BAB III
METODOLOGI PENELITIAN

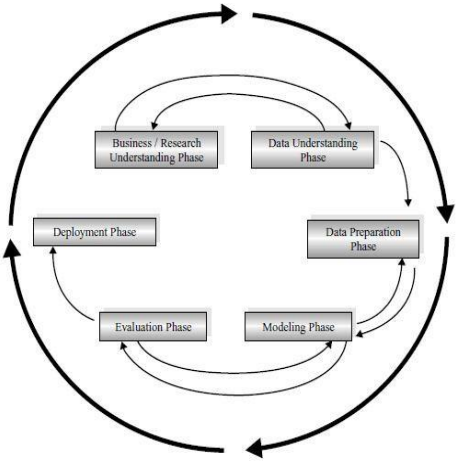
3.1 Metode Penelitian

Menurut (Dawson 2009, 87) ada empat metode penelitian yang umum digunakan yaitu tindakan penelitian, *eksperimen, studi kasus* dan *survey*. Dalam konteks penelitian ini menggunakan eksperimen, yaitu suatu metode yang dilakukan dengan mengacu kepada pemecahan masalah yang meliputi mengumpulkan data, merumuskan hipotesis, pengujian hipotesis, menafsirkan hasil, dan kesimpulan.

Jenis eksperimen dibagi dua, yaitu eksprimen absolut dan eksperimen komparatif. Eksperimen absolut mengarah kepada dampak yang dihasilkan dari eksperimen. Sedangkan eksperimen komparatif yaitu membandingkan dua objek yang berbeda, misalnya membandingkan dua algoritma yang berbeda dengan melihat hasil statistik masing-masing mana yang lebih baik penelitian (Kothari 2004) Pada penelitian ini, jenis penelitian yang diambil adalah eksperimen komparatif yaitu dengan membandingkan dua metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM).*

Penelitian ini jika dilihat dari bentuk data dan informasi yang dikelola, penelitian ini tergolong jenis penelitian kuantitatif. Penelitian kuantitatifadalah penelitian yang hipotesisnya dapat diuji dengan teknik-teknik statistik. Metode ini digunakan saat melakukan pengujian kualitas yaitu menggunakan *Confusion matrix* yang menampilkan nilai akurasi, *precision*, *recall*, ROC pada masing-masing metode yang dibandingkan.

Data mining adalah sebuah proses sehingga dalam melakukan prosesnya harus sesuai prosedur yaitu proses CRISP-DM (*Cross-Industry Standard Process for Data mining*) yaitu sebagai keseluruhan proses, *reprocessing* data, pembentukan model, model evaluasi, dan akhirnya penyebaran model seperti gambar 3.1 dibawah ini.



Gambar 3.1 Tahapan Metodologi CRISP-DM

Dalam gambar 3.1 dijelaskan bahwa metode penelitian eksperimen, digunakan model proses CRISP-DM (*Cross Standard Industry Process for Data mining*) yang terdiri dari 6 tahapan :

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modelling*
5. Evaluasi
6. *Deployment*

Penelitian ini menggunakan metode studi pustaka sebagai tahap awal dengan mempelajari landasan teori mengenai *data mining* menggunakan algoritma klasifikasi yaitu metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM)* pada beberapa literatur dan referensi lainnya. Referensi mencakup data-data dari jurnal, *internet, e-book,* dan dokumen lainnya yang berkaitan dengan penelitian ini.

Dan akan menggunakan *dataset* yang dijadikan *data training* maupun *data testing*. Dari data tersebut akan dibagi menjadi dua bagian, yaitu untuk *data training* dan untuk *data testing*. Dengan menggunakan *dataset* dan atribut tersebut akan dilakukan beberapa penyeleksian untuk menghasilkan data yang dibutuhkan, tahapannya yaitu:

1. Data *Cleaning*

   Untuk membersihkan nilai yang kosong.

2. Data *Integration*

   Berfungsi menyatukan tempat penyimpanan yang berbeda ke dalam satu data.

3. Data *Reduction*

   Jumlah atribut yang digunakan mungkin terlalu banyak dan atribut yang tidak diperlukan akan dihapus.

Data training akan digunakan untuk pembentukan pola algoritma klasifikasi *data mining* dari metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM). Data testing* digunakan untuk menguji pola algoritma yang telah dibentuk. Proses penentuan di terima atau ditolak pemberian beasiswa oleh pihak manajemen Universitas Muhamadiyah Pringsewu dengan mengacu pada 5C (*character, capacity, capital, colleteral, condition of economy*).

Dalam pengembangannya akan dibuat sebuah aplikasi dengan PHP. Hasil penelitian ini adalah perbandingan metode klasifikasi *data mining* untuk analisis pemberian beasiswa dengan metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM)* agar beasiswa dapat diberikan kepada mahasiswa dengan tepat sasaran.

