Prediksi Penerimaan SNMPTN Berdasarkan Nilai Rapor Menggunakan Metode C4.5

\*) Penulis korespondensi (Muhammad Iqbal Akbar)

Email: iqbal.elektro.um@gmail.com

Prediction of SNMPTN Acceptance Based on Report Card Value Using Method C4.5

Harits Ar Rosyid1), Muhammad Muza Al Fitra Pratama2), Muhammad Iqbal Akbar\*,3)

1 Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Malang

Jl. Semarang No.5, Malang, Indonesia 65145

|  |
| --- |
| **Cara sitasi**: N. P. Pertama and N. P. Kedua, "Petunjuk Penulisan dan Kirim Artikel Jurnal Teknologi dan Sistem Komputer Mulai Penerbitan Nomor 6(4) Tahun 2018," Jurnal Teknologi dan Sistem Komputer, vol. x no. x, 201x. doi: 10.14710/jtsiskom.x.x.xxxx.xx-xx, [Online]. |

Abstract - National University Entrance Entrance Selection (SNMPTN) is one of the selected paths for student admission to enter state universities (PTN) in Indonesia. This study aims to predict the opportunity to be accepted at the desired PTN. Data comes from www.halokampus.com subject input, the output will be in the form of information received or not received. The dataset is unbalanced, unequal class distribution affects the performance of the classification method. to eliminate the imbalance class using SMOTE. This study compares C4.5 without SMOTE and C4.5 with SMOTE. The results of the study are a predictive ability for SNMPTN acceptance. This study uses the C4.5 method to predict student report card grades that can be accepted on the SNMPTN pathway. C4.5 algorithm because the decision tree which was previously complex and very global, can be changed to be more simple and specific. The experimental results show that C4.5 with SMOTE has good accuracy compared to C4.5 without SMOTE. C4.5 with SMOTE showed good results with accuracy and precision of 95.22% and recall of 95.20%.

**Keywords** - SNMPTN; C4.5; SMOTE; classification; report card grades

Abstrak - Seleksi Nasional Masuk Perguruan Tinggi Negeri(SNMPTN) merupakan salah satu jalur seleksi penerimaan mahasiswa untuk memasuki perguruan tinggi negeri(PTN) di indonesia. Penelitian ini bertujuan untuk memprediksi kesempatan diterima di PTN yang diinginkan. Data berasal dari [www.halokampus.com](http://www.halokampus.com) input mata pelajaran , output akan berupa keterangan diterima atau tidak diterima. Dataset yang didapatkan tidak seimbang, distribusi kelas yang tidak seimbang mempengaruhi performa metode klasifikasi. untuk menghilangkan imbalance class menggunakan SMOTE. Penelitian ini membandingkan C4.5 tanpa SMOTE dan C4.5 dengan SMOTE. Hasil penelitian berupa kemampuan prediksi Penerimaan SNMPTN. Penelitian ini menggunakan metode C4.5 untuk memprediksi nilai rapor siswa yang dapat diterima pada pada jalur SNMPTN. Algoritma C4.5 karena pohon keputusan yang sebelumnya kompleks dan sangat global, dapat diubah manjadi lebih simpel dan spesifik. Hasil eksperimen menunjukkan bahwa C4.5 dengan SMOTE memiliki akurasi yang baik dibandingkan C4.5 tanpa SMOTE. C4.5 dengan SMOTE menunjukkan hasil yang baik dengan akurasi dan presisi sebesar 95,22% dan recall sebesar 95,20%.

**Kata kunci** – SNMPTN; C4.5; SMOTE; klasifikasi; nilai rapor

I. Pendahuluan

Setiap siswa SMA pasti memiliki keinginan untuk melanjutkan sekolah di perguruan tinggi negeri(PTN). Salah satu jalur seleksi masuk penerimaan mahasiswa untuk bisa diterima di perguruan tinggi negeri adalah SNMPTN. SNMPTN memiliki kepanjangan Seleksi Nasional Masuk perguruan tinggi negeri dan jalur SNMPTN dilakukan bersamaan di seluruh indonesia.berbeda dengan jalur SBMPTN atau jalur mandiri, SNMPTN merupakan jalur masuk penerimaan mahasiswa yang sangat diinginkan oleh siswa SMA karena jalur masuk ini tidak memerlukan ujian atau tes tulis agar bisa masuk di PTN. SNMPTN salah satu jalur masuk yang paling yang paling favorit dikarenakan pada jalur ini siswa tidak memerlukan ujian dan hanya menerima undangan untuk kuliah di PTN yang dipilih. Pada tahun 2011 dari semua siswa yang mendaftar hanya 20% yang diterima melalui SNMPTN dari seluruh indonesia, karena alasan tersebut SNMPTN merupakan seleksi masuk yang ketat dan bergengsi[1].

