**PENERAPAN FEATURE SELECTION INFORMATION GAIN RATIO PADA ALGORITMA NAIVE BAYES UNTUK PREDIKSI KELULUSAN MAHASISWA**

**(STUDI KASUS : TEKNIK INFORMATIKA UIN SUSKA RIAU)**

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat

Untuk Memperoleh Gelar Sarjana Teknik

Pada Jurusan Teknik Informatika

Oleh

**MUHAMMAD FAUZAN WIJANARKO**

**11651103693**



**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU**

**PEKANBARU**

**2020**

# PENDAHULUAN

## **Latar Belakang**

1. Perguruan Tinggi

|  |  |
| --- | --- |
| Sumber | hasil |
| (UU No. 20 Tentang Sistem Pendidikan Nasional, 2003) | Pendidikan tinggi merupakan jenjang pendidikan setelah pendidikan menengah yang mencakup program pendidikan diploma, sarjana, magister, spesialis, dan doktor yang diselenggarakan oleh perguruan tinggi. |
| Romadhona, A., Suprapedi , S. dan Himawan, 2017 | Perguruan tinggi perlu mendeteksi perilaku mahasiswa sehingga dapat diketahui faktor yang menjadi penyebab kegagalan mahasiswa sehingga tidak lulus atau lulus sesuai dengan masa studi yang telah ditetapkan,diantaranya adalah rendahnya kemampuan akademik, usia masuk perguruan tinggi, indeks prestasi ataupun faktor lainnya. Salah satu penyelenggara pendidikan di Sulawesi Tengah adalah STMIK Adhiguna pada tiap tahun akademik banyak peminatnya dari dua program studi. |
| Ban-PT, 2015 | Penilaian akreditasi diarahkan pada capaian kinerja tridharma perguruan tinggi (outcome-based accreditation), peningkatan daya saing, dan wawasan internasional (international outlook) pada program studi dan institusi perguruan tinggi. Outcome-based accreditation yang dimaksud di sini adalah, pada akreditasi program studi (APS) berfokus pada ketercapaian capaian pembelajaran lulusan, sedang pada akreditasi perguruan tinggi (APT) adalah ketercapaian visi, misi, dan tujuan perguruan tinggi. Standar Pendidikan Tinggi yang ditetapkan oleh perguruan tinggi, dan peraturan perundang-undangan tentang pengelolaan pendidikan tinggi, serta konformasi (conformance) yang diukur melalui kinerja mutu (performance) dalam konteks akuntabilitas publik; |

Perguruan tinggi adalah satuan penyelenggara pendidikan tinggi yang merupakan tingkat lanjutan dari jenjang pendidikan menengah di jalur pendidikan formal. Hal ini sesuai dengan pengertian perguruan tinggi menurut (UU No. 20 Tentang Sistem Pendidikan Nasional, 2003) yang menyatakan bahwa : perguruan tinggi merupakan jenjang pendidikan setelah pendidikan menengah mencakup program pendidikan diploma, sarjana, magister, spesialis, dan doktor yang diselenggarakan oleh perguruan tinggi. Perguruan tinggi juga perlu mendeteksi perilaku mahasiswa, sehingga dapat diketahui faktor yang menyebabkan kegagalan seorang mahasiswa untuk lulus atau lulus dengan masa studi yang telah ditetapkan, seperti rendahnya kemampuan akademik, usia masuk, indeks prestasi maupun faktor-faktor lainnya (Romadhona, A., Suprapedi , S. dan Himawan, 2017).

Menurut (BAN-PT, 2019) salah satu indikator yang menjadi tolak ukur keberhasilan perguruan tinggi dalam melakukan proses kegiatan belajar mengajar (KBM) adalah angka kelulusan. Angka kelulusan yang tinggi dianggap sebagai sebuah prestasi pada perguruan tinggi yang bersangkutan. Angka kelulusan yang tinggi bisa digunakan sebagai bahan promosi untuk menarik minat calon mahasiswa baru. Jika perguruan tinggi dapat mencapai tingkat kelulusan 100% maka dapat dikatakan perguruan tinggi tersebut sukses dan berhasil.