Penulis menggunakan jenis penelitian yang dilihat dari tujuannya adalah penelitian eksperimen. Penelitian eksperimen adalah penelitian yang menguji hipotesis mengenai hubungan sebab akibat.

Pendekatan pada penelitian ini adalah eksperimen dengan cara menguji metode melalui suatu *prototype* sistem yang keefektifannya akan diuji dengan menggunakan kelengkapan akurasi (*accuration*), (*recall*), dan ketepatan (*precision*).

3.2 Deskripsi Data

Deskripsi data adalah merupakan gambaran data yang digunakan dalam suatu penelitian. Data yang digunakan dari beberapa pengajuan beasiswa dari tahun 2019 sampai 2021yang diperoleh dari bagian kemahasiswaan. Data yang digunakan berjumlah 1091 terdiri dari 18 atribut dengan 1 atribut prediksi untuk lebih jelasnya dapat dilihat pada table 3.2 dibawah ini.

Table 3.1 Atribut Data yang digunakan

No	Atribut	Tipe	Keterangan
1	Tahun	*Integer*	[2019,2020,2021]
2	Nama Siswa	*Polynominal*	ID
3	Status DTKS	*Integer*	[0,1] 0. Belum Terdata, 1. Terdata
4	Pekerjaan Ayah	*Integer*	[1,2,3,4,5,6] 1. Tidak Bekerja, 2. Buruh, 3. Petani, 4. Wirausaha, 5. Peg. Swasta, 6. PNS
5	Penghasilan Ayah	*Integer*	[0-10000000]
6	Status Ayah	*Integer*	[1,2] 1. Meninggal, 2. Hidup
7	Pekerjaan Ibu	*Integer*	[1,2,3,4,5,6] 1. Tidak Bekerja, 2. Buruh, 3. Petani, 4. Wirausaha, 5. Peg. Swasta, 6. PNS
8	Penghasilan Ibu	*Integer*	[0-10000000]
9	Status Ibu	*Integer*	[0,1]1. Hidup, 0. Meninggal
10	Jumlah Tanggungan	*Integer*	[1,2,3,4,5,6]
11	Kepemilikan Rumah	*Integer*	[1,2]1. Kontrakan, 2. Sendiri
12	Sumber Listrik	*Integer*	[1,2]2. PLN, 1. Numpang
13	Luas Tanah	*Integer*	[0-1000]
14	Luas Bangunan	*Integer*	[0-250]
15	Sumber Air	*Integer*	[1,2]1. Kemasan, 2. Sumur
16	MCK	*Integer*	[1,2]1. Berbagi Pakai, 2. Kepemilikan Sendiri Didalam
17	Prestasi	*Integer*	[0,1]0. Tidak Ada, 1. Ada
18	BEASISWA	*Binominal*	[Tidak Diterima, Diterima]

3.3 Tahap Tahap Penelitian

3.3.1 Metode Pengumpulan Data

Dalam penelitian ini untuk mendapatkan data yang diharapkan maka peneliti mencari, mempelajari, serta mendalami berbagai literatur baik jurnal, buku, ataupun referensi-referensi lainnya yang berhubungan dengan topik penelitian ini.

1. Metode observasi

   Pengamatan langsung dilakukan penulis di Universitas Muhamadiyah Pringsewu untuk mengumpulkan data yang berhubungan dengan mahasiswa yang layak untuk menerima bantuan beasiswa dengan cara mengamati dan mencatat secara sistematis masalah-masalah yang diselidiki dan meneliti secara langsung terhadap objek yang akan diteliti.

2. Wawancara

   Penulis melakukan tanya jawab langsung kepada pihak-pihak yang berkompeten atau berkepentingan dalam menentukan pemberian beasiswa dan yang mengetahui kondisi actual mahasiswa tersebut. Dari hasil wawancara ini diharapkan dapat menambah kelengkapan data yang diperoleh dari hasil pengamatan.

