BAB I

PENDAHULUAN

1.1 Latar Belakang

Program studi merupakan ujung tombak dari sebuah perguruan tinggi dalam menghasilkan lulusan yang berkualitas karena erat hubungannya dengan penyelenggaraan pendidikan tinggi (Ary & Sanjaya, 2020). Dalam memilih sebuah program studi, calon mahasiswa akan mempertimbangkan citra program studi tersebut, seperti tingkat akreditasinya, fasilitas pendidikan yang ditawarkan, serta prospek dari program studi tersebut terhadap keterbukaan lapangan pekerjaan (Mahmudah & Faidah, 2020). Sebagai lembaga penyelenggara kegiatan pendidikan tinggi, universitas dituntut untuk dapat mengkaji perilaku calon mahasiswa dalam memilih program studi yang ditawarkannya, baik itu dari aspek kebudayaan, sosial, maupun kepribadian calon mahasiswa yang bersangkutan (Pinaraswati, 2020). Dengan kajian ini, universitas dapat memastikan bahwa calon mahasiswa sudah memilih program studi yang tepat sesuai dengan minat dan kemampuannya, sehingga calon mahasiswa tersebut dapat mengikuti proses perkuliahan dengan baik (Renatalia et al., 2020). Sebagai salah satu bagian dari Universitas Islam Sumatera Utara (UISU), fakultas ekonomi UISU menawarkan beberapa program studi yang dapat dipilih oleh calon mahasiswa seperti manajemen, akuntansi, ekonomi pembangunan, kewirausahaan dan magister manajemen.Permasalahan yang dialami fakultas ekonomi UISU adalah seringnya terjadi mahasiswa yang salah memilih jurusan, sehinggamempengaruhi proses perkuliahanserta tingkat kelulusan mahasiswa pada masing-masing program

Studi di fakultas tersebut.Kesalahan pemilihan jurusan ini dapat di atasi dengan cara memberikan rekomendasi jurusan yang paling sesuai untuk calon mahasiswa yang mendaftar, sehingga mahasiswa yang mendaftar dapat menyelesaikan masa kuliahnya dengan tepat waktu dan sesuai dengan bakat serta kemampuannya. Berdasarkan permasalahan ini, penulis tertarik untuk membangun sebuah model yang dapatmembantu fakultas ekonomi UISU dalam memberikan rekomendasipilihan program studiyang paling sesuai bagi calon mahasiwa yang mendaftar pada fakultas ekonomi di bawah naungannya, dengan memanfaatkan teknik *data mining*.

Data mining dapat digunakan untuk mengekstraksi informasi dari dalam data yang berjumlah yang besar (*big data*), dan kemudian mengeksplorasi data tersebut guna memperoleh pola-pola yang belum diketahui sebelumnya sehingga menghasilkan sebuah model prediksi (Alkhairi & Situmorang, 2022). Dengan menggunakan sebuah model prediksi,dapat dilakukan peramalan nilai variabel target berdasarkan hasil dari proses pelatihan menggunakan data yang sudah ada (Rumahorbo & Sekarwati, 2020). Di dalam *data mining*, dikenal dua teknik pembelajaran, yaitu *unsupervised learning* yang tidak membutuhkan adanya label target dan *supervised learning* yang membutuhkan adanya label target pada proses pembelajarannya (Chahal & Gulia, 2019).*Clustering* merupakan bagian dari *unsupervised learning*, yang dapat melakukan prediksi pada nilai variabel target data, tanpa adanya pemberian label kepada kelas target yang diinginkan(Suliman, 2021); sedangkan klasifikasi merupakan bagian dari *supervised learning*, dimana pengkategorian data sangat membutuhkan variabel-variabel target dalam

prosesnya (Wahyono et al., 2020).Dengan mengkombinasikan *clustering* dan klasifikasi, proses klasifikasi pada sekumpulan data berjumlah besar (*big data*) dapat dipersingkat waktunya dan lebih tahan terhadap *noise* yang ada di dalam kumpulan data tersebut (Ullah et al., 2019).Pada penelitian ini, penulis mengkombinasikan teknik *clustering* dan klasifikasidalam menghasilkan model yang dapat mengelompokkan data pilihan program studi sebelumnya, dan kemudian mentransfer hasilnya sebagai dataset klasifikasi untuk memecahkan masalah pemilihan program studi yang dialami fakultas Ekonomi UISU.

