BAB I
PENDAHULUAN

1.1. Latar Belakang

SMA Muhammadiyah 3 Jember merupakan sekolah swasta yang telah berdiri kurang lebih selama 38 tahun di jember. Selanjutnya, dalam rangka mematuhi program kebijakan Pemerintah, pada tahun 2005 SMA Muhammadiyah 3 Jember melaksanakan Akreditasi dengan 3 program/jurusan sehingga berubah status menjadi "Terakreditasi A". 3 program itu antara lain "IPA", "IPS" dan "Bahasa" (smamuh3jbr.sch.id, 2021). Dalam penentuan jurusan pada siswa, SMA Muhammadiyah menggunakan rekapitulasi nilai ujian nasional bahasa Indonesia, nilai ujian nasional bahasa inggris, nilai ujian nasional IPA, nilai ujian nasional matematika, ulangan harian IPA, Matematika, IPS, Bahasa Indonesia, Bahasa inggris, nilai tes verbal linguistik, logis matematis, spasial, kinestik, musikal, interpersonal, intrapersonal dan natural. Dalam menentukan jurusan siswa sekolah masih menggunakan hitung manual menggunakan *microsoft excel* serta memilah secara manual data siswa yang masih kosong. Kendala yang dihadapi dalam menentukan jurusan antara lain, jika terdapat nilai tugas atau nilai ulangan harian yang kosong pada mata pelajaran tertentu yang belum tentu siswa tersebut tidak berminat pada mata pelajaran tersebut. Nilai kosong pada atribut tertentu ini yang menjadi permasalahan pihak sekolah dalam mengklasifikasi jurusan siswa dan selanjutnya menjadi salah satu pembahasan dalam penelitian ini. Kekosongan nilai atau *Missing Value* ini akan diolah sehingga penentuan jurusan pada siswa yang memiliki nilai kosong pada atribut tertentu dapat diklasifikasi tanpa perlu proses lanjutan seperti melakukan tes ulang dan lain-lain. Dari data penjurusan siswa ini, penulis selaku peneliti ingin melakukan analisis menggunakan teknik *Data Mining*. Sesuai konsep *Data Mining*, data-data terdahulu yang akan diekstrak dan dimodelkan agar dapat digunakan untuk memproses data yang akan datang dalam hal ini mengklasifikasi data penjurusan siswa SMA Muhammadiyah 3 Jember. Dengan menggunakan *Data Mining* pola pada data siswa akan dapat dimodelkan dengan mudah. Teknik klasifikasi menggunakan *Data Mining* dapat menentukan penjurusan siswa dengan mudah dan lebih cepat.

Data Mining adalah konsep yang digunakan untuk menemukan pengetahuan yang tersembunyi dalam *database*. *Data Mining* adalah proses semi-otomatis yang menggunakan statistik, matematika, kecerdasan buatan, dan teknik pembelajaran mesin untuk mengekstrak dan mengidentifikasi informasi intelektual yang berpotensi berguna yang disimpan dalam *database*. (Turban et al., 2005).

Data Mining adalah bagian dari proses *Knowledge Discovery (KDD)* dalam *database* dan mencakup beberapa langkah seperti pemilihan data, preprocessing, transformasi, *Data Mining*, dan evaluasi hasil (Maimon dan Last, 2000)). *KDD* juga biasa disebut sebagai *database*. penggunaan *Data Mining* dapat dibagi menjadi dua kelompok yaitu *discovery* dan verifikasi. Metode verifikasi sering kali mencakup metode statistik seperti penyesuaian dan *ANOVA*. Sedangkan *discovery* dapat dibagi menjadi model prediktif dan model deskriptif. Metode prediksi menggunakan hasil yang diketahui dari berbagai data untuk memprediksi data. Model ini dapat dibuat berdasarkan penggunaan data historis lainnya. Pemodelan deskriptif, di sisi lain, bertujuan untuk mengidentifikasi pola atau hubungan antara data dan menyediakan cara untuk menemukan karakteristik data yang sedang diselidiki (Dunham, 2003).

Metode klasifikasi pada *Data Mining* bermacam-macam, diantaranya adalah *Gaussian Naive Bayes* dan *K Nearest Neighbor*. *Gaussian Naive Bayes* merupakan salah satu varian dari *Naive Bayes*, di mana metode ini bekerja melalui probabilitas. Pada metode ini data yang digunakan pada fitur yang digunakan adalah tipe angka. Metode ini bekerja dengan menggunakan nilai *mean* dan standar deviasi pada setiap fitur yang digunakan (Jason, 2016). Salah satu metode klasifikasi adalah *K Nearest Neighbor*. Metode ini bekerja dengan mengukur jarak terdekat yang dihasilkan dari vektor. Metode ini juga bekerja pada tipe data angka (Jason, 2016). Dikarenakan kedua metode ini memiliki karakteristik yang sama dari sudut tipe data yang digunakan, maka penulis mencoba membandingkan kedua metode tersebut dalam klasifikasi penjurusan siswa.

