataset* yang diambil berjudul

Chronic Kidney Disease Dataset yang terdiri sebanyak 25 atribut serta total jumlah *dataset* tersebut adalah 400 data dan *dataset* yang telah didapatkan akan dipisahkan sesuai dengan arahan dan saran dari dokter. *Dataset* tersebut cocok untuk dipakai dalam penelitian ini karena atribut tersebut berisikan atribut yang diperlukan saat pengambilan tes kesehatan untuk deteksi PGK.

3.3.3. *Data Preparation*

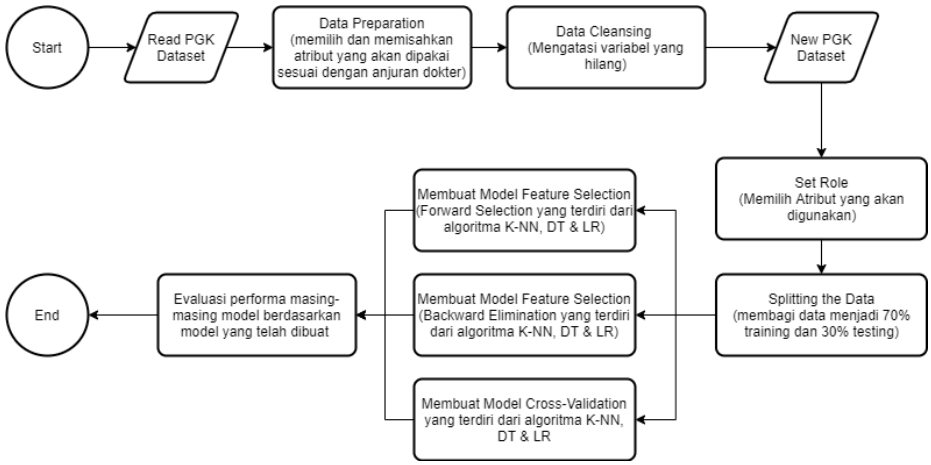
Fase *data preparation* ini bertujuan untuk mendapatkan data yang baik sehingga juga akan menghasilkan hasil yang baik pula. Maka dari itu, cara menyiapkan data untuk di lakukannya tahap selanjutnya yaitu *data preprocessing*. *Data preprocessing* dalam penelitian ini adalah *data cleansing* dengan cara mengisi *missing value* yang terdapat dalam *dataset* PGK. Salah satu cara untuk memperbaiki mengisi *missing value* dalam sebuah data adalah dengan menghapus baris yang memiliki *missing value* agar data yang dihasilkan semakin valid (Sharma et al., 2017).

3.3.4. *Modeling*

Fase pemodelan dalam penelitian juga akan melibatkan beberapa teknik *data mining* dan salah satu teknik *pre-processing*. Teknik *data mining* yang dipakai adalah menggunakan beberapa algoritma, algortima tersebut adalah *K-nearest neighbour* (K-NN), *Decision tree* serta *Logistic regression* (LR). Pemilihan algoritma-algoritma tersebut adalah bukan tanpa sebab, algoritma tersebut ditentukan berdasarkan dari penelitian sebelumnya oleh (Charleonnan et al., 2017) yang menggunakan masing-

masing algoritma untuk memprediksi PGK sedangkan teknik *pre-processing* yang digunakan adalah teknik *feature selection* yang memakai pattern detection yang memilih *subset* fitur yang paling relevan untuk masalah klasifikasi (He et al., 2018).

Penggunaan *feature selection* dalam penelitian ini adalah bertujuan untuk mengurangi kompleksitas algoritma klasifikasi, meningkatkan akurasi klasifikasi algoritma dan dapat mengetahui fitur yang paling berpengaruh dari tingkat akurasi (Zeniarja et al., 2020). Setelah itu akurasi dari algoritma K-NN, *decision tree* dan *logistic regression* akan digunakan sebagai pembanding dari nilai akurasi yang tidak terlibat dalam teknik *feature selection*. Pembuatan model ini akan dibuat dengan menggunakan *tools* Rapid Miner.



Gambar 3.2. *Flowchart* Fase *Modelling*

Berdasarkan pada model *flowchart* yang telah ditampilkan pada gambar 3.2 , berikut merupakan penjelasan tiap-tiap tahapan dari gambar tersebut :

a. Mulai dengan membaca *dataset* mentah atau original
b. Memilih dan memisahkan atribut yang akan dipakai dalam model sesuai dengan yang sudah divalidasi oleh dokter
c. Melakukan *data cleansing* dan menangani *missing value* yang ada
d. *Dataset* PGK baru sudah siap digunakan untuk membuat model
e. Memilih atribut yang akan dipakai.
f. Memulai pembagian *dataset* PGK baru menjadi 3 bagian, yaitu data latih *(training data)* sebesar 70% dan data uji *(testing data)* sebesar 30% sesuai dengan anjuran pembagian *dataset* pada penelitian sebelumnya.
g. Membuat model menggunakan teknik *feature selection* dengan metode *forward selection* yang terdiri dari algoritma K-NN, *Decision tree* dan *Logistic regression*.
h. Membuat model menggunakan teknik *feature selection* dengan metode *backward elimination* yang terdiri dari algoritma K-NN, *Decision tree* dan *Logistic regression*.
i. Membuat model *cross-validation* yang terdiri dari algoritma *Decision tree* dan *Logistic regression*.
j. Evaluasi performa akurasi dari masing-masing model yang telah dibuat.

3.3.5. Evaluation

Evaluasi terhadap model yang telah dibuat apakah sudah sesuai dengan *Business Understanding* dan memastikan semua proses tidak ada yang terlewatkan. Di fase ini, peneliti menggunakan data uji (*testing data*) sebagai pengujian dan evaluasi model *data mining*, serta menggunakan *10-fold cross-validation* sebagai teknik evaluasi. Jika model yang dibuat sudah sesuai dengan tujuan dari *business understanding* yang telah dibahas pada diawal, maka tahap selanjutnya adalah *deployment.*

3.3.6. *Deployment*

Dalam tahap ini, menurut (Fadillah, 2015) *deployment* dapat dilakukan dengan cara menyebarkan model secara paralel ke departemen lain. Pada penelitian ini, tahap *deployment* ditiadakan karena peneliti tidak mengimplementasikan model ke divisi dalam perusahaan dan hanya untuk keperluan studi saja. Penelitian ini juga hanya melakukan *deployment* dengan menggunakan *data science tools* Rapid Miner.

3.3 Validasi Hasil

Untuk melakukan hasil validasi dari model yang dihasilkan ini maka akan menggunakan teknik *pre-processing feature selection* dengan algoritma K-NN, *Decision tree* dan *Logistic regression* serta membandingkan hasil akurasinya pada penelitian terdahulu yang tidak menggunakan teknik ini. Dimana pada penelitian dengan judul implementasi *data mining* untuk deteksi penyakit ginjal kronis (PGK) menggunakan *k-nearest neighbor* (K-NN) dengan *backward data mining* mendapatkan akurasi sebesar 99.25% (Gamadarenda & Waspada, 2020). Setelah

model telah dibuat, 10% dari *data testing* akan diambil untuk divalidasi oleh dokter dengan atas kemauan dokter tersebut.