

BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

Penelitian adalah kegiatan yang dilakukan menurut kaidan dan metode ilmiah secara sistematis untuk memperoleh informasi, data dan keterangan yang berkaitan dengan pemahaman dan/atau pengujian suatu cabang pengetahuan dan teknologi (Fitri, 2020). Penelitian dimulai dari sebuah masalah dan berakhir di pemecahan masalah, dan penelitian ini adalah penelitian kuantitatif dengan pengumpulan data dan perhitungan statistik. Diperlukan kajian teoritis dan studi pendahuluan untuk menunjang penelitian ini.

2.1.1 Statistika

Statistika (*statistic*) adalah metode atau ilmu yang mempelajari cara merencanakan, mengumpulkan, menganalisis, dan menyajikan data untuk mengambil keputusan yang tepat. Istilah statistik digunakan untuk merujuk pada kumpulan pengetahuan tentang metode pengumpulan data, serta analisis dan interpretasi (Herlawati, 2019). Statistika memiliki banyak kegunaan untuk membuat keputusan yang tepat di berbagai bidang kehidupan. Seperti pengolahan data untuk menentukan transisi demografi, menentukan rata-rata dan standar deviasi dari nilai ujian mahasiswa di suatu perguruan tinggi, dan pengolahan data penjualan produk di suatu perusahaan.

Ketika kita menerapkan statistika untuk masalah ilmiah, industri atau sosial pertama kita mulai dengan studi populasi. Yang dimaksud studi populasi dalam statistika dapat berupa populasi makhluk hidup, benda mati, atau benda abstrak.

Pengumpulan data (collecting data) dari seluruh populasi disebut sensus. Sebuah sensus tentu memakan banyak waktu dan biaya. Untuk itu dalam statistika sampling sering dilakukan, yaitu sebagian kecil dari suatu populasi, yang dapat mewakili seluruh populasi. Jika sampel yang diambil cukup representatif, maka (pengambilan keputusan) dan kesimpulan yang diambil dari sampel tersebut dapat dipakai dalam menggambarkan populasi secara menyeluruh. Metode statistika bagaimana mendapatkan sampel yang tepat disebut teknik sampling. Analisis statistik terutama menggunakan probabilitas sebagai konsep intinya, dapat dilihat bahwa uji statistik banyak digunakan dan didasarkan pada distribusi peluang.

2.1.2.1 Pembagian Statistika

Statistika dapat dibagi atas beberapa macam seperti cara pengolahan data, ruang lingkup penggunaan atau disiplin ilmu yang digunakannya, dan bentuk parameternya.

1. Berdasarkan cara pengolahan datanya.

Berdasarkan cara pengolahan datanya, statistika dapat dibedakan menjadi dua, yaitu statistika Deskriptif dan statistika Inferensial

a Statistika Deskriptif

Statistika deskriptif atau statistika deduktif mempunyai tujuan untuk mendeskripsikan atau memberi gambaran objek yang diteliti sebagaimana adanya tanpa menarik kesimpulan atau generalisasi. Dalam statistika deskriptif ini dikemukakan cara-cara penyajian data dalam bentuk tabel maupun diagram, penentuan rata-rata (mean), modus, median, rentang serta

simpangan baku.

b Statistika Inferensial

Statistik inferensial atau induktif adalah statistik yang memberikan aturan atau metode yang dapat digunakan untuk membuat prediksi atau perkiraan dan untuk menarik kesimpulan umum dari sekumpulan data yang dipilih secara acak dalam subjek kumpulan data (populasi). Tujuan statistika adalah untuk menarik kesimpulan. Sebelum menarik kesimpulan, hipotesis dirumuskan berdasarkan statistik deskriptif. Contoh soal statistik inferensial: (a) Estimasi statistik, (b) Pengujian hipotesis (c) Prediksi dengan regresi/korelasi. Statistik inferensia sifatnya jauh lebih dalam dan ditindaklanjuti dengan statistik deskriptif. Oleh karena itu, untuk mempelajari dan memahami statistika inferensial, terlebih dahulu harus mempelajari statistik deskriptif. Oleh karena itu, statistik inferensial adalah statistik yang mempelajari bagaimana keputusan dibuat.

