SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE UNTUK MENGATASI *IMBALANCE CLASS* (Studi Kasus : Data Kelulusan Universitas Widya Husada Semarang)

JURNAL ILMIAH

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Statistika

Oleh

GALUH FRIDAYANTI PITALOKA
B2A219006

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS MUHAMMADIYAH SEMARANG
2021

PERSETUJUAN PEMBIMBING

Skripsi dengan judul **"*Synthetic Minority Oversampling Technique* untuk Mengatasi *Imbalance Class* (Studi Kasus : Data Kelulusan Universitas Widya Husada Semarang)"** yang disusun oleh:

Nama : Galuh Fridayanti Pitaloka

NIM : B2A219006

Program Studi : S1 Statistika

Telah disetujui oleh dosen pembimbing pada tanggal : 24 September 2021

Pembimbing Utama	Pembimbing Pendamping
Dr. Rochdi Wasono, M.Si	Fatkhurokhman Fauzi, M.Stat
NIK. 28.6.1026.119	NIK. CP. 1026.101

PENGESAHAN KELULUSAN

Skripsi dengan judul **"*Synthetic Minority Oversampling Technique* untuk Mengatasi *Imbalance Class* (Studi Kasus : Data Kelulusan Universitas Widya Husada Semarang)"** yang disusun oleh:

Nama : Galuh Fridayanti Pitaloka
NIM : B2A219006
Program Studi : S1 Statistika

Telah dipertahankan dalam Sidang Panitia Ujian Skripsi Program Sarjana, Universitas Muhammadiyah Semarang Pada Tanggal 24 September 2021

Panitia Ujian
Ketua Tim Penguji

Indah Manfaati Nur, M.Si
NIK. 28.6.1026.221

Anggota Tim Penguji I	Anggota Tim Penguji II	Anggota Tim Penguji III
Tiani Wahyu Utami, M.Si	Dr. Rochdi Wasono, M.Si	Fatkhurokhman Fauzi, M.Stat
NIK. 28.6.1026.225	NIK. 28.6.1026.119	NIK. CP. 1026.101

Mengetahui,
Ketua Program Studi

Indah Manfaati Nur, M.Si
NIK. 28.6.1026.221

SURAT PERNYATAAN
PUBLIKASI KARYA ILMIAH

Nama	:	Galuh Fridayanti Pitaloka
NIM	:	B2A219040
Fakultas/Jurusan	:	S1 Statistika
Jenis Penelitian	:	Skripsi
Judul	:	*Synthetic Minority Oversampling Technique* untuk Mengatasi *Imbalance Class* (Studi Kasus : Data Kelulusan Universitas Widya Husada Semarang)
Email	:	gpitaloka21gmail.com

Dengan ini menyatakan bahwa saya menyetujui untuk:

1. Memberikan hak bebas royalti kepada perpustakaan Unimus atas penulisan karya ilmiah saya, demi pengembangan ilmu pengetahuan.
2. Memberikan hak menyimpan, mengalih mediakan atau mengalih formatkan, mengelola dalam bentuk pengakalan data (*database*), mendistribusikannya, serta menampilkannya dalam bentuk *soficopy* untuk kepentingan akademis kepada perpustakaan Unimus, tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta.
3. Bersedia dan menjamin untuk menanggung secara pribadi tanpa melibatkan pihak perpustakaan Unimus, dari semua bentuk tuntutan hukum yang timbul atas pelanggaran hak cipta dalam karya ilmiah ini.

Demikian pernyataan ini saya buat dengan sesungguhnya dan semoga dapat digunakan sebagaimana mestinya.

Semarang, 24 September 2021

Yang Menyatakan,

Galuh Fridayanti Pitaloka

NIM. B2A219006

Synthetic Minority Oversampling Technique untuk Mengatasi *Imbalance Class* (Studi Kasus : Data Kelulusan Universitas Widya Husada Semarang)

Galuh Fridayanti Pitaloka[1], Rochdi Wasono[2], Fatkhurokhman Fauzi[3]

[123]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang

Alamat e-mail : gpitaloka21@gmail.com

ABSTRAK

Perguruan Tinggi mempunyai peran strategis dalam pembangunan, khususnya pembangunan Sumber Daya Manusia. Penelitian ini memperhatikan adanya aspek ketidakseimbangan data antara mahasiswa yang lulus tepat waktu dan tidak tepat waktu. Pengujian yang dipakai yaitu *k-fold cross validation 10-fold* dengan perbandingan data *training* dan *testing* 70:30, 80:20 dan 90:10. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan metode *oversampling* kelas minoritas dengan menciptakan data buatan untuk mengatasi masalah ketidakseimbangan kelas data. Tujuan penelitian ini adalah membandingkan metode CART dan CHAID dengan penerapan SMOTE untuk Klasifikasi Data Kelulusan Mahasiswa. Variabel dependen yang digunakan adalah lama studi dengan variabel independen faktor-faktor yang berpengaruh terhadap lama studi. Berdasarkan pengujian yang telah dilakukan pada CART dan CHAID dengan penerapan SMOTE, hasil tingkat akurasi dengan metode CART sebesar 83,23%, sensitivitas 93,19% dan spesivitas 72,83%. Ketepatan klasifikasi dengan penerapan SMOTE, metode CART lebih baik dibandingkan dengan menggunakan metode CHAID.

Kata kunci : *Imbalance dataset*, CART, CHAID, SMOTE.

ABSTRACT

University has an important strategy in development, especially human resource development. This study noticed an aspect of the data imbalance between students who graduated on time and not. The test used is k-fold cross validation 10-fold with a comparison of training and testing data 70:30, 80:20 and 90:10. Synthetic Minority Oversampling Technique (SMOTE) is a method of oversampling minority classes by creating artificial data to solve the problem of data class imbalance. The aim of the study was to compare CART and CHAID methods with the application of SMOTE for the Classification of Student Graduation Data. The dependent variable used is the length of the study with independent variable factors that affect the length of the study. Based on the tests conducted on CART and CHAID with the implementation of SMOTE, the accuracy rate with the CART method is 83.23%, sensitivity is 93.19% and specificity is 72.83%. Accuracy of classification with the application of SMOTE, CART method is better than using the CHAID method.

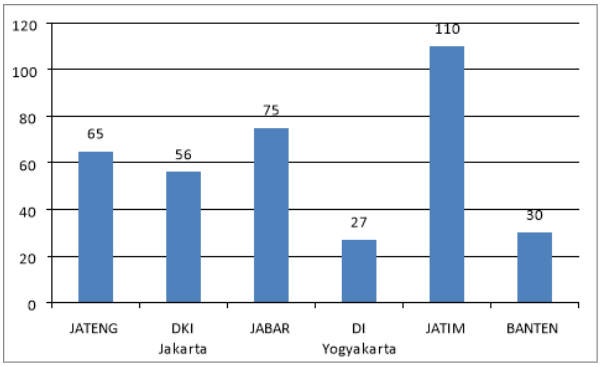
Keywords*: Imbalance dataset, CART, CHAID, SMOTE.*

PENDAHULUAN

Pendidikan merupakan kebutuhan mutlak yang harus dipenuhi sepanjang hayat. Tanpa pendidikan, manusia mustahil dapat hidup dan berkembang sejalan dengan cita-cita dan tujuan hidup. Begitu juga pentingnya peranan pendidikan dalam tata kehidupan pribadi maupun masyarakat, maka dalam pengembangan watak bangsa haruslah berpegang dan bertumpu pada landasan pendidikan yang kokoh. Sebagaimana termaktub pada pembukaan Undang-Undang Dasar negara Republik Indonesia tahun 1945, yaitu "…memajukan kesejahteraan umum, mencerdaskan kehidupan bangsa, dan ikut serta dalam perdamaian dunia." (Ansori, 2015)

Undang-Undang No. 20/2003 tentang Sistem Pendidikan Nasional menyatakan bahwa "Pembangunan nasional dalam bidang pendidikan merupakan upaya mencerdaskan kehidupan bangsa dan meningkatkan kualitas manusia Indonesia dalam mewujudkan masyarakat yang maju, adil dan makmur serta memungkinkan warga negaranya mengembangkan diri, baik berkenaan dengan aspek jasmaniah maupun rohaniah berdasarkan Pancasila dan UUD 1945". Selanjutnya, dijelaskan dalam UU No. 20/2003, bahwa "Pendidikan nasional bertujuan mengembangkan potensi peserta didik agar manusia yang beriman dan bertaqwa terhadap Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri dan menjadi warga negara yang demokratis serta bertanggungjawab dalam rangka mencerdaskan kehidupan bangsa".