1. Jurnal Kelulusan Mahasiswa

|  |  |  |  |
| --- | --- | --- | --- |
| Metode | judul | Penulis | hasil |
| Analisis Regresi | Analisis Regresi logistik ordinal terhadap faktor-faktor yang mempengaruhi predikat kelulusan mahasiswa S1 di ITS Surabaya | Siti Imaslihkah, Madu Ratna, dan Vita Ratnasari | Karakteristik mahasiswa ITS menunjukkan bahwa predikat kelulusan yang paling banyak adalah sangat memuaskan. Pre-dikat kelulusan dengan pujian telah menunjukkan angka sebe-sar 23%, hal ini berarti jumlah lulusan mahasiswa dengan mu-dah mencapai predikat tersebut. Namun, berdasarkan jalur masuk ke ITS, jalur masuk PMDK Beasiswa dan PMDK Pres-tasi menunjukkan paling banyak dengan predikat dengan pu-jian. Mahasiswa ITS yang memiliki IPK di atas 3,50 terbanyak 6 berada di fakultas TIF sebanyak 35,3%. Secara pengujian serentak, faktor yang berpengaruh adalah fakultas, jenis kelamin, jalur penerimaan, pekerjaan ayah, pekerjaan ibu dan pendapatan. Pengujian secara individu, faktor yang berpengaruh adalah fakultas, jalur penerimaan, pekerjaan ayah, pekerjaan ibu dan pendapatan. Ketepatan klasifikasi dari model serentak yang didapatkan sebesar 77,41%, yang berarti sudah cukup baik. |
| Linear Discriminant Analysis | Aplikasi Prediksi kelulusan Mahasiswa dengan Metode Linear Discriminant Analysis | Sulistio | Metode analisis diskriminan linier ini digunakan pada modul prediksi yang memisahkan sekelompok siswa ke kelas masing-masing (lulus tepat waktu atau terlambat) dan kemudian menunjukkan hasilnya kepada pengguna. Data yang digunakan untuk input adalah nilai subjek dari masing-masing siswa. Subjek yang digunakan dalam aplikasi ini dipilih oleh kepala jurusan dengan 22 subjek yang dipilih. Untuk tujuan evaluasi, metode cross validation digunakan, dengan hasil akurasi rata-rata sebesar 97%. Jumlah individu yang digunakan untuk proses *testing* sebanyak 100 dengan kegagalan  prediksi sebanyak 7 individu (baris dengan *shading* gelap) sehingga kemungkinan terjadi  kegagalan sebesar 7/100 atau 0.07 dengan demikian persentase keberhasilan adalah 93%. |

Terkait kelulusan mahasiswa telah banyak penelitian yang dilakukan, salah satunya pada penelitian (Imaslihkah et al., 2013) tentang faktor-faktor yang mempengaruhi predikat kelulusan mahasiswa menggunakan analisis *Regresi Logistik*. Pengujian dilakukan menggunakan dua cara yaitu pengujian secara serentak dan pengujian secara individu. Pengujian secara serentak, faktor yang berpengaruh antara lain jalur penerimaan, fakultas, pekerjaan orang tua, jenis kelamin dan pendapatan orang tua. Untuk pengujian individu, faktor yang berpengaruh seperti fakultas, jalur penerimaan, pekerjaan orang tua dan pendapatan. Akurasi yang diperoleh dari model pengujian serentak yaitu sebesar 77,41% yang dirasa sudah cukup baik. Lalu penelitian (Sulistio, 2017) pengimplementasian prediksi kelulusan mahasiswa menggunakan metode *discriminant analysis* berbasis web. Metode *discriminant analysis* ini mengklasifikasikan suatu objek dari objek lain menuju kelas masing-masing. Objek akan dianggap sebagai siswa, jadi metode ini akan memisahkan sekelompok siswa yang nantinya akan di letakkan pada kelas masing-masing. Hasil yang diperoleh dengan menggunakan 100 data *testing* tejadi kegagalan prediksi sebanyak 7, sehingga dapat disimpulkan akurasi yang diperoleh sebesar 93%.

