3. ANALISIS DAN DESAIN SISTEM

3.1 Analisis Permasalahan

Setelah dilakukan evaluasi dataset dari alumni Universitas Kristen Petra lulusan tahun 2015 sampai tahun 2020, diketahui bahwa rata-rata IPK keseluruhan dari rentang tahun tersebut berada pada angka 3,24. Jumlah lulusan Universitas Kristen Petra dari tahun 2015 sampai 2020 adalah 10.612 lulusan/alumni. Ada peningkatan rata-rata IPK dari lulusan tahun 2015 yang bernilai 3,28 jika dibandingkan dengan lulusan tahun 2020 yang bernilai 3,31. Meski begitu, peneliti merasa jika rata-rata IPK masih berpeluang untuk ditingkatkan lagi. Rata-rata IPK kelulusan merupakan salah satu aspek penting untuk mengetahui kualitas pendidikan di suatu universitas.

Perguruan tinggi swasta dapat mengurangi jumlah ataupun peluang mahasiswa putus kuliah dengan mengetahui calon mahasiswa yang berpotensi lulus dengan baik, seperti lulus dengan nilai memuaskan atau tepat waktu. Akan tetapi, ada banyak faktor yang perlu dipertimbangkan dalam menyeleksi dan menerima calon-calon mahasiswa. Faktor pertama yaitu pendidikan akhir calon mahasiswa, misalnya SMA/SMK. Nilai suatu mahasiswa tidak dapat disamakan dengan nilai mahasiswa lain yang berada pada SMA/SMK yang berbeda. Setiap SMA/SMK memiliki cara penilaian yang berbeda pada tugas atau ujian yang diberikan pada siswanya. Masing-masing SMA/SMK juga memiliki kualitas pendidikan yang berbeda dengan SMA/SMK lainnya. Selain itu, setiap SMA/SMK juga memiliki tingkat akreditasi masing-masing. Faktor kedua yaitu latar pendidikan orang tua calon mahasiswa. Latar pendidikan orang tua diperkirakan juga berdampak terhadap kemampuan calon mahasiswa untuk dapat menyelesaikan perkuliahannya dengan baik. Latar pendidikan orang tua, diperkirakan berpengaruh terhadap cara mendidik calon mahasiswa. Informasi, wawasan, ataupun saran yang diberikan dapat memiliki tingkat keluasan dan kedalaman yang berbeda. Faktor ketiga yaitu jurusan yang calon mahasiswa pilih saat di SMA/SMK. Jurusan yang dipilih juga dinilai berpengaruh terhadap potensi kelulusannya. Suatu jurusan dapat membantu calon mahasiswa untuk menyelesaikan perkuliahannya pada prodi-prodi tertentu. Hal ini karena ada beberapa materi berbeda yang diambil calon mahasiswa sewaktu mereka masih bersekolah.

3.2 Analisis Kebutuhan

Berdasarkan permasalahan yang ada, maka solusi yang ditawarkan adalah dilakukannya segmentasi pada alumni mahasiswa Universitas Kristen Petra untuk mengetahui karakteristik dari mahasiswa-mahasiswa yang berpotensi untuk menyelesaikan perkuliahannya dengan hasil yang baik. Suatu universitas akan dapat mengetahui calon-calon mahasiswa yang berpotensi untuk menjadi mahasiswa di universitas tersebut. Perguruan tinggi atau universitas akan dapat mengurangi angka putus kuliah pada mahasiswa di Indonesia dan mengurangi peluang terjadinya putus kuliah pada mahasiswa di Indonesia. Algoritma yang akan digunakan untuk segmentasi adalah *K-Prototype Clustering* karena data yang akan digunakan akan memiliki variabel-variabel numerik dan kategori. Pengujian hasil *cluster* akan dilakukan dengan menggunakan *Silhouette Score*.