Berdasarkan data buku statistik pendidikan tinggi Indonesia tercatat bahwa pulau Jawa mendominasi sebaran Perguruan Tinggi Negeri dan Perguruan Tinggi Swasta di lingkungan kemenristekdikti. Pada Pangkalan Data Pendidikan Tinggi laman https://forlap.kemdikbud.go.id/ yang diakses pada tanggal 30 Juni 2021 menunjukkan jumlah perguruan tinggi aktif baik negeri maupun swasta di pulau Jawa ada 363 yang terbagi dalam 6 provinsi. Provinsi Jawa Tengah berjumlah 65. Provinsi DKI Jakarta berjumlah 56. Jawa Barat berjumlah 75. DI Yogyakarta berjumlah 27. Jawa Timur 110. Banten berjumlah 30.



Gambar 1.1 Jumlah Perguruan Tinggi di Jawa

Dengan banyaknya perguruan tinggi negeri maupun perguruan tinggi swasta khususnya di Jawa, persaingan untuk mendapatkan jumlah mahasiswa baru semakin ketat. Dalam melihat perguruan tinggi mana yang akan dipilih, masyarakat sekarang sudah melihat status akreditasi perguruan tinggi. Masyarakat tentunya akan melihat perguruan tinggi yang nilai akreditasinya tinggi.

Saat ini dalam dunia pendidikan data yang berlimpah dan berkelanjutan bisa dimanfaatkan untuk data mining dalam rangka pengelolaan yang lebih baik dan pelaksanaan pembelajaran yang lebih efektif. Pada institusi pendidikan Perguruan Tinggi, data mahasiswa dan data jumlah kelulusan mahasiswa dapat menghasilkan informasi yang berlimpah berupa jumlah kelulusan setiap tahunnya, profil dan hasil akademik mahasiswa selama menempuh proses kegiatan belajar mengajar di perguruan tinggi. Adanya informasi mengenai lama studi mahasiswa tentu akan menjadi pendukung suatu pengambilan keputusan yang tepat bagi manajemen Perguruan Tinggi dalam mengambil langkah berikutnya.

Perguruan tinggi mempunyai peran strategis dalam pembangunan, khususnya pembangunan sumber daya manusia. Peran perguruan tinggi dalam pengembangan sumber daya manusia terutama sebagai penghasil agen-agen perubahan yang mampu merancang, mendorong dan mempelopori perubahan. Perguruan tinggi adalah pencipta dan pendukung gagasan-gagasan baru untuk disumbangkan bagi kemajuan intelektual dan sosial masyarakat. Sejalan dengan hal itu, maka perguruan tinggi perlu melakukan perubahan, baik dalam arah serta tujuan perguruan tinggi yang menyangkut aspek kualitas, yang tercermin pada para alumninya. Lulus dengan baik dan tepat waktu merupakan salah satu parameter keberhasilan proses pembelajaran.

Kebutuhan masyarakat terhadap pelayanan kesehatan yang optimal, semakin terbukanya peluang kerja baik di dalam maupun di luar negeri dan kebutuhan akan tenaga kesehatan yang profesional oleh penyedia layanan kesehatan, menjadi tantangan yang harus dijawab oleh penyedia tenaga kesehatan. Fakultas Kesehatan dan Keteknisian Medik (FKKM) merupakan salah satu fakultas di Universitas Widya Husada Semarang yang mencoba merealisasikan kebutuhan masyarakat tersebut. FKKM saat ini memiliki enam program studi diploma tiga yang terdiri dari Teknik Elektromedik, Teknik Rontgen, Kebidanan, Fisioterapi, Keperawatan dan Refraksi Optisi. Berbeda dengan jenjang pendidikan dasar dan menengah, pada tingkat perguruan tinggi terutama program diploma III memiliki syarat kelulusan bagi setiap mahasiswa yaitu telah menempuh minimal 108 SKS (Panduan Akademik, 2020). Lama studi mahasiswa ditetapkan dapat ditempuh dalam kurun waktu 3 tahun atau 6 semester dengan batas maksimal adalah 5 tahun atau 10 semester. Masa studi dihitung sejak pertama kali terdaftar sebagai mahasiswa.

Berbagai hal dalam dunia pendidikan mulai dari penerimaan mahasiswa baru hingga kelulusan dan evaluasi kinerja mahasiswa merupakan instrumen yang harus dilakukan agar terwujudnya pendidikan yang baik dan memenuhi standar pendidikan. Pendidikan tinggi dengan jumlah mahasiswa yang besar perlu memperhatikan kinerja mahasiswa baik dari penerimaan mahasiswa baru sebagai input maupun lulusan mahasiswa sebagai output. Pengambilan keputusan dalam evaluasi kinerja mahasiswa dapat dilakukan dengan penambangan data pendidikan (*educational data mining*) salah satunya dengan klasifikasi yang tepat untuk mempermudah mengevaluasi kinerja mahasiswa.

Pada perkembangan terbaru, teknik-teknik yang terdapat di dalam data mining mulai banyak digunakan. Khususnya teknik *decision tree* telah menjadi teknik yang populer karena *decision tree* yang dihasilkan mudah diinterpretasikan dan divisualisasikan (Chye, 2004). Metode klasifikasi telah banyak digunakan dalam berbagai bidang, seperti bidang pendidikan, pemerintahan, kesehatan, teknologi, maupun sosial. Klasifikasi sendiri didefinisikan sebagai pekerjaan mengelompokkan suatu objek ke dalam kategori tertentu. Klasifikasi dapat dilakukan pada data kategorik maupun bukan, jika data bukan kategorik maka harus diubah dalam bentuk kategorik terlebih dahulu.

Classification and Regression Tree (CART) dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone sekitar tahun 1980-an (Breiman et al., 1993). CART merupakan metodologi statistik nonparametrik yang dikembangkan untuk topik analisis klasifikasi, baik untuk variabel dependen kategorik maupun kontinu. CART menghasilkan suatu pohon klasifikasi jika variabel dependennya kategorik, dan menghasilkan pohon regresi jika variabel dependennya kontinu. (Komalasari, 2007)

CART dapat menyeleksi variabel-variabel dan interaksi-interaksi variabel yang paling penting dalam menentukan hasil atau variabel dependennya. Tujuan utama CART adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. Metode ini merupakan metode yang bisa diterapkan untuk himpunan data yang mempunyai jumlah besar, variabel yang sangat banyak dan dengan skala variabel campuran melalui prosedur pemilahan biner.

Chi-square Automatic Interaction Detection (CHAID) merupakan salah satu analisis pohon keputusan (*decision tree*). Keunggulan dari metode *decision tree* yaitu membutuhkan waktu yang cepat untuk membentuk diagram pohon, representasi visual, dan mudah diinterpretasikan (Swain, 2016). Metode pohon keputusan (*decision tree*) mempunyai beberapa keunggulan dibandingkan metode lainnya untuk klasifikasi atau prediksi, seperti *neural network* dan analisis diskriminan (Cha, G.W *et al*, 2017). CHAID adalah suatu teknik iteratif yang menguji satu persatu variabel independen yang digunakan dalam klasifikasi dan menyusunnya berdasarkan pada tingkat signifikansi statistik *chi-square* terhadap variabel dependennya (Gallagher *et al*, 2000). Dengan kata lain, CHAID mengklasifikasikan variabel dependen kategori ke dalam kategori tertentu berdasarkan statistik *chi-square* variabel independen terhadap variabel dependen.