1. Kelulusan dengan metode Klasifikasi

|  |  |  |  |
| --- | --- | --- | --- |
| Metode | judul | Penulis | hasil |
| C4.5, Naive Bayes, KNN dan SVM | Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4.5, Naive Bayes, KNN dan SVM | **Sri Widaningsih** | Metodologi penerapan data mining ini menggunakan tahapan Discovery Knowledge of Databases (KDD) dimulai dari tahap selecting, preprocessing, transformation, data mining dan evaluation/interpretation. Teknik yang akan digunakan untuk model data mining klasifikasi ini terdiri dari empat algoritma yaitu C4.5, Support vector machine (SVM), k-nearest neigbor (kNN,) dan Naïve Bayes. Metoda klasifikasi terdiri dari variabel-variabel prediktor dan satu variabel target. Variabel-variabel prediktor terdiri dari jenis kelamin dan indeks prestasi dari semester 3 hingga 6. Perangkat lunak yang digunakan untuk mengolah data yaitu software Rapid Miner. Hasil akhir dari keempat algoritma tersebut diperoleh bahwa algoritma Naïve Bayes merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu dan IPK ≥ 3 dengan nilai accuracy (76,79%), error (23,17%) , dan AUC (0,850). |

(Widaningsih, 2019) untuk membandingkan empat algoritma yaitu C4.5, *Support vector machine* (SVM), *k-nearest neigbor* (kNN,) dan Naïve Bayes untuk memprediksi nilai dan waktu kelulusan mahasiswa. Variable yang digunakan yaitu jenis kelamin dan nilai indeks prestasi dari semester 3 sampai semester 6. *Software* yang digunakan pada penelitian ini yaitu *Rapidminer*. Hasil yang diperoleh dari perbandingan antara empat algoritma tersebut, bahwa algoritma *Naive bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa tepat waktu dengan IPK lebih dari 3 dan nilai akurasi sebesar 76,79%.

1. Penelitian naive Bayes

|  |  |  |  |
| --- | --- | --- | --- |
| Metode | judul | Penulis | hasil |
| Naive bayes | Analisi Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naive Bayes : Sistematic Review | Lila Setiyani,  Mokhamad Wahidin ,  Dudi Awaludin dan  Sri Purwani | Beradasarkan hasil identifikasi literatur yang digunakan, didapat bahwa metode data mining naïve bayes dapat membuat suatu prediksi mengenai kelulusan mahasiswa tepat waktu dengan memperhitungkan atribut – atribut dari database perguruan tinggi yang digunakan. Sedangkan untuk tingkat akurasi ketiga literatur menghasilkan akurasi di atas 90% walaupun dengan menggunakan jumlah atribut dan aplikasi data mining yang berbeda. Atribut yang terdapat pada semua literatur yang dapat menentukan prediksi adalah atribut IPK |
| Naive bayes | Analisi Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Naive Bayes | Fajar Edi Prabowo dan Achmad Kodar | Berdasarkan hasil pengujian dapat disimpulkan bahwa algoritma Naïve Bayes bisa digunakan untuk memprediksi lama masa studi mahasiswa. Data yang digunakan adalah data kelulusan mahasiswa tahun 2011 hingga 2014 dengan 244 data sebagai data training dan 62 data sebagai testing. Atribut yang digunakan antara lain tahun masuk, jenis kelamin, sks semester, IPK dan asal sekolah mahasiswa. Perhitungan akurasi yang dilakukan menghasilkan nilai sebesar 82,26%. Hasil penelitian yang diperoleh dapat dijadikan sebagai pendukung pengambilan keputusan dalam peningkatan sarana dan prasarana untuk meningkatkan efektifitas kegiatan belajar mengajar di Universitas Mercu Buana. |
| Naive bayes | Optimasi Metode Naive Bayes Dalam Memprediksi Tingkat Kelulusan Mahasiswa Stekom Semarang | **Edy Siswanto** | Dari hasil penelitian yang dilakukan dari tahap awal sampai dengan tahap pengujian penerapan metode naïve bayes untuk proses prediksi kelulusan mahasiswa, didapatkan kesimpulan bahwa Penerapan metode algoritma naïve bayes menggunakan 14 ( empat belas ) parameter yaitu tanggal masuk, kelas, npm, nama, stekom, cabang, tanggal lulus, no. ijazah, sks, IPK, mulai bimbingan, selesai bimbingan, masa studi dan status mahasiswa memperoleh tingkat akurasi sebesar 95,14 %. |

