BAB II
LANDASAN TEORI

2.1. Tinjauan Studi

Tinjauan studi berguna bagi peneliti sebagai bahan referensi dari penelitian-penelitian sebelumnya. Berikut ini adalah beberapa referensi sebagai dasar pelaksanaan penelitian, antara lain:

Menurut Abdul Rohman dalam penelitiannya menerapkan Algoritma *K-Nearest Neighbor* untuk prediksi kelulusan mahasiswa menggunakan parameter usia, jenis kelamin, indeks prestasi semester satu sampai dengan empat. Data diperoleh dari *database* mahasiswa di Indonesia yang menempuh program Strata Satu (S1) tahun kelulusan 2011 sebanyak 1.633 data. Sebelum melakukan klasifikasi, peneliti melakukan pengolahan awal data menggunakan beberapa teknik, seperti: *data validation, data transformation and integration, data size reduction and dicrtization.* Setelah itu data diolah menggunakan metode *K-Nearest Neighbor* yang menghasilkan suatu model untuk kemudian dilakukan pengujian dengan *K-fold cross validation* dan dilakukan evaluasi serta validasi menggunakan *confusion matrix* dan *curve* ROC. Dengan nilai kluster *k*=5 maka diperoleh tingkat akurasi sebanyak 85,15% [5].

Ade Muchlis Maulana Anwar, Prihastuti Harsan dan Aries Maesya dalam penelitiannya yang berjudul penentuan daerah prioritas pelayanan akta kelahiran dengan metode *K*-NN dan *K*-Means. Pada penelitian ini memperoleh data sebanyak 10.000 untuk kemudian data diolah menggunakan metode *Knowledge Discovery and Data Mining* (KDD) dengan beberapa tahapan, yaitu pembersihan data untuk menghilangkan data *noise*, proses seleksi atribut untuk menghilangkan nilai yang tidak relevan dan berlebihan, proses transformasi data menjadi numerik, tahapan proses *data mining* untuk mengklasifikasikan data menggunakan *K*-NN yang kemudian hasilnya dijadikan atribut tambahan untuk proses *K*-Means dengan membagi data menjadi dua yaitu data pelatihan dan data uji, selanjutnya tahapan terakhir adalah proses evaluasi menggunakan perihutungan akurasi *confusion matrix* untuk *K*-NN dan *Index Davies Bouldin*

(IDB) untuk *K*-Means. Maka diperoleh tingkat akurasi sebanyak 74,00% untuk metode *K*-NN dan 1,179 untuk metode *K*-Means [6].

Rizki Muliono, Juanda Hakim Lubis dan Nurul Khairina dalam penelitiannya yang berjudul analisis Algoritma *K-Nearest Neighbor* dalam prediksi waktu kelulusan mahasiswa. Data diperoleh dari data mahasiswa angkatan 2015 dan dikategorikan sebagai mahasiswa yang lulus tepat waktu dan terlambat untuk kemudian dijadikan data *training*. Sedangkan data *testing* menggunakan data mahasiswa angkatan 2016 untuk kemudian dilakukan prediksi kelulusan mahasiswa. Data yang berjumlah 1530 tersebut diolah menggunakan metode *K*-NN dan menghasilkan suatu model, kemudian dilakukan pengujian dengan *K-fold cross validation* serta validasi menggunakan *confusion matrix* dan kurva ROC. Dari penelitian ini diperoleh 5 tingkat akurasi mulai dari 0,90 – 1,00 = *Excellent Classification* sampai akurasi 0,50 – 0,60 = *Failure* [7].

Berdasarkan penelitian diatas akan dilakukan penelitian menggunakan algoritma yang sama dengan data yang berbeda. Penelitian ini bertujuan untuk menentukan penerima beasiswa Peningkatan Prestasi Akademik (PPA) di Universitas Islam Nahdlatul Ulama (UNISNU) Jepara. Peneliti memilih metode ini karena relatif mudah, cepat dan akurat sehingga dapat meminimalisir kesalahan sasaran. Dalam pengujian terhadap model yang dihasilkan, peneliti memilih menggunakan *k-fold cross validation* serta evaluasi dan validasi hasil menggunakan *confusion matrix* dan kurva ROC (*Receiver Operating Characteristic*) yang akan menghasilkan nilai AUC (*Area Under Curve*).