Metode CHAID merupakan salah satu tipe dari metode *Automatic Interaction Detection* (AID). Metode AID adalah suatu teknik untuk menganalisis kelompok data berukuran besar dengan membaginya menjadi sub-sub kelompok yang tidak saling tumpang tindih (Kass, 1982, diacu dalam Soemartojo 2002) yang

diperuntukkan bagi data dengan peubah penjelas berskala ratio atau interval. Metode CHAID merupakan teknik eksplorasi nonparametrik untuk menganalisis sekumpulan data yang berukuran besar dan cukup efisien untuk menduga variabel independen yang paling signifikan terhadap peubah dependen. Interaksi antar peubah juga dapat dideteksi melalui metode ini (Du toit et al. 1977), sehingga diharapkan metode CHAID dapat menjadi alternatif analisis untuk menduga faktor-faktor yang menentukan suatu dependen kategorik.

Beberapa penelitian tentang metode CHAID dan CART telah dilakukan sebelumnya. Berdasarkan penelitian Siahaan, dkk (2016) dengan judul "Aplikasi Classification and Regression Tree (CART) dan Regresi Logistik Ordinal dalam Bidang Pendidikan". Variabel yang digunakan dalam penelitian ini adalah jenis kelamin, asal daerah, program studi, status sekolah menengah, lama studi. Hasil perbandingan kedua metode menunjukkan bahwa regresi ordinal lebih baik yaitu mempunyai tingkat keakuratan klasifikasi 65% dibandingkan CART yang memiliki tingkat keakuratan klasifikasi 54,9%.

Kemudian penelitian oleh Darmawan, dkk (2017) dengan judul "Klasifikasi Lama Masa Studi Mahasiswa Menggunakan Perbandingan Metode Algoritma C.45 dan Algoritma Classification and Regression Tree". Variabel yang digunakan dalam penelitian ini adalah IPK, jenis kelamin, asal daerah, program studi, asal sekolah dan lama studi. Hasil perbandingan kedua metode, kinerja algoritma CART lebih baik yaitu menghasilkan tingkat akurasi 60% dibandingkan algoritma C4.5 yang menghasilkan tingkat akurasi 40%.

Penelitian lain juga dilakukan Suniantara dan Muhammad Rusli (2017) dengan judul "Klasifikasi Waktu Kelulusan Mahasiswa STIKOM Bali menggunakan CHAID Regression Tree dan Regresi Logistik Biner". Variabel yang digunakan dalam penelitian ini adalah jurusan, IPK, IPS semester 6, lama menyusun skripsi, jenis kelamin, nilai ujian masuk dan lama studi. Hasil perbandingan kedua metode menunjukkan bahwa kinerja algoritma CHAID lebih baik yaitu menghasilkan tingkat akurasi 91,2% dibandingkan regresi logistik yang menghasilkan tingkat akurasi 90,2%.

Selanjutnya penelitian oleh Wibowo, dkk (2019) dengan judul "Komparasi Algoritma *Naive Bayes* dan *Decision Tree* Untuk Memprediksi Lama Studi Mahasiswa". Variabel yang digunakan dalam penelitian ini adalah gender, status mahasiswa, nilai, beasiswa dan lama studi. Dari komparasi dua metode menunjukkan bahwa *Decision Tree* lebih baik yaitu memiliki tingkat akurasi 55% dibandingkan Naive Bayes yang memiliki tingkat akurasi 30%.

Berbeda dengan penelitian sebelumnya, penelitian ini memperhatikan adanya aspek ketidakseimbangan data. *Class imbalance* atau ketidakseimbangan kelas merupakan salah satu permasalahan pada data mining. Hal ini terjadi pada saat kelas minoritas jauh lebih kecil atau lebih jarang dari kelas mayoritas (Syukron dan Subekti, 2018 dalam Bawono, Bonggo dan Rochdi Wasono. 2019). Data dikatakan tidak seimbang karena mahasiswa yang lulus tidak tepat waktu jumlahnya jauh lebih sedikit dibandingkan mahasiswa yang lulus tepat waktu. Kasus seperti ini dapat mengakibatkan salah klasifikasi pada kelas minoritas, yaitu kelas dengan jumlah amatan yang jauh lebih sedikit (Bunkhumpornpat et al, 2012). Sehingga mahasiswa yang seharusnya lulus tidak tepat waktu akan diprediksi lulus tepat waktu. Oleh karena itu penanganan pada data tidak seimbang perlu dilakukan. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan salah satu metode penanganan yang dapat digunakan. SMOTE merupakan metode *oversampling* kelas minoritas dengan menciptakan data buatan (sintetik) untuk mengatasi masalah ketidakseimbangan kelas data (Chawla et al. (2002)).

Penelitian mengenai SMOTE sebelumnya telah dilakukan Sulistiyowati dan Mohamad Jajuli (2020) dengan judul Integrasi Naive Bayes dengan Teknik Sampling SMOTE untuk Menangani Data Tidak Seimbang. Dari total 878 data dengan atribut yang digunakan terbagi menjadi 813 data mayoritas (93%) dan 65 data minoritas (7%). Data tersebut menunjukkan adanya ketidakseimbangan data diantara kedua kelas. Tahapan modify melakukan proses SMOTE 500%. Berdasarkan hasil evaluasi akurasi menunjukkan bahwa akurasi Naive Bayes dengan SMOTE lebih tinggi dibandingkan Naive Bayes tanpa SMOTE dengan selisih 0,855%. Sedangkan berdasar hasil evaluasi G-Mean menunjukkan bahwa nilai G-Mean Naive Bayes dengan SMOTE lebih tinggi dibandingkan Naive Bayes tanpa SMOTE dengan selisih 0,028.

Penelitian lain juga dilakukan Franseda (2020) dengan judul Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas. Berdasarkan hasil penelitian menunjukkan bahwa model Decision Tree dan SMOTE Split Data dengan akurasi 71,12%, presisi 89,71 dan AUC 0,773. Penelitian ini masuk ke dalam kategori fair classification (cukup).

Berdasarkan uraian tersebut penelitian ini membahas mengenai klasifikasi data kelulusan mahasiswa berdasarkan lama studi menggunakan metode *Chi-Squared Automatic Interaction Detection* (CHAID) dan *Classification and Regression Tree* (CART). Pramana (2018) menyebutkan bahwa CART dan CHAID merupakan contoh klasifikasi, bagian dari algoritma data mining untuk melakukan ekstraksi pengetahuan dari suatu data. Kedua metode tersebut menarik untuk digunakan dikarenakan kondisi variabel terikat yang kategorik dengan dua kategori atau biasa disebut biner. Sehingga topik tersebut dirasa perlu dikaji untuk menentukan metode mana yang lebih efisien dan sesuai dengan kebutuhan penelitian. Dari penelitian ini akan diketahui seberapa besar perbedaan output yang dihasilkan dari kedua metode.

TINJAUAN PUSTAKA

1. Lama Studi Mahasiswa

Dalam Undang-Undang Republik Indonesia nomor 12 tahun 2012, mahasiswa adalah peserta didik pada jenjang pendidikan tinggi. Menurut buku panduan akademik (2020) pada tingkat perguruan tinggi terutama program diploma III memiliki syarat kelulusan bagi setiap mahasiswa yaitu telah menempuh minimal 108 SKS. Lama studi mahasiswa ditetapkan dapat ditempuh dalam kurun waktu 3 tahun atau 6 semester dengan batas maksimal adalah 5 tahun atau 10 semester. Masa studi dihitung sejak pertama kali terdaftar sebagai mahasiswa.

2. Statistika Deskriptif

Statistika deskriptif adalah bagian dari statistika yang mempelajari cara pengumpulan data dan penyajian data sehingga mudah dipahami. Statistika deskriptif hanya berhubungan dengan hal menguraikan atau memberikan keterangan-keterangan mengenai suatu data atau keadaan (Hasan, 2004).

