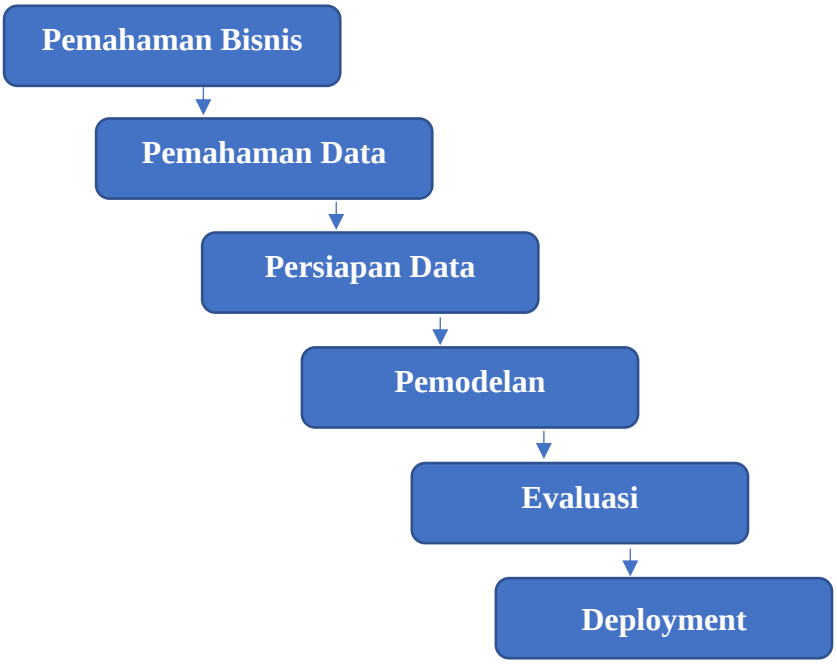
BAB III

METODOLOGI PENELITIAN

3.1 Metode Penelitian

Bab ini membahas langkah-langkah dari proses penelitian yang akan dilaksanakan. Metode yang digunakan dalam penelitian ini mengikuti tahapan *model Cross-Industry Standard Process for Data Mining* (CRISP-DM). Tahapan penelitian dapat dilihat pada Gambar 3.1 di bawah ini:



Gambar 3.1 Alur Penelitian

1. Fase Pemahaman Bisnis *(Bussiness Understanding Phase)*

   Pada tahap ini berfokus pada tujuan penelitian yaitu untuk mengetahui algoritma terbaik untuk memprediksi kelulusan mahasiswa dengan menerjemahkan data historis akademik mahasiswa dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu, sehingga didapatkan model terbaik untuk memenuhi dari tujuan penelitian.

2. Fase Pemahaman Data *(Data Understanding Phase)*

Data yang akan digunakan dalam penelitian merupakan data dari hasil pengumpulan data dokumentasi historis akademik mahasiswa dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu. Atribut yang akan digunakan dalam penelitian ini ada 9 (sembilan) yang dapat dilihat pada tabel 3.1 dibawah ini.

Tabel 3.1 Keterangan Atribut

No	Atribut	Keterangan
1	NIM	Nomor Induk Mahasiswa
2	Program Studi	Program Studi
3	Jenis Kelamin	Jenis Kelamin
4	Status Pernikahan	Status Pernikahan mahasiswa saat dilakukan penelitian
5	Status Pekerjaan	Status Pekerjaan mahasiswa saat dilakukan penelitian
6	IPK Semester 5	IPK Semester 5 (lima) mahasiswa
7	Asal Mahasiswa	Asal mahasiswa
8	SKS Total	Total SKS yang telah ditempuh oleh mahasiswa sampai dengan semester 5
9	Keterangan	Hasil yang akan dijadikan prediksi

3. Persiapan Data

Untuk memudahkan pemahaman instrumen maka atribut pada tabel 3.2 diuraikan dalam tabel 3.2 dibawah ini.

