e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 23 Agustus 2023

Penerapan Algoritma Klasifikasi untuk Menangani Data Tidak Seimbang pada Peningkatan Kualitas Siswa

Refido Arjunal Akmal[1], Aliyah Kurniasih[2]
Program Studi Ilmu Komputer
Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ
JL TB Simatupang, Cilandak Timur, Kec. Ps. Minggu, Kota Jakarta Selatan 12560
Refido.arjunal.a@students.esqbs.ac.id[1], aliyah.kurniasih@esqbs.ac.id [2]

Abstrak. Keberhasilan proses belajar siswa dapat dibuktikan melalui prestasi belajar siswa yang menggambarkan kemampuan yang dicapai seorang siswa. Peningkatan kualitas siswa ini menjadi hal yang penting karena juga meningkatkan kualitas pendidikan. Banyaknya data tentang kualitas siswa membuat distribusi data menjadi tidak seimbang. Untuk menangani ketidakseimbangan distribusi kelas data dapat dilakukan dengan teknik *oversampling* data menggunakan metode *Support Vector Machine* (SVM) dan *Synthetic Minority Over-Sampling Technique* (SMOTE). Pemilihan algoritma *oversampling* ini diharapkan dapat mengurangi jumlah data serta menambah pada dataset yang kurang pada *feature* minoritas. Penerapan algoritma *Support Vector Machine* (SVM) dan *Synthetic Minority Over-Sampling Technique* (SMOTE) menghasilkan akurasi sebesar 71%.

Kata Kunci: *Klasifikasi, Support Vector Machine, Smote*

1 Pendahuluan

Keberhasilan belajar seorang siswa dapat dibuktikan melalui prestasi belajar siswa yang menggambarkan kemampuan yang berhasil dicapai oleh siswa [1]. Tingkat kualitas pendidikan seorang siswa dilihat dari hasil belajarnya yang telah dia lalui dan ditampilkan dalam bentuk rapor. Namun terdapat banyak faktor yang mempengaruhi kualitas belajar siswa diantaranya faktor internal (dari dalam diri sendiri) dan faktor eksternal (dari luar diri). Faktor internal contohnya kematangan dari segi fisik maupun psikis, kondisi jasmani dan psikologi. Sedangkan faktor eksternal berasal dari luar siswa seperti lingkungan fisik, kehidupan sosial, budaya, dan lingkungan spiritual [2]. Dari banyaknya faktor yang mempengaruhi kualitas seorang siswa maka perlu dipelajari bagaimana cara peningkatan kualitas siswa dari variabel-variabel yang didapat dari data prestasi siswa. Data prestasi siswa akan diklasifikasikan menggunakan *data mining*. Klasifikasi merupakan teknik memperkirakan kelas dari suatu objek dari sebuah kumpulan data yang masih belum diketahui labelnya [3].

Penerapan metode klasifikasi cukup luas dalam berbagai bidang, namun dalam penerapannya terdapat masalah yang cukup sering dijumpai yaitu mengenai tidak seimbangnya distribusi data. Dalam kasus data siswa, ketidakseimbangan data kerap muncul dalam proses klasifikasi. Data yang tidak seimbang antara data mayoritas dengan data minoritas akan mengakibatkan kesalahan saat melakukan klasifikasi. Data yang tidak seimbang dapat menyebabkan perhitungan menjadi kurang akurat dan menimbulkan kesalahan interpretasi hasil klasifikasi. Apabila metode klasifikasi diterapkan langsung pada data yang tidak seimbang akan mengalami penurunan *performance* [2]. Konsekuensi dari penerapan data yang tidak seimbang pada data siswa akan menyebabkan hasil klasifikasi yang salah dan menghasilkan penanganan yang salah juga untuk peningkatan kualitas siswa.