2. Berdasarkan ruang lingkup dan penggunaannya.

Berdasarkan atas ruang lingkup penggunaannya atau disiplin ilmu yang menggunakannya, statistika dapat dibagi atas beberapa macam, yaitu sebagai berikut :

- a Statistika Sosial adalah statistika yang dipakai dalam ilmu-ilmu sosial.
- b Statistika Pendidikan adalah statistika yang dipakai dalam ilmu

dan bidang pendidikan.

- c Statistika Ekonomi adalah statistika yang dipakai dalam ilmu-ilmu ekonomi.
- d Statistika Perusahaan adalah statistika yang dipakai dalam bidang perusahaan.
- e Statistika Kesehatan adalah statistika yang dipakai dalam bidang kesehatan.
- f Statistika Pertanian adalah statistika yang dipakai dalam bidang pertanian.

Berdasarkan penggunaannya, statistika yang digunakan dalam pengambilan keputusan secara luas oleh berbagai bidang ilmu seperti pemasaran, akuntansi, produksi, dan lain-lain.

3. Berdasarkan bentuk parameternya

Berdasarkan bentuk parameternya (data yang sebenarnya), statistika dapat dibagi dua, yaitu statistika parametrik dan statistika nonparametrik.

a Statistik Parametrik

Statistika parametrik adalah bagian statistika yang parameter dari populasinya mengikuti suatu distribusi tertentu, seperti distribusi normal, dan memiliki varians yang homogen.

b Statistik Nonparametrik

Statistika nonparametrik adalah bagian statistika dimana parameter dari populasinya tidak mengikuti suatu distribusi tertentu atau memiliki distribusi yang bebas dari persyaratan, dan

variansnya tidak perlu homogen. Dengan perkataan lain uji statistik non-parametrik adalah uji yang modelnya tidak menetapkan syarat-syarat mengenai parameter populasi yang merupakan induk sampel penelitiannya. Perlu diketahui bahwa uji nonparametrik selayaknya tidak digunakan apabila uji parametrik dapat diterapkan, karena tingkat keampuhan uji nonparametrik lebih rendah dari pada uji parametrik. Namun kita sebagai pengambil keputusan jangan salah menafsirkan bahwa derajat kegunaan metode statistik nonparametrik dibawah metode statistic parametrik. Tentu saja tidak demikian, masing-masing metode dibuat dengan spesifikasi khusus sesuai dengan macam data yang digunakan. Peningkatan keampuhan uji nonparametrik harus dengan memperbesar sampel. Namun seperti kita ketahui memperbesar sampel berarti akan menambah biaya, waktu, dll.

4. Berdasarkan Skala Pengukurannya

Terdapat 4 skala pengukuran dalam statistik dan statistika, yaitu: skala nominal, skala ordinal, skala interval dan skala ratio.

a Skala Nominal

Skala nominal merupakan skala data yang hanya dapat digunakan untuk membedakan objek yang bersifat kualitatif atau kategoris saja. Seperti jenis kelamin, agama, warna kulit, dan lain sebagainya. Data nominal adalah data statistik dimana dalam menyusun angkanya berdasarkan beberapa kategori dengan tidak

memakai urutan tertentu. Jadi kedudukan kategori satu dengan yang lain adalah sama. Label, kode, symbol yang diberikan pada kategori untuk membedakan antara kategori satu dengan yang lain. Dengan demikian operasi aritmatika tidak berlaku bagi data nominal.

b Skala Ordinal

Skala ordinal merupakan skala pengukuran kualitatif dan nantinya akan diklasifikasikan dalam suatu kelompok tertentu dan diberi kode berhierarki. Seperti pendidikan, tingkat kepuasan pengguna, dan lain sebagainya. Data ordinal adalah data statistic yang angkanya disusun berdasarkan beberapa kategori dengan memperhatikan kedudukan ranking. Label, kode yang diberikan kepada masing-masing kategori menunjukkan peringkat dan urutan penilaian.

c Skala Interval

Skala interval merupakan skala pengukuran kualitatif dengan nilai nol mutlak maupun tidak sehingga menyebabkan bergesernya pengukuran tersebut sesuai dengan keinginan orang yang mengukur. Seperti suhu yang dinyatakan dalam celcius. Data interval adalah data yang disusun berdasarkan jarak yang sama antar golongan satu dengan yang lainnya. Data interval memiliki sifat data ordinal dan nominal serta ditambah satu sifat yaitu jarak yang sama antara kategori satu dengan yang lain.