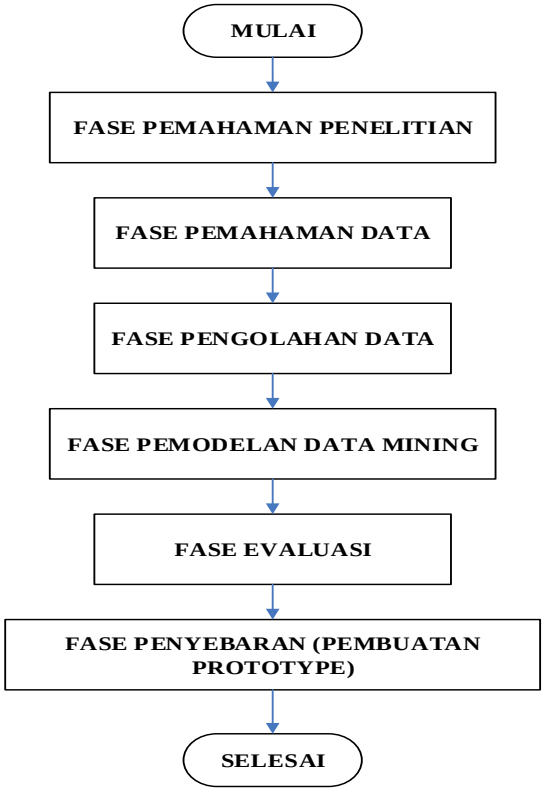
3. Studi Pustaka

   Suatu bentuk riset yang menggunakan proses pencarian data dengan cara mencari, membaca buku dan mengolah isi dari beberapa referensi buku yang dapat dijadikan tujuan dalam pencarian data. Data yang diperoleh dari studi pustaka inilah yang disebut dengan data sekunder, tujuan dari data sekunder ini adalah sebagai landasan teori untuk menganalisa pemecahan masalah di dalam tesis ini.

3.3.2 Langkah-langkah Penelitian

Setiap tahun dilakukan penilaian oleh bagian pengambil keputusan yang mencakup dua hal penting, yakni *review* terhadap biodata mahasiswa dan rencana pemberian kembali beasiswa dengan menggunakan data-data yang sudah ada sebelunya. Dalam hal ini, pihak manajemen Unversitas Muhamdiyah Pringsewu berdiskusi dengan staff dan jajarannya. Dalam peraturan program beasiswa KIP-Kuliah dapat ditentukan bahwa yang berwenang dalam membuat keputusan atau rekomendasi adalah pihak kampus. Penelitian ini dilaksanakan berdasarkan alur penelitian seperti yang ditunjukkan pada Gambar 3.2.

Langkah-langkah yang digunakan dalam penelitian ini menggunakan model CRISP-DM (*Cross Standard Industries Process for Data mining*), dalam metode ini terdapat 6 tahapan untuk lebih jelasnya dapat dilihat pada gambar 3.2 dibawah ini:



Gambar 3.2 Langkah-langkah Penelitian

Keterangan:

- *Business/Research Understanding Phase*

  Permasalahan yang ada di Universitas Muhamadiyah Pringsewu adalah tidak adanya metode yang digunakan dalam melakukan penentuan pemberian beasiswa. Berangkat dari permasalahan ini, pada fase pemahaman bisnis, peneliti mengumpulkan data yang terkait dengan penghitungan prediksi, langkah ini dilakukan dengan cara melakukan wawancara dan observasi di objek penelitian setempat, wawancara dilakukan dengan divisi yang berwenang menentukan pemberian beasiswa pada mahasiswa Universitas Muhamadiyah Pringsewu.

- *Data Understanding Phase* (Fase Pemahaman Data)

  Pada tahap ini mulai dilakukan menganalisis data master yang direlasikan dengan data sekunder. Data yang diambil hanya data Januari pada tahun 2021 yang akan dikumpulkan sebagai data training dan data testing dengan pembagian dataset sebesar 90% dan 10% dari semua data *record*. Pada fase pemahaman ini peneliti melakukan wawancara dengan staff database untuk mendalami data dan penentuan metode prediksi penjualan yang sedang berjalan.

- *Data Preparation Phase* (Fase Pengolahan Data)

  Dari fase pemahaman data, akan menjadi modal peneliti untuk masuk ke fase pengolahan data, tahap ini meliputi semua kegiatan untuk membangun dataset akhir data yang akan diproses pada tahap pemodelan *(modeling).*

- *Modeling Phase* (Fase Pemodelan)

  Proses pemodelan dilakukan dengan menguji metode-metode *data mining* yang akan dikomparasi yaitu metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM)*. terhadap jumlah atribut yang ada. Dalam proses modeling akan dilihat akurasi dari setiap metode. Pada tahapan ini juga dilakukan eksperimen terhadap atribut-atribut data mahasiswa berupa modifikasi ataupun menghapus atribut yang tidak memiliki pengaruh yang signifikan. Hal ini dilakukan untuk meningkatkan nilai akurasi. Proses pengujian setiap metode akan dilakukan menggunakan teknik pengujian *10-folds cross validation* yang menghasilkan nilai statistik pengujian berupa nilai akurasi, *precision, recall* dan *ROC Curve*. Metode dengan nilai akurasi terbaik akan diimplementasikan pada prototipe yang akan dirancang. Pengujian metode akan dilakukan dengan bantuan *tools Rapid Miner.*