Tabel 3.2 Atribut Data

No	Atribut	Keterangan
1	NIM	16010001
2	Program Studi	Program Studi
3	Jenis Kelamin	L/P
4	Status Pernikahan	Menikah / Belum Menikah
5	Status Pekerjaan	Sudah Bekerja / Belum Bekerja
6	IPK Semester 5	2,00-4,00
7	Asal Mahasiswa	Pringsewu / Luar Pringsewu
8	SKS Total	80-112 sks
9	Keterangan	Tidak Tepat Waktu / Tepat Waktu

4. Pemodelan *(Modeling Phase)*

   Algoritma yang digunakan dalam penelitian ini yaitu algoritma C4.5 dan Naive Bayes untuk mengklasifikasikan dalam memprediksi kelulusan mahasiswa di Universitas Muhammadiyah Pringsewu dan untuk memperoleh sebuah model atau fungsi untuk menggambarkan prediksi kelulusan dengan mengkomparasi algoritma C4.5 dan *Naive Bayes.*

5. Fase Evaluasi *(Evaluation Phase)*

   Pada tahap ini dilakukan evaluasi kinerja dari kedua algoritma yaitu Algoritma C4.5 dan *Naive Bayes* dengan membandingkan hasil nilai rata- rata akurasi, *recall,* dan *error rate* yang terdapat pada tabel *confusion matrix.*

6. Deployment Phase (Fase Penyebaran)

   Setelah tahap evaluasi dimana menilai secara detail hasil dari sebuah model maka dilakukan pengimplementasian dari keseluruhan model yang telah dibangun. Selain itu juga dilakukan penyesuaian terhadap model sehingga dapat menghasilkan suatu hasil yang sesuai dengan target awal tahap CRISP-DM ini.

3.2. Metode Pengumpulan Data

Metode pengumpulan data merupakan suatu hal yang penting dalam penelitian dan merupakan strategi atau cara yang digunakan oleh peneliti dalam mengumpulkan data yang diperlukan dalam penelitiannya. Metode pengumpulan data yang digunakan dalam penelitian ini yaitu :

1. Tinjauan Pustaka *(Research Library)*

   Tinjauan pustaka dilakukan dengan cara membaca, mengutip dan membuat catatan yang bersumber pada bahan-bahan pustaka yang mendukung dan berkaitan dengan penelitian dalam hal ini mengenai *data mining* Algoritma C4.5 dan *Naive Bayes.*

2. Studi Lapangan *(Field Research)*

   Observasi atau pengamatan langsung merupakan suatu teknik atau cara mengumpulkan data dengan dengan jalan pengamatan terhadap kegiatan yang sedang berlangsung (Sudaryono 2017). Pengamatan atau observasi adalah

suatu teknik atau cara untuk mengumpulkan data dengan jalan mengamati kegiatan yang sedang berlangsung. Pengamatan dapat dilakukan dengan partisipasi ataupun nonpartisipasi (Sudaryono 2015).

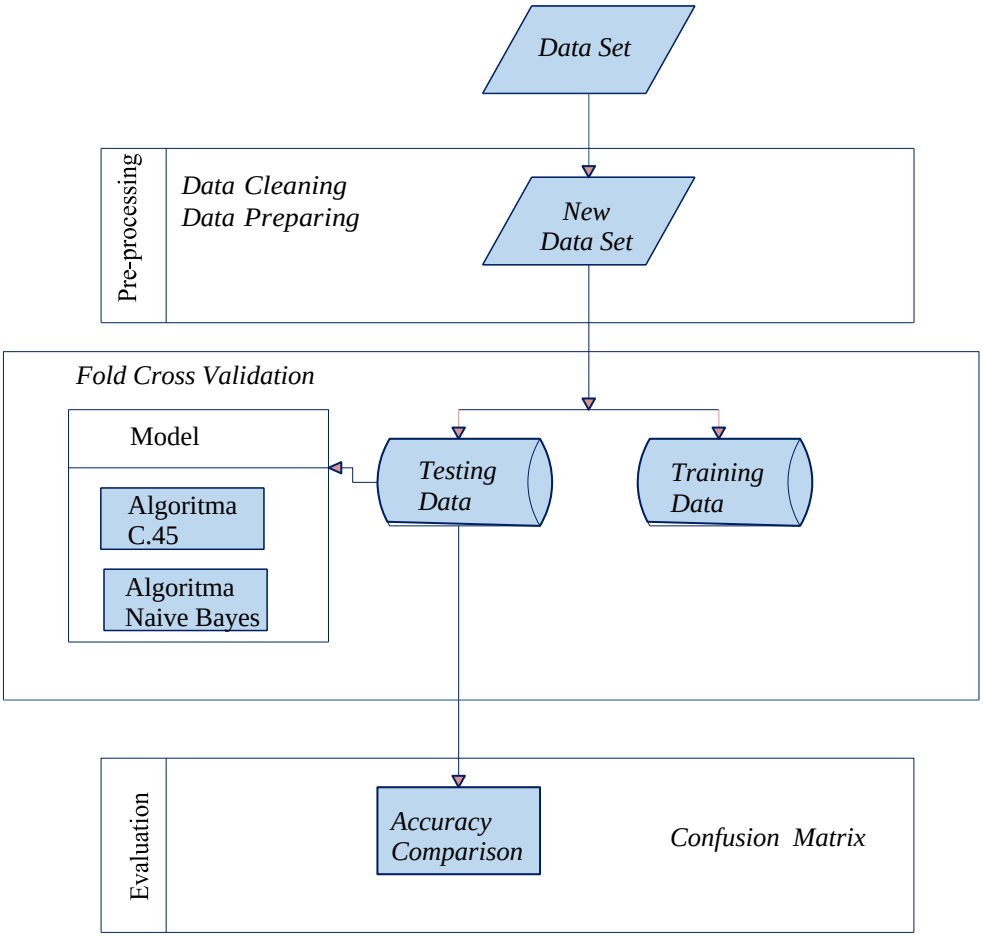
Studi lapangan *(Field Research)* adalah penelitian yang dilakukan dengan cara terjun langsung pada objek yang diteliti untuk mendapatkan data-data yang diperlukan baik dengan cara interview maupun observasi, dimana :