d Skala Ratio

Skala ratio merupakan salah satu jenis pengukuran yang memiliki nol alamiah atau nol absolute, sehingga memungkinkan kita membandingkan magnitudo angka-angka absolute. Seperti tinggi dan berat adalah contoh dari skala ratio. Seseorang yang memiliki berat badan 80kg bisa dikatakan dua kali lipat dari berat badan yang 40 kg. Data ratio ini diperoleh dengan membandingkan nilai variable yang satu dengan nilai absolut lainnya (variable pendamping). Data ratio memiliki sifat tiga jenis data yaitu : nominal, interval, ordinal ditambah satu sifat yaitu memiliki nilai nol artinya tidak ada. Penghasilannya nol artinya tidak mempunyai penghasilan sepeser pun.

2.1.2 Klasifikasi (*Clasification*)

Klasifikasi merupakan pengelompokkan sampel berdasarkan ciri-ciri persamaan dan perbedaan dengan menggunakan variabel terikat berupa kategori (Faisal et al., 2017). Dalam proses pengklasifikasian biasa terdapat dua proses yang harus dilakukan, yaitu learning dan test. Pada fase learning, sebagian data yang telah diketahui kelas datanya digunakan untuk membentuk model perkiraan Pada fase test, model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasi mencukupi, model ini dapat dipakai untuk memprediksi kelas data yang belum diketahui (Permana, 2011)

Klasifikasi merupakan salah satu tugas yang penting data mining. Sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah di

tentukan (Shadiq, 2009). Klasifikasi merupakan penempatan objek-objek dari beberapa kategori yang telah ditetapkan sebelumnya. Klasifikasi sekarang telah banyak digunakan dalam berbagai aplikasi, sebagai contoh pendeteksian pesan *e-mail*, *spam* berdasarkan *header* dan isi atau mengklasifikasikan galaksi berdasarkan bentuk bentuknya. Pada proses klasifikasi data yang diinputkan berupa *data record* atau *data sample*.

2.1.3 CHAID (Chi Squared Automatic Interaction Detection)

CHAID atau *Chi Squared Automatic Interaction Detection*. *CHAID* merupakan suatu teknik iteratif yang menguji satu persatu variabel independen yang digunakan dalam klasifikasi dan menyusunkan berdasarkan pada tingkat signifikansi statistik *chi-square* terhadap variabel dependennya (Ardiyansyah et al., 2018). Metode *CHAID* pertama kali diperkenalkan pada sebuah artikel yang berjudul “*An Exploratory Technique for investigating Large Quantities of Categorical Data*” oleh DR.G. V. Kass tahun 1980 pada buku *Applied Statistics*. Teknik tersebut merupakan teknik yang lebih awal sebagai *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 (Faisal et al., 2017). Metode *CHAID* secara umum bekerja dengan mempelajari hubungan antara variabel dependen dengan beberapa variabel independen kemudian mengklasifikasi sampel berdasarkan hubungan tersebut.(Hasibuan & Harahap, 2019).*CHAID* digunakan untuk membentuk segmentasi yang membagi sebuah sampel menjadi dua atau lebih kelompok yang berbeda berdasarkan sebuah kriteria tertentu (Siti Nurul Hasana & Yohanes

Sondang Kunto, 2006). Hal ini kemudian diteruskan dengan membagi kelompok tersebut menjadi kelompok yang lebih kecil berdasarkan variable-variabel independen yang lain. Proses terus berulang hingga tidak ditemukan lagi variable-variabel independen. Beberapa definisi *CHAID* menunjukkan bahwa *CHAID* adalah sebuah metode untuk mengklasifikasi data kategori dimana tujuan prosedurnya adalah untuk membagi rangkaian data menjadi subgrup – subgrup berdasarkan pada variabel dependennya. Menurut (Siti Nurul Hasana & Yohanes Sondang Kunto, 2006) *CHAID* memiliki tiga elemen kunci yaitu:

1. Uji Signifikan *chi – square*, uji ini dilakukan untuk mengidentifikasi variabel independen yang paling signifikan dalam data.
2. Koreksi *Bonferroni*.
3. Sebuah algoritma yang digunakan untuk menggabungkan kategori kategori variabel.