ketentuan umum snmptn menurut halaman website resmi SNMPTN [www.snmptn.ac.id](http://www.snmptn.ac.id/) bahwa SNMPTN dilakukan karena atas dasar semester satu sampai dengan semester lima bagi pendaftar dari SMA/MA/SMK dengan masa belajar tiga tahun, atau bagi pendaftar dari SMK masa belajar empat tahun menggunakan nilai rapor semester satu sampai dengan semester tujuh. Dengan rincian rapor mata pelajaran Matematika, Bahasa Indonesia, Bahasa Inggris, Kimia, Fisika, dan Biologi untuk siswa pendaftar pada jurusan IPA, lalu untuk jurusan IPS adalah Matematika, Bahasa Indonesia Bahasa Inggris, Sosiologi, Ekonomi, dan Geografi. Kemudian untuk siswa pendaftar dari SMK dengan menggunakan nilai rapor mata pelajaran Matematika, Bahasa Indonesia, Bahasa Inggris ditambah dengan nilai kompetensi keahlian berdasarkan Teori dan Praktik kejuruan yang diambil selama masa pembelajaran di SMK. Nilai rapor sangatlah menetukan penerimaan SNMPTN, dan juga faktor pertimbangan lain seperti faktor alumni yang diterima dan prestasi non-akademik juga menjadi pertimbangan diterima atau tidaknya pada seleksi nasional masuk perguruan tinggi negeri, ditambah lagi dengan tahapan seleksi menurut LTMPT siswa pendaftar diseleksi di PTN pilihan pertama berdasarkan urutan pemilihan jurusan atau progam studi. Jadi dua hal tersebut adalah faktor yang memungkinan diterima pada SNMPTN.

Pada penelitian sebelumnya penerapan algoritma *C4.5* untuk penentuan jurusan mahasiswa, dimana dalam penetuan jurusan yang sesuai dengan latar belakang, minat, dan kemampuanya sendiri[2]. Implementasi metode *Electre* pada system pendukung keputusan SNMPTN jalur undangan[3]. Dengan membuat sistem pendukung keputusan dengan menggunakan metode *Analytic Hierarchy Process*[4]. Dataset setelah dilakukan proses SMOTE tersebut berjumlah 414 yang terdiri dari 388 kelas positif dan 387 kelas negatif. Dalam penelitian tersebut menggunakan metode klasifikasi *C4.5* untuk membandingkan performa antara *C4.5* dengan SMOTE dan *C4.5* tanpa SMOTE. metode *C4.5* ketika dataset tidak seimbang memiliki performa buruk dan ketika dataset sudah seimbang dengan menggunakam SMOTE memiliki performa yang lebih baik. Pada penelitian ini bisa didapatkan bahwa SMOTE dapat membuat data yang tidak seimbang menjadi seimbang namun tidak overfitting, sehingga kerja algoritma yang digunakan menjadi maksimal[5].

Dalam penelitian ini untuk memprediksi jalur penerimaan SNMPTN dengan menggunakan nilai Matematika, Bahasa Indonesia, Bahasa Inggris, Kimia, Fisika, dan Biologi semester 1 sampai semester 5. Di indonesia sendiri masih sedikit penelitian yang membantu Siswa sekolah menengah atas(SMA) dalam memprediksi kesempatan siswa di jalur SNMPTN. Oleh karena itu penelitian ini dibuat untuk prediksi penerimaan di jalur SNMPTN, dalam penelitian ini juga menggunakan data yang didapat pada website halokampus[6]. Penelitian ini menggunakan algoritma *C4.5*  tanpa SMOTE dan *C4.5*  dengan SMOTE, hasil dari prediksi penelitian ini akan berupa label diterima atau tidak diterima.