a. Observasi yaitu berupa kegiatan pengamatan langsung terhadap objek yang diteliti.
b. Secondary data yaitu berupa data yang diperoleh peneliti dari sumber yang sudah ada.
c. Dokumentasi adalah sebuah cara yang dilakukan untuk menyediakan dokumen-dokumen dengan menggunakan bukti yang akurat dari pencatatan sumber-sumber informasi khusus dari karangan/tulisan, wasiat, buku, undang-undang, dan sebagainya.

Pada penelitian ini pengumpulan data melalui dokumen dilakukan dengan mempelajari fakta atau data yang ada pada file dokumentasi di Biro Administrasi Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu. Data yang diperoleh berupa *time series data* yang berupa data historis akademik mahasiswa periode 2016-2017 program sarjana.

3.3. Klasifikasi Algoritma

Penelitian ini akan dilakukan dengan menerapkan dua metode yaitu Algoritma C4.5 dan *Naive Bayes* untuk memprediksi kelulusan mahasiswa. Alur tahapan penelitian yang dibuat sebagai kerangka kerja memprediksi kelulusan mahasiswa dapat dilihat seperti yang ditunjukan pada gambar 3.2 dibawah ini.



Gambar 3.2 Bagan Alir Algoritma

Pada Gambar 3.2 menunjukan sistem akan melakukan memasukan *dataset* kelulusan yang kemudian dilakukan tahap *preprocessing data* untuk mengolah data ke dalam bentuk yang siap diproses oleh sistem. Pada tahap *preprocessing* dilakukan beberapa tahapan seperti membersihkan data, seleksi data, *balancing data*, dan transformasi data. Selanjutnya membagi data dengan *cross validation*. Sebagai contoh bila menggunakan 3 *fold cross validation* dengan 2/3 data akan digunakan sebagai *data training*. Selanjutnya *data training* untuk menghasilkan model C4.5 dan *Naive Bayes*. Pada tahap *testing* menggunakan 1/3 *dataset* akan dilakukan klasifikasi berdasarkan model C4.5 dan *Naive Bayes* yang telah dibuat. Dengan menggunakan *confusion matrix* yang akan membagi jumlah hasil prediksi

benar dengan jumlah seluruh data. Lalu didapatkan hasil akurasi dari rata - rata *confusion matrix* dari *k - fold cross validation*.

1. Algoritma C4.5

Pada tahap klasifikasi C4.5 setelah data terbagi menjadi *data training* dan *data testing* cara kerja algoritma C4.5 dapat menggunakan persamaan 2.1 dengan sebagai contoh sample pada gambar 3.2. Penulis mengambil beberapa sampel atribut yang terdapat pada tabel 3.3 dibawah ini.

