

BAB I PENDAHULUAN

1.1. Latar Belakang

Full Stack Web Development (FSW) merupakan program *course* selama enam bulan di Binar Academy, selama masa *bootcamp* siswa FSW dibantu oleh seorang pengajar yang disebut *Facilitator*, untuk urusan administrasi dan monitoring siswa dilakukan oleh tim *Academic Affairs* (AA). AA bekerja sama dengan *Academic Operation* (AO), *Facilitator Quality* (FQ), dan *Learning Quality* (LQ). AA melakukan tidak lanjut secara langsung terkait permasalahan yang timbul pada siswa, seperti siswa yang mengulang *chapter*, cuti, pindah *wave*, dan *Missing In Action* (MIA) (Binar Academy, 2017). Siswa yang mengambil pilihan untuk cuti dan atau mengulang *chapter* berdampak terhadap waktu kelulusan mereka.

Hasil studi lapangan menunjukkan program *bootcamp* Binar Academy menentukan kelulusan siswa berdasarkan nilai dari *challenge per chapter*, untuk kategori tepat waktu atau tidak tepat waktu didasarkan pada pilihan siswa yang mengambil cuti, dan mengulang *chapter*. Faktor lain seperti penilaian materi pembelajaran yang diajarkan, serta penilaian terhadap *facilitator* meliputi suasana kelas, membantu siswa dalam belajar, dan memfasilitasi forum diskusi (FD), mungkin saja mempengaruhi siswa untuk lulus tepat waktu, sehingga mereka memilih untuk mengambil cuti, dan/atau mengulang *chapter*. Oleh sebab itu perlu dilakukan proses prediksi terhadap waktu kelulusan siswa Binar Academy.

Machine Learning (ML) menjadi salah satu kemajuan sangat penting dalam matematika terapan, salah satu metode dalam ML yang bisa digunakan untuk melakukan proses prediksi adalah dengan klasifikasi (Osisanwo et al., 2017). Berbagai macam algoritma telah banyak digunakan untuk melakukan klasifikasi antara lain *Support Vector Machine* (SVM), *Decision Tree*, *Naïve Bayes*, *Random Forest*, dan *K-Nearest Neighbor* (KNN). Pembangunan model prediksi untuk data yang digunakan merupakan data numerik, sehingga algoritma yang bisa digunakan adalah KNN.

KNN menentukan hasil prediksi berdasarkan jarak/tetangga data baru dengan data lama (Muliono et al., 2020). Hasil prediksi dihitung berdasarkan

jarak dengan tetangga terdekat (Gou et al., 2018). Model KNN dipengaruhi oleh penentuan parameter K dan metode perhitungan jarak. Terdapat beberapa metode untuk menghitung jarak diantaranya *Euclidean*, *Manhattan*, dan *Minkowski* (Setiawan et al., 2022). Penelitian terdahulu membandingkan algoritma *Naive Bayes*, *Decision Tree* dan KNN pada klasifikasi kinerja siswa. Hasil penelitian menunjukkan bahwa akurasi tertinggi terdapat pada algoritma KNN yaitu mencapai 79,31% (Setiyorini et al., 2018).

Keterbatasan jumlah data menjadi tantangan dalam pembuatan model klasifikasi. Kumpulan data yang ada di satu kelas secara signifikan lebih sedikit dibandingkan dengan kelas lain, dikenal sebagai kelas minoritas dan sebaliknya disebut kelas mayoritas (Maciejewski et al., 2011). Untuk kumpulan data yang tidak seimbang, sebagian besar algoritma klasifikasi cenderung mengklasifikasikan kelas mayoritas dengan tingkat akurasi yang tinggi dan kelas minoritas dengan tingkat akurasi yang rendah (Qiong et al., 2016).

Pembangunan model klasifikasi menggunakan dataset yang diberikan oleh Binar Academy menghasil tingkat akurasi yang tinggi dan bahkan mencapai 100%, dalam hal ini tingkat prediksi data yang mencapai 100% menyebabkan adanya kesalahan dalam pembangunan model, hasil yang sangat akurat tersebut mengindikasikan bahwa persebaran data yang digunakan adalah tidak seimbang. Untuk mengatasi hal tersebut metode penyeimbangan data perlu diterapkan dalam penelitian ini, oleh karena itu digunakan metode *oversampling*, salah satunya SMOTE (*Synthetic Minority Over-sampling Technique*) (Siringoringo, 2018).