3. Data Mining

Data mining merupakan suatu komponen dari *knowledge discovery* dalam proses *database* dengan menggunakan alat algoritma dimana pola-polanya diekstrak dan disebutkan satu demi satu dari data yang ada (Growe, 1999). Secara singkat, *data mining* adalah sebuah proses penggalian pola dari data. *Data mining* menjadi hal yang sangat penting dalam mengubah data menjadi informasi. *Data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan (Pramana, 2018) yaitu:

a. Estimasi yaitu peran data mining untuk menghitung nilai kira-kira dari suatu objek baru.
b. Prediksi yaitu peran data mining untuk "meramalkan" di masa depan nilai kira-kira dari suatu data.
c. Klasifikasi yaitu peran data mining untuk memberikan label dari suatu data/objek baru.
d. Clustering yaitu dapat digunakan untuk mengidentifikasi klaster / daerah yang padat, pola-pola distribusi objek secara umum, dan keterkaitan yang menarik antar atribut objek.
e. Asosiasi yaitu hubungan antar atribut dengan atribut yang lain yang muncul bersamaan.

4. Klasifikasi

Supervised learning, disebut juga sebagai analisis klasifikasi, merupakan suatu teknik statistik yang bertujuan untuk mengelompokkan data ke dalam kelas-kelas yang telah memiliki label dengan membangun suatu model yang berdasarkan kepada suatu data training serta memprediksi kelas dari suatu data baru. Teknik klasifikasi dalam data mining dikelompokkan ke dalam beberapa kelompok, yaitu *linear discriminant analysis*, *k-nearest neighbour*, *decision trees*, *random forrest*, *support vector machine*, *naive bayes*, *logistic regression*, dan lain-lain (Pramana, 2018).

5. Evaluasi Klasifikasi

Performa dari setiap model klasifikasi dapat dievaluasi dengan menggunakan perhitungan statistik, yaitu keakuratan klasifikasi, sensitivitas, dan spesifitas. Ketiganya ditentukan oleh True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Ketika kita melakukan tes, hasil prediksinya adalah *true* dan ternyata nilai aktualnya adalah *true*, maka disebut *true positive*. Sedangkan ketika kita melakukan tes, hasil prediksinya adalah *false* dan ternyata nilai aktualnya adalah *false*, maka disebut *true negative*. Berikut ini merupakan sebuah matriks yang menunjukkan TP, TN, FP, dan FN, yang biasa dikenal dengan sebutan *confusion matrix*.

Tabel 2.1 *Confusion Matrix*

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Keterangan :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

A. Akurasi

Akurasi merupakan total keseluruhan seberapa sering model benar mengklasifikasi. Nilai akurasi dapat dihitung dengan rumus sebagai berikut.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Sensitivitas

Sensitivitas memperhitungkan proporsi true positive, dimana kemampuan sistem dalam memprediksi nilai yang benar ini ditunjukkan. Berikut adalah formula dalam menghitung sensitivitas :

$$Sensitivitas = \frac{banyaknya\ true\ positive}{banyaknya\ aktual\ positive} = \frac{TP}{TP + FN}$$

C. Spesivitas

Spesivitas memperhitungkan proporsi true negative, dimana kemampuan sistem dalam memprediksi nilai yang benar untuk keadaan yang berlawanan dengan keinginan. Berikut adalah formula dalam menghitung spesivitas :

$$Spesivitas = \frac{banyaknya\ true\ negative}{banyaknya\ aktual\ negative} = \frac{TN}{TN + FP}$$

6. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu metode oversampling yaitu teknik pengambilan sampel untuk meningkatkan jumlah data pada kelas positif dengan cara mereplikasi jumlah data pada kelas positif secara acak sehingga jumlahnya sama dengan data pada kelas negatif. Algoritma SMOTE pertama kali ditemukan oleh Chawla (2002). Pendekatan ini bekerja dengan membuat "*synthetic*" data, yaitu data replikasi dari data minor. Metode SMOTE bekerja dengan mencari *k nearest neighbors* (ketetanggaan data). Teknik ini termasuk dalam kelompok klasifikasi non parametrik. Mirip dengan *clustering*, teknik ini sangat sederhana dan mudah untuk diimplementasikan. Teknik ini bekerja dengan mengelompokkan data berdasarkan tetangga terdekat. Tetangga terdekat dipilih berdasarkan jarak *euclidean* antara kedua data. Misalkan diberikan dua data dengan p dimensi yaitu $x^T = \lfloor x_1,x_2,...,x_p \rfloor$ dan $y^T = \lfloor y_1,y_2,...,y_p \rfloor$ maka jarak euclidean $d(x,y)$ antara kedua vektor data adalah sebagai berikut,

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \quad (2.4)$$

Sedangkan synthetic data dilakukan dengan menggunakan persamaan berikut,

$$x_{syn} = x_i + (x_{knn} - x_i)\ x\ \beta, i = 1,2,...,n \quad (2.5)$$

dengan,

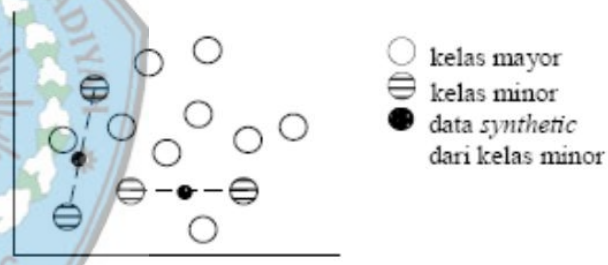
x_{syn} = data hasil replikasi

x_i = data yang akan direplikasi

x_{knn} = data yang memiliki jarak terdekat dari data yang akan direplikasi

β = bilangan random antara 0 sampai 1

Ilustrasi distribusi data setelah diterapkan metode SMOTE dapat dilihat pada Gambar 2.1.



Gambar 2.1 Ilustrasi Algoritma SMOTE

7. *Classification and Regression Trees* (CART)

Clasification and Regression Trees (CART) merupakan salah satu metode atau algoritma dari teknik pohon keputusan (*decision tree*). Metode yang dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone ini merupakan teknik klasifikasi dengan menggunakan algoritma penyekatan rekursif secara biner (*binary recursive partitioning*) (Roger dan Lewis, 2000). Istilah "*binary*" berarti pemilahan dilakukan pada sekelompok data yang terkumpul dalam suatu ruang yang disebut simpul (*node*) menjadi dua kelompok yang disebut simpul anak (*child nodes*). Istilah "*recursive*" berarti prosedur penyekatan secara biner dilakukan secara berulangulang. Setiap simpul anak yang diperoleh dari penyekatan simpul awal kemudian bisa dipilah kembali menjadi dua simpul anak lagi, dan begitu seterusnya hingga

memenuhi kriteria tertentu. Sedangkan istilah "*partitioning*" memiliki arti bahwa proses klasifikasi dilakukan dengan cara memilah suatu kumpulan data menjadi beberapa bagian atau partisi.

Menurut Breiman, et al (1993), CART akan menghasilkan pohon klasifikasi jika variabel respon mempunyai skala kategorik dan akan menghasilkan pohon regresi jika variabel respon berupa data kontinu. Tujuan utama CART adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. Metode pengklasifikasian CART memiliki beberapa kelebihan. Pertama, CART merupakan metode nonparametrik sehingga tidak ada asumsi distribusi variabel prediktor yang perlu dipenuhi. Kedua, CART tidak hanya memberikan klasifikasi, namun juga estimasi probabilitas kesalahan pengkalsifikasian. Ketiga, metode ini memudahkan dalam hal eksplorasi dan pengambilan keputusan pada struktur data yang kompleks dan multivariabel karena struktur data dapat dilihat secara visual. Keempat, hasil klasifikasi akhir berbentuk sederhana dan mengklasifikasikan data baru secara efisien. Kelima, kemudahan dalam menginterpretasi hasil.

Menurut Sartono dan Syafitri (2010), metode Classification dan Regression Tree (CART) merupakan metode yang memiliki kemampuan dalam memberikan kemudahan untuk menginterpretasikan hasil analisis dan memberikan dugaan dengan tingkat kesalahan kecil.