Tabel 3.3 Sampel Dataset

NIM	Program Studi	Jenis Kelamin	Status Pernikahan	Status Pekerjaan	IPK Sem 5	SKS Total	Asal Mahasiswa	Keterangan
16010001	S1 Manajemen	L	Belum	Tidak	3.22	104	Luar Pringsewu	Tidak Tepat Waktu
16010002	S1 Manajemen	P	Belum	Tidak	3.22	104	Pringsewu	Tepat Waktu
16010003	S1 Manajemen	P	Belum	Tidak	3.53	107	Pringsewu	Tepat Waktu
16020002	S1 Bimbingan dan Konseling	P	Belum	Tidak	3.26	113	Luar Pringsewu	Tepat Waktu
16020004	S1 Bimbingan dan Konseling	P	Belum	Tidak	3.26	113	Luar Pringsewu	Tidak Tepat Waktu
16020006	S1 Bimbingan dan Konseling	P	Belum	Tidak	3.36	113	Luar Pringsewu	Tepat Waktu
16030001	S1 Pendidikan Matematika	L	Belum	Tidak	3.14	109	Luar Pringsewu	Tepat Waktu
16030002	S1 Pendidikan Matematika	P	Belum	Tidak	3.24	109	Pringsewu	Tepat Waktu
16030003	S1 Pendidikan Matematika	P	Belum	Tidak	3.21	111	Pringsewu	Tepat Waktu

Hitung nilai *entropy* per atribut terlebih dahulu dengan persamaan sama dengan di atas

$$Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi$$

1. Program Studi
   a. S1 Manajemen
      $Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi$
      = (-3/5· log2 (3/5)+(- 2/5· log2 (2/5))
      = 0,9709
   b. S1 Bimbingan dan Konseling
      $Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi$

= (-3/5· log2 (3/5)+(- 2/5· log2 (2/5))
= 0,9709

c. S1 Pendidikan Matematika

$Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi$

= (-4/5· log2 (4/5)+(- 1/5· log2 (1/5))
= 0,7219

Hitung nilai *gain* untuk tiap atribut, lalu tentukan nilai *gain* tertinggi. Yang mempunyai nilai *gain* tertinggi itulah yang akan dijadikan akar dari pohon.

$$Gain\ (S, A) = \ entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si)$$

1. Program Studi

$$Gain\ (S, A) = \ entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si)$$

= 0,9182 – ((5/15)*0,9709) – ((5/15)*0,9709)-((5/15)*0,7219)

= 0,0303

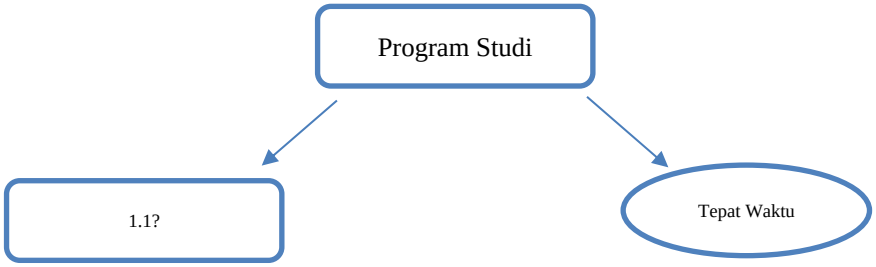
Dengan cara yang sama, akan didapatkan nilai rasio *gain* untuk opsi yang lain. Hasil dari beberapa perhitungan *Split Info* di sajikan pada Gambar Tabel 3.4 dibawah ini.

		Jumlah	Yes	No	Entrophy	Gain	Rasio Gain
Total		15	10	5	0,918296		
Prodi							
	S1 Manaje	5	3	2	0,970951	0,030353	1,724139538
	S1 Bimbin	5	3	2	0,970951		
	S1 Pendid	5	4	1	0,721928		
Jenis Kelamin							
	L	4	1	3	0,811278	0,200327	0,292859004
	P	11	9	2	0,684038		
Status Pernikahan							
	Belum	15	10	5	0,918296	0	0
	Sudah	0	0	0	0		
Status Pekerjaan							
	Belum	15	10	5	0,918296	0	0
	Sudah	0	0	0	0		
IPK Sem 5							
	2.00-3.00	1	0	1	0	0,112717	0,130591995
	3.00-4.00	14	10	4	0,863121		
SKS Total							
	100-110	9	5	4	0,991076	0,063641	0,097906207
	110-120	6	5	1	0,650022		
Asal Mahasiswa							
	Pringsewu	8	6	2	0,811278	0,025841	0,026228481
	Luar Pring	7	4	3	0,985228		