3.3 *Data Preprocessing*

Tabel 3.1

Dataset Mahasiswa Awal

#	Column	Non-Null Count	Dtype
0	nim	10221 non-null	object
1	periodemasuk	10221 non-null	object
2	periodelulus	10221 non-null	object
3	kodeunit	10221 non-null	object
4	namaunit	10221 non-null	object
5	ipk	10221 non-null	object
6	kodesma	9747 non-null	object
7	namasma	9747 non-null	object
8	namakota	9746 non-null	object
9	namapropinsi	9746 non-null	object
10	keterangan	9731 non-null	object
11	pendidikanayah	9900 non-null	object
12	pendidikanibu	219 non-null	object
13	r_ing1	818 non-null	object
14	r_ing2	818 non-null	object
15	r_mat1	818 non-null	object
16	r_mat2	818 non-null	object
17	nilai_ta	9952 non-null	object
18	nilai_ept	9481 non-null	object
19	tgl_ept	7419 non-null	object
20	skkk	0 non-null	object

Pada dataset ini, dapat terlihat beberapa atribut yang banyak memiliki nilai *NULL* atau tidak ada isinya. Kolom pendidikanibu, r_ing1, r_ing2, r_mat1, r_mat2, dan skkk dihapus karena

jumlah data yang memiliki nilai *NULL* sangat banyak. Kolom nim, kodeunit, kodesma, namapropinsi, dan tgl_ept juga dihapus karena tidak diperlukan untuk proses *clustering*. Melalui tabel diatas dapat terlihat juga tidak setiap data memiliki nama SMA/SMK. Data yang tidak memiliki nama SMA/SMK dihapus sehingga terdapat sisa 9.746 data. Pada kolom ipk, nilai_ta, dan nilai_ept terlihat tipe data dari ketiga kolom tersebut masih berupa *object/string*. Ketiga kolom tersebut diubah tipe datanya menjadi *float64* yang merupakan tipe data numerik. Kolom keterangan yang merupakan jurusan yang dipilih mahasiswa saat SMA/SMK juga memiliki nilai *NULL*. Data yang kosong ini diisi dengan nilai yaitu SMA LAINNYA untuk mahasiswa yang berasal dari SMA dan nilai SMK / SMEA untuk mahasiswa yang berasal dari SMK. Kolom pendidikanayah memiliki beberapa nilai kosong dan diisi dengan nilai TAMAT SMU karena merupakan nilai terbanyak dengan jumlahnya 4.413. Kolom nilai_ta yang memiliki nilai kosong serta memiliki nilai 0 diisi dengan nilai 3,5 karena nilai ini merupakan median dan modus dari kolom nilai_ta. Rata-rata dari kolom nilai_ta 3,46 yang juga mendekati median ataupun modus. Pada kolom nilai_ept, nilai *NULL* serta nilai 0 akan diganti dengan nilai yang berkorelasi dengan kolom ipk. Hal ini karena kolom ipk dan kolom nilai_ept memiliki korelasi yang cukup dengan nilai 0,32.

3.4 Analisis Data

Tabel 3.2

Contoh Data Mahasiswa Lulusan

	namaunit	ipk	keterangan	pendidikanayah	nilai_ta	nilai_ept	akreditasi	guru	durasi	rasio	namasmakota
0	ARSITEKTUR	3.34	IPA	S1	3.0	467.0	93.0	65.0	4.0	18.661538	SMA KATOLIK KOLESE SANTO YUSUP MALANG
1	ILMU KOMUNIKASI	3.45	IPA	Tamat SD	3.5	533.0	93.0	65.0	4.5	18.661538	SMA KATOLIK KOLESE SANTO YUSUP MALANG
2	HOTEL MANAGEMENT	3.26	IPS	Tamat SMU	3.5	498.3	93.0	65.0	4.5	18.661538	SMA KATOLIK KOLESE SANTO YUSUP MALANG
3	BUSINESS ACCOUNTING	3.35	IPA	Tamat SMU	4.0	460.0	93.0	65.0	3.5	18.661538	SMA KATOLIK KOLESE SANTO YUSUP MALANG
4	TEKNIK SIPIL	3.30	IPA	S1	4.0	463.0	93.0	65.0	4.0	18.661538	SMA KATOLIK KOLESE SANTO YUSUP MALANG

Terdapat 11 kolom pada dataset mahasiswa lulusan. Kolom pertama yaitu namaunit merupakan prodi mahasiswa. Kolom ketiga yaitu keterangan merupakan jurusan yang diambil mahasiswa di SMA/SMK. Kolom akreditasi merupakan nilai akreditasi dari SMA/SMK. Kolom guru merupakan jumlah guru di SMA/SMK. Kolom durasi merupakan waktu perkuliahan yang ditempuh mahasiswa. Kolom rasio yaitu perbandingan antara jumlah guru dan jumlah siswa pada SMA/SMK. Kolom namasmakota merupakan gabungan dari nama SMA/SMK dan kota dari SMA/SMK tersebut.

