BAB II

TINJAUAN PUSTAKA

2.1 Pendidikan

Pendidikan diartikan sebagai proses pembelajaran bagi individu untuk mencapai pengetahuan dan pemahaman yang lebih tinggi mengenai obyek-obyek tertentu dan spesifik. Pengetahuan tersebut diperoleh secara formal yang berakibat individu mempunyai pola pikir dan perilaku sesuai dengan pendidikan yang telah diperolehnya (KBBI, 1991 dalam Ambar, 2017 hal.4).

Menurut buku Pedoman Pendataan Pendidikan (2003), pendataan pendidikan merupakan suatu kegiatan atau proses pembuktian yang ditemukan dari hasil penelitian yang dapat dijadikan dasar kajian atau pendapat yang dikaitkan dengan otonomi pendidikan pada khususnya dan otonomi daerah pada umumnya.

Pendataan pendidikan mempunyai dua kegiatan utama yaitu :

a. Produksi Data

   Merupakan kegiatan tersimpannya hasil pengumpulan data dalam sistem komputer, dan tersusunnya laporan-laporan periodik dan tersusunnya berkas laporan untuk umpan balik pada sumber data.

b. Pemberdayaan/Pelayanan

   Merupakan suatu cara untuk melakukan eksplorasi terhadap data berdasarkan pada perkembangan pembangunan pendidikan, kebutuhan pimpinan atau permintaan data.

https://repository.unimus.ac.id

2.2 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.

2.3 Faktor-faktor yang Berpengaruh Terhadap Lama Studi Mahasiswa

Beberapa faktor yang diduga berpengaruh terhadap lama studi mahasiswa FKKM Universitas Widya Husada Semarang adalah sebagai berikut :

1. Jenis Kelamin

   Perbedaan antara perempuan dengan laki-laki secara biologis sejak seseorang lahir (Hungu 2007 dalam Suhardin 2016). Semua jenis kelamin baik laki-laki maupun perempuan memiliki peluang untuk berpengaruh terhadap lama studi mahasiswa, tidak terkecuali mahasiswa FKKM Universitas Widya Husada Semarang.

2. Asal Daerah

   Asal daerah terbagi menjadi dua yaitu Jawa dan luar Jawa. Dewasa ini, pendidikan masih berpusat di Pulau Jawa, sehingga alasan dipilihnya dua kelompok ini dikarenakan ingin diketahui apakah mahasiswa yang berasal daerah dari luar Jawa memiliki kesempatan untuk menyelesaikan masa

https://repository.unimus.ac.id

studi secara tepat waktu atau tidak tepat waktu dibandingkan dengan mahasiswa yang berasal dari Jawa.

3. Indeks Prestasi Kumulatif (IPK)

   Indeks prestasi kumulatif diduga berpengaruh terhadap lama studi mahasiswa. Indeks Prestasi Kumulatif adalah alat ukur berupa angka yang menunjukkan prestasi atau keberhasilan studi mahasiswa secara kumulatif, yang diperoleh mahasiswa setiap semesternya, mulai dari semester pertama sampai semester paling akhir yang ditempuh. Berdasarkan pada buku pedoman akademik, IPK dikategorikan "memuaskan" apabila ada pada interval 2,76 ≤ IPK ≤ 3,00. Selanjutnya IPK dikategorikan "sangat memuaskan" apabila ada pada interval 3,01≤ IPK ≤ 3,50. Kemudian IPK dikategorikan "dengan pujian/cumlaude" apabila ada pada interval IPK > 3,50.

4. Usia

   Mahasiswa merupakan masa memasuki masa dewasa yang pada umum berada pada rentang usia 18-25 tahun (Djibran, 2018).

5. Program Studi

   Program Studi adalah kesatuan kegiatan pendidikan dan pembelajaran yang memiliki kurikulum dan metode pembelajaran tertentu dalam satu jenis pendidikan akademik, pendidikan profesi, dan/atau pendidikan vokasi. Program studi yang ada pada penelitian ini antara lain Teknik Rontgen, Kebidanan, Fisioterapi, Keperawatan, Refraksi Optisi dan Teknik Elektromedik.

https://repository.unimus.ac.id

6. Organisasi

Menurut Siswanto (2007: 73) Organisasi dapat didefinisikan sebagai sekelompok orang yang saling berinteraksi dan bekerja sama untuk merealisasikan tujuan bersama. Mahasiswa yang secara aktif menggabungkan diri dalam suatu kelompok atau organisasi tertentu untuk melakukan suatu kegiatan dalam rangka mencapai tujuan organisasi, menyalurkan bakat, memperluas wawasan dan membentuk kepribadian mahasiswa seutuhnya. Semua mahasiswa yang mengikuti organisasi atau tidak mengikuti organisasi memiliki peluang untuk berpengaruh terhadap lama studi mahasiswa, tidak terkecuali mahasiswa FKKM Universitas Widya Husada Semarang.

