BAB 3

METODE PENELITIAN

3.1 Desain Penelitian

Metode yang digunakan dalam penelitian tugas akhir ini adalah secara eksperimen, yaitu peneliti menerapakan metode Data Mining dengan algoritma K-Means Clustering pada data mahasiswa UNISNU Jepara angkatan 2016 sampai dengan angkatan 2018 di Universitas Islam Nahdlatul Ulama Jepara untuk dianalisa dan dikelompokkan sesuai dengan persebaran wilayah dan sesuai asal sekolah berdasarkan indeks prestasi kumulatif selama dua semester awal yaitu pada semester 1 dan semester 2. Yang nantinya diuji atau diolah dengan microsoft excel dan rapidminer studio. Kemudian hasilnya akan dibandingkan berdasarkan kategori atau pengklasteran yang sudah ditentukan, yaitu tiga kategori: pertama kategori wilayah terbanyak, kedua kategori wilayah sedang, dan ketiga wilayah kategori sedikit yang nantinya dijadikan bahan pertimbangan dalam media promosi Universitas Islam Nahdlatul Ulama Jepara.

3.2 Pengumpulan Data

Untuk melakukan proses *K-Means Clustering* tentu membutuhkan sebuah data yang cukup banyak dan sesuai dengan yang dibutuhkan, di dalam penelitian ini peneliti menggunakan data mahasiswa UNISNU Jepara angkatan 2016 sampai dengan angkatan 2018 di Universitas Islam Nahdlatul Ulama Jepara. Pengambilan data di UPT Pusat Data dan Pengembangan IT Universitas Islam Nahdlatul Ulama Jepara cukup dengan melampirkan surat izin penelitian dari pihak UNISNU dan melampirkan proposal penelitian ke bagian akademik Universitas Islam Nahdlatul Ulama Jepara. Setelah mendapat balasan dari pihak Universitas Islam Nahdlatul Ulama Jepara, data bisa diambil ke bagian UPT Pusat Data dan Pengembangan IT dengan cara

mengkopi file berupa *file excel*. File yang didapatkan sejumlah 2 file yang terdiri dari file mahasiswa angkatan selama tiga angkatan terakhir, dan file data Indeks Prestasi Kumulatif (IPK) mahasiswa.

File pertama adalah file data mahasiswa Universitas Islam Nahdlatul Ulama Jepara angkatan 2016 sampai angkatan 2018. Jumlah data yang didapatkan berjumlah 3910 data mahasiswa yang terdiri dari beberapa kolom atau beberapa atribut yaitu:

1. Nomor Induk Mahasiswa (NIM)
2. Nama
3. Alamat
4. Program Studi
5. Tahun Akademik
6. Asal Sekolah
7. Jenis Sekolah
8. Jurusan Sekolah

File kedua adalah file data Indeks Prestasi Kumulatif (IPK) mahasiswa dari angkatan 2016 sampai angkatan 2018. Jumlah data mahasiswa yang didapat dalam file ini berjumlah 3910 data mahasiwa. File kedua ini berisikan beberapa kolom yaitu kolom NIM, tahun semester, dan IPK.

3.3 Lokasi Penelitian

Lokasi dari penelitian adalah Universitas Islam Nahdlatul Ulama Jepara yang disingkat UNISNU beralamat di Jalan Taman Siswa, Pekeng, Jl. Kauman, Kec. Tahunan, Kabupaten Jepara, Jawa Tengah 59451.

3.4 Pengolahan Data

Dalam melakukan penelitian ini, dibutuhkan peralatan perangkat keras (Hardware) dan perangkat lunak (Software). Untuk lebih lengkapnya di jelaskan berikut ini:

3.4.1 Perangkat Keras

Perangkat keras (hardware) yang digunakan dalam penelitian ini yaitu menggunakan 1 buah laptop yang digunakan selama penelitian. Spesifikasi laptop sebagai berikut: LENOVO YOGA 520, Intel CORE i5-7200U CPU 2.70 GHz, Ram 4gb, Hardisk Storage 1 Terabyte, Windows 10 Pro 64bit.