2.1.2.2 Variabel - Variabel dalam analisis CHAID

Variabel yang digunakan dalam analisis *CHAID* dibedakan atas variabel dependen dan variabel independen. Klasifikasi dalam *CHAID* dilakukan berdasarkan pada hubungan yang ada antara kedua variabel tersebut, oleh karena itu *CHAID* termasuk dalam metode dependensi dalam menentukan segmentasi. Menurut (Siti Nurul Hasana & Yohanes Sondang Kunto, 2006) menunjukkan bahwa *CHAID* akan membedakan variabel -riabel independennya menjadi tiga bentuk berbeda, yaitu:

- a. Monotonik

Kategori-kategori pada variabel ini dapat dikombinasikan atau digabungkan oleh *CHAID* jika keduanya berdekatan satu sama lain, yaitu variable-variabel Contohnya adalah : usia dan pendapatan.

b. Bebas

Kategori-kategori pada variable ini dapat dikombinasikan atau digabungkan pada saat keduanya berdekatan ataupun tidak (data nominal). Contohnya: Pekerjaan, kelompok etnik, dan area geografis.

c. Mengambang (*Floating*)

Kategori-kategori pada variable ini dapat diperlukan seperti monotonik kecuali untuk kategori *missing value*, yang dapat berkombinasi dengan kategori manapun.

2.1.2.1 Metode Algoritma *CHAID*

CHAID bekerja untuk melakukan pemisahan dan penggabungan data-data dalam variable yang dipakai pada analisisnya. *CHAID* dapat dibagi dalam tiga tahap yaitu Penggabungan (*Merging*), Pemisahan (*Splitting*) dan Penghentian (*Stopping*).

1. Penggabungan (*Merging*)

Tahap pertama dalam algoritma *CHAID* adalah penggabungan (*merging*). Pada tahap ini akan diperiksa signifikansi dari masing-masing kategori. Tahap Penggabungan untuk setiap variable Independen dalam menggabungkan kategori – kategori non-signifikan adalah sebagai berikut.

- a. Bentuk tabel kontingensi dua arah untuk masing-masing variabel independen dengan variabel dependennya.
- b. Hitung statistik *chi-square* untuk setiap pasang kategori yang dapat

dipilih untuk digabungkan menjadi satu, untuk menguji kebebasannya dalam sebuah sub tabel kontingensi yang dibentuk oleh sepasang kategori tersebut dengan variabel dependennya yang mempunyai sebanyak kategori. Rumus hitung statistik *chi-square* adalah:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{Rumus: } E_{ij} = \left(\frac{n_i \cdot n_j}{n} \right)$$

O_{ij} : banyaknya pengamatan pada baris ke-i dan kolom ke-j

E_{ij} : Nilai harapan pengamatan pada baris ke-i dan kolom ke-j

n_i : Total banyaknya pengamatan pada baris ke -i

n_j : Total banyaknya pengamatan pada baris ke – j

n : Total banyaknya responden

Keputusan yang di ambil dari uji *chi-square* ini adalah H_0 ditolak

jika nilai $\chi^2_{\text{hit}} > \chi^2_{\text{tabel}}$ atau $p\text{-value} < \alpha$

- c. Untuk masing-masing nilai *chi-square* berpasangan, hitung *p-value* berpasangan bersamaan. Diantara pasangan-pasangan yang tidak signifikan, gabungkan sebuah pasangan kategori yang paling mirip menjadi sebuah kategori tunggal.
- d. Periksa kembali kesignifikansian kategori baru setelah digabungkan dengan kategori lainnya dalam variabel independen. Jika masih ada pasangan yang belum signifikan, ulangi langkah C. Jika semua signifikan lanjutkan ke langkah selanjutnya.
- e. Hitung *p-value* terkoreksi Bonferroni didasarkan pada tabel yang

telah digabung.

2. Pemisahan (*Splitting*)

Tahap *splitting* memilih variable bebas yang mana akan digunakan sebagai *split node* (pemisah node) yang terbaik. Pemilihan dikerjakan dengan membandingkan *p-value* (dari tahap *merging*) pada setiap variable bebas.

3. Penghentian (*Stopping*)

Tahap *stopping* dilakukan jika proses pertumbuhan pohon harus dihentikan jika tidak ada lagi variable bebas yang signifikan menunjukkan perbedaan terhadap variable terkait.