2.4 Statistika Deskriptif

Metode statistik adalah prosedur-prosedur yang digunakan dalam pengumpulan, penyajian, analisis dan penafsiran data. Kemudian metode tersebut dibagi menjadi dua, yaitu statistika deskriptif dan statistika inferensial. Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu data sehingga memberikan informasi yang berguna (Walpole dkk, 1995). Statistik deskriptif adalah statistika yang berfungsi untuk mendeskripsikan atau memberi gambaran terhadap objek yang diteliti melalui data sampel atau populasi (Sugiyono, 2007). 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

https://repository.unimus.ac.id

memberikan keterangan-keterangan mengenai suatu data atau keadaan (Hasan, 2004).

2.5 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:

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

https://repository.unimus.ac.id

2.6 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 mempredikasi 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).

2.7 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)

https://repository.unimus.ac.id

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.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

B. Sensitivitas

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

$$Sensitivitas = \frac{banyaknya\ true\ positive}{banyaknya\ aktual\ positive} = \frac{TP}{TP + FN} \quad (2.2)$$

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 :

$$Spesivitas = \frac{banyaknya\ true\ negative}{banyaknya\ aktual\ negative} = \frac{TN}{TN + FP} \quad (2.3)$$

2.8 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu metode oversampling yaitu teknik pengambilan sampel untuk meningkatkan

https://repository.unimus.ac.id

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

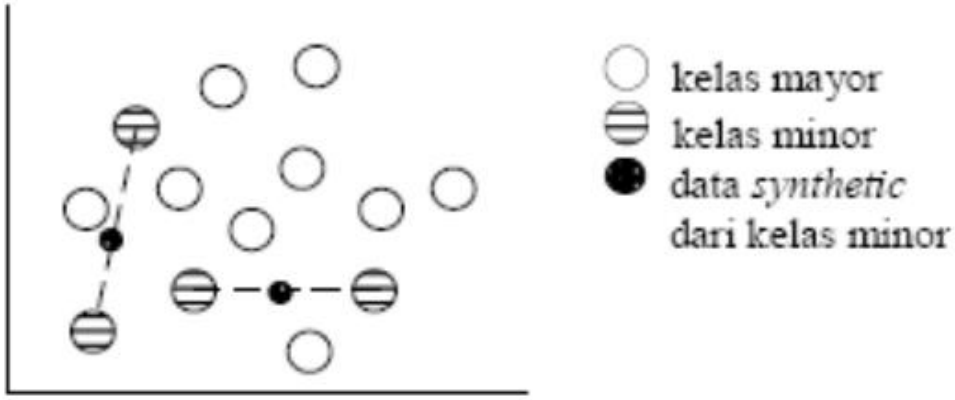
x_i = data yang akan direplikasi

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.

https://repository.unimus.ac.id



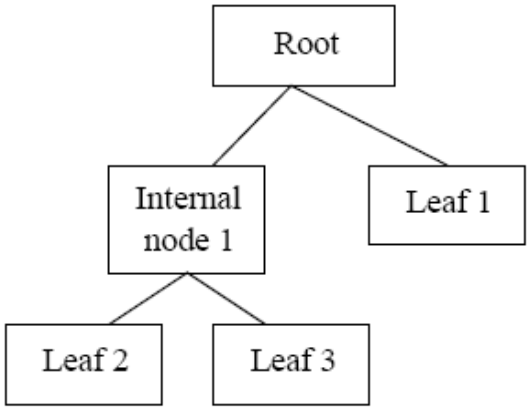
Gambar 2.1 Ilustrasi Algoritma SMOTE

2.9 Decision Tree

Decision tree merupakan salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Konsep dasar dari *decision tree* adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan *(rule)* (Pramana,2018). 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).

Decision Tree adalah sebuah struktur pohon, dimana setiap node pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakan suatu pembagian hasil uji, dan node daun (*leaf*) merepresentasikan kelompok kelas tertentu. Level node teratas dari sebuah d*ecision tree* adalah node akar (*root*) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu.

https://repository.unimus.ac.id



Gambar 2.2 Konsep Dasar Pohon Keputusan

Beberapa algoritma dapat digunakan dalam membangun pohon keputusan antara lain IDS, ID3, C4.5, CHAID dan CART (Defiyanti, 2013). Klasifikasi decision tree merupakan salah satu teknik terkenal dalam data mining dan merupakan salah satu metode yang popular dalam menentukan keputusan suatu kasus. Metode ini tidak memerlukan proses pengelolaan pengetahuan terlebih dahulu dan dapat menyelesaikan kasus-kasus yang memiliki dimensi yang besar (Widodo, dkk, 2013)