3.4.2 Perangkat Lunak

Perangkat lunak memiliki peran penting pada peneilitian ini karena hasil dari analisis data dapat diketahui dari pengolahan menggunakan perangkat lunak dalam mengetahui hasilnya. Pada penelitian ini, perangkat lunak yang dipakai adalah:

a) Sistem Operasi

Dalam penelitian ini sistem operasi yang dipakai adalah sistem operasi windows 10 Pro 64bit.

b) Microsoft Word

Microsoft Word dalam penelitian disini digunakan untuk menyusun laporan penelitian, *Microsoft Word* yang dipakai adalah versi 2019.

c) Microsoft Excel

Microsoft Excel dalam penelitian disini digunakan untuk mengolah data mahasiswa Universitas Islam Nahdlatul Ulama Jepara, *Microsoft Excel* yang dipakai adalah versi 2019.

3.5 Tahapan Metode

Dalam penelitian ini, peneliti menggunakan salah satu metode dalam Data Mining yaitu metode K-Means Clustering. Dalam metode ini memiliki beberapa tahapan. Tahapan yang dilakukan adalah:

3.5.1 Pre-processing Data

Tahap *pre-processing* data adalah tahap dimana data yang sudah didapatkan, dipilah, dan dipisahkan agar mendapatkan data yang dibutuhkan untuk proses selanjutnya. Tahapan ini mempunyai beberapa proses dimana setiap proses tersebut saling berhubungan satu sama lainnya. Proses dalam tahapan *pre-processing* adalah sebagai berikut:

1. Data Reduction

   Data reduction adalah proses untuk mereduksi atau mengurangi dimensi, atribut, ataupun sejumlah data yang tidak dibutuhkan dalam suatu file data. *Data reduction* sangat berguna untuk mendapatkan atribut dan sejumlah data yang akan digunakan di dalam penelitian ini.

2. Data Cleaning

   Data cleaning adalah proses dalam tahapan *preprocessing* untuk mengisi data kosong atau blank apabila memungkinkan, duplikasi data, memperbaiki data yang tidak sesuai dengan ketentuan atau salah ketik seperti kurang huruf dan kelebihan huruf, mengubah dan memodifikasi data agar data yang akan diolah adalah data yang konsisten, mengatur data yang kurang rapi dalam penulisan huruf besar dan kecil, dan mengganti format penulisan angka dan huruf sesuai dengan yang dibutuhkan.

3. Data Transformation

   Karena metode K-Means Clustering adalah metode yang bisa dilakukan apabila data yang dipakai adalah data berupa angka, maka proses transformation ini sangat dibutuhkan. Proses transformation adalah tahap untuk mengubah data atribut yang selain angka ke dalam nilai angka agar data tersebut dapat diolah menggunakan algoritma K-Means Clustering.

4. Data Integration

   Data integration adalah suatu proses untuk menggabungkan atau mengintegrasikan data dari beberapa file sumber. Data integration hanya dilakukan jika data berasal dari tempat yang berbeda-beda (sumber data tidak hanya dari 1 tempat). Langkah yang dilakukan antara lain mengintegrasikan skema, mengidentifikasi masalah entitas, dan mendeteksi sekaligus menyelesaikan konflik pada nilai data.

3.5.2 K-Means Clustering

Dari eksperimen ini di ambil sampel data Universitas Islam Nahdlatul Ulama Jepara, berikut ini adalah sampel data mahasiswa angkatan 2016 sampai angkatan 2018.

Table 3.1 Data Mahasiswa 2016 sampai 2018

No	NIM	Nama	Kecamatan Asal	Asal Sekolah	IPK
1	161240000470	Akhmad Toha	Batealit	SMK	3,66
2	161240000471	Bima Muluk Maulana Ishaq	Mlonggo	SMK	3,52
3	161240000473	Taufiq Hidayat	Pakis Aji	SMK	3,54
4	161240000474	Miftahul Huda	Mijen	SMK	2,96