Dalam *machine learning, K-means clustering* merupakan algoritma *clustering* yang termasuk ke dalam golongan *unsupervised learning*, yang melakukan klasifikasi maupun prediksi berdasarkan data yang belum memiliki label (Halbouni et al., 2022). *K-means clustering* merupakan algoritma yangberbasis pada pengelompokandata berdasarkan nilai rata-rata (*means*) dari data terdekat berdasarkan jumlah *cluster* yang sudah ditentukan sebelumnya(Sagala et al., 2021). *K-means clustering* penulis pilih sebagai algoritma *clustering* yang digunakan di dalam penelitian ini karena dapat mengelompokkan data berdasarkan tingkat kemiripannya, tanpa dibutuhkan adanya pelabelan target pada data tersebut (Sitompul et al., 2022);selain itu, algoritma ini tergolong sederhana dalam implementasinya namun tetap memiliki tingkat efektifitas yang cukup tingg (Hartono et al., 2018). Dalam algoritma ini dengan nilai K yang dipilih mempengaruhi jumlah partisi dari kelompok data yang dipisahkan, sehingga akan mempengaruhi pula pada kinerjanya (Awad & Hamad, 2022). Beberapa penelitian sudah membuktikan bagaimana nilai K

mempengaruhi kinerja dari algoritma ini, dimana dengan nilai K yang berbeda akan menghasilkan kinerja yang berbeda pula, seperti pada pengelompokan dokumen skripsi dengan nilai K terbaik adalah 2 (Adhe et al., 2020), pengelompokan data perawatan kesehatan dengan nilai K terbaik adalah 5 (Ambigavathi & Sridharan, 2020), pengelompokan data penggunaan media *smartphone* dengan nilai K terbaik adalah 3 (Awad & Hamad, 2022), dan pengelompokan respon masyarakat terhadap kampanye pemilu dengan nilai K terbaik adalah 7 (Cahyanto et al., 2020). Dalam penelitian ini, K-Means digunakan untuk mengelompokkan data-data pilihan program studi Fakultas Ekonomi UISU pada periode-periode sebelumnya, sehingga menghasilkan cluster-cluster yang bersesuaian dengan jumlah program studi pada fakultas tersebut.Hasil *clustering* ini digunakan sebagai target pada proses klasifikasi menggunakan algoritma *K-nearest neighbors*.

Algoritma*k-nearest neighbors* (K-NN) merupakan algoritma klasifikasi jenis *supervised learning* yang mengklasifikasikan data dengan cara mengelompokkan data latih ke dalam kelompok-kelompok yang memiliki jarak terdekat dengan kelas target (Mardhyath et al., 2020). Tujuan algoritma k-NN adalah mengkategorikan objek baru berdasarkan karakteristik dan sampel pelatihan, dengan hasil sampel uji diklasifikasikan berdasarkan mayoritas kategori kemiripan data yang satu dengan data yang lain berdasarkan nilai K yang sudah ditentukan (Gunawan et al., 2021). K-NN penulis pilih sebagai algoritma klasifikasi dalam penelitian ini karena mudah diimplementasikan serta sering digunakan sebagai *benchmark* dalam komparasi terhadap algoritma lain karena

kinerjanya yang cukup baik, seperti pada klasifikasi bunga Iris dengan akurasi tertinggi sebesar 98%(Lubis et al., 2020) dan pada prediksi penerimaan pegawai dengan akurasi sebesar 81.8% (Sinaga et al., 2022).

Pada algoritma ini, nilai K (jumlah tetangga terdekat) mempengaruhi hasil akhir klasifikasi, khususnya dalam hal akurasinya (Yuan & Yang, 2019).Beberapa penelitian sudah membuktikan hal ini, dimana dengan memilih jumlah tetangga terdekat yang berbeda, diperoleh kinerja yang berbeda pula baik itu dari segi akurasi, *purity*, tingkat kesalahan (*error*) maupun jumlah iterasi dari proses klasifikasi, seperti pada klasifikasi penentuan karir siswa SMK dengan K = 3 (Nunsina et al., 2020), klasifikasi spesies bunga Iris dengan K = 1 (Rao et al., 2022), klasifikasi pemerataan tenaga pengajar dengan K = 9 (Umargono et al., 2019), dan klasifikasi pengguna narkoba dengan K = 6.Selain jumlah tetangga terdekat, metode untuk menghitung jarak terdekat antara *cluster* juga dapat mempengaruhi kinerja algoritma K-NN, seperti yang ditunjukkan oleh perubahan akurasi yang dihasilkan algoritma ini pada proses klasifikasi dari beberapa penelitian yang menggunakan metode pengukuran jarak *euclidean distance* dan *manhattan distance*, seperti pada klasifikasi 5 dataset UCL *machine learning* dimana *euclidean distance* lebih baik daripada *manhattan distance* (Ghazal et al., 2021), klasifikasi kredit macet dimana *euclidean distance* lebih baik daripada *manhattan distance* (Margolang et al., 2022),klasifikasi data tekstual dimana *euclidean distance* lebih baik daripada *manhattan distance*(Wahyono et al., 2020),klasifikasi 28 dataset UCL *machine learning* dimana *manhattan distance* lebih baik daripada *euclidean distance* (Abu Alfeilat et al., 2019), dan klasifikasi gambar radiologi dimana