Algoritma CART melalui tiga tahapan, yaitu pembentukan pohon klasifikasi, pemangkasan pohon klasifikasi dan penentuan pohon klasifikasi optimum.

8. *Chi-square Automatic Interaction Detection* (CHAID)

Metode CHAID (*Chi-square Automatic Interaction Detection*) diperkenalkan oleh Dr. G. V. Kass pada tahun 1980, melalui sebuah artikel yang berjudul "*An Exploratory Technique for Investigating Large Quantities of Categorical Data*". Metode CHAID merupakan pengembangan dari metode yang sudah ada sebelumnya, yaitu *Automatic Interaction Detection* (AID). CHAID adalah sebuah analisis berdasarkan variabel kategori (Perez & Cejas, 2016). Menurut Gallagher (2000), CHAID merupakan suatu teknik iteratif yang menguji satu-persatu variabel independen yang digunakan dalam klasifikasi, dan menyusunnya berdasarkan pada tingkat signifikansi statistik uji *chi-square* terhadap variabel dependen.

CHAID digunakan untuk membentuk segmentasi yang membagi data menjadi dua atau lebih kelompok yang berbeda berdasarkan sebuah kriteria (variabel independen). Pada setiap tahap, CHAID memilih variabel independen yang mempunyai interaksi paling kuat dengan variabel dependen. kategori dari setiap variabel independen digabungkan jika mereka tidak signifikan berbeda terhadap variabel dependen (Cinca & Nieto, 2016). Hal ini kemudian diteruskan dengan membagi kelompok-kelompok tersebut menjadi kelompok yang lebih kecil berdasarkan variabel independen yang lain. Proses tersebut terus berlanjut sampai tidak ditemukan lagi variabel independen yang signifikan secara statistik (Kunto & Hasana, 2006).

9. *K-Fold Cross Validation*

K-Fold Cross Vaidation adalah metode validasi dengan membagi data ke dalam k-subset, kemudian melakukan pengulangan sebanyak k kali untuk training dan testing. Pada setiap pengulangan, digunakan satu subset sebagai data testing dan subset lainnya sebagai data training. Keuntungan dari metode ini adalah setiap data, minimal akan menjadi data uji sebanyak satu kali dan akan menjadi data learning juga minimal satu kali (Widjaya, 2017).

Gambar 2.4 Contoh iterasi data dengan *cross validation*

Kinerja dari *K-fold cross validation* yaitu :

1. Total *instance* dibagi menjadi N bagian
2. *Fold* ke-1 adalah ketika bagian ke-1 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). Selanjutnya, hitung akurasi atau kesamaan atau kedekatan suatu hasil pengukuran dengan angka atau data yang sebenarnya berdasarkan porsi data tersebut. Perhitungan akurasi tersebut menggunakan persamaan sebagai berikut.

$$akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} X\ 100\%$$

3. *Fold* ke-2 adalah ketika bagian ke-2 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai fold ke-k. Hitung rata-rata akurasi dari k buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

Pada gambar di atas terlihat bahwa tiap percobaan akan menggunakan satu data testing dan k-1 bagian akan menjadi data testing, kemudian data testing itu akan ditukar dengan satu buah data training sehingga setiap percobaan akan didapatkan data testing yang berbeda-beda.

METODE PENELITIAN

1. Sumber Data

Pada penelitian ini, data yang digunakan adalah data alumni Fakultas Kesehatan dan Keteknisian Medik Universitas Widya Husada Semarang tahun 2018 sampai dengan 2020 untuk diklasifikasikan menggunakan metode analisis CART dan CHAID dengan penerapan SMOTE. Dalam penelitian ini data sekunder diperoleh dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) .

2. Variabel Penelitian

Pada penelitian kali ini variabel dependen Y dikategorikan menjadi 2 yaitu bernilai 0 untuk kategori lama studi tidak tepat waktu (alumni yang lulus > 6 semester) dan 1 untuk kategori masa studi tepat waktu (alumni yang lulus ≤ 6 semester).

Adapun variabel independen pada penelitian ini meliputi:

a. Jenis Kelamin (X_1)
b. Asal Daerah (X_2)
c. IPK (X_3)
d. Usia (X_4)
e. Program Studi (X_5)
f. Organisasi (X_6)

3. Langkah Penelitian

A. Langkah-langkah yang dilakukan pada CART sebagai berikut :

1. Menginput data
2. Melakukan analisis deskriptif
3. Pra-pemrosesan data menggunakan SMOTE pada kelas minor
4. Melakukan validasi dengan 10-*fold cross validation*
5. Membagi data menjadi data training dan testing (perbandingan data training dan testing 90:10, 80:20, 70:30)
6. Pembentukan pohon klasifikasi
7. Pemangkasan pohon klasifikasi
8. Penentuan Pohon Klasifikasi Optimal
9. Melakukan pengujian nilai akurasi, spesifitas, sensitivitas menggunakan *Confussion Matrix*

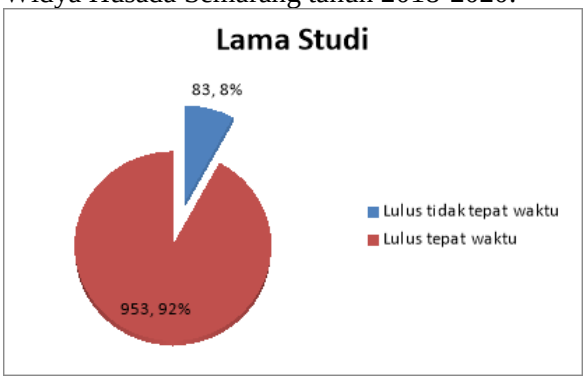
B. Langkah-langkah yang dilakukan pada CHAID sebagai berikut :

1. Menginput data
2. Melakukan analisis deskriptif
3. Pra-pemrosesan data menggunakan SMOTE pada kelas minor
4. Melakukan validasi dengan 10-*fold cross validation*
5. Membagi data menjadi data training dan testing (perbandingan data training dan testing 90:10, 80:20, 70:30)
6. Penentuan variabel dependen dan independen
7. Uji *chi-square*
8. Koreksi Bonferonni
9. Menerapkan tiga langkah analisis CHAID yaitu langkah penggabungan, pemisahan, dan peberhentian.
10. Melakukan pengujian nilai akurasi, spesifitas, sensitivitas menggunakan *Confussion Matrix*

HASIL DAN PEMBAHASAN

1. Analisis Deskriptif

Analisis deskriptif digunakan untuk memperoleh gambaran data secara umum. Berikut adalah gambaran umum dari data kelulusan Universitas Widya Husada Semarang tahun 2018-2020.



Gambar 4.1 Persentase Kelulusan Berdasarkan Lama Studi

Pada gambar 4.1 diketahui jumlah data kelulusan Universitas Widya Husada Semarang tahun 2018 sampai dengan tahun 2020 sebanyak 1036 mahasiswa dengan 83 mahasiswa lulus tidak tepat waktu (8%) dan 953 mahasiswa lulus tepat waktu (92%). Hal ini menunjukkan bahwa status

mahasiswa yang lulus tidak tepat waktu dan mahasiswa yang lulus tepat waktu memiliki jumlah yang tidak seimbang(*imbalance*).

2. Uji Independensi

Uji independensi dilakukan untuk mengetahui adanya hubungan antara variabel independen dengan variabel dependen. Pengujian independensi menggunakan uji *Chi-Square* dengan hipotesis uji sebagai berikut :

H_0 : Tidak terdapat hubungan antara variabel independen dengan variabel dependen

H_1 : Terdapat hubungan antara variabel independen dengan variabel dependen

Hasil pengujian independensi antara variabel prediktor dengan variabel respon dapat dilihat pada tabel 4.1

Tabel 4.1 Hasil Uji Independensi

variabel independen	sig.	chi hitung	Keputusan
JK (X_1)	0,527	0,399	Terima Ho
Asal Daerah (X_2)	0,416	0,66	Terima Ho
IPK (X_3)	0,000	175,326	Tolak Ho
Usia (X_4)	0,858	0,032	Terima Ho
prodi (X_5)	0,000	87,505	Tolak Ho
organisasi (X_6)	0,000	22,15	Tolak Ho

Tabel 4.1 menunjukkan bahwa sebanyak 3 variabel independen memiliki nilai signifikansi kurang dari taraf signifikansi (nilai *alpha*) yang ditentukan sebesar 0,05, sehingga berdasarkan pengujian hipotesis, ketiga variabel tersebut menolak hipotesis awal, maka dapat dikatakan bahwa ketiga variabel tersebut memiliki hubungan dengan variabel dependen. Berdasarkan pengujian independensi IPK (X_3), program studi (X_5) dan organisasi (X_6) memiliki hubungan yang signifikan dengan lama studi.