5	161240000475	Muhammad Hidayatul Mustafid	Donorojo	SMK	2,62
6	161240000476	Anis Safitri	Bangsri	MA	3,48
7	161240000477	Dimas Cornellya Agatta	Jepara	SMK	3,47
8	171110002054	Emilia Inta Argadea	Jepara	SMA	3,17
9	171110002089	Heri Fajar Saputra	Batealit	SMA	2,10
10	171110002058	Fania Eka Kumala	Jepara	SMK	3,54
11	171110002187	Ade Rahmawati	Bangsri	SMK	2,08
12	171110002060	Aldo Ilham Hadzafi	Tahunan	SMK	2,22
13	171110002057	Venny Aulia Rohmah	Batealit	MA	2,71
14	171110002062	Akhmad Safii	Tahunan	MA	3,15
15	181110002277	Anwar Ramadan	Jepara	SMK	3,32
16	181110002279	Dinda Laili Savitri	Batealit	SMK	3,17
17	181110002429	Unsaa Nailul Munaa	Jepara	SMK	3,43
18	181120002245	Reyhan Ade Tirany	Jepara	SMA	3,55
19	181130001615	Rizki Nor Amalia	Pecangaan	SMA	3,51
20	181240000747	Muhammad Agung Prayogi	Jepara	SMK	3,40
21	181250000256	Siti Marhamah	Jepara	SMK	3,82

1. Transformasi Data

Transpormasi data dilakukan untuk mengubah data agar data dapat diolah dengan menggunakan metode *K-Means Clustering*. Data yang berjenis nominal seperti Kecamatan Asal dan Asal Sekolah harus dilakukan proses inisialisasi data terlebih dahulu ke dalam bentuk angka/numerikal.

Table 3.2 Inisialisasi Data Kecamatan Asal

Kecamatan Asal	Frekuensi	Inisial
Jepara	8	1
Batealit	4	2
Bangsri	2	3
Tahunan	2	4
Donorojo	1	5
Mijen	1	6
Mlonggo	1	7
Pakis Aji	1	8
Pecangaan	1	9

Table 3.3 Inisialisasi Data Asal Sekolah

Asal Sekolah	Frekuensi	Inisial
SMK	14	1
SMA	4	2
MA	3	3

2. Pengolahan Data

Setelah semua data mahasiswa ditransformasi ke dalam bentuk angka, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma *K-Means Clustering*. Untuk dapat melakukan pengelompokan data-data tersebut menjadi beberapa cluster perlu dilakukan beberapa langkah, yaitu:

1. Tentukan jumlah cluster yang diinginkan. Dalam eksperimen ini data-data yang ada akan dikelompokkan mejadi tiga cluster.
2. Tentukan titik pusat awal dari setiap cluster. Dalam eksperimen ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap cluster dapat dilihat pada tabel 3.4.

Table 3.4 Titik Pusat Awal Setiap *Cluster*

Titik Pusat	Kecamatan Asal	Asal Sekolah	IPK
Cluster 1	Mijen	SMK	2,956
Cluster 2	Bangsri	SMK	2,077
Cluster 3	Jepara	SMA	3,554

3. Tempatkan setiap data pada cluster. Dalam eksperimen ini digunakan metode hard k-means untuk mengalokasikan setiap data ke dalam suatu cluster, sehingga data akan dimasukan dalam suatu cluster yang memiliki jarak paling dekat dengan titik pusat dari setiap cluster. Untuk mengetahui cluster mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap cluster. Sebagai contoh, akan dihitung jarak dari data mahasiswa pertama ke pusat cluster pertama:

$$D\,(1,1) = \sqrt{(2-6)^2 + (1-1)^2 + (3{,}663-2{,}956)^2} = 4{,}062$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster pertama adalah 4,062. Jarak data mahasiswa pertama ke pusat cluster kedua:

$$D\,(1,2) = \sqrt{(2-3)^2 + (1-1)^2 + (3{,}663-2{,}077)^2} = 1{,}875$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster kedua adalah 1,875. Jarak data mahasiswa pertama ke pusat cluster ketiga:

$$D\,(1,3) = \sqrt{(2-1)^2 + (1-2)^2 + (3{,}663 - 3{,}554)^2} = 1{,}418$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster ketiga adalah 1,418. Berdasarkan hasil ketiga perhitungan di atas dapat disimpulkan bahwa jarak data mahasiswa pertama yang paling dekat adalah dengan *cluster* 3, sehingga data mahasiswa pertama dimasukkan ke dalam *cluster* 3. Hasil perhitungan selengkapnya untuk 21 data mahasiswa pertama dapat di lihat pada tabel 3.5.