III. Metode Penelitian

Machine Learning merupakan bidang penelitian yang mengembangkan dan mempelajari algoritma yang dapat belajar dan membuat prediksi dari data[7]. Dalam Machine Learning berdasarkan metode penerapannya Machine Learning memiliki tiga devisi antara lain yaitu Supervised learning ,unsupervised learning dan Semi-supervised learning. Supervised learning merupakan metode Machine Learning yang harus diamati bagaimana jalannya algortima. Untuk unsupervised learning merupakan metode Machine Learning yang tidak perlu diamati atau hasilnya dari tindakan komputer itu sendiri . sedangkan semi-supervised learning merupakan metode Machine Learning yang dimana tidak semua data diberi label atau ada beberapa data yang diberi label[8]. Machine learning juga dapat diterapkan dalam game edukasi[9] dan pendidikan[10]. Masalah yang dapat diselesaikan Machine learning antara lain regresi, clustering dan klasifikasi. Klasifikasi merupakan metode pengelompokan data yang sudah ditentukan kelasnya [11]. Untuk proses klasifikasi sendiri dapat menggunakan algoritma seperti C4.5. Algoritma C4.5 tanpa SMOTE dan C4.5 dengan SMOTE akan digunakan dalam penelitian ini.

## **Preprocessing Data**

Sebelum melakuan proses klasifikasi pada dataset, langkah yang dilakukan adalah dengan melakukan *preprocessing* terhadap data terlebih dahulu. Tahapan *preprocessing data* yang dilakukan adalah, (1) Itegrasi data, Pemilihan Data, (3) *Data* *Transformation*, (4) *Replace Missing* *Value*.

1. *Data Integration*

Ntegrasi database menyediakan akses terintegrasi ke berbagai sumber data. Integrasi database memiliki dua kegiatan utama: integrasi skema (membentuk pandangan global dari konten data yang tersedia di sumber) dan integrasi data (mengubah data sumber menjadi format yang seragam)[12]. Integrasi data adalah proses penggabungan data dari sumber yang berbeda kedalam satu dataset baru. Penelitian ini menggabungkan 9 data kampus tahun 2016 dan 7 data kampus 2017 dengan total 784 baris data dengan menggunakan bantuan *Microsoft excel*.

1. *Data Cleaning*

pembersihan data adalah proses mengidentifikasi dan menghapus kesalahan dalam gudang data[13]. Pemilihan data atau seleksi atribut data yang digunakan. Dalam penelitian ini, pemilihan data digunakan untuk memfokuskan penelitian pada prediksi kemungkinan diterima atau tidaknya pada jurusan di snmptn berdasarkan urutan pemilihan jurusan atau program studi, faktor alumni (jumlah alumni yang diterima pada jurusan), prestasi non-akademik dan juga berdasarkan nilai rapor siswa pendaftar.

Maka dari itu penelitian ini menggunakan atribut urutan pemilihan jurusan/prodi yaitu pilihan satu dan pilihan kedua , jumlah alumni, prestasi tambahan dan juga nilai rapor semester satu sampai dengan semester 5 dengan objek studi siswa IPA dengan pilihan jurusan Ilmu IPA pada Seleksi Nasional Masuk Perguruan Tinggi Negeri. Jumlah setelah dilakukan proses seleksi atribut yaitu 416 baris data.

1. *Data Transformation*

Transformasi data mengubah seperangkat nilai data dari format data sistem data sumber ke dalam format data sistem data tujuan[14]. *Data Transformation* ini data akan diubah atau dikonsolidasi sehingga proses mining yang dihasilkan lebih efisien, dan pola yang ditemukan lebih mudah dipahami. Pada tahapan ini, pengelompokan berfokuskan pada siswa IPA pada Seleksi Nasional Masuk Perguruan Tinggi Negeri dengan acuan yang telah ditetapkan oleh Kementerian Riset, Teknologi, dan Pendidikan Tinggi.

|  |  |  |
| --- | --- | --- |
| **Data awal** | **Hasil konversi** | **Tipe Data** |
| Juara OSN Matematika se-Kota | Kota | Nominal/ Kategorikal |
| Olimpiade Kedokteran tingkat Internasional | Internasional |
| Olimpiade Sains tingkat Provinsi | Provinsi |
| Olimpiade Teknik Kimia ITS 100 Nasional | Nasional |
| (No data) | Kosong |
| - | Kosong |

**Tabel 1.** *Data Transformation*

## **SMOTE**

Strategi paling sederhana yang dapat digunakan untuk kasus data tidak seimbang adalah *Random over-sampling*, yang dimana cara kerja metode tersebut menyeimbangkan kelas melalui replikasi kelas minoritas agar sama dengan kelas mayoritas. Meskipun metode *Random over-sampling* ini terlihat efektif tetapi metode ini dapat meningkatkan kemungkinan overfitting karena salinan yang dibuat sama persis yang berasal dari data kelas minoritas. Untuk menghindari overfitting dibuatlah teknik SMOTE [15].