2.10 *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

https://repository.unimus.ac.id

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.

https://repository.unimus.ac.id

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

2.10.1 Pembentukan Pohon Klasifikasi

Pembentukan pohon klasifikasi diawali dengan menentukan variabel dan nilai dari variabel tersebut (*threshold*) untuk dijadikan pemilah tiap simpul. Dalam prosesnya, pembentukan pohon klasifikasi dibutuhkan data *learning* sampel L yang terdiri atas N pengamatan. Menurut Breiman, et al (1993), proses pembentukan pohon klasifikasi terdiri atas 3 tahapan yaitu sebagai berikut.

a. Pemilihan Pemilah

Pada tahap ini, data yang digunakan adalah sampel data *learning* L yang kemudian dipilah berdasarkan aturan pemilahan dan kriteria *goodness of split*. Himpunan bagian yang dihasilkan dari proses pemilahan harus lebih homogen dibandingkan pemilahan sebelumnya. Hal ini dilakukan dengan mendefinisikan keheterogenan simpul (*impurity* atau *i(t)*). Menurut Breiman, et al (1993), fungsi keheterogenan yang sangat mudah dan sesuai diterapkan dalam berbagai kasus adalah Indeks Gini. Indeks Gini akan selalu memisahkan kelas dengan anggota paling besar atau kelas terpenting dalam simpul tersebut terlebih dahulu. Pemilahan yang memberikan nilai penurunan keheterogenan tertinggi merupakan pemilahan terbaik. Fungsi Indeks Gini dituliskan dalam persamaan berikut.

$$i(t) = \sum_{i,j=1} p(j \mid t)p(i \mid t), i \neq j \qquad (2.6)$$

Dengan $p(j \mid t)$ adalah proporsi kelas j pada simpul t dan $p(i \mid t)$ adalah proporsi kelas i pada simpul t.

https://repository.unimus.ac.id

Pemilahan simpul dimulai dengan memeriksa nilai–nilai variabel independen dan dilakukan secara rekursif pada setiap simpul dengan dua tahapan. Tahapan yang pertama adalah mencari semua kemungkinan pemilah pada variabel prediktor. Menurut Breiman, et al (1993), proses pemilahan simpul menjadi dua simpul anak dilakukan dengan mengikuti aturan sebagai berikut.

1. Setiap pemilahan hanya bergantung pada nilai yang berasal dari satu variabel prediktor saja.
2. Apabila variabel prediktor berskala kontinu, maka pemilahan yang diperbolehkan adalah $x_j \leq c_i$ dan $x_j > c_i$, dengan $i = 1,2,\ldots, n-1$ dan c_i adalah nilai tengah dari dua nilai amatan sampel berurutan yang berbeda dari variabel X_j. jika suatu ruang sampel berukuran *n* dan terdapat *n* nilai amatan sampel yang berbeda pada variabel X_j, maka akan terdapat sebanyak *n*-1 kemungkinan pemilahan yang berbeda.
3. Apabila variabel prediktor berskala kategorik, maka pemilahan berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua simpul yang saling lepas (*disjoint*). Apabila variabel prediktor berskala nominal bertaraf *L*, maka akan diperoleh sebanyak 2^{L-1}-1 pemilahan yang mungkin. Akan tetapi, apabila kategori variabel prediktor berskala ordinal bertaraf *L*, maka akan diperoleh sebanyak *L*-1 pemilahan yang mungkin.

Pemilahan yang terpilih akan membentuk suatu himpunan kelas yang disebut sebagai simpul. Simpul tersebut akan melakukan pemilahan secara rekursif sampai diperoleh simpul akhir (*terminal nodes*).

https://repository.unimus.ac.id

Setelah dilakukan pemilahan dari semua kemungkinan pemilah, maka tahapan berikutnya adalah menentukan kriteria *goodness of split* ($\phi(s,t)$) untuk mengevaluasi pemilah dari pemilah *s* pada simpul *t*. *Goodness of split* ($\phi(s,t)$) merupakan penurunan heterogenitas, yaitu.