2.2. Tinjauan Pustaka

2.2.1. Data Mining

Data Mining merupakan suatu proses menggunakan teknik kecerdasan buatan, matematika, statistik dan *machine learning* untuk mengekstraksi serta mengidentifikasi informasi dan pengetahuan yang bermanfaat dari berbagai *database* dengan jumlah yang besar [8].

Menurut Budi Santosa dan Ardian Umam (2018:1), Data Mining adalah kegiatan mengekstrak informasi atau pengetahuan (*knowledge*) penting dari suatu

dataset berukuran besar dengan menggunakan teknik tertentu. Informasi atau *kowledege* yang dihasilkan bisa dipakai untuk memperbaiki pengambilan keputusan [4].

Sedangkan istilah lain menyebutkan, bahwa Data Mining merupakan rangkaian proses untuk mendapatkan pengetahuan (*knowledge*) atau pola dari kumpulan beberapa data [9].

Dari beberapa pengertian diatas peniliti menyimpulkan bahwa Data Mining adalah proses pengumpulan data, baik berskala kecil maupun besar untuk menemukan informasi dan pengetahuan di masa mendatang menggunakan aturan tertentu. Berikut adalah gambaran besar proses data mining:

Menurut Budi Santosa dan Ardian Umam (2018:3) data mining memiliki beberapa tugas yang biasa dilakukan sebagai berikut [4]:

1. Estimasi/Regresi

   Regresi atau yang disebut juga estimasi hampir sama dengan klasifikasi, yaitu memerlukan data *training* yang sudah diberi label. Perbedaannya terletak pada keluaran, klasifikasi memiliki keluaran berupa nilai diskrit sedangkan regresi bernilai kontinyu. Contoh regresi adalah memprediksi nilai kurs Rupiah terhadap Dollar.

2. Klastering

   Pengelompokkan obyek berdasarkan kemiripan, dimana dalam satu klaster harus berisi obyek yang saling mirip dan mempunyai ketidakmiripan dengan obyek klaster lain. Berbeda dengan klasifikasi, klastering tidak memerlukan data pelatihan (*training*) yang sudah diberi label.

3. Klasifikasi

   Berbeda dengan klastering, klasifikasi memerlukan data pelatihan yang sudah diberi label. Klasifikasi adalah mengelompokkan obyek berdasarkan kelompok yang ada. Pengelompokkan dilakukan dengan membangun model

terlebih dahulu melalui proses pelatihan dengan data yang sudah disiapkan. Setelah model terbentuk dari proses tersebut, data baru bisa dikelompokkan menggunakan model tersebut.

4. Asosiasi

   Melakukan asosiasi antar obyek dalam satu set data, biasanya data transaksional. Asosiasi dilakukan dengan menghitung beberapa kali dalam suatu set data suatu transaksi yang mengandung dua item atau lebih yang berhubungan. Asosiasi sering disebut *Market Basket, Analysis*.