Teknik SMOTE digunakan untuk memecahkan masalah kelas ketidakseimbangan. SMOTE membuat contoh kelas prioritas sintesis yang menggunakan ruang fitur daripada ruang data. Teknik SMOTE menghasilkan sintesis baru dengan memanfaatkan jarak antara sampel minoritas dan tetangga terdekat dari sampel minoritas, diantara jarak dari kedua sampel tersebut dibuatlah sintesis baru sebanyak yang dibutuhkan sehingga data menjadi seimbang[16].

Asumsikan bahwa dataset kelas minoritas adalah sampel, tingkat oversampling adalah N dan titik tetangga terdekat adalah K. Langkah-langkah perhitungan dengan SMOTE [17]:

1. Tentukan nilai K sampel tetangga terdekat untuk setiap sampel kelas minoritas i dalam dataset sampel kelas minoritas
2. Pilih N sampel secara acak dari masing-masing tetangga terdekat.
3. Hitung sampel baru menggunakan rumus (1) dari sampel kelas minoritas dan setiap sampel N merupakan sintesis baru dan kemudian sistesis baru tersebut ditambahkan kedalam data sampel kelas minoritas.

SyntheticSampel[indekbaru]=Sampel[i]+Random\*(SampelTetangga[i] -Sample[i]) (1)

Keterangan :

SyntheticSampel = sampel sisntesi baru.

Sampel[i] = sampel dataset dari kelas minoritas.

i = jumlah sampel kelas minoritas.

random = memiliki nilai angka acak antara [0,1].

sampel tetangga[i] = sampel tetangga terdekat dari sampel[i].

1. Ulangi proses diatas sampai semua sampel kelas minoritas memenuhi sesuai yang diminta.

Dengan menerapkan cara tersebut nantinya pemilihan titik acak disepanjang segmen garis antara dua sample akan menjadi sentesis baru. Dengan menggunakan teknik tersebut SMOTE dapat memperluas area keputusan untuk kelas minoritas [16]. Dikarenakan dalam kasus SNMPTN dataset yang didapatkan tidak seimbang antara yang diterima dan tidak diterima, sehingga tahap preprocessing data untuk menyeimbangkan distribusi kelas label minoritas dan mayoritas.dibutuhkan dalam kasus ini. Langkah selanjutnya hasil SMOTE di hitung menggunakan *C4.5*.

## **Dataset**

Data yang digunakan dalam proses penelitian adalah data Mahasiswa yang telah diterima SNMPTN. Data ini diperoleh dari laman website halo kampus [www.halokampus.com.](http://www.halokampus.com/) Terdapat 9 data dari tahun 2016 dan 7 data dari 2017 dengan total 784 baris data dan 94 atribut. Berikut Tabel 2 dan Tabel 3 adalah salah satu contoh dari dataset yang digunakan pada penelitian ini.

**Tabel 2.** Dataset Survey SNMPTN Tahun 2016

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nama** | **Jurusan sekolah** | **Pilihan 1** | **Pilihan 2** | **Jurusan**  **diterima** | **MTK**  **Semester 1** | **…** | **Prestasi**  **lain** |
| Sari | IPA | Pend. Ners | Budidaya  perairan | Budida-ya perairan | 81 - 85 | **…** | Olimpiade Kimia |
| Miko | MM | Ilmu Komp-uter | Teknik  Inform-atika | Ilmu  Komputer | 76 -80 | … | - |

**Tabel 3.** Dataset Survey SNMPTN Tahun 2017

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nama** | **Jurusan sekolah** | **Pilihan 1** | **Pilihan 2** | **Jurusan**  **diterima** | **MTK**  **Semester 1** | **…** | **Prestasi**  **lain** |
| Nadya | IPA | Statiska | Matematika | Statiska | 86 - 90 | **…** | Piagam  Lomba MTQ |
| Zahnia | IPA | Mana-Jemen | Akutansi | Manajemen | 91 - 95 | … | Parlemen Remaja (Nasional) |

Pada tahap ini data yang awalnya memiliki distribusi atribut label yang tidak seimbang, diseimbangkan menggunakan metode preprocessing SMOTE. Hasil SMOTE ditunjukkan pada Table 4.