3. *Synthetic Minority Oversampling Technique* (SMOTE)

Pada gambar 4.1 menunjukkan bahwa jumlah mahasiswa lulus tepat waktu lebih banyak daripada yang lulus tidak tepat waktu, yakni sebanyak 953 mahasiswa lulus tepat waktu sedangkan 83 mahasiswa lulus tidak tepat waktu. Sehingga perlu dilakukan pra-pemrosesan dengan menggunakan metode SMOTE guna menyeimbangkan jumlah anggota kelas minor, dalam penelitian ini adalah mahasiswa lulus tidak tepat waktu.

Tabel 4.2 Perbandingan Data Awal dengan Data SMOTE

	Lulus tidak tepat waktu	Lulus tepat waktu	Jumlah
Data Awal	83 (8%)	953 (92%)	1036 (100%)
SMOTE 11X	913 (49%)	953 (51%)	1866 (100%)

Pada tabel 4.2 terjadi penambahan data kelas lulus tidak tepat waktu dari yang awalnya 83 data menjadi 913 data hampir mendekati keadaan *balanced data.* Persentase awal jumlah data pada kelas lulus tidak tepat waktu sebesar 8% ditambahkan data buatan melalui tahap SMOTE dengan *oversampling* sebanyak 11 kali, sehingga persentase kelas lulus tidak tepat waktu menjadi 49%. Jumlah data keseluruhan menjadi 1866 setelah melalui tahap SMOTE.

4. Klasifikasi CHAID dan CART dengan penerapan SMOTE

Tingkat akurasi diperoleh dari jumlah data yang kelasnya diprediksi dengan tepat oleh setiap model. Tingkat akurasi pada penelitian ini berbeda-beda untuk setiap model yang digunakan. Persentase perbandingan antara data latih data data uji yang digunakan setiap model yaitu 70:30, 80:20 dan 90:10.

4.1 CHAID dengan penerapan SMOTE pada data latih : data uji 70 : 30

CHAID dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*. Berikut ini adalah tabel 4.3 *confusion matrix* untuk data latih : data uji dengan persentase 70:30

Tabel 4.3 *CHAID* dengan penerapan *SMOTE data latih:data uji 70:30*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	665	66	90,97%
Pred 1	248	887	78,15%
Class Recall	72,84%	93,07%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{887+665}{887+665+248+66} = \frac{1552}{1866} = 0{,}8317 \approx 83{,}17\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{887}{887+66} = \frac{887}{953} = 0{,}9308 \approx 93{,}08\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{665}{665+248} = \frac{665}{913} = 0{,}7283 \approx 72{,}83\%$$

Berdasarkan tabel 4.3 perhitungan dengan data latih 1306 data dan data uji 560 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 66 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 248 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.3 diperoleh hasil perhitungan ketepatan klasifikasi sebesar 83,17%, nilai sensitivitas sebesar 93,08% dan nilai spesivitas sebesar 72,83%.

4.2 CHAID dengan penerapan SMOTE pada data latih : data uji 80 : 20

CHAID dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*. Berikut ini adalah tabel 4.4 *confusion matrix* untuk data latih : data uji dengan persentase 80:20

Tabel 4.4 *CHAID* dengan penerapan *SMOTE data latih:data uji 80:20*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	665	69	90,60%
Pred 1	248	884	78,09%
Class Recall	72,84%	92,76%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{884+665}{884+665+248+69} = \frac{1549}{1866} = 0{,}8301 \approx 83{,}01\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{884}{884+69} = \frac{884}{953} = 0{,}9277 \approx 92{,}77\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{665}{665+248} = \frac{665}{913} = 0{,}7283 \approx 72{,}83\%$$

Berdasarkan tabel 4.4 perhitungan dengan data latih 1493 data dan data uji 373 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 69 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 248 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.4 diperoleh hasil perhitungan ketepatan klasifikasi sebesar 83,01%, nilai sensitivitas sebesar 92,77% dannilai spesivitas sebesar 72,83%.

4.3 CHAID dengan penerapan SMOTE pada data latih : data uji 90 : 10

CHAID dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*. Berikut ini adalah tabel 4.5 *confusion matrix* untuk data latih : data uji dengan persentase 90:10

Tabel 4.5 *CHAID* dengan penerapan *SMOTE data latih:data uji 90:10*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	661	64	91,17%
Pred 1	252	889	77,91%
Class Recall	72,40%	93,28%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{889+661}{889+661+252+64} = \frac{1550}{1886} = 0{,}8307 \approx 83{,}07\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{889}{889+64} = \frac{889}{953} = 0{,}9329 \approx 93{,}29\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{661}{661+252} = \frac{661}{913} = 0{,}7239 \approx 72{,}39\%$$

Berdasarkan tabel 4.5 perhitungan dengan data latih 1679 data dan data uji 187 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 64 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 252 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.5 diperoleh hasil perhitungan ketepatan klasifikasi sebesar 83,07%, nilai sensitivitas sebesar 93,29% dan nilai spesivitas sebesar 72,39%.

4.4Klasifikasi CART dengan Penerapan SMOTE

Tingkat akurasi diperoleh dari jumlah data yang kelasnya diprediksi dengan tepat oleh setiap model. Tingkat akurasi pada penelitian ini berbeda-beda untuk setiap model yang digunakan. Persentase perbandingan antara data latih data data uji yang digunakan setiap model yaitu 70:30, 80:20 dan 90:10.

4.5 CART dengan penerapan SMOTE pada data latih : data uji 70 : 30

CART dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*.

Berikut ini adalah tabel 4.6 *confusion matrix* untuk data latih : data uji dengan persentase 70:30

Tabel 4.6 *CART* dengan penerapan *SMOTE data latih:data uji 70:30*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	666	68	90,74%
Pred 1	247	885	78,18%
Class Recall	72,95%	92,86%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{885+666}{885+666+247+68} = \frac{1551}{1866} = 0{,}8312 \approx 83{,}12\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{885}{885+68} = \frac{885}{953} = 0{,}9287 \approx 92{,}87\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{666}{666+247} = \frac{666}{913} = 0{,}7294 \approx 72{,}94\%$$

Berdasarkan tabel 4.6 perhitungan dengan data latih 1306 data dan data uji 560 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 68 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 247 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.6 diperoleh hasil perhitungan ketepatan klasifikasi sebesar 83,12%, nilai sensitivitas sebesar 92,87% dan nilai spesivitas sebesar 72,94%.

4.7 CART dengan penerapan SMOTE pada data latih : data uji 80 : 20

CART dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*. Berikut ini adalah tabel 4.7 *confusion matrix* untuk data latih : data uji dengan persentase 80:20

Tabel 4.7 *CART* dengan penerapan *SMOTE data latih:data uji 80:20*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	665	70	90,48%
Pred 1	248	883	78,07%
Class Recall	72,84%	92,65%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{883+665}{883+665+248+70} = \frac{1548}{1866} = 0{,}8296 \approx 82{,}96\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{883}{883+70} = \frac{883}{953} = 0{,}9266 \approx 92{,}66\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{665}{665+70} = \frac{665}{735} = 0{,}7283 \approx 72{,}83\%$$

Berdasarkan tabel 4.7 perhitungan dengan data latih 1493 data dan data uji 373 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 70 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 248 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.7 diperoleh hasil perhitungan ketapatan klasifikasi sebesar 82,96%, nilai sensitivitas sebesar 92,66% dan nilai spesivitas sebesar 72,83%.