$$\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (2.7)$$

Dengan

$i(t)$: fungsi heterogenitas pada simpul t

p_L: proporsi pengamatan menuju simpul kiri

p_R: proporsi pengamatan menuju simpul kanan

$i(t_L)$: fungsi heterogenitas pada simpul anak kiri

$i(t_R)$: fungsi heterogenitas pada simpul anak kanan

Pemilah yang menghasilkan $\phi(s,t)$ lebih tinggi merupakan pemilah terbaik karena mampu mereduksi heterogenitas lebih tinggi. Pengembangan pohon ini dilakukan dengan pencarian pemilah yang mungkin pada simpul t_1 yang kemudian akan dipilah menjadi t_2 dan t_3 oleh pemilah *s*, dan seterusnya. t_L dan t_R merupakan partisi dari simpul *t* menjadi dua himpunan bagian saling lepas dimana p_L dan p_R adalah proporsi masing-masing peluang simpul. Karena $t_L \cup t_R = t$ maka nilai $\Delta i(s,t)$ merepresentasikan perubahan dari kehetoregenan dalam simpul *t* yang semata-semata disebabkan oleh pemilah *s*. jika simpul yang diperoleh merupakan kelas yang tidak homogen, prosedur yang sama diulangi sampai pohon klasifikasi menjadi suatu konfigurasi dan memenuhi:

https://repository.unimus.ac.id

$$\Delta i\left(s^{*},t_1\right) = max_{s \subset s}\Delta i\left(s,t_1\right) \quad (2.8)$$

b. Penentuan Simpul Terminal

Suatu simpul *t* akan menjadi simpul terminal atau tidak, akan dipilah kembali apabila pada simpul *t* tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum sebesar *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak. Menurut Breiman, et al (1993), pengembangan pohon akan berhenti apabila pada simpul terdapat pengamatan berjumlah kurang dari atau sama dengan 5 ($n \leq 5$). Selain itu, proses pembentukan pohon juga akan berhenti apabila sudah mencapai batasan jumlah level yang telah ditentukan atau tingkat kedalaman (*depth*) dalam pohon maksimal.

c. Penandaan Label Kelas

Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p\left(j_0 \mid t\right) = max_j p(j \mid t) = max_j \frac{N_j(t)}{N(t)} \quad (2.9)$$

Dengan :

$p(j \mid t)$: proporsi kelas j pada simpul t

$N_j(t)$: jumlah pengamatan kelas j pada *terminal node t*

$N(t)$: jumlah total pengamatan pada *terminal node t*

Label kelas untuk simpul terminal t adalah j_0 yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul t yang paling kecil sebesar $r(t) = 1 - max_j p(j \mid t)$.

https://repository.unimus.ac.id

Proses pembentukan pohon klasifikasi berhenti ketika terdapat hanya satu pengamatan dalam tiap-tiap simpul anak atau adanya batasan minimum *n*, semua pengamatan dalam tiap simpul anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal. Setelah pembentukan pohon maksimal, tahap selanjutnya adalah pemangkasan pohon untuk mencegah terbentuknya pohon klasifikasi yang berukuran besar dan kompleks.

2.10.2 Pemangkasan Pohon Klasifikasi

Pohon yang dibentuk dengan aturan pemilah dan kriteria *goodness of split* berukuran sangat besar karena penghentian pohon berdasarkan banyaknya amatan pada simpul terminal atau besarnya tingkat kehomogenan. Jika semakin banyak pemilahan yang dilakukan, maka dapat mengakibatkan kecilnya tingkat kesalahan prediksi, akan tetapi akibatnya pohon klasifikasi yang dibentuk berukuran besar. Ukuran pohon yang besar dapat dapat memunculkan adanya *overfitting*, akan tetapi apabila pengamatan pohon dibatasi dengan ketepatan batas tertentu, maka dapat terjadi kasus *underfitting*. Oleh karena itu, untuk mendapatkan pohon yang layak, maka perlu dilakukan pemangkasan *pruning* yaitu suatu penilaian ukuran pohon tanpa mengorbankan akurasi yang berarti. Pemangkasan ini dilakukan dengan melakukan pengurangan simpul pohon sehingga dicapai ukuran pohon yang layak dan tidak terlalu melebar. Menurut Breiman dkk (1993), ukuran pohon yang layak dapat dilakukan dengan pemangkasan pohon dengan ukuran *cost complexity minimum*.

$$R_{\alpha}(T) = R(T) + \alpha|\tilde{T}| \qquad (2.10)$$

Dengan :

https://repository.unimus.ac.id

$R(T)$: *Resubtitusion Estimate* (proporsi kesalahan pada pohon T)

α: kompleksitas parameter (complexity parameter)