**Tabel 4.** Perbandingan Jumlah Data Asli Dengan Data+SMOTE

|  |  |  |  |
| --- | --- | --- | --- |
|  | Jumlah diterima | Jumlah tidak diterima | Total Data |
| Data Asli | 388 | 26 | 414 |
| Data+SMOTE | 388 | 387 | 775 |

## **Algoritma Decision Tree C4.5**

## Algoritma *Decision Tree* C4.5 merupakan algoritma yang memiliki kelebihan dapat mengolah suatu data numerik (kontinyu), kategori (diskret), dapat menangani nilai atribut yang hilang serta menghasilkan aturan-aturan yang mudah diinterpretasikan [18]. Algoritma C4.5 ini merupakan pengembangan dari algoritma ID3[19].

Prinsip kerja dari algoritma ID3 dan algoritma C4.5 hampir sama, namun terdapat beberapa perbedaan yang membuat algoritama C4.5 memiliki hasil lebih baik dibanding algoritma ID3. Algoritma C4.5 mampu menangani atribut dengan tipe diskrit atau kontinu. Pemilihan atribut pada algoritma ini menggunakan ukuran berdasarkan Entropi yang dikenal dengan *information gain* sebagai sebuah heuristic untuk memilih atribut yang merupakan bagian terbaik dari contoh ke dalam kelas. Semua atribut bersifat kategori yang bernilai diskrit. Atribut dengan nilai *continuous* harus didiskritkan. Diskritisasi atribut bertujuan untuk mempermudah pengelompokan nilai berdasarkan kriteria yang telah ditetapkan. Hal ini juga bertujuan untuk menyederhanakan permasalahan dan meningkatkan akurasi dalam proses pembelajaran [20].

Pemilihan atribut pada algoritma C4.5 menggunakan *Gain* menggantikan *Information gain*. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan *Decision Tree* yang paling kecil ukurannya. Atau atribut yang bisa memisahkan obyek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling bersih. Ukuran paling bersih dinyatakan dengan tingkat impurity, dan untuk menghitungnya, dapat dilakukan dengan menggunakan konsep Entropi, Entropi menyatakan *impurity* suatu kumpulan objek [21]. Berdasarkan [22] terdapat empat tahapan dalam melakukan langkah klasifikasi menggunakan Algoritma C 4.5, antara lain :

1. Memilih atribut sebagai akar.
2. Membuat cabang untuk masing-masing nilai.
3. Membagi setiap kasus dalam cabang.
4. Mengulangi proses dalam setiap cabang sehingga semua kasus dalam cabang memiliki kelas yang sama.

Perhitungan dimulai dari menghitung banyaknya jumlah atribut dan menentukan atribut mana yang akan digunakan sebagai akar dari pohon keputusan. Selanjutnya akan dilakukan perhitungan Entropi dan *gain* untuk membentuk *leaf* dari pohon keputusan tersebut. Setelah semua perhitungan selesai dilakukan, pohon keputusan dapat dibentuk berdasarkan nilai *gain* yang telah dihitung sebelumnya. Atribut dengan nilai *gain* tertinggi akan terletak pada prioritas yang lebih tinggi dan memiliki kedudukan yang lebih tinggi juga pada pohon keputusan. Rumus dalam mencari Entropi sebagai berikut:

* Entropi

(2)

dimana:

S  = Himpunan (dataset) kasus

K = Jumlah partisi S

= Probabilitas yang didapat dari sum dibagi total kasus

* Gain

(3)

dimana:

S = Himpunan kasus

A = Jumlah partisi atribut A

= Jumlah sample untuk atribut ke-i

= Jumlah seluruh sample data

= Entropi untuk sample-sample yang memiliki nilai i

* Gain Ratio

(4)

dimana:

a  = Atribut  
gain(a) = *Information gain* pada atribut a

split(a) = Split information pada atribut a

* SplitInfo

(5)

dimana:

S = Ruang sampel yang digunakan untuk training

A = Atribut

= Jumlah sample untuk atribut i

## **Confusion Matrix**

Tahap Evaluasi hasil klasifikasi pada penelitian ini menggunakan Confusion Matrix. Confusion Matrix merupakan tabel yang berisi banyaknya data yang diuji pada sebuah penelitian, serta mencatat seberapa sering data diklasifikasikan bernilai benar atau bernilai salah [23]. Berdasarkan hasil model klasifikasi tersebut nantinya bisa menunjukkan hasil prediksi dan klasifikasi penelitian ini . Model Confusion Matrix seperti berikut :