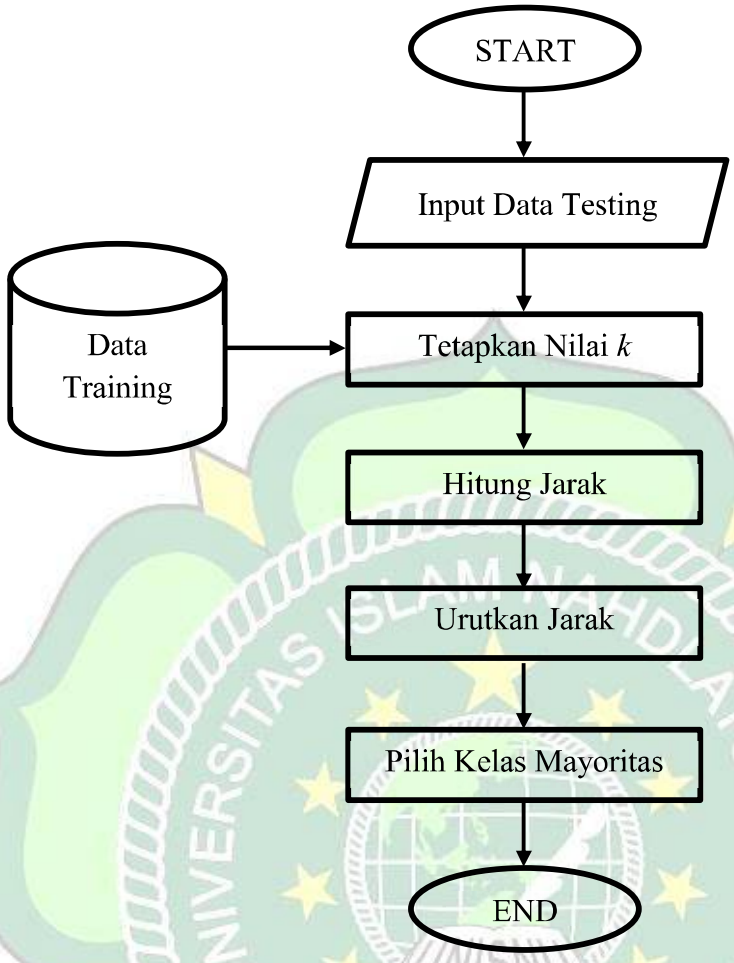
2.2.2. Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi yang mengelompokkan data baru berdasarkan jarak data baru itu ke beberapa data/tetangga terdekat [10]. Algortima *K-Nearest Neighbor* tergolong dalam algoritma *supervised learning* berfungsi untuk melakukan klasifikasi berdasarkan data *training* yang sudah terklasifikasi sebelumnya, kemudian diambil nilai *k* berdasarkan tetangga terdekat (*nearest neighbor*).

Algortima *K*-NN merupakan metode yang sederhana dan banyak digunakan. Pengklasifikasian dilakukan dengan cara menentukan nilai *k* untuk kemudian dihitung jarak terdekat antara data *training* dengan data *testing*. Untuk menghindari hasil klasifikasi memenuhi dua label atau lebih, nilai *k* harus berupa angka ganjil. Dalam pengukuran jarak peneliti memilih menggunakan *euclidean distance* yang dirumuskan pada persamaan berikut:

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \qquad (1)$$

Dimana *D*(a,b) adalah jarak skala dari dua vektor data a dan b yang berupa matrik berukuran d dimensi [11]. Lebih jelasnya dapat dilihat *pseudo code K-Nearest Neighbor* pada gambar berikut:



1. Menentukan nilai *k*.
2. Menghitung jarak antara data *training* dengan data *testing*.
3. Mengurutkan jarak.
4. Menentukan jarak terdekat hingga ke-*k*.
5. Menentukan kelas mayoritas [12].

Seperti algoritma pada umumnya, *K-Nearest Neighbor* mempunyai kelebihan dan kelemahan. Adapun kelebihan dan kelemahan pada *K-Nearest Neighbor* adalah sebagai berikut:

1. Kelebihan
   a. Tangguh pada data *training* yang banyak noisenya.
   b. Efektif untuk jumlah data *training* yang cukup besar.

2. Kelemahan
   a. Diperlukan menentukan nilai optimal parameter *k* (jumlah tetangga terdekat).
   b. Biaya komputasi yang tinggi karena perhitungan jarak harus dilakukan pada setiap *query instance* pada keseluruhan *training instance*.
   c. Mempunyai *training* berdasarkan jarak yang tidak jelas mengenai jenis jarak dan atribut yang digunakan guna memperoleh hasil yang maksimal [13].