Beberapa penelitian sebelumnya telah membahas tentang distribusi *imbalance* data, salah satunya penelitian yang menggunakan algoritma *Naive Bayes* yang menghasilkan akurasi paling baik sebesar 96,43% [4]. Adapun pada penelitian lainnya menguji algoritma K-means +C4.5 dengan akurasi 94.01% [5]. Penelitian lainnya menggunakan algoritma K-NN dengan menerapkan nilai k tetangga yang bervariasi yaitu 1, 3, 5, 7 dan 9 serta diterapkannya SMOTE menghasilkan peningkatan akurasi algoritma K-NN pada nilai k=1 dan k=3 [6]. Dari latar belakang di atas maka pada penelitian ini, akan dilakukan penanganan data yang tidak seimbang, metode klasifikasi digunakan pada penelitian terkait prediksi peningkatan kualitas siswa.

2 Dasar Teori

2.1 Kualitas Pendidikan Siswa

Pendidikan merupakan proses pengembangan potensi diri manusia melalui proses belajar. Kualitas pendidikan salah satunya ditentukan oleh sistem pendidikan yang diterapkan di suatu negara. Di Indonesia, kualitas pendidikan dapat dikatakan tergolong masih rendah karena beberapa faktor, di antaranya lemahnya sektor manajemen pendidikan antara pusat dan daerah yang menimbulkan kesenjangan yang cukup lebar [7].

2.2 Data Mining

Data mining merupakan suatu teknik yang digunakan dalam mengumpulkan data, membersihkan data, mengolah dan menganalisis data serta memperoleh wawasan ataupun kesimpulan yang bermanfaat dari data tersebut [8]. *Data mining* akan mengekstrak informasi yang berguna dari dataset. *Data mining* juga menganalisis sejumlah data untuk menemukan pola yang bermakna [9].

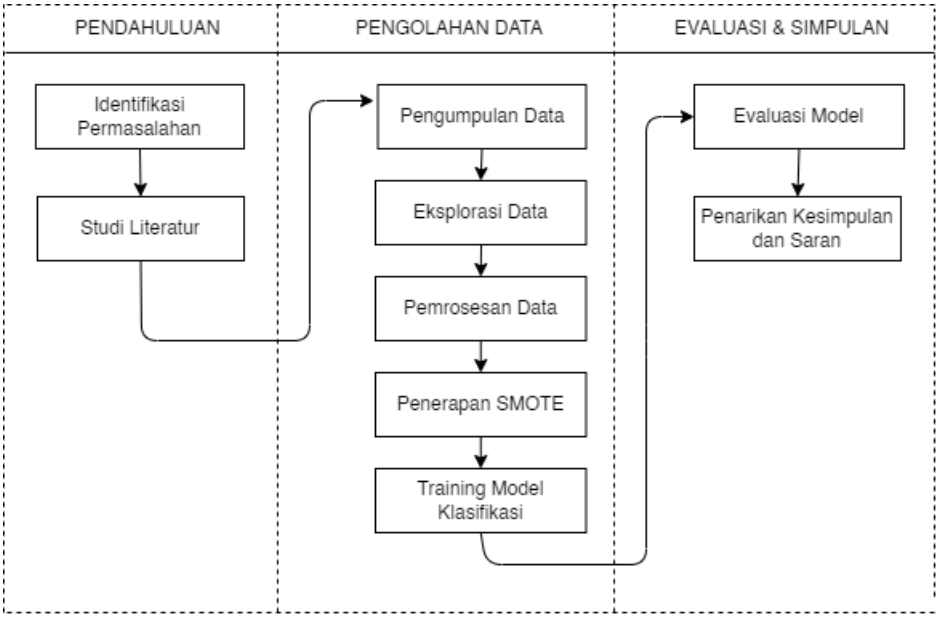
2.3 Klasifikasi

Klasifikasi merupakan proses analisis data dengan cara objek/sampel yang ingin diklasifikasi ditempatkan dalam set kategori atau label berdasarkan sifat dari objek yang diteliti [10]. Setiap objek data akan dipetakan ke dalam sebuah kelompok kelas yang telah ditetapkan. Teknik ini bertujuan untuk memberikan perkiraan kelas dari suatu objek yang belum diketahui kelasnya [5]. Terdapat beberapa metode klasifikasi yang umum digunakan di antaranya algoritma Genetika, *Decision Tree, K-Nearest Neigboor*, *Naïve Bayes Classification*, dan *Support Vector Machine.*