Tabel 3.3

Informasi Mengenai Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9703 entries, 0 to 9702
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   namaunit        9703 non-null   object
 1   ipk             9703 non-null   float64
 2   keterangan      9703 non-null   object
 3   pendidikanayah  9703 non-null   object
 4   nilai_ta        9703 non-null   float64
 5   nilai_ept       9703 non-null   float64
 6   akreditasi      9703 non-null   float64
 7   guru            9703 non-null   float64
 8   durasi          9703 non-null   float64
 9   rasio           9703 non-null   float64
 10  namasmakota     9703 non-null   object
dtypes: float64(7), object(4)
memory usage: 834.0+ KB
```

Dataset memiliki 9.703 data mahasiswa lulusan dengan setiap baris data memiliki 11 atribut yang berhubungan dengan mahasiswa tersebut. Kolom namaunit, keterangan, pendidikanayah, dan namasmakota merupakan atribut bertipe kategorikal. Kolom ipk, nilai_ta, nilai_ept, akreditasi, guru, durasi, dan rasio merupakan atribut bertipe numerikal.

Tabel 3.4

Eksplorasi Atribut-Atribut Numerik

	ipk	nilai_ta	nilai_ept	akreditasi	guru	durasi	rasio
count	9703.000000	9703.000000	9703.000000	9703.000000	9703.000000	9703.000000	9703.000000
mean	3.220713	3.498712	500.131104	92.897661	39.978460	4.212924	19.621658
std	0.322272	0.485858	59.460750	2.826192	17.115787	0.599474	6.269542
min	2.210000	2.000000	137.000000	71.000000	1.000000	3.500000	1.189189
25%	2.980000	3.000000	467.000000	92.000000	29.000000	4.000000	16.600000
50%	3.220000	3.500000	503.000000	93.000000	40.000000	4.000000	18.661538
75%	3.470000	4.000000	537.000000	94.000000	51.000000	4.500000	21.153846
max	3.990000	4.000000	677.000000	99.000000	169.000000	8.000000	69.000000

Nilai IPK memiliki rata-rata 3,22 dari 9.703 mahasiswa lulusan, dengan nilai ipk paling rendah yaitu 2,21 dan nilai IPK paling tinggi yaitu 3,99. Nilai TA memiliki rata-rata 3,49 dengan nilai paling rendah 2 dan nilai paling tinggi 4. Nilai EPT memiliki rata-rata 500,1 dengan nilai paling rendah 137 dan nilai paling tinggi 677. Kolom akreditasi memiliki rata-rata 92,89 dengan nilai minimum 71 dan nilai maksimum 99. Nilai kuartil 1 yaitu 92, hal ini menandakan sekitar 75 persen data akreditasi memiliki nilai 92 ke atas. Jumlah guru memiliki rata-rata 39.97 dengan jumlah minimum 1 dan jumlah maksimum 169. Jumlah guru yang sangat sedikit biasanya terdapat pada PKBM dan jumlah guru yang cukup banyak biasanya terdapat pada SMA/SMK Negeri. Durasi perkuliahan mahasiswa menggunakan satuan dalam tahun dengan rata-rata 4,21 tahun serta durasi berkuliah paling cepat yaitu 3,5 tahun dan durasi berkuliah paling lama yaitu 8 tahun. Rasio guru dan siswa memiliki rata-rata 19,62 dengan rasio terendah 1,18 dan rasio tertinggi 69,00.

Terdapat 9 jenis latar belakang pendidikan ayah pada atribut pendidikan ayah, yaitu S1, Tamat SD, Tamat SMU, Tidak SD, SARJNA MUD, Tamat SMP, Diploma, S3, dan S2. Tamat SMU memiliki jumlah terbanyak yaitu 4.432, lalu diikuti dengan S1 sebanyak 3.611 dan SARJNA MUD sebanyak 537.