Gambar 3.4 Sample dataset

Pada Tabel 3.4 *Entropy* total atribut hasil adalah 0,9182 dengan jumlah sampel 15 *record* yaitu 10 “Diterima” dan 5 “Tidak Diterima”. Dari hasil perhitungan atribut dengan nilai information gain sebagai nilai terbesar adalah 1,7241 yaitu atribut

Program Studi. Selanjutnya Program Studi dijadikan sebagai *root node* (akar). Berikut bentuk pohon keputusan *root node* dapat dilihat pada Gambar 3.5.



Gambar 3.5 *Root Node*

2. *Algoritma Naive Bayes*

Pada tahap klasifikasi *Naive Bayes* setelah data terbagi menjadi *data training* dan *data testing* cara kerja algoritma *Naive Bayes* dapat menggunakan persamaan 2.6 sebagai contoh sample pada gambar 3.3. Penulis mengambil beberapa sampel atribut yang terdapat pada tabel 3.6 dibawah ini.

Tabel 3.6 *Sample Dataset*

NIM	Program Studi	Jenis Kelamin	Status Pernikahan	Status Pekerjaan	IPK Sem 5	SKS Total	Asal Mahasiswa	Keterangan
16010001	S1 Manajemen	L	Belum	Tidak	3.22	104	Luar Pringsewu	Tidak Tepat Waktu
16010002	S1 Manajemen	P	Belum	Tidak	3.22	104	Pringsewu	Tepat Waktu
16010003	S1 Manajemen	P	Belum	Tidak	3.53	107	Pringsewu	Tepat Waktu
16010004	S1 Manajemen	P	Belum	Tidak	3.35	107	Pringsewu	Tepat Waktu
16010005	S1 Manajemen	L	Belum	Tidak	3.15	104	Pringsewu	Tidak Tepat Waktu
16010006	S1 Manajemen	L	Belum	Tidak	3.25	104	Pringsewu	Tepat Waktu
16010007	S1 Manajemen	P	Belum	Tidak	3.70	107	Pringsewu	Tepat Waktu
16010008	S1 Manajemen	P	Belum	Tidak	3.48	107	Luar Pringsewu	Tepat Waktu
16010009	S1 Manajemen	P	Belum	Tidak	3.63	107	Pringsewu	Tepat Waktu
16010010	S1 Manajemen	P	Belum	Tidak	3.52	107	Pringsewu	Tepat Waktu
16010011	S1 Manajemen	P	Belum	Tidak	3.37	107	Luar Pringsewu	Tepat Waktu
16010012	S1 Manajemen	L	Belum	Tidak	3.51	107	Luar Pringsewu	Tepat Waktu
16010013	S1 Manajemen	P	Belum	Tidak	3.64	107	Pringsewu	Tepat Waktu
16010014	S1 Manajemen	L	Belum	Tidak	3.21	107	Pringsewu	Tepat Waktu
16010015	S1 Manajemen	L	Belum	Tidak	3.15	104	Pringsewu	Tepat Waktu

Hitung jumlah *class*.

Jumlah dari masing-masing *class* hasil dibagi dengan total data yang terdapat pada *data training*.

- H1: P(Class Hasil =”Tepat Waktu”) = 13/15 = 0,866666667
- H2: P(Class Hasil =”Tidak Tepat Waktu”) = 2/15 = 0,071428571

Hitung P(X|Ci), yaitu probabilitas dari setiap atribut pada data X, kemudian dibagi dengan banyaknya jumlah class Tepat Waktu dan Tidak Tepat Waktu: Data X= NIM = “16010001” Program Studi = “S1 Manajemen” Jenis Kelamin = ”Laki-laki”, Status Pernikahan = ”Belum”, Status Pekerjaan = ”Belum” IPK Sem 5 = ”2.00..4.00” SKS Total = ”100..1.20”. Asal= ”Pringsewu”.