Beberapa penelitian menggunakan *naïve bayes* untuk kasus memprediksi kelulusan mahasiswa, diantaranya : (Setiyani et al., 2020) menghasilkan akurasi diatas 90% dengan jumlah atribut yang berbeda pada setiap literatur, atribut yang terdapat pada semua literatur adalah IPK (indeks prestasi kumulatif). pada penelitian (Prabowo & Kodar, 2019) dengan menggunakan 244 data latih dan 62 data uji pada data kelulusan mahasiswa tahun 2011 sampai 2014, menghasilkan akurasi sebesar 82,26%. Dan juga penelitian yang dilakukan oleh (Siswanto, 2019) penerapan algoritma *naive bayes* menggunakan 14 atribut memperoleh akurasi sebesar 95,14%.

1. Kelemahan Naive Bayes

|  |  |
| --- | --- |
| Sumber | hasil |
| Samuel Natalius, 2011 | Kekurangan dari metoda Naïve Bayes classifier ini adalah banyaknya celah untuk mengurangi keefektifan metoda ini dan akibatnya meloloskan dokumen ke dalam kelas tertentu padahal jelas-jelas dokumen tersebut tidak layak berada di kelas tersebut. Dalam kasus spam filtering, kelemahan ini banyak digunakan oleh spammers berpengalaman untuk meloloskan spam ke dalam kelas bukan spam (menganggap surat elektronik bukan spam padahal sebenarnya adalah spam : Galat tipe II). |
| Triowali Rosandy, 2016 | Salah satu algoritma klasifikasi yang sering digunakan dan mendapat banyak perhatian para peneliti dalam memprediksi pembiayaan bermasalah pada perbankkan adalah Naive Bayes Clasifier dan Decision Tree (C4.5). Tetapi kelemahan yang dihadapi pada kedua algoritma tersebut adalah lamanya waktu dan tingkat akurasi prediksi yang digunakan untuk melakukan prediksi. Masalah ini juga menjadi perhatian banyak peneliti untuk memperbaikinya agar kinerja waktu dan akurasi prediksi menjadi lebih singkat tetapi kinerja akurasi tetap baik. |

Penelitia yang dilakukan oleh (Natalius, 2011) mengungkapkan terdapat kekurangan pada metode *Naïve Bayes Classifier*,dimana pada metode ini memiliki banyak celah yang mengakibatkan pengurangan keefektifitasannya. Seperti meloloskan atribut-atribut yang tidak layak untuk dilakukan proses mining. Lalu penelitian yang dilakukan oleh (Rosandy, 2016) menemukan kelemahan lainnya pada algoritma *naive bayes* yaitu lama waktu dan tingkat akurasi prediksi yang digunakan untuk melakukan prediksi.

1. Feature selection

|  |  |
| --- | --- |
| Sumber | hasil |
| Adnyana, 2019 | Tahapan pertama dari penelitian ini adalah melakukan proses feature selection atau seleksi fitur. Seleksi fitur adalah salah satu teknik data mining yang umum digunakan pada tahapan pre-processing. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan dikelola pada processing dan analisis. Teknik ini dilakukan untuk mengetahui subset fitur yang paling signifikan dari data set nilai mahasiswa. Pemilihan fitur sering digunakan untuk pengurangan dimensi model. Pemilihan fitur membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran/pemodelan [3]. Pada penelitian ini digunakan tiga teknik seleksi fitur, yaitu Correlation Based, Information Gain Based, dan Learner Based. |

Salah satu teknik yang digunakan untuk mengurangi kompleksitas atribut adalah menggunakan seleksi fitur (*Feature selection*). Teknik ini dilakukan untuk memberitahu *subset* fitur yang paling berpengaruh dalah suatu *dataset*, seleksi fitur juga membentu perngurangan dimensi model, mengurangi fitur domain dan menghilangkan fitur yang berlebihan. Dengan cara ini dapat mempercepat proses pemodelan/pembelajaran