2.4 *Imbalance Class*

Imbalance class atau ketidakseimbangan distribusi data merupakan suatu kondisi di mana beberapa kelas data memiliki jumlah data yang lebih banyak atau lebih sedikit daripada kelas data lainnya di dalam sebuah dataset. Kelas data yang memiliki jumlah data yang lebih banyak dikenal sebagai kelas mayoritas, dan sebaliknya kelas data dengan jumlah lebih sedikit merupakan kelas minoritas [6]. Ketidakseimbangan data memungkinkan hasil klasifikasi mengalami penurunan akurasi dikarenakan kelas minoritas tidak mampu menarik aturan yang sama dengan kelas mayoritas sehingga dapat menyebabkan adanya aturan yang hilang dan mengurangi kemampuan generalisasi untuk menghasilkan prediksi yang lebih akurat [5].

3 Metodologi Penelitian



Gambar 1. Alur Penelitian

Penjelasan tahapan penelitian sebagai berikut.

3.1 Identifikasi Masalah

Masalah yang terjadi pada penelitian ini yaitu adanya data yang *imbalance class*, dan bagaimana *data mining* menangani masalah tersebut untuk menghasilkan model klasifikasi *machine learning* dalam memprediksi prestasi siswa. Di mana atribut yang mempengaruhi bukan hanya dari pada atribut nilai siswa seperti *match score*, *reading score*, dan *writing score*, akan tetapi bagaimana pengaruh atribut lain seperti jenis kelamin, ras atau etnis siswa, pendidikan orang tua, dan dan jenis makan siang siswa dalam mempengaruhi hasil model.

3.2 Pengumpulan Data

Pada tahapan ini peneliti melakukan proses pencarian data publik yang berkaitan dengan atribut – atribut yang mempengaruhi kualitas prestasi siswa. Peneliti mendapatkan data yang bersumber dari situs Kaggle (https://www.kaggle.com/code/spscientist/student-performance-in-exams/input). Data tersebut terdiri dari 1.000 *record* dan delapan *features*. Penjelasan dari masing – masing *features* disajikan pada Tabel 1.

Tabel 1. Atribut Data

Atribut	Keterangan
gender	Jenis kelamin
race/ethnicity	Ras atau etnis siswa yang dikelompokan menjadi beberapa kelompok data
parental level of education	Tingkat pendidikan orang tua
lunch	Jenis makan siang siswa
test preparation course	hasil final test yang dinyatakan selesai atau tidak
math score	Score ujian siswa
reading score	Nilai membaca siswa
writing score	Nilai menulis siswa

3.3 Eksplorasi Data

Eksplorasi data dilakukan dengan tujuan untuk mendapatkan pemahaman yang lebih terhadap data yang digunakan, agar dapat melakukan pemrosesan data yang sesuai terhadap kebutuhan permasalahan pada data atau penggunaan metode – metode pada *data mining*. Eksplorasi data yang dilakukan seperti mengecek dimensi data, tipe data dan isi data, data yang kosong dan data duplikat, *outlier* pada data, dan unik data dari kelas target model beserta distribusi data.