Untuk memahamai lebih lanjut tentang algoritma *K-Nearest Neighbor*, berikut adalah contoh perhitungannya:

Tabel 2.1. Contoh Dataset

No	A1	A2	A3	A4	A5	A6	Keterangan
1	7	150	4	900	1	4000	TM
2	4	100	3	900	1	2000	M
3	9	120	3	1300	2	5000	TM
4	5	100	2	450	1	3000	M
5	8	100	4	900	2	3000	TM
6	4	150	3	450	1	2000	M
7	10	150	3	1300	2	6000	?

Pata Tabel 2.1. merupakan contoh data set yang sudah ditransformasikan dari data kategori menjadi numeric dengan ketentuan sebagai berikut:

A1 = jenis dinding (penilaian ekuivalen 1-10)
A2 = luas bangunan (parameter ekuivalen 1-3)
A3 = jumlah anggota keluarga
A4 = daya listrik (parameter ekuivalen 1-4)
A5 = penggunaan air (1. Sumur bor, 2. PAM)
A6 = pendapatan keluarga (parameter ekuivalen 1-4)
Keterangan = TM (tidak miskin), M (miskin)

Dataset pada tabel 2.1. tersebut akan diimplementasikan dengan metode *K-Nearest Neighbor*. Adapun tahapan perhitungannya adalah sebagai berikut:

1. Menetapkan nilai *k*, disini ditetapkan parameter *k*=3
2. Menghitung jarak *euclidean distance* masing-masing objek terhadap data *training*. Berikut adalah perhitungan jarak menggunakan rumus *euclidean distance* pada objek pertama:

   $=\sqrt{(7-10)^2+(150-150)^2}+(4-3)^2+(900-1300)^2+(1-2)^2+(4000-6000)^2$

   $=\sqrt{9+0+1+16000+1+4000000}$

   $=\sqrt{4016010}$

   = 2039,610502

   Hasil perhitungan *euclidean distance* pada data *training* dapat dilihat di tabel berikut ini:

Tabel 2.2. Hasil *Euclidean Distance* Dataset

Ranking	Hasil	Keterangan
2	2039,610502	TM
5	4020,265787	M
1	1000,450399	TM
4	3118,497555	M
3	3026,962999	TM
6	4089,31987	M

3. Mengurutkan jarak berdasarkan hasil perhitungan *euclidean distance* pada tabel 2.2. kemudian menentukan kategori berdasarkan tetangga terdekat (*nearest neighbor*) dengan parameter *k*=3, maka diperoleh hasil pada tabel di bawah ini:

Tabel 2.3. Urutan *Euclidean Distance* Dataset

Ranking	Hasil	Keterangan
1	1000,450399	TM
2	2039,610502	TM
3	3026,962999	TM
4	3118,497555	M
5	4020,265787	M
6	4089,31987	M

Dimana kolom yang berwarna kuning merupakan kategori tetangga terdekat hingga *k=3*.

4. Dari tabel 2.3. dapat diperoleh kelas mayoritasnya adalah Tidak Miskin (TM), jadi dapat disimpulkan data *testing* yang diuji merupakan **Tidak Miskin (TM)**.

Tabel 2.4. Penentuan Kategori Data *Testing*

No	A1	A2	A3	A4	A5	A6	ED	Ranking	KET
1	7	150	4	900	1	4000	2039,610502	2	TM
2	4	100	3	900	1	2000	4020,265787	5	M
3	9	120	3	1300	2	5000	1000,450399	1	TM
4	5	100	2	450	1	3000	3118,497555	4	M
5	8	100	4	900	2	3000	3026,962999	3	TM
6	4	150	3	450	1	2000	4089,31987	6	M
7	10	150	3	1300	2	6000	-	-	TM