4.8 CART dengan penerapan SMOTE pada data latih : data uji 90 : 10

CART dengan penerapan SMOTE yang dilakukan dengan menggunakan *software Rapidminer* menghasilkan nilai *confusion matrix*. Berikut ini adalah tabel 4.8 *confusion matrix* untuk data latih : data uji dengan persentase 90:10

Tabel 4.8 *CART* dengan penerapan *SMOTE data latih:data uji 90:10*

Lama Studi	*True 0*	*True 1*	*Class Precision*
Pred 0	665	65	91,10%
Pred 1	248	888	78,17%
Class Recall	72,84%	93,18%	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{888+665}{888+665+248+65} = \frac{1553}{1866} = 0{,}8323 \approx 83{,}23\%$$

$$Sensitivitas = \frac{TP}{TP+FN} = \frac{888}{888+65} = \frac{888}{953} = 0{,}9319 \approx 93{,}19\%$$

$$Spesivitas = \frac{TN}{TN+FP} = \frac{665}{665+248} = \frac{665}{913} = 0{,}7283 \approx 72{,}83\%$$

Berdasarkan tabel 4.8 perhitungan dengan data latih 1679 data dan data uji 187 data. Dari data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 65 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk data yang diprediksi tepat waktu terdapat 248 data yang salah prediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model.

Berdasarkan tabel 4.8 diperoleh hasil perhitungan ketepatan klasifikasi sebesar 83,23%, nilai sensitivitas sebesar 93,19% dan nilai spesivitas sebesar 72,83%.

4.9 Perbandingan Evaluasi Klasifikasi *CHAID* dan *CART* dengan Penerapan *SMOTE*

Berikut hasil tingkat akurasi yang dihasilkan oleh tabel klasifikasi dari *software Rapidminer* sebagai berikut:

Tabel 4.9 Perbandingan Hasil Akurasi, *Sensitivitas* dan *Spesifita CHAID* dan *CART* dengan Penerapan *SMOTE*

Data Latih : Data Uji	*CHAID+SMOTE*			*CART+SMOTE*		
	akurasi	*sensitivitas*	*spesivitas*	akurasi	*sensitivitas*	*spesivitas*
70:30	83,17%	93,08%	72,83%	83,12%	92,87%	72,94%
80:20	83,01%	92,77%	72,83%	82,96%	92,66%	72,83%
90:10	83,07%	93,29%	72,39%	83,23%	93,19%	72,83%

Berdasarkan tabel 4.9 hasil perbandingan tingkat akurasi kedua algoritma pada data latih:data uji 70:30 yaitu *CHAID* 83,17% lebih unggul dibandingkan CART dengan penerapan SMOTE yaitu mempunyai tingkat akurasi 83,12%. Untuk data latih:data uji 80:20, *CHAID* dengan penerapan *SMOTE* mempunyai tingkat akurasi 83,01% lebih unggul dari algoritma *CART* dengan penerapan *SMOTE* yang mempunyai tingkat akurasi 82,96%. Untuk data latih:data uji 90:10, *CART* dengan penerapan *SMOTE* mempunyai tingkat akurasi 83,23% lebih unggul dari algoritma *CHAID* dengan penerapan *SMOTE* yang mempunyai tingkat akurasi 83,07%.

Pada tabel 4.9 menunjukkan bahwa *CART* dengan penerapan *SMOTE* pada upsampling 11 kali pada studi kasus data kelulusan di Universitas Widya Husada Semarang data latih:data uji 90:10 lebih baik dibandingkan *CHAID* dengan penerapan *SMOTE*. Nilai akurasi yang menunjukkan tingkat ketepatan hasil pengukuran dengan nilai sebenarnya yaitu sebesar 83,23%. Nilai *sensitivitas* menunjukkan mahasiswa yang lulus tepat waktu dari seluruh populasi yang benar-benar lulus tepat waktu sebesar 93,19%. Sedangkan nilai *spesivitas* menunjukkan mahasiswa yang lulus tidak tepat waktu dari populasi yang benar-benar tidak lulus tepat waktu sebesar 72,83%.

4.10 Analisis CART

Derajat impurity merupakan ukuran kehomogenan suatu simpul, dimana sebuah simpul dengan derajat *impurity* tinggi menunjukkan simpul tersebut tidak homogen, sedangkan simpul dengan derajat *impurity* rendah menunjukkan simpul tersebut homogen. Ukuran derajat impurity tersebut salah satunya dengan perhitungan *gini index*. Dengan menggunakan persamaan 2.6 berikut tabel perhitungan *gini index* untuk setiap variabel:

Tabel 4.10 Matriks Perhitungan *Gini Index* IPK

Kelas	IPK		
	Kategori		
	1	2	3
Tidak tepat waktu	484	414	15
Tepat waktu	58	761	134
Gini Index	**0,357**		

Tabel 4.11 Matriks Perhitungan *Gini Index* Program Studi

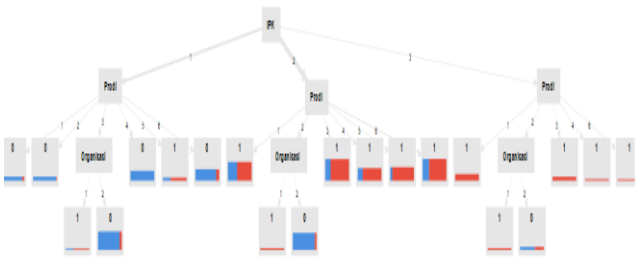
Kelas	Program Studi					
	Kategori					
	1	2	3	4	5	6
Tidak Tepat waktu	118	225	240	137	23	170
Tepat waktu	203	41	248	115	152	194
Gini Index	**0,432**					

Tabel 4.12 Matriks Perhitungan *Gini Index* Program Studi

Kelas	Organisasi	
	Kategori	
	1	2
Tidak Tepat waktu	16	893
Tepat waktu	257	696
Gini Index	**0,439**	

Dari matriks perhitungan *gini index* di atas, atribut yang mempunyai nilai index gini paling kecil yaitu IPK. Sehingga IPK dipilih sebagai atribut akar pertama. Selanjutnya yang dipilih menjadi atribut yaitu program studi dan organisasi.

Hasil proses klasifikasi data kelulusan mahasiswa dengan CART dengan penerapan SMOTE pada data latih : data uji 90:10 juga ditunjukkan seperti pada gambar 4.8 dapat dilihat bahwa atribut yang memiliki pengaruh paling tinggi untuk menentukan hasil klasifikasi adalah IPK yang mana atribut ini menjadi akar pertama. Selanjutnya yang menjadi *node* 1 yaitu variabel program studi dan *node* 2 yaitu organisasi.



Gambar 4.8 *Graph* CART 90:10

Tabel 4.10 Segmentasi Hasil Analisis CART :