- P(NPM=”16010001” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(NPM=”16010001” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(Program Studi=” S1 Manajemen” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(Program Studi =” S1 Manajemen” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(Jenis Kelamin=” L” | class_hasil =” Tepat Waktu”) = 4/13 = 0,3076923077
- P(Jenis Kelamin=” L” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1
- P(Status Pernikahan=”Belum” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(Status Pernikahan=” Belum” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(Status Pekerjaan=”Belum” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(Status Pekerjaan=” Belum” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(IPK SEM 5=”3.00..4.00” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(IPK SEM 5=” 3.00..4.00” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(SKS TOTAL=”100..110” | class_hasil =” Tepat Waktu”) = 13/13 = 1
- P(SKS TOTAL=” 100..110” | class_hasil =” Tidak Tepat Waktu”) = 2/2 = 1

- P(Asal=”Pringsewu” | class_hasil =” Tepat Waktu”) = 10/13 = 0,7692307692

- P(Asal=”Pringsewu” | class_hasil =” Tidak Tepat Waktu”) = 1/2 = 0,5

Hitung total masing-masing nilai pada setiap atribut pada data X :

• P(X| class_hasil =”Tepat Waktu”)

= 1 x 1 x 0,3076923077 x 1 x 1 x 1 x 1 x 0,7692307692

= 0,2366863905

• P(X| class_hasil =”Tidak Tepat Waktu”)

= 1 x 1 x 1 x 1 x 1 x 1 x 1 x 0,5

= 0,5

3.4 Alat dan Bahan

Adapun alat dan bahan yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Hardware

   Kebutuhan perangkat keras (*hardware*) yang digunakan : Laptop Huawei D14 2022 Edition 11th Gen Intel(R) Core(TM) i3-1115G4 @ 3.00GHz RAM 8 GB, SSD dengan kapasitas 256 GB.

2. Software

   Kebutuhan perangkat lunak (*software*) yang digunakan :

   a. Sistem Operasi Windows 11

   b. Aplikasi Rapid Miner Studio 9.10.001.

3. Data

   Data yang akan digunakan dalam penelitian ini adalah data akademik yang diperoleh dari Biro Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu yang terdiri dari Data mahasiswa angkatan 2016 dan 2017 jenjang sarjana sebanyak 881.

3.5 Pengujian *Cross Validation*

Cross Validation merupakan sebuah teknik yang digunakan untuk memvalidasi keakuratan sebuah model yang dibangun berdasarkan data set tertentu. Dalam proses pembentukan model data yang digunakan merupakan data *training* sedangkan data *testing* digunakan untuk memvalidasi model. Pendekatan pada metode *cross validation*, data set dibagi menjadi sejumlah *k-fold* eksperimen dengan masing-masing eksperimen menggunakan data partisi *ke-k* sebagai *data testing* dan menggunakan sisa partisi lainya sebagai data *training*. Dari beberapa percobaan peneliti setuju bahwa menggunakan data uji dari *dataset* untuk menghitung *test error* merupakan cara yang lebih baik untuk mendapatkan estimasi yang lebih handal pada akurasi model di masa mendatang. Menggunakan data uji juga pendekatan yang efisien untuk memvalidasi model. Tetapi pada praktiknya masih ada potensi masalah yang timbul: bagaimana mengetahui data uji tersebut tidak terlalu mudah untuk model? Bisa jadi sampel acak yang dipilih tidak begitu acak, terutama jika hanya memiliki *dataset* yang sedikit. Dalam kasus tersebut, *test error* yang dihasilkan mungkin kurang mewakili akurasi model.

Dari permasalahan tersebut, ide yang muncul adalah mengulangi sampling data uji beberapa kali dan menggunakan sampel yang berbeda untuk setiap kali pengujian. Misalnya dengan membuat 10 sampel data uji berbeda yang digunakan untuk mencari 10 kali nilai *test error*. Dengan cara ini dapat diketahui rata-rata nilai dari 10 *test error*. Prosedur ini merupakan cara standar untuk memvalidasi model tetapi memerlukan waktu yang lebih lama. Meskipun pada prinsipnya dengan mencari rata-rata nilai *test error* lebih dari satu kali pengujian lebih unggul daripada *test error* tunggal, tetapi cara ini masih juga memiliki satu kelemahan yaitu: beberapa baris data yang digunakan dalam beberapa sampel data uji mungkin belum digunakan untuk pengujian sama sekali. Sebagai akibatnya, kesalahan yang dibuat pada baris yang diulang memiliki dampak yang lebih tinggi pada kesalahan pengujian yang lain. Untuk itu dilakukanya menggunakan *k-fold cross validation.*