$|\tilde{T}|$: ukuran banyaknya simpul terminal pohon T

$R_\alpha(T)$ merupakan kombinasi linear biaya dan kompleksitas pohon yang dibentuk dengan menambahkan *cost penalty* bagi kompleksitas terhadap biaya kesalahan klasifikasi pohon. *Cost complexity pruning* menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(T)$ pada seluruh pohon bagian atau untuk setiap nilai α. Selanjutnya, dilakukan pencarian pohon bagian $T(\alpha) < Tmax$ yang meminimumkan $R_\alpha(T)$ yaitu

$$R_\alpha(T(\alpha)) = min_{T<Tmax} R_\alpha(T) \quad (2.11)$$

Pemangkasan pohon dimulai dengan mengambil $t_R \, dan \, t_L$ dari T_{max} yang dihasilkan dari simpul induk t. Jika diperoleh dua simpul anak dan simpul induk yang memenuhi persamaan $R(t) = R(t_R) + R(t_L)$, maka simpul anak $t_R \, dan \, t_L$ dipangkas. Sehingga diperoleh pohon T_1 yang memenuhi kriteria $R(T_1) = R(T_{max})$. Jika $R(T)$ digunakan sebagai kriteria penentuan pohon optimal maka akan cenderung dipilih pohon terbesar T_i. Sebab semakin besar pohon, maka semakin kecil nilai $R(T)$ nya.

2.10.3 Penentuan Pohon Klasifikasi Optimal

Pohon klasifikasi yang berukuran besar akan memberikan nilai *cost complexity* yang tinggi karena struktur data yang digambarkan cenderung kompleks sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi

https://repository.unimus.ac.id

memberikan nilai penduga pengganti yang cukup kecil. Apabila $R(T)$ dipilih sebagai penduga terbaik, maka cenderung akan dipilih pohon yang besar, sebab pohon yang semakin besar akan membuat nilai $R(T)$ semakin kecil. $R(T)$ atau *Resubstitution Estimate*/penduga pengganti merupakan proporsi amatan yang mengalami kesalahan pengklasifikasian.

Penduga pengganti ini sering digunakan apabila pengamatan yang ada tidak cukup besar. Pengamatan dalam L dibagi secara random menjadi V bagian yang saling lepas dengan ukuran kurang lebih sama besar untuk setiap kelas. Pohon $T^{(V)}$ dibentuk dari sampel *learning* ke-v dengan v=1,2,…,V. Dimisalkan $d^{(v)}(x)$ adalah hasil pengklasifikasian, maka penduga sampel uji untuk $R(T_t^{(v)})$ adalah sebagai berikut.

$$R(T_t^{(v)}) = \frac{1}{N_2} \sum_{(x_n, j_n) \subset L_v}^{N} X(d^{(v)}(x_n) \neq j_n) \qquad (2.12)$$

Dengan $N_v \cong {}^{N}/_{V}$ adalah jumlah pengamatan dalam L_v.

Selanjutnya dilakukan prosedur yang sama dengan menggunakan semua pengamatan dalam L untuk membentuk deret pohon T_t. Penduga $cross\ validation\ v-fold$ untuk $T_t^{(v)}$ adalah

$$R^{cv}(T_t) = \frac{1}{V} \sum_{v=1}^{V} {}_{v} R^{cv}\left(T_t^{(v)}\right) \qquad (2.13)$$

Pohon klasifikasi yang optimum dipilih T^* dengan $R^{cv}(T^*) = \min R^{cv}(T_t)$

https://repository.unimus.ac.id

2.11 *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).

https://repository.unimus.ac.id

2.11.1 Variabel-variabel Metode CHAID

Variabel yang digunakan dalam metode CHAID adalah data kategori (nominal atau ordinal), baik variabel dependen maupun variabel independen. Menurut Gallagher (2000), Variabel independen dalam metode CHAID dapat dibedakan menjadi 3 jenis. Variabel-variabel tersebut adalah sebagai berikut.

a. Variabel Monotonik

Variabel monotonik adalah variabel independen di mana kategori-kategori di dalamnya dapat digabungkan jika berurutan (data ordinal).

b. Variabel Bebas

Variabel bebas adalah variabel independen di mana kategori-kategori di dalamnya dapat digabungkan meskipun tidak berurutan (data nominal).

c. Variabel Mengambang

Variabel mengambang adalah variabel independen yang dapat diperlakukan sebagai variabel monotonik, kecuali untuk kategori yang *missing value*, yang dapat dikombinasikan dengan kategori manapun.

2.11.2 Uji Chi-square

Sesuai dengan namanya, statistik uji yang digunakan dalam metode CHAID adalah statistik uji *chi-square*. Statistik uji *chi-square* dapat digunakan untuk mengetahui independensi (kebebasan) antara dua variabel.

Misalkan dua variabel akan diuji independensinya, yang mana variabel pertama mempunyai r kategori dan variabel kedua mempunyai c kategori. Maka struktur data uji *chi-square* dapat dilihat pada Tabel 2.2 (Daniel, 1989).