Segmen	Karakteristik	Persentase
Ke-1	Mahasiswa dengan IPK 2,75-3,00 dari prodi TRO	1,90
Ke-2	Mahasiswa dengan IPK 2,75-3,00 dari prodi BID	1,90
Ke-3	Mahasiswa dengan IPK 2,75-3,00 dari prodi FIS dengan kriteria mengikuti organisasi	0,18
Ke-4	Mahasiswa dengan IPK 2,75-3,00 dari prodi FIS dengan kriteria tidak mengikuti organisasi	11,07
Ke-5	Mahasiswa dengan IPK 2,75-3,00 dari prodi PER	5,65
Ke-6	Mahasiswa dengan IPK 2,75-3,00 dari prodi RO	1,31
Ke-7	Mahasiswa dengan IPK 2,75-3,00 dari prodi TEM	6,67
Ke-8	Mahasiswa dengan IPK 3,01-3,50 dari prodi TRO	11,31
Ke-9	Mahasiswa dengan IPK 3,01-3,50 dari prodi BID dengan kriteria mengikuti organisasi	0,48
Ke-10	Mahasiswa dengan IPK 3,01-3,50 dari prodi BID dengan kriteria tidak mengikuti organisasi	10,18
Ke-11	Mahasiswa dengan IPK 3,01-3,50 dari prodi FIS	13,10
Ke-12	Mahasiswa dengan IPK 3,01-3,50 dari prodi PER	7,32
Ke-13	Mahasiswa dengan IPK 3,01-3,50 dari prodi RO	7,98
Ke-14	Mahasiswa dengan IPK 3,01-3,50 dari prodi TEM	13,21
Ke-15	Mahasiswa dengan IPK > 3,50 dari prodi TRO	3,57
Ke-16	Mahasiswa dengan IPK > 3,50 dari prodi BID dengan kriteria mengikuti organisasi	0,48
Ke-17	Mahasiswa dengan IPK > 3,50 dari prodi BID dengan kriteria tidak mengikuti organisasi	1,25
Ke-18	Mahasiswa dengan IPK > 3,50 dari prodi FIS	1,85
Ke-19	Mahasiswa dengan IPK > 3,50 dari prodi PER	0,36
Ke-20	Mahasiswa dengan IPK > 3,50 dari prodi TEM	0,24

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan pada bab 4, maka dapat disimpulkan sebagai berikut.

1. *CHAID* dengan penerapan *SMOTE* pada data kelulusan Universitas Widya Husada Semarang tahun 2018-2020 dengan menggunakan data latih : data uji yaitu 70:30, 80:20, 90:10 memperoleh tingkat akurasi yaitu 83,17%, 83,01% dan 83,07%
2. *CART* dengan penerapan *SMOTE* pada data kelulusan Universitas Widya Husada Semarang tahun 2018-2020 dengan menggunakan data latih : data uji yaitu 70:30, 80:20, 90:10 memperoleh tingkat akurasi yaitu 83,12%, 82,96% dan 83,23%
3. Tingkat akurasi CART dengan penerapan SMOTE pada data kelulusan Universitas Widya Husada Semarang tahun 2018-2020 dengan data latih : data uji 90:10 lebih baik dibandingkan dengan tingkat akurasi CHAID dengan penerapan SMOTE yaitu menghasilkan tingkat akurasi sebesar 83,23%

DAFTAR PUSTAKA

Ansori, Ari Hasan (2015) Strategi Peningkatan Sumber Daya Manusia. *Jurnal Qathruna*, Vol.2, No.2 : 19-56.

Antipov, E. & E. Pokryshevskaya. 2009. *Applying CHAID for logistic regression diagnostics and classification accuracy improvement*. Munich Personal RePEc Archieve (MPRA) No. 21499. Munich : Ludwig Maximilians Universitat Munchen.

Bawono, Bonggo dan Rochdi Wasono. 2019. Perbandingan Metode Random Forest dan Naive Bayes untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit. *Prosiding Seminar Nasional Edusaintek FMIPA UNIMUS 2019. Literasi Teknologi Saintifik & Big Data Melalui Pembelajaran 4C'S* : 343-348. Semarang, 28 September 2019: Universitas Muhammadiyah Semarang.

Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (1993). *Classification And Regression Trees*. New York: Chapman And Hall.

Chye, K.H., Chin, T.W., dan Reng, G.C. 2004. *Credit Scoring Using Data Mining Techniques*. Singapore Management Review, 26 (2) pp. 25-47.

Daniel, Wayne, W. 1989. *Statistik Nonparametrik Terapan*. Jakarta : PT Gramedia.

Darmawan, dkk. 2017. Klasifikasi Lama Masa Studi Mahasiswa Menggunakan Perbandingan Metode Algoritma C.45 dan Algoritma Classification and Regression Tree. Jurnal Eksponensial, Vol.8, No.2 hal 151-160.

Gallagher, C.A., H. M. Monroe, & J.L. Fish. 2000. *An Iterative Approach to Classification Analysis*.

Growe, G. A., 1999. *Comparing Algorithm and Clustering Data: Component of the Data Mining Process.* MSc Thesis. Department of Computer Science and Information System Grand Valley State University.

Gujarati, D.N. 2006. *Dasar-Dasar Ekonometrika Jilid 2*. Penerjemah : Julius A. Mulyadi. Jakarta : Penerbit Erlangga.

Hasan, Iqbal. 2004. *Analisa Data Penelitian dengan Statistik*. Jakarta : PT Bumi Aksara.

Hosmer, D. W. & S. Lemeshow. 2000. Applied Logistic Regression. New York : John Wiley & Sons Inc.

Johnson, R.A. dan Winchern, D.W. 2007. *Applied Multivariate Statistical Analysis, 6th Edition.* New Jersey : Person Prentice Hall.

Josephat, P. & A. Ismail. 2012. *A Logistic Regression Model of Customer Satisfaction of Airline.* International Journal of Human Resource Studies 2(2012) : 255-265.

Komalasari, Wieta, B. (2007). Metode Pohon Regresi untuk Eksploratori Data dengan Peubah yang Banyak dan Kompleks. Jurnal Informatika Pertanian Volume 16 No.1, Juli 2007.

Kunto, Y.S. dan Hasana, S.N. 2006. Analisis CHAID sebagai Alat Bantu Statistika untuk Segmentasi Pasar. *Jurnal Manajemen*, vol. 1 No. 2. Surabaya : Universitas Kristen Petra.

Lewis, M.D dan Roger, J. (2000). *An Introducion to Classification and Regression Trees (CART) Analysis*. Annual Meeting of Society For Academic Emergency. California, UCLA Medical Center

Miftahuddin. 2012. Penggunaan Metode CHAID (Chi Square-Automatic Interaction Detection) Pada Pohon Klasifikasi Menggunakan Satu Peubah Respon Dengan Perbandingan Taraf Nyata. Jurnal Matematika, Statistika, dan Komputasi. Vol. 9, No.1, 11-22.

Perez, F.M.D., & M. B. Cejas. 2016. CHAID Algorithm As An Appropriate Analytical Method for Tourism Market Segmentation. *Journal of Destination Marketing & Management* 5(2016) : 275-282.

Pramana, Setia, dkk. 2018. *Data Mining dengan R Konsep serta Implementasi.* Jakarta: In Media.

Sartono, Bagus dan Utami Dyah Safitri. 2010. Metode Pohon Gabungan: Solusi Pilihan untuk Mengatasi Kelemahan Pohon Regresi dan Klasifikasi Tunggal. Forum Statistika dan Komputasi. Vol 15, No.1, 1-7.

Sharp, A., J. Romaniuk dan S. Cierpicki. 2002. The Performance of Segmentation Variables : A Comparative Study.

Siswanto. (2007). Pengantar Manajemen. Jakarta : Bumi Aksara.

Siahaan, dkk. 2016. Aplikasi Classification and Regression Tree (CART) dan Regresi Logistik Ordinal dalam Bidang Pendidikan. Jurnal Eksponensial, Vol. 7, No.1 hal 95-104.

Sugiyono. 2007. *Metode Penelitian Pendidikan, Pendekatan, Kuantitatif, Kualitatif, dan R&D.* Bandung: Alfabeta.

Suniantara, I Ketut Putu dan Muhammad Rusli. 2017. Klasifikasi Waktu Kelulusan Mahasiswa STIKOM Bali menggunakan CHAID Regression Tree dan Regresi Logistik Biner. Statistika, Vol.5, No.1 hal 27-32.

Wibowo, dkk. 2019. Komparasi Algoritma Naive Bayes dan Decision Tree Untuk Memprediksi Lama Studi Mahasiswa. *ILKOMNIKA*, Vol.1, No.2 hal 65-74.

Widjaya, dkk. 2017. Prediksi Masa Studi Mahasiswa dengan Voting Feature Interval 5 Pada Aplikasi Konsultasi Akademik Online. *Computatio*, Vol.1, hal 25-33.

Widodo, P. P., Handayanto, R. T. dan Herlawati. 2013. Penerapan Data Mining dengan Matlab. Bandung: Rekayasa Sains.