Penggunaan *Data Mining* pada dunia pendidikan sudah banyak dilakukan penelitiannya, diantaranya yaitu pada penelitian dengan judul "Implementasi *Data Mining* Dalam Pengelompokan Jurusan yang Diminati Siswa SMK Negeri 1 Lolowa'u menggunakan Metode *Clustering*" (Nduru & Limbong, 2018). Pada

penelitian ini teknik *Data Mining* yang digunakan adalah metode *clustering*. Teknik *Data Mining* memiliki beberapa metode yang dapat digunakan antara lain klasifikasi, pengelompokan dan asosiasi. Teknik klasifikasi banyak digunakan dalam implementasi *Data Mining* diantaranya *Naive Bayes*, *K Nearest Neighbor* dan *Decision Tree*. Beberapa penelitian melakukan penelitian terhadap data pendidikan atau siswa menggunakan metode-ini. Misalnya pada penelitian dengan judul "Implementasi *Naïve Bayessian* dengan *Laplacian Smoothing* untuk Peminatan dan Lintas Minat Siswa SMAN 5 Pamekasan", pada penelitian ini menggunakan teknik *laplacian smoothing* untuk menghindari hasil akhir bernilai 0. Nilai akurasi yang diperoleh metode *Naive Bayes* adalah 92.11% dengan menggunakan pengujian 5 kali (Listiowarni, 2019). Pada penelitian dengan judul "Pemilihan Jurusan Dengan Metode *K-Nearest Neighbor* Untuk Calon Siswa Baru" diperoleh hasil penelitian yaitu tingkat akurasi yang dihasilkan dari metode *K Nearest Neighbor* adalah 92,8% (Saepudin & Muslih, 2019). Berdasarkan fakta-fakta di atas, penulis dalam penelitian ini akan menggunakan perbandingan dua metode yaitu *Naive Bayes* dan *K Nearest Neighbor* untuk menemukan metode terbaik dalam melakukan klasifikasi pemilihan jurusan pada siswa SMA Muhammadiyah 3 Jember.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, berikut poin yang dijadikan rumusan masalah pada penelitian adalah berapakah nilai akurasi, presisi dan *recall* yang dihasilkan metode *Gaussian Naive Bayes* dan *K Nearest Neighbor* pada klasifikasi penjurusan data siswa SMA Muhammadiyah 3 Jember?

1.3. Tujuan

Berdasarkan rumusan masalah yang telah disusun di atas, berikut tujuan yang ingin dicapai pada penelitian ini adalah mengetahui nilai akurasi yang dihasilkan metode *Gaussian Naive Bayes* dan *K Nearest Neighbor* pada klasifikasi penjurusan data siswa SMA Muhammadiyah 3 Jember.

1.4. Manfaat

Penulis selaku peneliti berharap penelitian ini bermanfaat bagi banyak pihak, berikut manfaat penelitian ini:

1. Bagi penulis, penelitian ini bermanfaat dalam pengembangan *Data Mining*.
2. Bagi pembaca, penelitian ini berguna untuk pengembangan penelitian atau sebagai landasan untuk melakukan penelitian serupa.
3. Bagi pihak sekolah, penelitian ini berguna sebagai bahan atau landasan untuk melakukan otomatisasi dalam melakukan penjurusan siswa berdasarkan data.

1.5. Batasan Masalah

Peneliti sadar bahwa lingkup penelitian ini memiliki batasan agar penelitian ini dapat berfokus terhadap titik masalah. Berikut batasan-batasan masalah yang digunakan pada penelitian ini:

1. Data yang digunakan adalah siswa SMA Muhammadiyah 3 Jember.
2. Data yang digunakan adalah data siswa pada proses penjurusan pada tahun ajaran 2018-2019 dengan total data 343 siswa.
3. Vektor yang digunakan untuk mengukur jarak pada *K Nearest Neighbor* adalah *Euclidean Distance* dengan nilai tetangga terdekat adalah 3, 5, 7 dan 9 .
4. Skenario uji yang digunakan adalah *K Fold Cross Validation*.
5. Fitur yang digunakan dalam penjurusan siswa adalah nilai ujian nasional bahasa Indonesia, nilai ujian nasional bahasa inggris, nilai ujian nasional IPA, nilai ujian nasional matematika, ulangan harian IPA, Matematika, IPS, Bahasa Indonesia, Bahasa inggris, nilai tes verbal linguistik, logis matematis, spasial, kinestik, musikal, interpersonal, intrapersonal dan natural.