- *Evaluation Phase* (Fase Evaluasi)

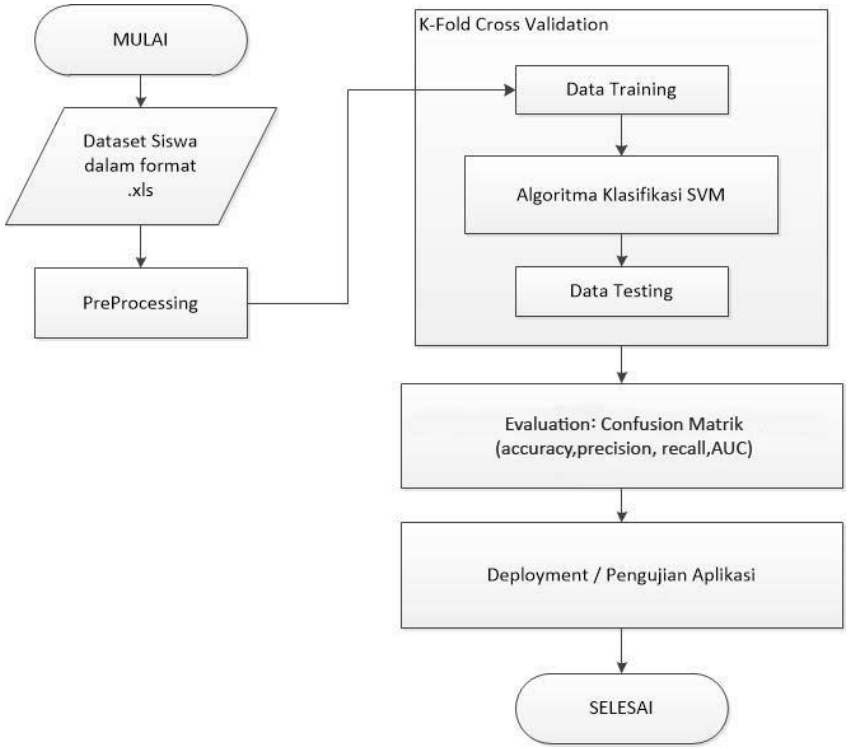
  Metode dengan hasil statistik terbaik akan diimplementasikan pada prototipe yang dirancang. Tahap evaluasi akan dilakukan dengan

menguji prototipe menggunakan data testing baru selain dari data training yang telah digunakan saat proses modeling untuk mengetahui tingkat kesalahan dalam model yang dipakai. Data testing yang dipakai tersebut akan ditentukan keputusan mahasiswa yang layak diberikan beasiswa.

- *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.

3.3.3 Klasifikasi Algoritma SVM

Teknik *data mining* untuk klasifikasi penerimaan beasiswa dengan metode *Support vector machine (SVM)* dapat digunakan untuk membantu pengambilan keputusan pemberian beasiswa. Algoritma SVM digunakan untuk mencari nilai probabilitas pior masing-masing kriteria. Bagan alir perhitungan algoritma SVM seperti ditunjukkan pada gambar 3.3 di bawah ini.



Gambar 3.3 Bagan Alir Algoritma SVM

Pada Gambar 3.1 menunjukan sistem akan melakukan memasukan dataset beasiswa 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. Membersihkan data untuk menghapus nilai NaN yang merupakan nilai missing value. Pada seleksi data akan menyeleksi data atribut untuk memilih atribut penunjang yang berpengaruh dengan atribut ICU. Transformasi data yaitu melakukan normalisasi data dengan normalisasi min-max. 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 SVM dengan kernel linear, kernel gaussian RBF, dan kernel polynominal. Pada tahap testing menggunakan 1/3 dataset akan dilakukan klasifikasi berdasarkan model SVM 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*. Pada tahap klasifikasi

support vector machine (SVM) setelah data terbagi menjadi data training dan data testing cara kerja algoritma SVM dapat menggunakan rumus persamaan 2.6 dengan sebagai contoh sample pada Tabel 3.2. Penulis mengambil beberapa sample atribut yang terdapat pada tabel 3.2 dibawah ini

Tabel 3.2 Data sample model klasfikasi

Nama Siswa	Status Ayah	Kepemilikan Rumah	Sumber Listrik	Sumber Air
REYKA ANISSA PUTRI	1	1	2	1
AHMAD WAHYU NASRULLOH	2	2	2	2
Nur Habibah Azzahra	1	1	1	2
LENI AGUSTIN	1	1	2	1

Berdasarkan tabel 3.1 maka atribut “Status Ayah” sebagai x1, atribut “kepemilikan rumah” sebagai x2, atribut “Sumber Listrik” sebagai x3 dan “Sumber Air” sebagai b. Dalam SVM hanya mempunyai dua kelas penentu yaitu kelas positif dan negatif. Dalam atribut sumber air terdapat variable 1 sebagai kelas positif dan variable 2 sebagai kelas negatif. Setelah itu menggunakan rumus persamaan 2.6 dapat diterapkan sebagai berikut :