Table 3.5 Contoh Hasil Perhitungan Setiap Data ke Setiap *Cluster*

No	NIM	Kecamatan Asal	Asal Sekolah	IPK	Jarak Ke			Jarak terdekat ke *Cluster*
					C1	C2	C3	
1	161240000470	2	1	3,66	5,51	2,19	0,88	3
2	161240000471	7	1	3,52	0,58	3,40	5,69	1
3	161240000473	8	1	3,54	0,58	4,36	6,68	1
4	161240000474	6	1	2,96	1,58	2,35	4,70	1
5	161240000475	5	1	2,62	2,63	1,45	3,75	2
6	161240000476	3	3	3,48	4,83	1,64	2,31	2
7	161240000477	1	1	3,47	6,51	3,01	0,57	3
8	171110002054	1	2	3,17	6,55	2,84	0,68	3
9	171110002089	2	2	2,10	5,70	1,91	1,47	3
10	171110002058	1	1	3,54	6,51	3,03	0,60	3
11	171110002187	3	1	2,08	4,69	1,30	2,10	2

12	171110002060	4	1	2,22	3,70	0,96	2,90	2
13	171110002057	2	3	2,71	5,81	2,16	1,81	3
14	171110002062	4	3	3,15	3,92	1,29	3,10	2
15	181110002277	1	1	3,32	6,51	2,98	0,54	3
16	181110002279	2	1	3,17	5,51	2,02	0,79	3
17	181110002429	1	1	3,43	6,51	3,00	0,56	3
18	181120002245	1	2	3,55	6,55	2,93	0,73	3
19	181130001615	9	2	3,51	1,68	5,27	7,69	1
20	181240000747	1	1	3,40	6,50	2,99	0,55	3
21	181250000256	1	1	3,82	6,52	3,12	0,76	3

4. Setelah semua data ditempat ke dalam cluster yang terdekat, kemudian hitung kembali pusat cluster yang baru berdasarkan rata-rata angggota yang ada pada cluster tersebut.
5. Setelah didapatkan titik pusat yang baru dari setiap cluster, lakukan kembali dari langkah ketiga hingga titik pusat dari setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain.

Dalam sampel data ini, iterasi clustering data mahasiswa terjadi sebanyak 5 kali iterasi. Pada iterasi ke-4 ini, titik pusat dari setiap cluster sudah tidak berubah dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain.

3.6 Evaluasi

Dalam eksperimen sampel data Universitas Islam Nahdlatul Ulama Jepara, dari data mahasiswa angkatan 2016 sampai angkatan 2018, pasti memiliki nilai error, semakin kecil nilai error yang dimiliki pada hasil

perhitungan maka semakin bagus pula hasil yang akan didapatkan. Pada penelitian ini menghitung nilai error menggunakan persamaan yang pada sebelumnya sudah dijelaskan, perhitungan nilai error terdapat pada proses berikut ini:

3.6.1 Pengujian Metode BCV dan WCV

1. Menentukan iterasi keberapa akan dihitung

Untuk menentukan iterasi diambil pada iterasi terakhir karena iterasi terakhir memiliki kualitas centeroid yang lebih baik dari sebelumnya, pada penelitian ini menggunakan sampel data perhitungan Universitas Islam Nahdlatul Ulama Jepara, dari data mahasiswa angkatan 2016 sampai angkatan 2018, untuk lebih jelasnya dapat Tabel di bawah ini.