1. Penelitian Terkait naive bayes menggunakan Seleksi fitur

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metode | judul | | Penulis | hasil | |
| Naive bayes dan information gain | Penerapan feature selection information gain pada algoritma naive bayes untuk prediksi kelulusan mahasiswa | | Doni Fahrudi | Pengujian akurasi dilakukan dengan metode evaluasi *split validation* menghasilkan akurasi pada metode *Naïve Bayes* yaitu 79.25% dengan rasio perbandingan 90:10. Sedangkan akurasi pada *Naïve Bayes* berbasis *Information Gain* dengan 4 atribut terbaik pada hasil seleksi diantaranya ip semester 1, ip semester 6, tempat lahir dan ip semester 2 yaitu 86.79% dengan rasio perbandingan 90:10. Sehingga terjadinya peningkatan akurasi sebesar 7.54%. | |
| Decision Tree, information gain dan gini index | Decision Tree learning untuk penentuan jalur kelulusan mahasiswa | Winda Widya Ariestya, Yulia Eka Praptiningsih, Wahyu Supriatin3 | | | Metode yang digunakan adalah Iterative Dichotomiser 3 (ID3) dengan menggunakan alat bantu perangkat lunak Rapidminer. Dari hasil penelitian diperoleh 7 rule atau aturan dalam penentuan jalur kelulusan yaitu 3 rule untuk jalur skipsi dan 4 rule untuk jalur non skripsi dan dapat dikatakan algoritma ID3 dapat dimanfaatkan dalam penentuan jalur kelulusan dengan nilai akurasi antara 0,85-1,00. |
| Naive Bayes Dan Gain Ratio | Optimasi Naive Bayes Dengan pemilihan Fitur dan Pembobotan Gain Ratio | I. Gusti. A. Socrates , Afrizal L. Akbar, dan M. Sonhaji Akbar | | | Metode Weighted Naïve Bayes dapat mengoptimalkan nilai akurasi metode Naïve Bayes biasa. Hal ini dapat dilihat dari hasil akurasi Weighted Naïve Bayes sebesar 94% dibandingkan dengan Naïve Bayes biasa sebesar 92%. Weighted Naïve Bayes dapat menghasilkan tingkat akurasi yang lebih tinggi dikarenakan setiap probabilitas dari atribut diberi bobot yang menghasilkan nilai yang lebih tinggi. Ketika dilakukan pemilihan fitur mengunkan 10 term terbaik didapatkan akurasi sebesar 94% untuk metode usulan dan 91% untuk metode Naïve Bayes biasa. Hal ini dapat disimpulkan bahwa pemilihan fitur dapat mengatasi kesalahan klasifikasi. |

Penelitian yang berkaitan dengan kelulusan mahasiswa menggunakan seleksi fitur *information gain* pada *naive bayes* telah dilakukan oleh (doni, 2020) dimana akurasi metode *naive bayes* sebesar 79,25% dapat ditingkatkan menggunakan seleksi fitur *information gain* dengan hasil 86,79%. Terdapat pengembangan dari *information gain* disebut *gain ratio*, *gain ratio* merupakan modifikasi dari *information gain* yang mengurangi biasnya. Penelitian yang dilakukan oleh (Ariestya et al., 2016) dimana melakukan perbandingan antara seleksi fitur *information gain, gain ratio* dan *gini index* pada *decision tree* untuk menentukan jalur kelulusan mahasiswa, memberikan hasil *gain ratio* tertinggi dengan akurasi sebesar 100% diikuti dengan *information gain* sebesar 90% dan *gini index* sebesar 85%. Dan penelitian (Socrates et al., 2016) untuk mengoptimalkan nilai akurasi *naive bayes* dengan menggunakan fitur seleksi *gain ratio*. Hasil yang diperoleh terbukti dengan menggunakan fitur seleksi *gain ratio* dapat meningkatkan akurasi metode *naive bayes* dari 91% ke 94%.