**Tabel 5.** *Confusion Matrix*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  | PREDIKSI | |
|  | |  | + | - |
| AKTUAL | + | | TP | FN |
| - | | FP | TN |

Keterangan :

* TP (True Positif) : prediksi pada kasus menyatakan TRUE dan kenyataanya TRUE.
* TN (True negatif) : prediksi pada kasus menyatakan FALSE dan kenyataannya FALSE.
* FP (False positif) : prediksi pada kasus menyatakan TRUE dan kenyataannya FALSE.
* FN (False negatif) : prediksi pada kasus menyatakan FALSE dan kenyataannya TRUE.

Dari hasil Confusion Matrix tersebut nantinya dapat dihitung akurasi (6), precision (7), dan recall (8) dengan rumus sebagai berikut :

(6)

(7)

(8)

Akurasi digunakan untuk menghitung efektifitas dan mengevaluasi kinerja metode klasifikasi [24]. Untuk precision digunakan untuk menghitung proporsi kasus prediksi positif yang itu benar TP, sedangkan recall digunakan untuk menghitung proporsi kasus TP yang diprediksi dengan benar [25]. Yang terakhir error rate merupakan rasio dari jumlah data yang diklasifikasikan dengan benar dengan jumlah semua data [24].

III. Hasil dan Pembahasan

## **Hasil penelitian**

Seteleh melalui berbagai tahapan, tahap akhir yang harus dilakukan adalah proses evaluasi hasil dari klasifikasi. Dalam proses evaluasi ini, akan dibandingkan hasil klasifikasi algoritma *C4.5* tanpa SMOTE dengan hasil klasifikasi *C4.5* dengan SMOTE. Hasil klasifikasi yang memiliki performa lebih tinggi akan mengindikasikan algoritma yang lebih baik untuk klasifikasi penerimaan SNMPTN berdasarkan nilai rapor. Perbandingan dapat dilihat pada Table 6 dan Table 7.

**Tabel 6.** Hasil Confusion Matrix C4.5 Tanpa SMOTE

|  |  |  |
| --- | --- | --- |
|  | Diterima | Tidak Diterima |
| Diterima | 388 | 0 |
| Tidak Diterima | 26 | 0 |

**Tabel 7.** Hasil Confusion Matrix C4.5 Dengan SMOTE

|  |  |  |
| --- | --- | --- |
|  | Diterima | Tidak Diterima |
| Diterima | 370 | 18 |
| Tidak Diterima | 19 | 368 |

Pada Tabel 6 dan 7 merupakan hasil *confusion matrix* dari hasil klasifikasi *C4.5* tanpa SMOTE. berdasarkan table 6 hasil yang berasil diprediksi dengan adalah 388 *instances* untuk kelas Diterima*,* dan 0 *instances* untuk kelas Tidak Diterima*.* Jumlah *instances* yang berhasil diprediksi dengan benar pada skenario ini sebesar 388 dari total *instances* sebanyak 414. Sementara hasil *instances* yang tidak berhasil diprediksi dengan benar berjumlah 26 *instances.* Pada kelas Tidak Diterima terjadi *nois*,Karena distribusi data antar kelas yang tidak seimbang. Sehingga mempengaruhi *classifier* untuk mengklasifikasi data, khususnya pada kelas minoritas. Sedangkan hasil klasifikasi *C4.5* menggunakan SMOTE. Berdasarkan table 7 hasil *instances* yang berhasil diprediksi dengan benar adalah 370 *instances* untuk kelas Diterima*,* dan 368 *instances* untuk kelas Tidak Diterima*.* Jumlah *instances* yang berhasil diprediksi dengan benar pada skenario ini sebesar 738 dari total *instances* sebanyak 775. Sementara hasil *instances* yang tidak berhasil diprediksi dengan benar berjumlah 37 *instances*. Table berikut ini merupakan hasil akurasi, presisi,dan recall.