Table 3.6 Nilai *Centroid* pada iterasi terakhir

Titik Pusat	Kecamatan Asal	Asal Sekolah	IPK
Cluster 1	7,50	1,25	3,38
Cluster 2	3,80	1,80	2,71
Cluster 3	1,33	1,42	3,28

Kemudian hitung nilai *Centroid* dengan persamaan *Between-Class Variation* (BCV).

$$\text{BCV} = \sqrt{(7{,}50-3{,}80)^2+(1{,}25-1{,}80)^2+(3{,}38-2{,}71)^2}+$$

$$= \sqrt{(7{,}50-1{,}33)^2+(1{,}25-1{,}42)^2+(3{,}38-2{,}28)^2}+$$

$$= \sqrt{(3{,}80-1{,}33)^2+(1{,}80-1{,}42)^2+(2{,}71-2{,}28)^2}$$

$$= 12{,}53$$

2. Menentukan jarak minimum centeroid

Pada proses ini menggunakan jarak minimum pusat centeroid yang didapat pada iterasi terakhir, dapat dilihat pada tabel 3.5. Setelah mendapatkan jarak minimum dengan nilai pusat centroid maka langkah selanjutnya menghitung seluruh jarak minimum dengan persamaan *Within-Class Variation* (WCV) sebagai berikut:

$$\text{WCV} = 0{,}88^2 + 0{,}58^2 + 0{,}58^2 + 1{,}58^2 + \ldots + 0{,}55^2 + 0{,}76^2$$

Sehingga hasil yang didapat adalah WCV = 25,01

3. Menghitung perbandingan BCV dengan WCV

Pada langkah terakhir adalah menghitung nilai perbandingan BCV dengan WCV sehingga menghasilkan nilai error hitung dengan persamaan Rasio seperti terlihat pada hasil dibawah ini.

$$\text{Rasio} = \frac{12{,}53}{25{,}01} = 0{,}050$$

Untuk menentukan bagus atau tidaknya hasil pengujian dari nilai rasio yang didapat maka harus memperhatikan kriteria pengukuran rasio, dapat dilihat pada tabel 2.6. Hasil pengujian menggunakan perbandingan *Between-Class Variation* (BCV) dan *Within-Class Variation* (WCV) mendapatkan nilai rasio yang tidak tinggi yaitu 0,50 dan artinya tingkat penggunaan nilai sample data *Centroid* memiliki kualitas yang baik.

3.6.2 Pengujian Metode ROC

Metode ROC digunakan untuk menghitung nilai akurasi hasil *clustering* yang telah diproses oleh sistem. Selain nilai akurasi, nilai sensifitas dan nilai spesifitas dapat dihitung juga. Adapun untuk mencari nilai *akurasi* dapat dicari dengan persamaan akurasi, untuk mencari nilai *sensifitas* dengan persamaan sensifitas, dan mencari nilai *spesifitas* dengan persamaan spesifitas. Pada penelitian ini digunakan sampel data dari hasil *clustering* data

mahasiswa angkatan 2016 sampai angkatan 2018 yang berupa data nilai *centeroid* awal dan nilai *centeroid* pada iterasi terakhir. Data tersebut ditampilkan dalam tabel 3.7 di bawah ini.

Table 3.7 Nilai *Centroid* pada iterasi terakhir

Titik Pusat	Centroid Awal	Centroid Iterasi Terakhir
Cluster 1	2,96	3,38
Cluster 2	2,08	2,71
Cluster 3	3,17	3,28

$$\text{Akurasi} = \frac{2,96+3,38}{2,96+3,38+3,17+3,28} = 0,50$$

$$\text{Sensifitas} = \frac{2,96}{2,96+3,28} = 0,47$$

$$\text{Spesifitas} = \frac{2,96}{2,96+3,17} = 0,48$$

Sesuai dengan perhitungan nilai akurasi yang didapat adalah 0,50. Nilai akurasi ini berada dalam kategori baik berdasarkan referensi pada tabel 3.8 dibawah ini.

Table 3.8 Standar Receiver Operating Characteristic (ROC)

Nilai Rasio	Kategori
0,80-1,00	Sangat baik
0,60-0,80	Baik
0,40-0,60	Cukup Baik
0,20-0,40	Kurang Baik
0,00-0,20	Tidak Baik