(1) $(1w_1 + 1w_2 + 2w_3 + 1b) \geq 1 \rightarrow (1w_1 + 1w_2 + 2w_3 + b) \geq 1$

(2) $(2w_1 + 2w_2 + 2w_3 + 2b) \geq 1 \rightarrow (-2w_1 - 2w_2 - 2w_3 - b) \geq 1$

(3) $(1w_1 + 1w_2 + 1w_3 + 2b) \geq 1 \rightarrow (- 1w_1 - 1w_2 - 1w_3 - b) \geq 1$

(4) $(2w_1 + 1w_2 + 2w_3 + 1b) \geq 1 \rightarrow (2w_1 + 1w_2 + 2w_3 + b) \geq 1$

Maka dapat dituliskan sebagai matrix :

$1\ 1\ 2\ 1 \mid 1$

$-2\ -2\ -2\ -1 \mid 1$

$-1\ -1\ -1\ -1 \mid 1$

$2\ 1\ -2\ 1 \mid 1$

Lalu selanjutnya menggunakan persamaan rumus

$R_2 + 2R_1 \rightarrow R_2$; $R_3 + 1R_1 \rightarrow R_3$; $R_4 - 2R_1 \rightarrow R_4$ sebagai berikut :

$1\ 1\ 2\ 1 \mid 1$

$0\ 0\ 2\ 1 \mid 3$

$0\ 0\ 1\ 0 \mid 2$

$0\ -1\ -6\ -1\ |\ -1$

$R_2 \leftrightarrow R_4$ sebagai berikut :

$1\ 1\ 2\ 1\ |\ 1$

$0\ -1\ -6\ -1\ |\ -1$

$0\ 0\ 1\ 0\ |\ 2$

$0\ 0\ 2\ 1\ |\ 3$

$R_2/-1 \rightarrow R_2$ sebagai berikut :

$1\ 1\ 2\ 1\ |\ 1$

$0\ 1\ 6\ 1\ |\ 1$

$0\ 0\ 1\ 0\ |\ 2$

$0\ 0\ 2\ 1\ |\ 3$

$R_1 - 1R_2 \rightarrow R_1$ sebagai berikut :

$1\ 0\ -4\ 0\ |\ 0$

$0\ 1\ 6\ 1\ |\ 1$

$0\ 0\ 1\ 0\ |\ 2$

$0\ 0\ 2\ 1\ |\ 3$

$R_1 + 4R_3 \rightarrow R_1$; $R_2 - 6R_3 \rightarrow R_2$; $R_4 - 2R_3 \rightarrow R_4$ sebagai berikut:

$1\ 0\ 0\ 0\ |\ 8$

$0\ 1\ 0\ 1\ |\ -11$

$0\ 0\ 1\ 0\ |\ 2$

$0\ 0\ 0\ 1\ |\ -1$

$R_2 - 1R_4 \rightarrow R_2$ sebagai berikut :

$1\ 0\ 0\ 0\ |\ 8$

$0\ 1\ 0\ 0\ |\ -10$

$0\ 0\ 1\ 0\ |\ 2$

$0\ 0\ 0\ 1\ |\ -1$

Jadi di dapatkan nilai variable :

$w1 = 8$

$w2 = -10$

$w3 = 2$

$b = -1$

Lalu menerapkannya dalam data sample baru pada tabel 3.3 untuk pengujian model yang sudah didapatkan.

Tabel 3.3 Data sample baru

ID	Status Ayah	Kepemilikan Rumah	Sumber Listrik
1	1	1	2
2	2	2	2
3	1	1	1
4	1	1	2

Dengan menggunakan rumus 2.16 maka :

f(x) = $w_1 . x_1 + w_2 . x_2 + w_3 . x_3 + b$

Data 136318 = (8 . 1) + (-10 . 1) + (2 . 2) + (-1)

= 8 – 10 + 4 -1

= 1

Data 0b5976 = (8 . 1) + (-10 . 1) + (2 . 2) + (-1)

= 8 – 10 + 4 -1

= 1

Data 1761b2 = (8 . 2) + (-10 . 2) + (2 . 2) + (-1)

= 16 – 20 + 4 – 1

= -1

Data 03b7be = (8 . 2) + (-10 . 2) + (2 . 2) + (-1)

= 16 – 20 + 4 -1

= -1

Keterangan :