**Tabel 8.** Perbandingan Akurasi C4.5 dengan C4.5+SMOTE

|  |  |  |
| --- | --- | --- |
|  | *C4.5* | *C4.5* + SMOTE |
| Akurasi | 93.71% | 95.22% |
| *Preccision* | 93.70% | 95.22% |
| *Recall* | 93.70% | 95.20% |

Pada hasil Table 8 menunjukkan bahwa *C4.5* menggunakan SMOTE memiliki akurasi yang lebih baik dari pada *C4.5* tanpa SMOTE. *C4.5* dengan SMOTE memiliki akurasi terbaik sebesar 95.22%, sedangkan *C4.5* memiliki akurasi dengan nilai sebesar 93.71%**.** Pada hasil *preccision* menunjukkan bahwa *C4.5* dengan SMOTE memiliki *preccision* yang lebih baik dari pada *C4.5* tanpa SMOTE. *C4.5* dengan SMOTE memiliki precision terbaik sebesar 95.22%, sedangkan *C4.5* tanpa SMOTE memiliki precision dengan nilai sebesar 93.70%.Pada hasil *recall* menunjukkan bahwa *C4.5* dengan SMOTE memiliki *recall* yang lebih baik dari pada *C4.5* tanpa SMOTE. *C4.5* dengan SMOTE memiliki recall terbaik sebesar 95.20%, sedangkan *C4.5* tanpa SMOTE memiliki recall terbaik dengan nilai sebesar 93.70%

## **Pembahasan**

## Penelitian ini mendapatkan hasil bahwa metode *C4.5* dengan SMOTE mendapatkan hasil yang sangat baik dari pada metode *C4.5* tanpa SMOTE. Dari table 6 dapat dilihat bahwa pada klasifikasi *C4.5* tanpa SMOTE hasil akurasi yang baik tetapi terjadi nois pada kelas nimoritas(Tidak Diterima). Karena ketidak seimbangan kelas minoritas yang ada membuat hasil klasifikasi cenderung ke kelas mayoritas. Dan sedangkan dengan menggunakan *C4.5* dengan SMOTE tidak terjadi nois. Karena metode SMOTE mengatasi ketidak seimbangan kelas dan mengatasi nois, dengan cara menambah data baru pada kelas minoritas bardasarkan nilai tetangga terdekat sehingga memiliki sifat yang hampir sama dengan kelas minoritas. Data baru ditambahakan pada tahap SMOTE sebanyak jumlah kelas mayoritas, sehingga kelas label menjadi seimbang.

Dari hasil eksperimen yang didapat pada bebrapa table yang diatas dapat dilihat bahwa algoritma *C4.5* dengan SMOTE memiliki hasil yang lebih baik daripada *C4.5* tanpa SMOTE. Pada hasil klasifikasi *C4.5* dengan SMOTE mendapatkan nilai akurasi sebesar 95.22, *precision* 95.22% dan *recall* 95.20% dan sedangkan *C4.5* tanpa SMOTE mendapatkan nilai akurasi sebesar 93.71, *precision* 93.70% dan *recall* 93.70%.

Dari hasil diatas dapat disimpulkan bahwa *C4.5* dengan SMOTE memiliki akurasi yang lebih baik dibandingkan *C4.5* tanpa SMOTE, tetapi performa *C4.5* tanpa SMOTE sangat buruk dikarenakakn terjadi nois pada kelas minoritasnya. Hal itu disebabkan karena *C4.5* tanpa SMOTE memiliki distribusi data yang tidak seimbang dalam kelas label, sehingga lebih sering mengklasifikasikan kelas label Diterima dibandingkan menebak hasil yang kelas label Tidak Diterima dalam penerimaan SNMPTN. berbeda dengan hasil *C4.5* dengan SMOTE yang dimana hasilnya memiliki nilai akurasi, precision dan recall yang lebih tinggi, dikarenakan metode SMOTE yang membuat distribusi data menjadi seimbang dengan meningkatkan kelas minoritas sehingga metode *C4.5* dapat mengklasifikasikan data dengan baik.

IV. Kesimpulan

Untuk Hasil dari penelitian ini dapat disimpulkan bahawa penerapan metode SMOTE dapat meningkatkan kinerja algoritma yang digunakan sehingga mendapatkan nilai evaluasi yang meksimal. Perbandingan antara *C4.5* dengan SMOTE dengan *C4.5* tanpa SMOTE untuk kasus memprediksi jalur penerimaan SNMPTN pada penelitian ini cocok menggunakan algoritma *C4.5* hasil yang didapatkan bagus dikarenakan akurasi dan *Precision* 95.22% dan Recall 95.20%. Untuk penelitian selanjutnya dapat dikembangkan menggunakan algoritma yang lain dan diharapkan mampu medapatkan hasil akurasi yang lebih tinggi, sehingga implementasi untuk memprediksi jalur penerimaan SNMPTN bisa dimanfaatkan oleh para siswa sebagai bahan pertimbangan ketika ingin memilih perguruan tinggi negeri (PTN).