Koreksi Bonferroni adalah suatu proses koreksi yang digunakan ketika beberapa uji statistik untuk kebebasan atau ketidakbebasan dilakukan secara bersamaan (Siti Nurul Hasana & Yohanes Sondang Kunto, 2006). Koreksi *Bonferroni* hanya dapat digunakan dalam perbandingan berganda. pada metode *CHAID* dan *Exhaustive CHAID* dibedakan menjadi 2 jenis yaitu monotonik yang digunakan pada variabel independen awal berdata ordinal dan jenis bebas yang digunakan pada variabel independen awal berdata nominal (Nazar, 2018). Persamaan koreksi *bonferroni* ditunjukkan dalam tabel 2.1

Tabel 2.1 Koreksi bonferroni pada metode *CHAID*

Metode	Jenis Monotonik	Jenis Bebas
<i>CHAID</i> (M)	$\binom{c-1}{r-1}$	$\sum_{i=0}^{r-1} (-1)^i \frac{(r-1)^c}{i!(r-i)!}$

2.1.3 Rapidminer

Rapidminer adalah sebuah *software* pengolahan data yang menggunakan prinsip dan algoritma *data mining*. *Rapidminer* mengekstrak pola dari *big data*, menggabungkan metode statistik, kecerdasan buatan dan *database*. *Rapidminer* bisa memudahkan penggunanya untuk menganalisis sebuah data yang besar menggunakan operator. Operator tersebut digunakan untuk mengubah data dan digabungkan dengan node di operator, dan kemudian dihubungkan ke node hasil untuk melihat hasilnya (Ainurrohmah, 2021) *Rapidminer* ditulis menggunakan bahasa pemrograman *Java* untuk melakukan tugasnya yang menggunakan operator. Operator yang ada dalam *rapidminer* lebih dari 500 operator, selain itu ada *market* untuk mencari operator yang belum tersedia dalam *rapidminer*. Operator memiliki fungsi mencakup aspek utama dari analisis data seperti pemuatan dan transformasi data *preprocessing* dan visualisasi data, pemodelan dan evaluasi model (Ainurrohmah, 2021) Cara menggunakan *rapidminer* adalah *import* data kedalam aplikasi terlebih dahulu, setelah data *set* terpasang di *rapidminer* maka bisa disambungkan dengan beberapa operator, misal data ada yang kurang atau hilang bisa disambungkan dengan *operator missing*.

2.2 State of The Art

2.2.1 Penelitian Terdahulu

Penelitian terdahulu memiliki peran penting pada penelitian, dengan adanya penelitian terdahulu sangat berpengaruh dalam memperkaya teori dan referensi wawasan pengetahuan bagi penelitian ini. Ulasan dari penelitian terkait, dilakukan

dengan maksud dan tujuan untuk menganalisis penelitian yang telah dilakukan sebelumnya. Penelitian- penelitian terdahulu dapat dilihat sebagai judul berikut:

1. Penerapan *Data Mining* Untuk Evaluasi Kinerja Akademik Mahasiswa di Universitas Klabat Dengan Metode Klasifikasi

Nama penulis dari jurnal tersebut adalah Green Ferry Mandias dengan terbitan Konferensi Nasional Sistem & Informatika 2015 STMIK STIKOM Bali, 9 – 10 Oktober 2015.

Penelitian yang dilakukan pada jurnal tersebut adalah melakukan klasifikasi kinerja akademik mahasiswa Klabat yang memiliki tujuan untuk mengevaluasi proses pendidikan di Universitas Klabat berdasarkan data sejarah akademik mahasiswa. Maka untuk melakukan klasifikasi data penelitian ini menggunakan metode eksperimen dan deskriptif dan menggunakan algoritma C.45 . Hasil dari penelitian ini menunjukkan bahwa Untuk mencapai semester akhir dalam 4 tahun seorang mahasiswa diawal semester berdasarkan pohon keputusan yang terbentuk yaitu mahasiswa yang mempunyai indeks prestasi 3.5 – 3.87 (magda) serta penempatan bahasa inggris yang berada pada level *adv* dan *inter*. Bila nilai bahasa inggris general maka subjek matematika awal semester “*pass*” atau tidak diambil maka mahasiswa tersebut bisa mencapai akhir semester tepat waktu.