2.2.3. Cross Validation

Merupakan teknik *validation* dengan cara membagi data secara *random* kedalam *k* bagian dan masing-masing bagian akan dilakukan klasifikasi [14]. *Crosss validation* merupakan sebuah metode statistik yang berfungsi untuk mengevaluasi kinerja suatu algoritma atau model. Dalam penelitian ini, peneliti

memilih menggunakan *K-fold cross validation* karena mampu mengurangi waktu komputasi tanpa mengurangi hasil akurasi estimasi. Sebagai contoh metode 3-*fold cross validation*, dengan membagi dataset secara acak menjadi 3 bagian dan dilakukan percobaan sebanyak 3 kali untuk *training* dan *testing*. Pada setiap bagian disisakan satu dataset untuk *testing* dan dataset lainnya untuk *training*. Berikut adalah pembagian data pada 3-*fold cross validation* [15]:

Tabel 2.5. Contoh Pembagian Dataset

Data Training	**Data Testing**
Dataset 2 dan Dataset 3	Dataset 1
Dataset 1 dan Dataset 3	Dataset 2
Dataset 1 dan Dataset 2	Dataset 3

Sebagai contoh terdapat 200 data, maka akan dilakukan pembagian jumlah data sebagai berikut:

Tabel 2.6. Contoh Pembagian Jumah Dataset

Jumlah Data	**Data Testing**
66	Dataset 1
67	Dataset 2
67	Dataset 3

Setelah dilakukan pengujian, maka diperoleh hasil akurasi sebagai berikut:

Tabel 2.7. Contoh Hasil Perhitungan Akurasi

Percobaan	**Jumlah Data**		**Akurasi**
	Training	**Testing**	
1	134	66	88%
2	133	67	84%
3	133	67	83%

Tingkat akurasi tersebut dihitung dengan membagi jumlah keseluruhan klasifikasi yang benar dengan jumlah semua *instance* pada data awal.

2.2.4. Confusion Matrix

Menurut Han dan Kamber (2006) *Confusion matrix* adalah sebuah matriks yang digunakan untuk mengevaluasi hasil dari suatu prediksi [14]. Evaluasi menggunakan *confusion matrix* menghasilkan nilai *Accuracy, Precision, dan Recall. Accuracy* pada klasifikasi merupakan presentase ketepatan rekaman data yang diklasifikasikan secara benar setelah dilakukan pengujian. Sedangkan *Precision* merupakan proporsi kasus yang diprediksi positif serta positif benar dalam data yang sebenarnya. Berikutnya yaitu *Recall*, merupakan proporsi kasus positif yang sebenarnya dan diprediksi positif secara benar [16]. *Confusion matrix* sangat popular digunakan dalam klasifikasi dua kelas yang diilustrasikan pada tabel berikut ini:

Tabel 2.8. *Confusion Matrix*

	***Prediction Class* 1**	***Prediction Class* 2**
Actual Class 1	*True Positive* (TP)	*False Negative* (FN)
Actual Class 2	*False Positive* (FP)	*True Negative* (TN)

Tabel diatas menjelaskan identifikasi dari suatu prediksi. TP dan TN adalah hasil klasifikasi yang benar di masing-masing kelas, sedangkan FP adalah hasil yang salah dimana seharusnya data masuk pada *class* 1 akan tetapi diidentifikasikan ke dalam *class* 2, begitupun FN yang seharusnya data masuk pada *class* 2 akan tetapi diidentifikasikan ke dalam *class* 1 [17]. Berikut adalah rumus perhitungan pada *confusion matrix*:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (2)$$