# LANDASAN TEORI

### Data Mining

|  |  |
| --- | --- |
| Sumber | hasil |
| Larose, 2014 | Data mining is an interdisciplinary field bringing togther techniques from  machine learning, pattern recognition, statistics, databases, and visualization to  address the issue of information extraction from large data bases” |
| Jiawei han, 2012 | Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams. |

Data mining menurut (Larose & Larose, 2014) merupakan korelasi, pola dan arah yang baru dengan menggunakan teknologi pengenalan pola serta statistika dan teknik matematika pada penyaringan sejumlah data yang besar dalam repositori. lalu menurut (Asriningtias & Mardhiyah, 2014) data mining adalah suatu kegitan untuk menemukan pola yang menarik dalam jumlah data yang besar, data dapat disimpan didalam *data base*, *data warehouse* atau penimpanan lainnya. Dan pendapat lain (Jiawei Han, 2012) data mining adalah proses menemukan pola yang menarik dan sejumlah besar pengetahuan dari data yang besar. Sumber data tersebut bisa berupa *database*, gudang data, web, dan repositori informasi maupun data langsung ke sistem.

### Tahapan Data Mining Dalam KDD

|  |  |
| --- | --- |
| Sumber | hasil |
| Handoko & lesmana, 2018 | Saat ini, konsep Data Mining semakin dikenal sebagai tools penting dalam manajemen informasi karena jumlahinformasi yang semakin besar jumlahnya. Data Mining sendiri sering disebut sebagai Knowledge Discovery in Database (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam set data berukuran besar |

Data Mining sendiri sering disebut sebagai *Knowledge Discovery in Database* (KDD) merupakan suatu kegiatan yang melakukan penngumpulan, pemakaian data masa lampau untuk menemukan informasi seperti hubungan suatu pola dalam *dataset* yang berukuran besar (Handoko & Lesmana, 2018). beberapa Tahapan-tahapan pada *data mining* ialah seperti pembersihan data (*data cleaning*), seleksi data (*data selection*), transformasi data (*data transformation*), proses mining, dan evaluasi pola (*pattern evaluation*).

### Teknik Data Mining

|  |  |
| --- | --- |
| Sumber | hasil |
| Maulana & fajrin, 2018 | Pengelompokan *Data Mining* dibagi menjadi beberapa kelompok [1] yaitu :  a. Deskripsi  Deskripsi merupakan cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data yang dimiliki.  b. Estimasi  Estimasi hampir sama dengan klasifikasi, kecuali *variable* target estimasi lebih ke arah numerik daripada ke arah kategori. Model yang dibangun menggunakan *record* lengkap yang menyediakan nilai *variable* target sebagai nilai prediksi.  c. Prediksi  Prediksi menerka sebuah nilai yang belum diketahui dan juga memperkirakan nilai untuk masa mendatang.  d. Klasifikasi  Dalam klasifikasi terdapat target *variable* kategori, misal penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang, dan rendah.  e. Pengklasteran  Merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.  f. Asosiasi  Asosiasi bertugas menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. |

pengelompokan *data mining* dibagi menjadi beberapa kelompok yaitu :

* + - 1. Deskripsi

Deskripsi merupakan cara untuk menggambarkan pola dan memiliki kecenderungan yang terdapat dalam data yang tersedia.

* + - 1. Estimasi

Memiliki kemiripan dengan klasifikasi, tetapi dalam target variabel lebih kearah numerik ketimbang arah kategori

* + - 1. Prediksi

Prediksi merupakan memperkirakan atau menerka suatu nilai yang belum diketahui pada masa mendatang

* + - 1. *Clustering*

Clustering adalah suatu metode pengelompokan *record*, pengamatan, atau pembentukan kelas objek-objek yang memiliki kemiripan.

* + - 1. Asosiasi

Asosiasi merupakan metode yang mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

* + - 1. Klasifikasi

Dalam klasifikasi terdapat target kategori variabel, seperti penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang, dan rendah.