Tabel 2.2 Struktur Data Uji Chi-square

https://repository.unimus.ac.id

Baris	Kolom						Total
	1	2	...	j	...	c	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_1
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_2
.	.	.	...	.	...	.	.
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_i
.	.	.	...	.	...	.	.
r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	n_r
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	N

Keterangan

n_{ij} : banyaknya pengamatan yang termasuk dalam kategori ke-i dari variabel pertama dan kategori ke-j dari variabel kedua

$n_{i.}$: banyaknya pengamatan yang termasuk dalam kategori ke-i dari variabel pertama

$n_{.j}$: banyaknya pengamatan yang termasuk dalam kategori ke-j dari variabel kedua

Hipotesis yang digunakan pada pengujian *chi-square* adalah sebagai berikut.

H_0 : kedua kriteria klasifikasi adalah saling bebas (tidak terdapat hubungan antara variabel pertama dan variabel kedua atau independen)

H_1 : kedua kriteria klasifikasi adalah tidak saling bebas (terdapat hubungan antara variabel pertama dan variabel kedua atau dependen)

Taraf signifikansi : α

Statistik Uji

https://repository.unimus.ac.id

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (2.14)$$

Keterangan

n_{ij} : banyaknya pengamatan yang termasuk dalam kategori ke-i dari variabel pertama dan kategori ke-j dari variabel kedua

E_{ij} : frekuensi harapan pengamatan yang termasuk dalam kategori ke-i dari variabel pertama dan kategori ke-j dari variabel kedua

r : jumlah kategori dalam variabel pertama

c : jumlah kategori dalam variabel kedua

Untuk menghitung frekuensi harapan masing-masing sel digunakan rumus (Daniel, 1989).

$$E_{ij} = \frac{n_{i.}n_{j}}{n} \qquad (2.15)$$

Keterangan

$n_{i.}$: banyaknya pengamatan yang termasuk dalam kategori ke-i dari variabel pertama

$n_{.j}$: banyaknya pengamatan yang termasuk dalam kategori ke-j dari variabel kedua

E_{ij} : frekuensi harapan pengamatan yang termasuk dalam kategori ke-i dari variabel pertama dan kategori ke-j dari variabel kedua

n : banyaknya seluruh pengamatan

https://repository.unimus.ac.id

Menurut Daniel (1989), kriteria pengambilan keputusan dalam uji *chi-square* yaitu H_0 ditolak jika $X^2_{hitung} > X^2_{\alpha;(r-1)(c-1)}$ atau dengan membandingkan nilai signifikansi dengan taraf signifikansi (α).

Statistik uji *chi-square* digunakan dalam dua cara dalam analisis CHAID. Pertama, untuk menentukan apakah kategori-kategori dalam sebuah variabel independen bersifat seragam dan bisa digabungkan menjadi satu. Kedua, ketika semua variabel independen sudah diringkas menjadi bentuk yang signifikan dan tidak mungkin digabung lagi, maka statistik uji *chi-square* digunakan untuk menentukan variabel independen mana yang paling signifikan untuk membagi kategori-kategori dalam variabel dependen.

2.11.3 Koreksi Bonferroni

Menurut Sharp *et al* (2002), koreksi Bonferroni adalah suatu proses koreksi yang digunakan ketika beberapa uji statistik untuk kebebasan atau ketidakbebasan dilakukan secara bersamaan. Koreksi Bonferroni biasanya digunakan dalam perbandingan berganda. Pengurangan pada tabel kontingensi pada algoritma CHAID dibutuhkan untuk uji signifikansi. Jika tidak ada pengurangan pada tabel kontingensi asal, maka statistik uji X^2 dapat digunakan. Ketika terjadi pengurangan yaitu c kategori dari variabel asal menjadi r kategori (r < c), maka tingkat kesalahan tunggal untuk uji signifikansi antara variabel dependen dan variabel independen yang tereduksi tersebut dikalikan dengan pengali Bonferroni sesuai dengan jenis variabelnya.

https://repository.unimus.ac.id

Kass (1980) menyebutkan bahwa pengali Bonferroni dihitung sesuai dengan jenis variabel independen.