Penelitian yang telah dilakukan oleh (Hidayati et al., 2021) terkait kelulusan tepat waktu menggunakan data tidak seimbang dengan metode penentuan jarak yaitu *Euclidean* dan *Manhattan*. Hasil menunjukkan bahwa *Euclidean* dan *Manhattan* menunjukkan performa yang sama pada nilai $K=7$ dengan akurasi 85.28%. Pada penelitian lain yang dilakukan oleh (Muliono et al., 2020) menggunakan data seimbang dengan metode perhitungan jarak *Euclidean*, hasil menunjukkan tingkat akurasi dipengaruhi oleh pola persebaran data, dan nilai K yang digunakan.

Pada penelitian ini akan dilakukan proses pembuatan model menggunakan dua metode dengan dataset tanpa SMOTE dan dataset dengan SMOTE dalam mengatasi tingkat akurasi yang disebabkan oleh persebaran data. Masing-masing model akan di terapkan metode perhitungan jarak ketetanggaan menggunakan *Eulidean* dan *Manhattan* dengan parameter nilai K 1, 3, 5, 7 dan 9 untuk mengetahui hasil akurasi terhadap data seimbang dan tidak seimbang.

Metodologi penelitian ini menggunakan *Smart Learning Environment Establishment Guideline* (SLEEG) merupakan panduan yang telah dikembangkan berdasarkan standar *International Organization for Standardization* (ISO) dan *Analyze, Design, Develop, Implement, and Evaluate* (ADDIE). Metodologi ini memiliki empat siklus yang dapat diterapkan dalam system manajemen organisasi pendidikan: *Plan, Do, Check, and Art* (PDCA), yang mencakup kepemimpinan untuk menentukan visi, misi, tujuan, dan sasaran instansi. Pendekatan ADDIE dikenal sebagai desain instruksional untuk mengembangkan pendidikan yang efektif (Rosmansyah et al., 2022).

Berdasarkan penjelasan diatas, penelitian ini akan membahas pembuatan model klasifikasi dengan metode SMOTE-KNN terhadap data tidak seimbang.

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas maka dapat dirumuskan masalah sebagai berikut:

- 1) Bagaimana perhitungan jarak ketetanggaan terbaik dalam algoritma KNN dalam memprediksi waktu kelulusan siswa Binar Academy?
- 2) Bagaimana pemilihan K tetangga terbaik dalam algoritma KNN dalam memprediksi waktu kelulusan siswa Binar Academy?
- 3) Bagaimana hasil over-sampling dengan metode SMOTE untuk mengatasi dataset yang tidak seimbang dalam memprediksi waktu kelulusan *siswa* Binar Academy?
- 4) Bagaimana tingkat akurasi dari model klasifikasi dalam memprediksi waktu kelulusan siswa Binar Academy?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka tujuan dilakukan penelitian ini adalah sebagai berikut:

- 1) Menentukan metode terbaik perhitungan jarak dalam algoritma KNN dalam memprediksi waktu kelulusan siswa Binar Academy.
- 2) Melakukan evaluasi pemilihan K tetangga terbaik dalam algoritma KNN dalam memprediksi waktu kelulusan siswa Binar Academy
- 3) Melakukan evaluasi penggunaan metode over-sampling SMOTE dalam mengatasi dataset yang tidak seimbang dalam memprediksi waktu kelulusan siswa Binar Academy.
- 4) Melakukan analisis model klasifikasi SMOTE-KNN terhadap tingkat akurasi dari model klasifikasi dalam memprediksi waktu kelulusan siswa Binar Academy

1.4. Manfaat Penelitian

Dengan dilakukan penelitian ini maka diharapkan dapat memberikan manfaat terhadap tim AA dalam memonitoring perkembangan siswa selama *bootcamp*.