3.4 Pemprosesan Data

Pada proses ini melakukan penghapusan data *outlier* pada atribut *match score*, *reading score*, dan *writing score*, dengan menggunakan matematika statistik persentile 25 (Q1) dan persentil 75 (Q3) untuk mendapatkan nilai IQR=Q3-Q1, dimana batas bawah =Q1-1.5*IQR dan batas atas =Q3+1.5*IQR. Kemudian mengubah data dari *string* kategorikal menjadi data numerik pada atribut *gender, race/ethnicity, parental level of education*, dan *lunch*. Selanjutnya melakukan pembagian data ke dalam data latih untuk proses *training* model dan data uji untuk proses evaluasi model dengan persentase perbandingan 80:20. Setiap data latih dan data uji terdapat dua komponen data yaitu komponen X sebagai *input* model yang terdiri dari atribut *match score*, *reading score*, *writing score, gender, race/ethnicity, parental level of education*, dan *lunch,* serta komponen Y sebagai target model yaitu atribut *test preparation course*. Proses selanjutnya melakukan normalisasi data menjadi data dalam rentang nilai antara 0 dan 1 menggunakan *function* MinMaxScaler pada *input* model data latih dan data uji.

3.5 SMOTE

SMOTE (*Synthetic Minority Oversampling Technique*) digunakan untuk menangani *imbalance class* pada data latih dengan cara membuat data sintesis pada sampel data kelas minoritas mengikuti sejumlah data pada sampel kelas mayoritas. SMOTE digunakan pada penelitian ini karena jumlah dataset pada penelitian ini hanya sedikit, dan hanya dilakukan pada data latih dengan tujuan agar hasil prediksi model lebih dapat diterima karena tidak bias.

3.6 Model Klasifikasi

Model klasifikasi *machine learning* dibuat dengan membandingkan algoritma dari keluarga *Naïve Bayes Classifier* dengan algoritma *Support Vector Machine* (SVM) pada kernel yang berbeda. *Naïve Bayes Classifier* dengan algoritma *Gaussian Naïve Bayes, Complement Naïve Bayes, Bernoulli Naïve Bayes*, dan *Logistic Regressin*, serta kernel pada SVM diantaranya kernel poly, rbf, sigmoid dan linear dencan 10-*fold cross-validation*. Variasi ini digunakan untuk mengetahui algoritma terbaik dalam pemodelan klasifikasi prestasi siswa.

3.7 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja model klasifikasi *machine learning* secara keseluruhan dalam menentukan model terbaik. *Metrix evaluation* yang digunakan yaitu *accuracy*, serta *precision, recall*, dan *f1-score* menggunakan *everage macro*. *Average macro* bekerja dengan tidak mengabaikan kelas minoritas pada data yang akan di uji, artinya *average macro* akan memberikan bobot yang sama pada kelas minoritas dan kelas mayoritas.

4 Hasil dan Pembahasan

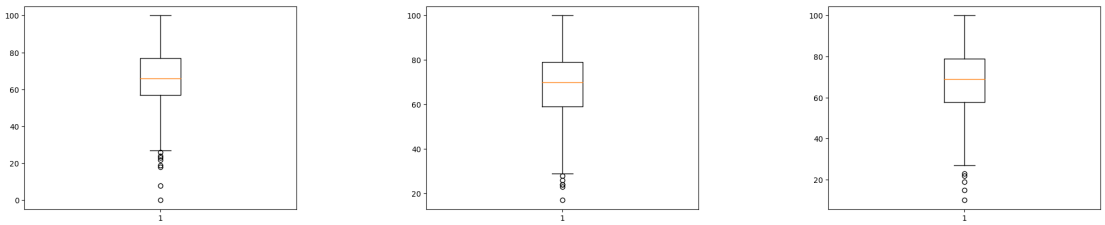
4.1 Hasil Eksplorasi Data

Hasil dari pada eksplorasi data mengecek dimensi data bahwa memang terdapat jumlah data sebanyak 1.000 baris data dan 8 atribut. Tabel 2 merupakan distribusi atribut berdasarkan tipe data dan telah terindikasi adanya