2. Perbandingan Kinera Metode Klasifikasi *Chi-Square Automatic Interaction Determination (CHAID)* dengan Metode Klasifikasi

Algoritma C4.5 pada Studi Kasus: Penderita Diabetes Melitus Tipe 2 Di Samarinda Tahun 2015

Nama penulis dari jurnal tersebut adalah Muhammad Faisal, Yuki Novia Nasution, dan Fidia Deny T A. Terbitan Jurnal EKSPONENSIAL Volume 8, Nomor 2, Nopember 2017

Penelitian yang dilakukan pada jurnal tersebut adalah perbandingan kinerja metode *CHAID* dengan metode Algoritma C4.5 pada Studi kasus Penderita Diabetes Melitus tipe 2. Diabetes Melitus adalah penyakit kronis yang terjadi manusia memiliki dampak yang buruk untuk kesehatan, klasifikasi Diabetes Melitus dibagi menjadi beberapa golongan. Maka untuk menentukan klasifikasi diabetes melitus penelitian menggunakan metode algoritma *CHAID* dan juga algoritma C4.5 untuk melakukan klasifikasi, serta melakukan perbandingan akurasi algoritma mana yang paling besar akurasi ketepatannya. Data yang diperoleh adalah data pasien diabetes melitus Rumah Sakit Abdul Wahab Sjahranie Samarinda tahun 2015. Hasil dari penelitian ini adalah

3. Pembentukan Pohon Keputusan (*Decision Tree*) Menggunakan Metode *CHAID* dan *CART* Pada Studi Kasus Pernikahan Usia Dini.

Nama Penulis dari jurnal tersebut adalah Selvy Agustin, Islamil AB, dan Rahmi Susanti, terbitan jurnal Hospital Mahapahit Vol 13 No. 2 November 2021.

Penelitian yang dilakukan pada jurnal tersebut adalah melakukan klasifikasi menggunakan metode *Cross Sectional*, dengan algoritma *CHAID*

dan *CART* terhadap data pernikahan dini yang terjadi di Indonesia khususnya di pulau Kalimantan. Proses klasifikasi terhadap pernikahan yang terjadi, peneliti menggunakan data sekunder yang didapat dari *surver* dasar kesehatan Indonesia Provinsi Kalimantan Timur pada tahun 2017. Hasil dari penelitian ini adalah wanita yang mampu menyelesaikan pendidikan tertingginya di tingkat dasar dan memiliki ekonomi menengah yaitu wanita yang menikah di usia dini. Sedangkan wanita yang mampu menyelesaikan pendidikan tertingginya di tingkat menengah dan memiliki ekonomi menengah ke atas yaitu wanita yang menikah di usia >20 tahun.

4. *Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students*

Nama penulis dari jurnal tersebut adalah Elakia, Gayathri, Aarthi, dan Naren J terbitan Elakia et al, / (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, Vol. 5 (3) , 2014, 4649-4652.

Penelitian yang dilakukan pada jurnal tersebut adalah pengaplikasian *Educational Data Mining* untuk memprediksi kebiasaan pada pelajar dengan menggunakan metode *Data Mining* dengan algoritma C4.5 dan *tools Rapidminer*. Pemanfaatan *EDM* biasanya hanya dilakukan untuk memprediksi kualitas dan performa pendidikannya saja. padahal bisa juga digunakan untuk memprediksi masa depan pelajar dengan memberikan pilihan yang telah dilakukan oleh algoritma C4.5. Oleh karna itu, penelitian ini memanfaatkan *EDM* memprediksi masa depan pelajar menggunakan

algoritma C4.5 yang bisa memberikan pilihan yang tepat untuk masa depan pelajar. Data yang digunakan pada penelitian ini adalah data pelajar pada suatu sekolah. Hasil yang didapatkan untuk penelitian ini adalah bisa memprediksi kemungkinan masa depan yang akan dipilih oleh pelajar. Serta algoritma C4.5 bisa digunakan untuk klasifikasi data ini.

5. Akurasi Algoritma Klasifikasi pada *Software Rapidminer* dan *Weka*

Nama penulis dari jurnal tersebut adalah Ainurrohmah terbitan PRISMA, Prosiding Seminar Nasional Matematika 4, 493-499.