1.5. Batasan Masalah

Fokus permasalahan dalam penelitian ini adalah sebagai berikut:

- 1) Penelitian dilakukan di Binar Academy *course Full Stack Web Development (FSW)*
- 2) Dataset dari wave 14-18 *course Full Stack Web Development (FSW)* sebanyak 58 siswa, data ini diberikan langsung oleh tim AA melalui email, setelah sebelumnya telah dilakukan penandatanganan surat asas kerahasiaan. Proses pengumpulan data oleh tim AA berjalan selama dua bulan.
- 3) Penelitian ini menggunakan algoritma klasifikasi KNN dengan metode perhitungan jarak menggunakan *Eulidean* dan *Manhattan*.

1.6. Sistematika Pelaporan Skripsi

BAB I PENDAHULUAN

BAB ini menjelaskan awal dari penelitian. Didalamnya terdiri dari latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika pelaporan skripsi. Pada BAB I akan

dijelaskan permasalahan utama penelitian ini dilakukan yaitu memprediksi waktu kelulusan siswa Binar Academy menggunakan algoritma KNN dengan menambahkan metode SMOTE untuk mengatasi permasalahan data yang tidak seimbang, memberikan manfaat terhadap tim AA untuk memonitoring perkembangan siswa selama menjalani *bootcamp* di Binar.

BAB II KAJIAN PUSTAKA

BAB II berisi teori – teori yang relevan dengan kajian penelitian ini. Teori yang dibahas yaitu terkait definisi data mining dan *Educational Data Mining* (EDM), serta *Computer Support Predictive Analytic* (CSPA), berbagai jenis algoritma klasifikasi, dan metode pengolahan data tidak seimbang. Pada bab ini pula akan dipaparkan Binar Academy selaku objek penelitian.

BAB III KAJIAN PUSTAKA

BAB III berisi penjelasan mengenai metodologi penelitian yang digunakan dalam penelitian ini. Metodologi *Smart Learning Environment Establishment Guideline* (SLEEG) yang memiliki empat tahapan utama yakni *Plan*, *Do*, *Check*, dan *Act* (PDCA). *Plan* terdiri dalam dua tahap yakni *Analyze* dan *Design*, *Do* terdiri dari *Develop* dan *Implement*, terakhir *Check* dan *Art* berisi *Evaluate*.

BAB IV HASIL DAN PEMBAHASAN

BAB IV menjabarkan hasil serta pembahasan dari penelitian. Pembahasan akan dimulai dengan proses *Analyze* terkait data yang digunakan dalam penelitian. Berikutnya *Design* akan dijelaskan terkait *flowchart*, *Entity Relationship Diagram* (ERD), serta *Storyboard* system yang akan dibuat. Tahap *Develop* dimulai dengan empat tahap *pre-processing* data pengumpulan data, pembersihan data, pemilihan data, dan perubahan data pembuatan model dari dataset menggunakan aplikasi *mining* yakni *rapid miner* serta menerapkan parameter algoritma KNN yakni perhitungan jarak dan nilai K, dalam hal ini yang digunakan adalah *Euclidean* dan *Manhattan* dengan nilai

$K=1, 3, 5, 7,$ dan 9 , dilanjutkan dengan menghitung akurasi model terhadap data yang digunakan. *Implement* tahap pengaplikasian data ke dalam sistem yang telah dibuat yakni berbentuk web menggunakan *framework codeigniter* (CI). Terakhir *Evaluate* tahap perhitungan hasil prediksi menggunakan *confussion matrix* dan penarikan simpulan.

BAB V SIMPULAN DAN SARAN

BAB V berisi tentang simpulan yang didapatkan dari penelitian, simpulan merupakan jawaban atas rumusan masalah pada BAB I, yakni bagaimana pembangunan model SMOTE-KNN dan bagaimana pengolahan hasil dari model algoritma yang dipakai. Pada bab ini juga akan dilampirkan saran terkait penelitian yang telah dilakukan guna membantu pada penelitian berikutnya. Saran juga akan menjelaskan letak kekurangan penelitian ini agar berikutnya dapat diperbaiki.