Daftar Pustaka

[1] A. T. Wibowo and D. Fitrianah, “A K-NEAREST ALGORITHM BASED APPLICATION TO PREDICT SNMPTN ACCEPTANCE FOR HIGH SCHOOL,” vol. 5, no. 01, pp. 9–20, 2018.

[2] L. Swastina, “Penerapan Algoritma C4 . 5 Untuk Penentuan Jurusan Mahasiswa,” vol. 2, no. 1, 2013.

[3] M. Fahmi, S., Fatma, I., “Implementasi Metode Electre Pada Sistem Pendukung Keputusan SNMPTN Jalur Undangan,” vol. 02, no. 02, pp. 88–101, 2015.

[4] S. E. Putri, R. Sovia, and A. F. Hadi, “Implementasi Sistem Pendukung Keputusan Untuk Pemilihan Jurusan Pada Jalur SNMPTN Dengan Metode Analitycal Hierarchy Process ( Studi Kasus : SMAN 3 Batusangkar ),” pp. 783–788, 2019.

[5] R. Siringoringo, “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan K-Nearest Neighbor,” *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[6] P. A. Santoso, A. P. Wibawa, and U. Pujianto, “Internship recommendation system using simple additive weighting,” vol. 2, no. 1, pp. 15–21, 2018.

[7] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, “Can Machine Learning explain Human Learning ?,” *Neurocomputing*, 2015.

[8] H. Y. Chen, C. H. Chuang, Y. J. Yang, and T. P. Wu, “Exploring the risk factors of preterm birth using data mining,” *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5384–5387, 2011.

[9] H. Ar, M. Palmerlee, and K. Chen, “Deploying learning materials to game content for serious education game development : A case study,” *Entertain. Comput.*, vol. 26, no. December 2017, pp. 1–9, 2018.

[10] M. D. Jaelani, A. P. Wibawa, and U. Pujianto, “Technology acceptance model of student ability and tendency classification system,” *Bull. Soc. Informatics Theory Appl.*, vol. 2, no. 2, pp. 47–57, 2018.

[11] A. S. B. Asmoro, W. S. G. Irianto, and U. Pujianto, “Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa,” *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, pp. 84–89, 2018.

[12] K. C. Davis, K. Janakiraman, A. Minai, and R. B. Davis, “Data Integration Using Data Mining Techniques,” no. January 2002, 2016.

[13] R. R. Deshmukh and V. Wangikar, “Data Cleaning: Current Approaches and Issues,” no. January 2011, 2015.

[14] K. Swati and S. Kumar, “A Comparative Study of Various Data Transformation Techniques in Data Mining,” vol. 148, no. 4, pp. 146–148, 2015.

[15] R. A. Mollineda, “Surrounding neighborhood-based SMOTE for learning from imbalanced data sets,” pp. 347–362, 2012.

[16] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE : Synthetic Minority Over-sampling Technique,” vol. 16, pp. 321–357, 2002.

[17] H. Li, D. Pi, and C. Wang, “The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE,” vol. 2014, 2014.

[18] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, “A comparative study of decision tree ID3 and C4 . 5,” no. 2, pp. 13–19.

[19] L. Rokach and O. Maimon, “Decision Trees,” no. January, 2005.

[20] R. Al-otaibi and R. B. C. Prudˆ, “Versatile Decision Trees for Learning,” no. September, 2015.

[21] Chen Jin, Luo De-lin, and Mu Fen-xiang, “An improved ID3 decision tree algorithm,” *2009 4th Int. Conf. Comput. Sci. Educ.*, no. 3, pp. 127–130, 2009.

[22] S. Hansun, “Penerapan Algoritma C4 . 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE,” no. March, 2017.

[23] P.-N. Tan, M. Steinbach, and Vipin Kumar, *Introduction to data mining*. 2006.

[24] M. Junker, R. Hoch, and A. Dengel, “On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy,” *Proc. Fifth Int. Conf. Doc. Anal. Recognit.*, 1999.

[25] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.