manhattan distance lebih baik daripada *euclidean distance* (Arora et al., 2022). *Eculidean distance* merupakan metode pengukuran jarak antar data yang populer digunakan, karena kemampuannya untuk mengukur titik (*point*) antar data pada ruang (*space*) yang sama; metode ini memiliki kelemahan dimana penggunaan akar kuadrat pada formulanya membutuhkan waktu proses yang lebih lama saat menangani data berjumlah besar (M. K. Hossain & Abufardeh, 2019). *Manhattan distance* mengukur jarak antar data dengan memperhatikan total perbedaan jarak seluruh atribut pada data-data tersebut, sehingga dapat memproses data pada ruang selain dua dimensi (Suwanda et al., 2020); metode ini memiliki kelemahan dalam hal memproses data yang bersifat polynomial, karena hanya menggunakan nilai absolut tiap emelen data di dalam sebuah vector (Selvida et al., 2020). Berdasarkan referensi dari penelitian-penelitian ini, penulis tertarik untuk mengkombinasikan variasi nilai K pada algoritma K-NN dengan metode pengukuran jarak *euclidean* dan *manhattan distance* pada model-model yang akan dibangun.

Pada penelitian ini, algoritma *k-means* digunakan untuk menghasilkan kelompok program studi pilihan mahasiswa berdasarkan dataset periode penerimaan mahasiswa sebelumnya. Pada pengelompokan ini, digunakan nilai K = 3 pada algoritma *clustering,* sesuai dengan jumlah program studi pilihan pada dataset yang digunakan. Selanjutnya, hasil dari pengelompokan program studi tersebut digunakan sebagai target untuk mengklasifikasikan data mahasiswa sehingga dapat digunakan sebagai rekomendasi program studi yang paling sesuai bagi calon mahasiswa baru berikutnya. Klasifikasi program studi pilihan

menggunakan algoritma *k-nearest neighbors* pada penelitian ini menerapkan pengukuran jarak *euclidean* dan *manhattan,* untuk kemudian dibandingkan berdasarkan nilai akurasi, presisi, dan *recall* yang dihasilkan.

Berdasarkan permasalahan penelitian di atas, penulis bertujuan untuk menghasilkan satu model pemilihan program studi yang dapat diterapkan pada Fakultas Ekonomi UISU dalam membantu calon mahasiswa baru untuk memilih program studi yang paling sesuai bagi mereka. Dari hasil pengumpulan data berupa teori dan hasil-hasil penilitian sebelumnya,proses dan hasil penelitian ini kemudian penulis tuangkan dalam bentuk proposal tesis yang berjudul "**Analisa Algoritma K-Means Clustering Dan K-Nearest Neighbors Dalam Pemilihan Program Studi Di Fakultas Ekonomi UISU**".

1.2 Rumusan Masalah

Adapun rumusan masalah pada penelitian ini, berdasarkan latar belakang masalah di atas, adalah:

1. Bagaimana membangun model prediksi pilihan program studi di fakultas ekonomi UISU dengan kombinasi algoritma *K-Means Clustering* dan *K-Nearest Neighbors*?
2. Bagaimana hasil pengelompokan data dalam penentuan kelas target program studi pilihan yang dibangun menggunakan algoritma *k-means clustering*, dengan menggunakan nilai K = 3?

3. Bagaimana pengaruh pengukuran jarak *euclidean distance,* dan *manhattan distance,* dan variasi nilai Kdalam model yang dibangun menggunakan algoritma *k-nearest neighbors*?
4. Bagaimana kinerja model pemilihan program studi yang dihasilkan dari kombinasi algoritma *k-means clustering* dan *k-nearest neighbors* yang diterapkan pada masalah pemilihan program studi pada fakultas ekonomi UISU?