$R1$: Baris ke 1

$R2$: Baris ke 2

$R3$: Baris ke 3

$R4$: Baris ke 4

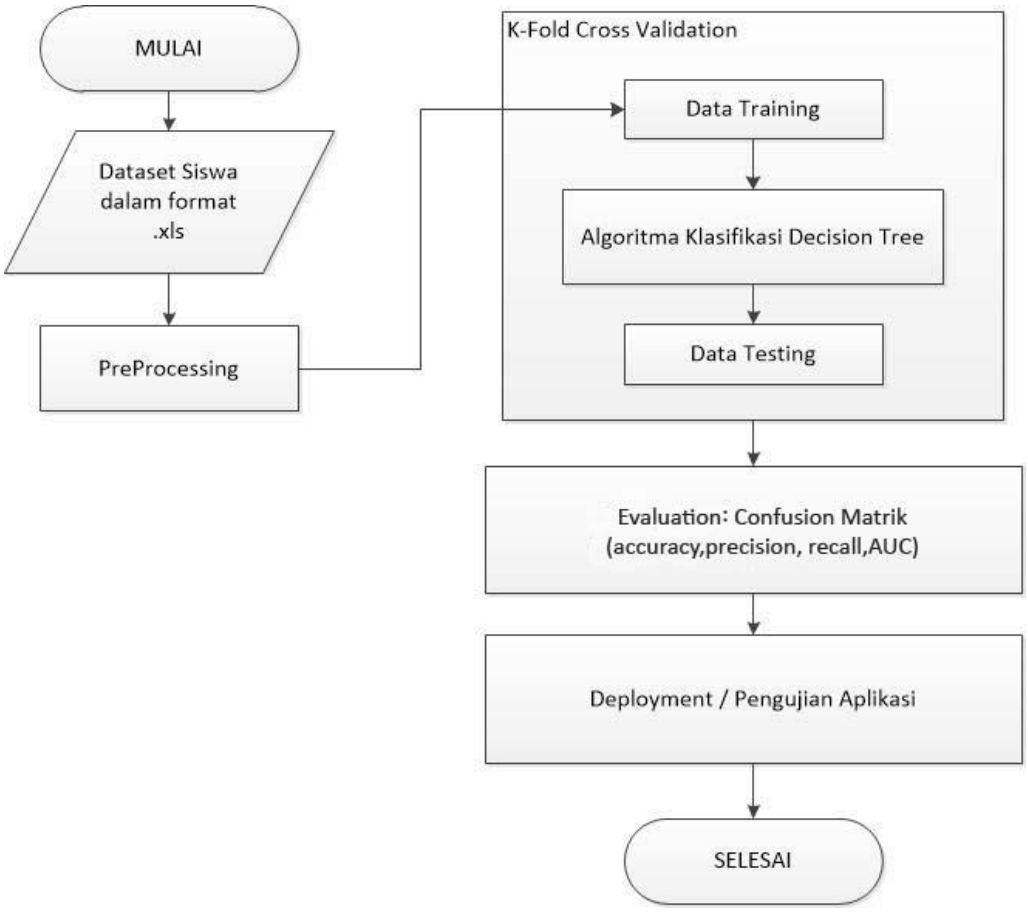
$x1$: Data baru atribut “Status Ayah”

$x1$: Data baru atribut “Kepemilikan Rumah”

$x1$: Data baru atribut “Sumber Listrik"

3.3.4 Klasifikasi Algoritma Decision Tree

Algoritma *Decision tree* digunakan untuk mencari nilai *entrop*i per-atribut dan menghitung nilai gain per-atribut dimana kriteria dengan nilai gain tertinggi menjadi akar dari *tree* (pohon). Bagan alir algoritma *decision tree* seperti ditunjukkan pada gambar 3.4 di bawah ini:



Gambar 3.4 Bagan Alir Algoritma *Decision tree C4.5*

Pada Gambar 3.1 menunjukan sistem akan melakukan memasukan dataset beasiswa 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. Membersihkan data untuk menghapus nilai NaN yang merupakan nilai missing value. Pada seleksi data akan menyeleksi data atribut

untuk memilih atribut penunjang yang berpengaruh dengan atribut ICU. Transformasi data yaitu melakukan normalisasi data dengan normalisasi min-max. 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 dengan kernel linear, kernel gaussian RBF, dan kernel polynominal. Pada tahap testing menggunakan 1/3 dataset akan dilakukan klasifikasi berdasarkan model C4.5 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*. Pada tahap klasifikasi C4.5 setelah data terbagi menjadi data training dan data testing cara kerja algoritma C4.5 dapat menggunakan rumus persamaan 2.1 dengan sebagai contoh sample pada gambar 3.5. Penulis mengambil beberapa sample atribut yang terdapat pada tabel 3.5 dibawah ini.