1. Variabel Monotonik

$$M = \binom{c-1}{r-1} \tag{2.16}$$

Keterangan

M : pengali Bonferroni

c : banyaknya kategori variabel independen awal

r : banyaknya kategori variabel independen setelah penggabungan

2. Variabel Bebas

$$M = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!} \tag{2.17}$$

3. Variabel Mengambang

$$M = \binom{c-2}{r-2} + r\binom{c-2}{r-2} \tag{2.18}$$

2.11.4 Algoritma CHAID

Algoritma CHAID digunakan untuk menghasilkan diagram pohon CHAID yang dapat digunakan untuk memprediksi nilai variabel dependen. Secara garis besar algoritma ini dapat dibagi menjadi 3 tahap, yaitu tahap penggabungan (*merging*), tahap pemisahan (*splitting*), dan tahap penghentian (*stopping*). Diagram pohon diperoleh melalui tiga tahap tersebut, dimulai dari simpul akar dan dilakukan secara berulang pada setiap simpul yang terbentuk.

https://repository.unimus.ac.id

2.11.4.1 Tahap Penggabungan *(Merging)*

Pada tahap ini diperiksa signifikansi dari masing-masing kategori variabel independen terhadap variabel dependen. Tahap penggabungan untuk setiap variabel independen dalam menggabungkan kategori-kategori yang tidak signifikan adalah sebagai berikut.

1. Membentuk tabel kontingensi dua arah untuk masing-maisng variabel independen dengan variabel dependen.
2. Menghitung statistik uji *chi-square* untuk setiap pasang kategori yang dapat dipilih untuk digabung menjadi satu, untuk menguji kebebasannya dalam sebuah sub tabel kontingensi 2 x d yang dibentuk oleh sepasang kategori tersebut dengan variabel dependen yang mempunyai sebanyak d kategori.

   Misalnya, sebuah variabel independen Xi adalah variabel monotonik dengan c kategori, di mana i = 1,2,...,c. Variabel dependen Y memiliki r kategori. Untuk mengetahui kategori variabel independen mana yang tidak signifikan, maka dipasangkan masing-masing kategori pada variabel independen dengan variabel dependen. Banyaknya pasangan yang mungkin adalah kombinasi r dari c.

Tabel 2.3 Ilustrasi Penggabungan Pasangan Kategori Variabel Independen

Kategori 1	Kategori 2	*p-value*
X_1	X_2	$P_{1,2}$
X_1	X_3	$P_{1,3}$
.	.	.
.	.	.

https://repository.unimus.ac.id

X_c	X_1	$P_{c,1}$
.	.	.
.	.	.
X_c	X_{c-1}	$P_{c,c-1}$

3. Untuk masing-masing nilai *chi-square* berpasangan, hitung *p-value* berpasangan bersamaan. Di antara pasangan-pasangan yang tidak signifikan, gabungkan sebuah pasangan kategori yang paling mirip (yaitu pasangan yang mempunyai nilai *chi-square* berpasangan terkecil dan *p-value* terbesar) menjadi sebuah kategori tunggal, dan kemudian dilanjutkan ke langkah nomor 4. Dari ilustrasi Tabel 2.3, jika terdapat pasangan dengan *p-value* lebih besar dari taraf signifikansi, maka pasangan tersebut akan digabungkan. Misalnya pasangan kategori X_1 dan X_2 pada Tabel 2.3 tidak signifikan, maka pasangan tersebut akan digabungkan menjadi satu variabel baru yaitu $X_{1,2.}$
4. Periksa kembali kesignifikanan kategori baru setelah digabung dengan kategori lainnya dalam variabel independen. Jika masih ada pasangan yang belum signifikan, ulangi langkah 3. Jika semua sudah signifikan lanjutkan langkah berikutnya. Misalnya, pada ilustrasi sebelumnya didapat gabungan variabel baru $X_{1,2}$. Variabel tersebut akan dipasangkan dengan variabel lainnya, misalnya X_3, X_4,...,X_5 kemudian dilihat apakah pasangan tersebut sudah signifikan, ketika semua signifikan bisa dilanjutkan ke langkah 5, namun jika masih ada yang belum signifikan kembali ke langkah 3.

https://repository.unimus.ac.id

5. Hitung *p-value* terkoreksi Bonferroni didasarkan pada tabel yang telah digabung.

2.11.4.2 Tahap Pemisahan *(Splitting)*

Tahap pemisahan memilih variabel independen yang mana yang akan digunakan sebagai pemisah simpul terbaik. Pemilihan dikerjakan dengan membandingkan *p-value* (dari tahap penggabungan) pada setiap variabel independen. Langkah tahap pemisahan adalah sebagai berikut.

1. Pilih variabel independen yang memiliki *p-value* terkecil (paling signifikan) yang akan digunakan sebagai pemisah simpul.
2. Jika *p-value* kurang dari atau sama dengan taraf signifikansi (α), pemisah simpul menggunakan variabel independen ini.

2.11.4.3 Tahap Penghentian *(Stopping)*

Ulangi langkah penggabungan untuk subkelompok berikutnya, tahap penghentian dilakukan jika proses pertumbuhan pohon harus dihentikan sesuai dengan peraturan penghentian di bawah ini.