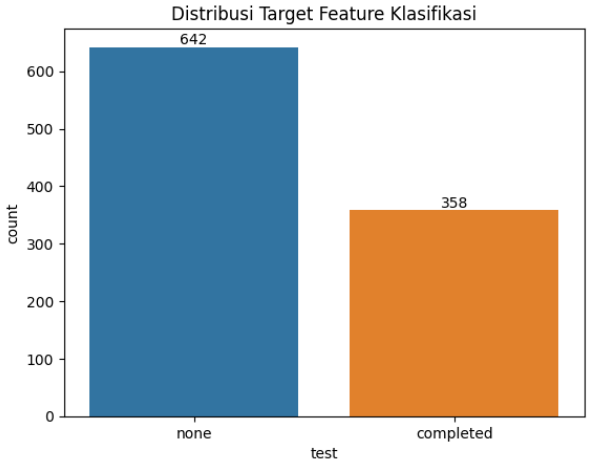
kesesuaian isi data dengan tipe data. Kemudian dari hasil pengecekan data yang kosong dan data duplikat bahwa tidak adanya data yang mengandung *missing values* dan dupikat data. Akan tetapi terdapat *outlier* data pada atribut *match score*, *reading score*, dan *writing score* seperti yang disajikan pada Gambar 2. Unik data dari atribut target model yaitu data dengan label 'none' dan 'completed', distribusi data disampaikan pada Gambar 3 yaitu label 'none' sebanyak 642 data dan label 'completed' sebanyak 358 data, di mana sudah sangat jelas berdasarkan jumlah data yang berbeda dari setiap kelas adanya indikasi data yang *imbalance*.

Tabel 2. Tipe Data

Tipe data	Atribut
String kateogirikal	Gender, race/ethnicity, parental level of education, lunch, test preparation course
Interger	Math score, reading score, writing score



Gambar 2. Indikasi Data *Outlier*



Gambar 3. Distribusi Jumlah Data Target

4.2 Hasil Pemprosesan Data

Hasil dari pada proses setelah menghapus data *outlier* dengan atribut *match score*, *reading score*, dan *writing score* yaitu dihasilkan jumlah data menjadi sebanyak 984 baris data, dan jumlah data inilah yang akan digunakan untuk *trainig* dan evaluasi model. Kemudian hasil dari mengubah data dari *string* kategorikal ke data numerik pada atribut *gender, race/ethnicity, parental level of education*, dan *lunch* menggunakan fungsi *label encoder* ditampilkan pada Tabel 3. Hasil dari pembagian data setelah penghapusan data *outlier* atau dari sejumlah data 984 baris data yaitu latih sebesar 80% dan data uji sebesar 20% adalah data latih yang dihasilkan menjadi sebanyak 787 baris data, dan data uji yang dihasilkan sebanyak 197 baris data.

Tabel 3. Hasil Label Encoder

Atribut	Sebelum	Sesudah
gender	Female	0
	Male	1
race/ethnicity	Group A	0
	Group B	1
	Group C	2
	Group D	3
	Group E	4
parental level of education	Associate's degree	0
	Bachelor's degree	1
	High school	2
	Master's degree	3
	Some collage	4
	Some high school	5
lunch	Free/reduced	0
	Standard	1
test preparation course	Completed	0
	None	1

4.3 Hasil SMOTE

Hasil dari penangan *imbalance class* hanya pada data latih disajikan pada Tabel 4. Yaitu label kelas minoritas data sebanyak 282 disintesis menjadi 505, sebanyak jumlah data kelas mayoritas. Kemudian hasil dari pada jumlah data latih setelah di SMOTE menjadi sebanyak 10.10 data.