Tabel 3.5 Sampel Dataset

Jumlah Tanggungan	Kepemilikan Rumah	Sumber Listrik	Luas Tanah	Luas Bangunan	Sumber Air	Prestasi	BEASISWA
2	Sendiri	PLN	50	25	Sumur	Tidak Ada	DITERIMA
4	Kontrakan	Menumpang	0	0	Sumur	Tidak Ada	DITERIMA
3	Sendiri	PLN	200	100	Sumur	Ada	DITERIMA
2	Kontrakan	Menumpang	0	0	Sumur	Ada	DITERIMA
2	Sendiri	PLN	25	25	Sumur	Tidak Ada	DITERIMA
1	Sendiri	PLN	100	50	Sumur	Tidak Ada	DITERIMA
2	Kontrakan	Menumpang	0	0	Sumur	Tidak Ada	DITERIMA
3	Kontrakan	Menumpang	0	0	Sumur	Ada	DITERIMA
1	Sendiri	PLN	220	180	Sumur	Tidak Ada	TIDAK DITERIMA
2	Sendiri	PLN	220	180	Sumur	Tidak Ada	TIDAK DITERIMA
3	Sendiri	PLN	99	50	Kemasan	Ada	DITERIMA
3	Sendiri	PLN	99	50	Kemasan	Tidak Ada	DITERIMA
3	Sendiri	PLN	220	180	Sumur	Tidak Ada	TIDAK DITERIMA

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

Hitung nilai entropy per atribut terlebih dahulu dengan rumus sama dengan di atas

1. Kepemilikan rumah

   a. Kontrakan

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

   = (-5/5· log2 (5/5)+(- 0/5· log2 (0/5))

   = 0.0000

   b. Sendiri

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

   = (-6/9· log2 (6/9)+(- 3/9· log2 (3/9))

   = 0,918295834

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. Kepemilikan rumah

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

= 0,7496 – ((5/14)*0)-((9/14)*0,9183)

= 0,159262221

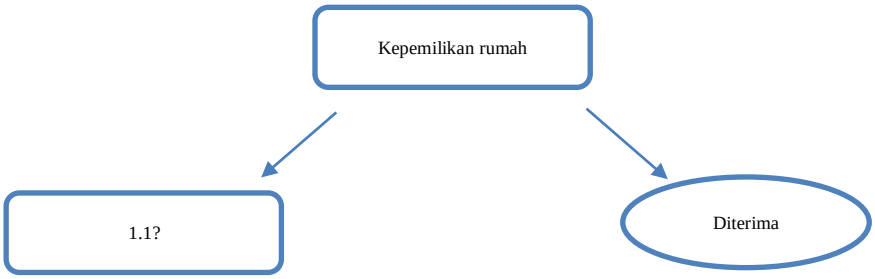
Dengan cara yang sama, akan didapatkan nilai rasio gain untuk opsi yang lain. Hasil dari beberapa perhitungan *Split Info* di sajikan pada Tabel 3.4 dibawah ini.

Tabel 3.4 Nilai Tiap Atribut

		Jumlah	**Yes**	**No**	**Entrophy**	**Gain**	**Rasio Gain**
Total		14	11	3	0,749595		
Kepemilikan Rumah							
	1. Kontrakan	5	5	0	0	0,159262	0,173432
	2. Sendiri	9	6	3	0,918296		
Sumber Listrik							
	1. Numpang	5	5	0	0	0,159262	0,173432
	2. PLN,	9	6	3	0,918296		
Luas Tanah							
	0	4	4	0	0	0,120102	0,136279
	0-1000	10	7	3	0,881291		
Luas Bangunan							
	0	4	4	0	0	0,120102	0,136279
	0-250	10	7	3	0,881291		

Sumber Air							
	1. Kemasan	2	2	0	0	0,054214	0,066825
	2. Sumur	12	9	3	0,811278		
MCK							
	1. Berbagi Pakai	5	5	0	0	0,159262	0,173432
	2. Kepemilikan Sendiri Didalam	9	6	3	0,918296		
Prestasi							
	0. Tidak Ada	12	9	3	0,811278	0,054214	0,066825
	1. Ada	2	2	0	0		

Pada Tabel 3.3 Entropy total atribut hasil adalah 0,74959 dengan jumlah sampel 14 record yaitu 11 “Diterima” dan 3 “Tidak Diterima”. Dari hasil perhitungan atribut dengan nilai information gain sebagai nilai terbesar adalah 0,173432 yaitu atribut Kepemilikan rumah. Selanjutnya Kepemilikan rumah dijadikan sebagai root node (akar). Berikut bentuk pohon keputusan root node dapat dilihat pada Gambar 3.6 dibawah ini.