Penelitian yang dilakukan pada jurnal tersebut adalah mengukur akurasi algoritma-algoritma klasifikasi pada *rapidminer* dan *weka*. Penilaian akurasi ini memiliki tujuan untuk membandingkan *software data mining* yang akurat antara *rapidminer* dan *weka*. Algoritma yang dipakai pada penelitian ini adalah *Decision tree*, *Random Forest*, *K-NN*, *Naive Bayes*, dan *MLP*. Penelitian ini dilakukan untuk mengukur dan mencari tau *software* mana yang akurasinya lebih tinggi dan lebih efisien. Data yang digunakan pada penelitian ini adalah data *dummy*, karena tujuan penelitian ini hanya untuk mencari tau tingkat akurasi. Hasilnya adalah pada setiap algoritma dan *software* berbeda beda namun tingkat akurasi lebih tinggi ada pada *software rapidminer*.

6. *Data Mining* Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode *Decision Tree C4.5*

Nama penulis dari jurnal tersebut adalah Dwi Untari terbitan Program Studi Teknik Informasika Fakultas Ilmu Komputer Universitas Dian Nuswantoro, Semarang.

Penelitian yang dilakukan pada jurnal tersebut adalah melakukan prediksi mahasiswa yang memiliki potensi akan tetapi mahasiswa tersebut tidak lagi melakukan perkuliahan atau mahasiswa *non-aktif*. Mahasiswa *non-aktif* akan mempengaruhi mahasiswa lulus tidak tepat waktu, dan memiliki potensi *drop out* lebih tinggi. Penelitian ini menerapkan *data mining* dengan metode C4.5 untuk mencari karakteristik mahasiswa yang berpontensi *non-aktif*. Desain penelitian ini menggunakan CRISP-DM. Data yang digunakan pada penelitian ini menggunakan data mahasiswa S1 Fakultas Ilmu Komputer Universitas Dian Nuswantoro. Hasil dari penelitian ini menunjukkan nilai akurasi terbaik adalah 97.60% dengan *ratio data training* 90%, dan IPS mahasiswa semester 3 adalah 2.60, semester 4 adalah kosong.

7. Klasifikasi Faktor-Faktor Yang Mempengaruhi Masa Penyelesaian Skripsi Mahasiswa Dengan Metode CHAID Di Fakultas Tarbiyah Dan Ilmu Keguruan IAIN Bukittinggi

Nama penulis dari jurnal tersebut adalah Haidan Fitri dengan terbitan Jurnal Sains Matematika dan Statistika, Vol. 6,no 1, Januari 2020 ISSN 2460-4542, e-ISSN 2615-8663.

Penelitian yang dilakukan pada jurnal tersebut adalah melakukan klasifikasi faktor faktor yang dapat mempengaruhi masa penyelesaian

skripsi pada mahasiswa dengan metode *CHAID*. Dalam pengerjaan skripsi biasanya selesai selama 6 bulan atau 1 semester pembelajaran, namun mahasiswa memiliki masalah tersendiri yang mengakibatkan pengerjaan skripsi menjadi lebih lambat. Penelitian ini menggunakan metode deskriptif dengan studi korelasi terhadap faktor-faktor yang mempengaruhi masa penyelesaian skripsi. Algoritma yang digunakan pada penelitian ini adalah algoritma *CHAID*. Hasil dari penelitian ini adalah data yang digunakan dalam penelitian ini adalah 1537 orang, dengan 47% atau 723 orang mampu menyelesaikan skripsi tepat waktu, dan 53 % melebihi waktu idel yang ditetapkan. 25,2% berjenis kelamin laki-laki dan 74,8% berjenis kelamin perempuan. Untuk jenis kelamin dari dosen pembimbing antara dosen berjenis kelamin laki-laki dengan perempuan cukup berimbang yaitu berkisar 57% dan 43% nya berjenis kelamin perempuan. Sedangkan untuk bidang keahlian dosen pembimbing utama didominasi oleh dosen dengan latar belakang keilmuan pendidikan umum yaitu berkisar 40,8% sedangkan bidang keahlian dosen pembimbing pendamping cukup lebih beragam dimana antara bidang keahlian yang sesuai dengan prodi dan keahlian yang relevan dengan prodi cukup berimbang yaitu sekitar 38%.