1.3 Tujuan

Adapun tujuan dari penelitian berdasarkan latar belakang masalah di atas adalah sebagai berikut:

1. Menganalisa kinerja algoritma *k-means clustering*, dalam mengelompokkan data menjadi kelas target program studi pilihan untuk proses klasifikasi selanjutnya
2. Menganalisa kinerja algoritma *k-nearest neighbors*, dengan menggunakan pengukuran jarak *euclidean distance,* dan *manhattan distance* serta variasi nilai K.
3. Menganalisa kinerja model pemilihan program studi yang dibangun dengan kombinasi algoritma *k-means clustering* dan *k-nearest neighbors* dalam masalah pemilihan program studi pada fakultas ekonomi UISU.

1.4 Manfaat

Adapun manfaat yang diharapkan dari penelitian berdasarkan tujuan penelitian di atas adalah sebagai berikut:

1. Memperoleh hasil analisa berupa penerapan *clustering* menggunakan algoritma *K-means clustering,* dalam mengelompokkan data berdasarkan jumlah program studi pada Fakultas Ekonomi UISU.
2. Memperoleh hasil analisa pengaruh metode pengukuran jarak *euclidean distance,manhattan distance,* serta variasi nilai Kpada kinerja algoritma *k-nearest neighbors* dalam model pemilihan program studi yang dibangun.
3. Menghasilkan model pemilihan program studi pada fakultas ekonomi UISU dengan hasil analisis kinerjanya dalam bentuk tingkat akurasi, presisi, dan *recall*.

1.5 Batasan Masalah

Agar penelitian ini tidak melebar terlalu jauh pembahasannya, ada beberapa batasan masalah yang penulis berikan dalam penelitian ini. Adapun batasan masalah tersebut antara lain:

1. Dataset yang digunakan dalam penelitian dibatasi pada data penerimaan mahasiswa baru pada fakultas ekonomi UISU yang diperolehdari periode 2020 sampai 2022.
2. Indikator klasifikasi yang digunakan dibatasi pada rentang nilai usia calon mahasiswa, jenis kelamin, asal sekolah, asal daerah, dan program studi yang diminati.

3. Nilai K yang digunakan pada algoritma *k-means clustering* adalah 3 (tiga), sesuai dengan program studi pilihan yang ditemukan pada dataset yang digunakan.
4. Jumlah K pada algoritma *k-nearest neighbors*dibatasi pada variasi 3, 5, 7, dan 9.
5. Metode pengukuran jarak pada algoritma *k-nearest neighbors* yang digunakan adalah *euclidean distance* dan *manhattan distance*.
6. Output yang digunakan dalam prediksi pemilihan program studi ini adalah kategori program studi yang ditawarkan oleh fakultas ekonomi UISU untuk jenjang S1, seperti manajemen, akuntansi, dan ekonomi pembangunan.
7. Analisa yang dilakukan untuk mengevaluasi kinerja model adalah metode *10-fold* dan *20-fold cross validation* untuk menghasilkan nilai akurasi, presisi, dan *recall*.

1.6 Sistematika Pembahasan

Adapun sistematika pembahasan yang digunakan dalam penelitian ini adalah sebagai berikut:

BAB I : PENDAHULUAN

Bab ini menjelaskan tentang Latar Belakang, Rumusan Masalah, Tujuan, Manfaat serta Batasan Masalah penelitian.

BAB II : TINJAUAN PUSTAKA

Bab ini berisikan tentang teori-teori yang berkaitan langsung dengan penelitian seperti teori mengenai program studi,*data mining*,

clustering, *K-neareast neighbors*, dan *K-means clustering* dari literaturyang diperoleh melalui buku-buku, jurnal dan penelitian-penelitian sebelumnya sebagai rujukan dalam penelitian ini.

BAB III : METODOLOGI PENELITIAN

Bab ini membahas tentang bentuk penerapan algoritma *k-means clustering* dan *k-nearest neighbors*ke dalam sebuah model prediksi pemilihan program studi pada fakultas ekonomi UISUdan bentuk analisa yang digunakan untuk membandingkan hasil prediksimodel tersebut dengan data penerimaan mahasiswa baru sebenarnya pada periode sebelumnya