Gambar 3.6 Pohon Keputusan Root Node

Pada Gambar 5. merupakan pohon keputusan bagian node akar. Atribut yang gainnya terbesar yang menjadi root adalah Kepemilikan rumah yang menghasilkan dua node yaitu pada node pertama menghasilkan Sumber Air.

1. Sumber Air
   a. Kemasan

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

= (-2/2· log2 (2/2)+(- 0/2· log2 (0/2))

= 0.0000

   b. Sumur

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

= (-9/12· log2 (9/12)+(- 3/12· log2 (3/12))

= 0,811278

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

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

2. Sumber Air

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

= 0,7496 – ((2/14)*0)-((12/14)*0,9183)

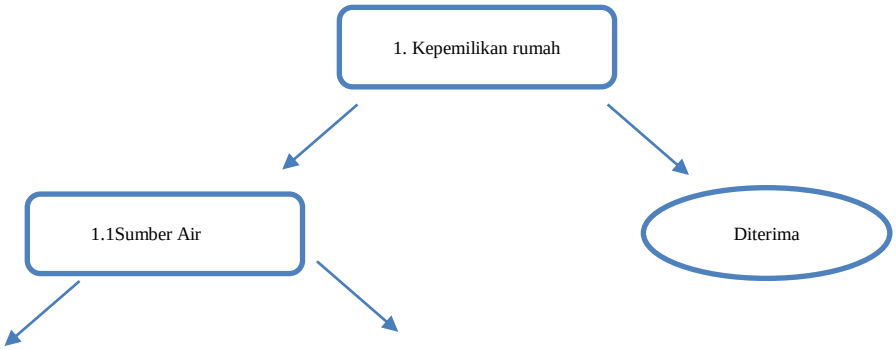
= 0,159262221

Dengan cara yang sama, akan didapatkan nilai rasio gain untuk opsi yang lain. Hasil dari beberapa perhitungan *Split Info* di sajikan pada Tabel 3.5 dibawah ini.

Tabel 3.5 Nilai Tiap Atribut

		Jumlah	Yes	No	Entrophy	Gain	Rasio Gain
Total		14	11	3	0,749595		
Sumber Air							
	1. Kemasan	2	2	0	0	0,054214	0,066825
	2. Sumur	12	9	3	0,811278		
MCK							
	1. Berbagi Pakai	5	5	0	0	0,159262	0,173432
	2. Kepemilikan Sendiri Didalam	9	6	3	0,918296		
Prestasi							
	0. Tidak Ada	12	9	3	0,811278	0,054214	0,066825
	1. Ada	2	2	0	0		

Pada Tabel 3.4 Karena Sumber air Memiliki Gain Lebih Besar Maka, Jumlah Sumber air Menjadi Node [1.1]. Pada kriteria sumber air sumur memiliki 9 orang Diterima (SUM(Total)/SUM(Yes) = 12/9=1.3, Maka Kriteria kriteria sumber air sumur Menjadi daun Diterima seperti pada gamabr 3.7 dibawah ini.



Gambar 3.7 Pohon Keputusan Root Node

3.4 Teknik Analisis, Desain, dan Pengujian

3.4.1 Teknik Analisis

Teknik analisa deskriptif dilakukan untuk menganalisa data yang akan dilakukan terhadap hasil pengumpulan data dengan studi pustaka, wawancara dan observasi untuk mendapatkan spesifikasi kebutuhan sistem yang akan dikembangkan.

Teknik analisis yang akan dilakukan menggunakan metode-metode klasifikasi *data mining* antara lain: metode *Decisioan Tree algoritma C4.5* dan *Support vector machine (SVM)*. Metode- metode tersebut digunakan untuk mengolah data mahasiswa dalam diprogram pemberian beasiswa guna menghasilkan prediksi yang akurat dalam analisis pemberian beasiswa.

Data yang didapat akan dibagi menjadi dua *set* yaitu: data *training* dan data *testing*. Hasil dari masing-masing metode dengan data latih akan dibandingkan hasil pengujiannya dengan menggunakan *k-fold cross validation* dengan k=10 untuk mendapatkan hasil berupa nilai akurasi, *precision*, *recall*, *ROC's curve*.

3.4.2 Teknik Pengujian

Teknik pengujian terhadap metode yang akan dilakukan menggunakan *k- folds cross validation* dengan k=10. Metode ini membagi data latih secara acak menjadi 10 bagian dengan jumlah yang hampir sama pada masing-masing kelompok. Pada setiap perulangan dalam proses *training*, maka 1 bagian data digunakan sebagai data uji dan 9 bagian data lainnya sebagai data latih. Proses *training* dilakukan sebanyak 10 kali data *testing*. Hasil pengujian akan didapatkan dengan menghitung rata-rata nilai-nilai statistik pengujian pada keseluruhan perulangan.