Terdapat 14 jurusan berbeda yang diambil oleh mahasiswa lulusan saat di SMA/SMK, yaitu IPA, IPS, Bahasa, SMK/SMEA, SMK MESIN, SMA LAINNYA, SMK INFORMATIKA, SMK PARIWISATA, SMK ELEKTRO, SMK BANGUNAN, SMK PERHOTELAN, SMK LISTRIK, SMK TEKNIK, dan SMK KEPERAWATAN. IPA merupakan jurusan terbanyak dengan jumlah mahasiswa 5.699, diikuti oleh IPS dengan jumlah mahasiswa 3.668 dan Bahasa sebanyak 121.

3.5 Data Normalization

Dataset dinormalisasi supaya setiap kolom memiliki nilai dalam skala yang sama untuk memudahkan proses *clustering*. Kolom yang akan dinormalisasi yaitu kolom-kolom dengan tipe data numerik. Normalisasi akan dilakukan dengan metode *min-max* dimana nilai dari atribut-atribut numerik akan diubah ke range 0 sampai 1. Metode untuk *clustering* akan dilakukan dua tahap. Pada tahap pertama, clustering akan difokuskan untuk atribut-atribut saat mahasiswa masih SMA/SMK. Di tahap pertama ini, atribut-atribut tersebut akan diberi bobot yang berbeda. Kolom akreditasi akan diberi bobot 6 kali sehingga range nilainya menjadi 0 sampai 6. Kolom guru akan diberi bobot 4 kali dan kolom rasio akan diberi bobot 2 kali. Pada tahap kedua, *clustering* akan difokuskan untuk atribut-atribut saat mahasiswa pada tahap perkuliahan. Kolom

ipk akan diberi bobot 4 kali, kolom durasi dan nilai_ta akan diberi bobot 3 kali, dan kolom nilai_ept akan diberi bobot 2 kali.

3.6 Pembuatan Model

Metode yang akan digunakan dalam *clustering* atau segmentasi pada skripsi ini adalah *K-Prototype Clustering*. Metode ini merupakan kombinasi dari algoritma *K-Means Clustering* dan algoritma *K-Modes Clustering*. *K-Means Clustering* adalah salah satu metode yang biasa digunakan untuk melakukan *clustering* atau segmentasi. Algoritma *K-Means Clustering* hanya dapat bekerja dengan fitur atau variabel numerikal saja. *K-Modes Clustering* merupakan algoritma untuk *clustering* atau segmentasi yang berdasar pada *K-Means Clustering* dengan beberapa perubahan atau modifikasi. Perubahan inilah yang membuat algoritma *K-Modes Clustering* dapat bekerja dengan fitur-fitur atau variabel yang bersifat kategorikal. Dalam skripsi ini data yang akan digunakan memiliki fitur-fitur numerikal dan juga kategorikal sehingga diperlukan metode yang dapat bekerja dengan kedua jenis fitur sekaligus. *K-Prototype Clustering* menggunakan titik pusat hibrida "Prototype" yang merupakan kombinasi dari titik pusat *Means* pada *K-Means Clustering* dan titik pusat *Modes* pada *K-Modes Clustering*. Algoritma *K-Prototype Clustering* menggunakan perhitungan *Cost Function* dan *Dissimilarity Coefficient* yang akan dihitung secara terpisah baru setelah itu digabungkan. Perhitungan untuk bagian kategorikal akan menggunakan *Hamming Distance* dan perhitungan untuk bagian numerikal akan menggunakan *Euclidean Distance*.

3.7 Pengujian Model

Pengujian performa model dilakukan dengan menggunakan *Silhouette Score*. *Silhouette Score* adalah metrik untuk mengevaluasi performa dari *cluster* yang dihasilkan oleh algoritma *K-Prototype Clustering*. Metode yang akan digunakan pada pengujian yaitu *K-Protoype Clustering*. Pengujian dilakukan pada dua cakupan berbeda, yaitu pada satu universitas dan beberapa jurusan atau prodi. Hasil dari tiap-tiap segmentasi akan dianalisis lebih lanjut.