1. Tidak ada lagi variabel independen yang signifikan menunjukkan perbedaan terhadap variabel dependen.

2. Jika pohon sekarang mencapai batas nilai maksimum pohon dari spesifikasi maka proses pertumbuhan pohon akan berhenti. Misalnya, ditetapkan kedalaman pertumbuhan pohon klasifikasi adalah 3, ketika pertumbuhan pohon sudah mencapai kedalaman 3 maka pertumbuhan pohon klasifikasi dihentikan.

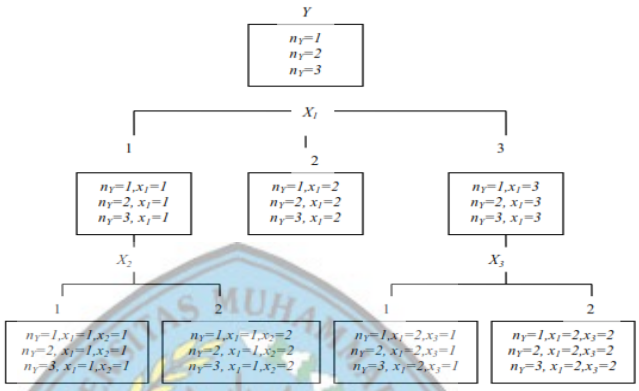
3. Jika ukuran dari simpul anak kurang dari nilai ukuran simpul anak minimum yang telah ditentukan, atau berisi pengamatan-pengamatan dengan jumlah yang

https://repository.unimus.ac.id

terlalu sedikit maka simpul tidak akan dipisah. Misalnya, ditetapkan ukuran minimum simpul anak adalah 50, ketika pemisahan menghasilkan ukuran simpul anak kurang dari 50, maka simpul tidak akan dipecah.

2.11.5 Pohon Klasifikasi CHAID

Hasil proses pembelahan dalam CHAID akan ditampilkan dalam sebuah diagram pohon.



Gambar 2.3 Diagram Pohon CHAID

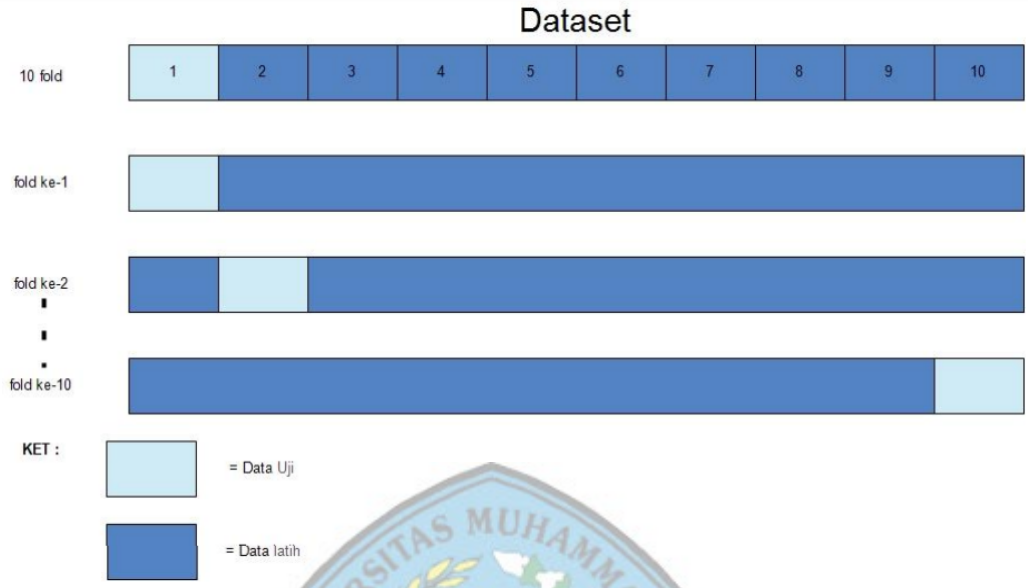
Setiap simpul (node) dari diagram pohon mewakili setiap subgrup dari sampel. Akar simpul (root node) mengandung seluruh sampel dan frekuensi tertentu n_i untuk setiap kategori dari peubah respon Y. Untuk tingkat selanjutnya ada tiga sampel yang dibagi oleh X_1 sebagai penduga (predictor) terbaik dari peubah respon. Anak simpul mengandung informasi tentang frekuensi dari kriteria peubah respon Y untuk subgrup yang bersesuaian. Begitu untuk pembagian selanjutnya yang dijelaskan oleh peubah X_2 dan X_3 (Miftahuddin, 2012).

2.12 *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

https://repository.unimus.ac.id

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.

   $$akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} X\ 100\% \qquad (2.25)$$

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.

https://repository.unimus.ac.id

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.

https://repository.unimus.ac.id