$$Precision = \frac{TP}{FP+TP}\ x\ 100\% \quad (3)$$

$$Recall = \frac{TP}{FN+TP}\ x\ 100\% \qquad (4)$$

2.2.5. Beasiswa Peningkatan Prestasi Akademik

Adalah salah satu jenis bantuan dari Pemerintah Republik Indonesia (RI) melalui Kementrian Riset, Teknologi, dan Pendidikan Tinggi (Kemenristekdikti) untuk diberikan kepada mahasiswa/i. Beasiswa tersebut diberikan secara rutin setiap tahunnya, dimana penyaluran dananya terbagi menjadi dua gelombang. Pemberian beasiswa tersebut merupakan bentuk apresiasi dan kepedulian pemerintah terhadap dunia pendidikan di Indonesia. Mekanisme pemberian beasiswa PPA dimulai dari pendaftaran hingga seleksi calon penerima. Berdasarkan data dari Kemenristekdikti yang diberikan kepada pihak perguruan tinggi terdapat skala prioritas penentu penerima beasiswa, yaitu mahasiswa/i yang memiliki Indeks Prestasi Kumulatif (IPK) yang tinggi, memiliki Satuan Kredit Semester (SKS) yang paling banyak dalam satu angkatan, memiliki prestasi ko-kurikuler maupun ekstrakurikuler serta memiliki keterbatasan kemampuan ekonomi keluarga. Tahapan penyeleksian beasiswa tersebut diserahkan sepenuhnya kepada pihak perguruan tinggi [18].

2.2.6. RapidMinner

RapidMinner merupakan sebuah perangkat lunak (*software*) aplikasi yang digunakan untuk menganalisis data mining. Dalam mengolah data menjadi sebuah pengetahuan RapidMinner menggunakan berbagai teknik deskriptif maupun prediksi untuk membuat keputusan yang tepat. RapidMinner mempunyai kurang lebih 500 operator data mining untuk menunjang pengguna dalam mengolah data, termasuk masukan, keluaran, preprocessing data dan visualisasi.

Sebelumnya RapidMinner dikenal dengan nama YALE (*Yet Another Learning Environment*) dan mulai dikembangkan di tahun 2001 oleh Simon Fischer, Ralf Klinkenberg dan Ingo Mierswa dari Unit *Artificial Intelligence* (Kecerdasan Buatan) [19]. Pada RapidMinner tersedia tampilan *Graphic User Interface* (GUI) yang digunakan untuk merancang sebuah *pipeline* analitis. GUI tersebut menghasilkan file *Extensible Markup Language* (XML) yang

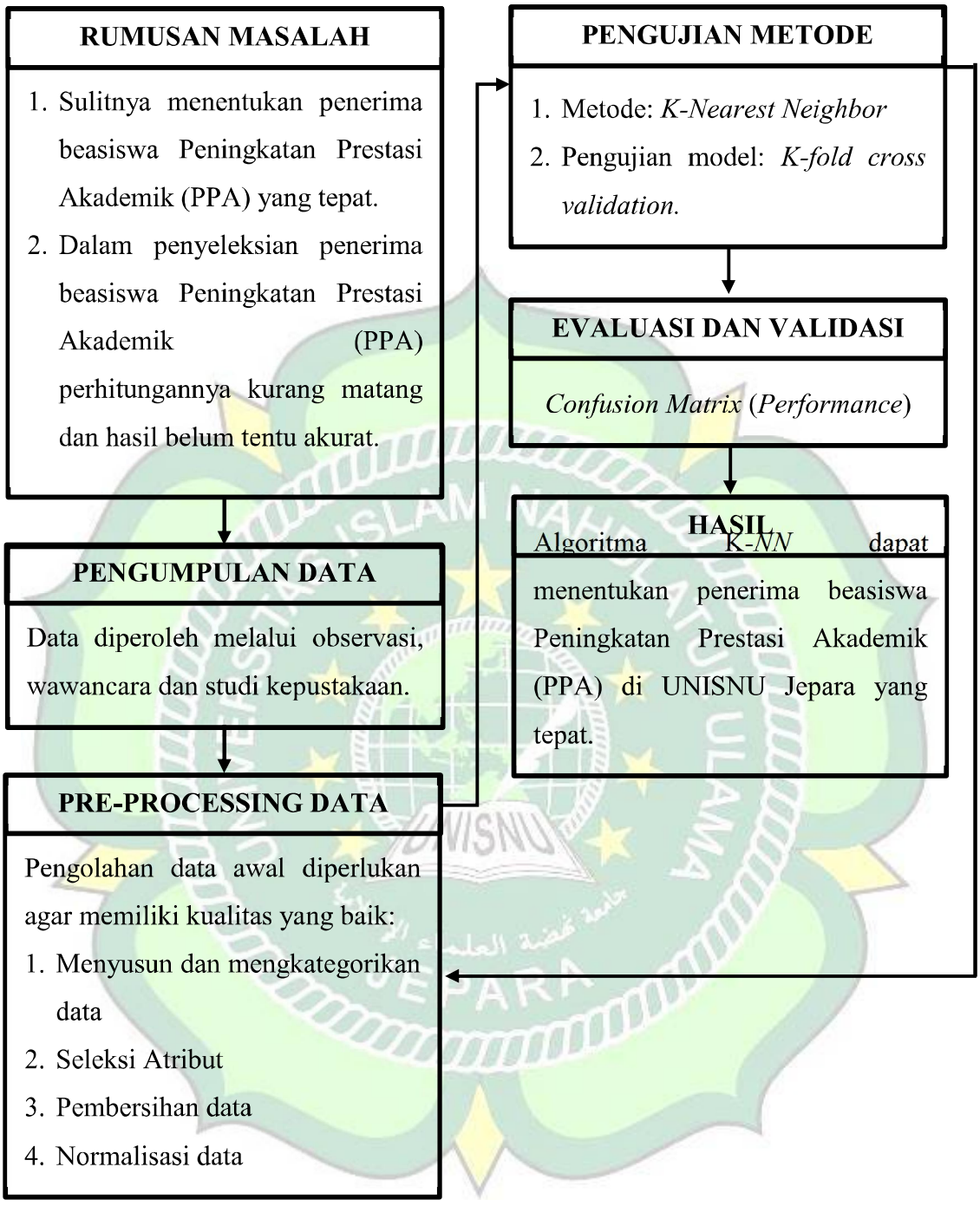
menggambarkan proses analitis keinginan pengguna untuk diaplikasikan ke data. File tersebut kemudian dibaca dan diproses oleh RapidMinner untuk menganalisa secara otomatis.