Tabel 4. Hasil SMOTE

Label	Sebelum	Sesudah
completed (0)	282	505
none (1)	505	505

4.4 Hasil Evaluasi Model

Hasil evaluasi model terbaik dalam mengklasifikasikan prestasi siswa ke dalam kelas 'completed' atau 'none' seperti yang disampaikan pada Tabel 5, yaitu dihasilkan model terbaik pada algoritma SVM dengan kernel *Poly* dengan akurasi sebesar 74%, nilai precision 74%, recall 75% dan f1-score 74%. Hasil ini mengungguli algoritma SVM dengan kernel lainnya seperti RBF, Sigmoid, dan Linear, serta jauh lebih besar dari nilai evaluasi model dari algoritma keluarga *Naïve Bayes Classifier*, karena kernel Polynomial mampu untuk menangani hubungan data yang lebih kompleks atau tingkat kompleksitas data yang tinggi di antara fitur-fitur dan SVM dapat bekerja baik dengan jumlah data yang relatif kecil dan dimensi fitur cukup rendah.

Tabel 5. Hasil Evaluasi Model

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
Gaussian Naïve Bayes	57%	59%	60%	57%
Complement Naïve Bayes	53%	55%	55%	53%
Bernoulli Naïve Bayes	51%	47%	47%	47%
Logistic Regression	67%	67%	68%	67%
SVM Kernel Poly	**74%**	**74%**	**75%**	**74%**

SVM Kernel RBF	67%	67%	68%	67%
SVM Kernel Sigmoid	68%	68%	69%	67%
SVM Kernel Linear	59%	59%	59%	58%

5 Kesimpulan

Dari hasil percobaan terbukti bahwa proses klasifikasi dengan menggunakan algoritma klasifikasi menunjukan hasil yang cukup baik dengan teknik sampling smote, terbukti dengan metode *Support Vector Machine* dengan *kernel poly* menghasilkan nilai akurasi yang cukup tinggi sekitar 71 % dengan menggunakan semua *feature* data. Untuk penelitian yang akan datang, jika menggunakan dataset yang sama atau serupa, agar dapat mencoba dengan menggunakan teknik oversampling lainnya untuk menangani *imbalance class*. Kemudian apabila masih menggunakan algoritma yang sama, agar dapat melakukan *fine tunning hyperparameter* pada SVM dan *Logistic Regression* untuk dapat meningkatkan hasil evaluasi model, atau dapat menggunakan *feature extraction* lainnya untuk mengubah data kategorikal menjadi numerik. Selanjutnya membandingkan hasil model ketika pemodelan klasifikasi dibangun hanya dengan *features* dari nilai siswa.

Referensi

[1] A. Syafi'i, T. Marfiyanto, dan S. K. Rodiyah, "Studi Tentang Prestasi Belajar Siswa Dalam Berbagai Aspek Dan Faktor Yang Mempengaruhi," *Jurnal Komunikasi Pendidikan*, vol. 2, no. 2, hlm. 115, Jul 2018, doi: 10.32585/jkp.v2i2.114.

[2] A. Ahmadi, *Teknik Belajar Yang Efektif* . Jakarta: Rineka Cipta, 2004.

[3] P. , S. M. , & K. V Tan, *Introduction to Data Mining*. Boston: Pearson Education., 2016.

[4] M. Sulistiyono, Y. Pristyanto, S. Adi, dan G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *SISTEMASI*, vol. 10, no. 2, hlm. 445, Mei 2021, doi: 10.32520/stmsi.v10i2.1303.

[5] C. Agus Sugianto, "Analisis Komparasi Algoritma Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan," 2015.

[6] A. N. Kasanah, M. Muladi, dan U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, hlm. 196–201, Agu 2019, doi: 10.29207/resti.v3i2.945.

[7] S. Fadia dan N. Fitri, "Problematika Kualitas Pendidikan di Indonesia," *Jurnal Pendidikan Tambusai*, 2021.

[8] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[9] J. Ledolter, *Data Mining and Business Analytics with R*, 1st edition. Wiley, 2013.

[10] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," *Faktor Exacta*, vol. 11, no. 3, Okt 2018, doi: 10.30998/faktorexacta.v11i3.2777.