Dalam melakukan penelitian ini, peneliti menggunakan RapidMinner Studio versi 9.6. Berikut adalah perbedaan RapidMinner studio versi 9.6 dengan versi lainnya yang diterjemahkan peneliti kedalam Bahasa Indonesia:

1. Mengubah Java JRE pada RapidMinner dari oracle menjadi OpenJDK 8.
2. Dapat membuat koneksi dari ekstensi yang hanya didukung menggunakan mekanisme berbasis repositori baru.
3. Dapat menambahkan parameter *timezone* ke JDBC.
4. Dapat mendeteksi ekstraksi *fitur* untuk data *time series*.
5. Terdapat peningkatan kinerja beserta fungsi umum model ops.
6. Terdapat perbaikan untuk penyempurnaan bug [20].

2.3. Kerangka Pemikiran

Kerangka pemikiran merupakan gambaran pola pikir penelitian yang akan dilakukan. Dalam tahapan ini peneliti membuat suatu kerangka pemikiran seperti gambar berikut:


RUMUSAN MASALAH
1. Sulitnya menentukan penerima beasiswa Peningkatan Prestasi Akademik (PPA) yang tepat.
2. Dalam penyeleksian penerima beasiswa Peningkatan Prestasi Akademik (PPA) perhitungannya kurang matang dan hasil belum tentu akurat.
PENGUMPULAN DATA
Data diperoleh melalui observasi, wawancara dan studi kepustakaan.
PRE-PROCESSING DATA
Pengolahan data awal diperlukan agar memiliki kualitas yang baik:
1. Menyusun dan mengkategorikan data
2. Seleksi Atribut
3. Pembersihan data
4. Normalisasi data
PENGUJIAN METODE
1. Metode: K-Nearest Neighbor
2. Pengujian model: K-fold cross validation.
EVALUASI DAN VALIDASI
Confusion Matrix (Performance)
HASIL
Algoritma K-NN dapat menentukan penerima beasiswa Peningkatan Prestasi Akademik (PPA) di UNISNU Jepara yang tepat.