### Feature Selection

|  |  |
| --- | --- |
| Sumber | hasil |
| Adnyana, 2019 | Tahapan pertama dari penelitian ini adalah melakukan proses feature selection atau seleksi fitur. Seleksi fitur adalah salah satu teknik data mining yang umum digunakan pada tahapan pre-processing. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan dikelola pada processing dan analisis. Teknik ini dilakukan untuk mengetahui subset fitur yang paling signifikan dari data set nilai mahasiswa. Pemilihan fitur sering digunakan untuk pengurangan dimensi model. Pemilihan fitur membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran/pemodelan . Pada penelitian ini digunakan tiga teknik seleksi fitur, yaitu Correlation Based, Information Gain Based, dan Learner Based. |

Seleksi fitur *(Feature selection*) digunakan untuk mengurangi kompleksitas atribut yang nantinya akan dikelola. Teknik ini dilakukan untuk memberitahu *subset* fitur yang paling berpengaruh dalah suatu *dataset*, seleksi fitur juga membentu perngurangan dimensi model, mengurangi fitur domain dan menghilangkan fitur yang berlebihan. Dengan cara ini dapat mempercepat proses pemodelan/pembelajaran (Adnyana, 2019).

### Pembagian Fitur Selection

|  |  |
| --- | --- |
| Sumber | hasil |
| Adnyana, 2019 | Algoritma Feature selection dibagi menjadi tiga kelompok: filters, wrappers, dan embedded selectors.   * Filters mengevaluasi setiap feature secara bebas dari pengklasifikasi, memberikan peringkat pada feature setelah mengevaluasi dan mengambil yang unggul . * Wrappers mengambil subset dari feature set, mengevaluasi kinerja pengklasifikasian pada subset ini, dan kemudian subset lainnya dievaluasi oleh pengklasifikasi. Subset yang memiliki kinerja paling maksimum pada pengklasifikasian yang akan dipilih. Jadi wrappers bergantung pada pengklasifikasi yang dipilih. Bahkan wrappers lebih dapat diandalkan karena algoritma klasifikasi mempengaruhi tingkat akurasi . * Teknik Embedded disisi lain melakukan feature selection selama proses mempelajari data sama seperti yang dilakukan jaringan syaraf tiruan. |

* + - 1. Wrappers

Teknik wrapper akan mengambil subset dari suatu set fitur, dengan mengevaluasi kinerja klasifikasi pada subset, lalu subset lain akan di evaluasi menggunakan pengklasifikasi. Subset yang memiliki kinerja paling tinggi pada klasifikasian akan dipilih, dengan artian wrapper lebih dapat diandalkan untuk klasifikasi dengan kepentingan akurasi.

* + - 1. Embedded

Algoritma *Decision Tree* mewakili di antara model *Embedded*, yang memilih atribut dengan kemampuan klasifikasi potensial terbesar di setiap *node* untuk membagi subruang.Teknik Embedded melakukan feature selection selama proses mempelajari data sama seperti yang dilakukan jaringan syaraf tiruan.

* + - 1. Filter

Metode filter akan mengevaluasi setiap fitur secara bebas dari klasifikasi, Metode *Filter* menggunakan kriteria penilaian yang tepat yang mencakup jarak, informasi, ketergantungan dan konsistensi. lalu akan memberikan peringkat pada feature yang telah dievaluasi dan mengambil yang memiliki bobot tinggi. Beberapa contoh metode filter yaitu seperti *Information Gain, Symmetrical Uncertainty , Gain Ratio* dan lain-lain.

### Information Gain

|  |  |
| --- | --- |
| Sumber | hasil |
| (Chormunge & Jena, 2016) | Information gain is a simplest attribute ranking method, widely used in text categorization  applications and now days for microarray data analysis and image data analysis. If A is an attribute and C is the class, Equations A1 and A2 show the entropy of the class before and after observing the attribute |
| (Bimantoro & Uyun, 2017) | Peran Information gain pada decision tree adalah untuk menyusun struktur tree.Hanya saja dalam penyusunan jaringan tree tidak dilakukan perhitungan jarak antara entropy keseluruhan data.Nilai information gain diperoleh dari nilai entropy sebelum pemisahan dikurangi dengan nilai entropy setelah pemisahan. Pengukuran nilai ini hanya digunakan sebagai tahapan awal untuk penentuan atribut yang nantinya akan digunakan atau dibuang. Atribut yang memenuhi kriteria pembobotan yang nantinya akan digunakan dalam proses pengenalan pola. |

Information gain merupakan metode seleksi fitur yang paling sederhana, dimana algoritma ini melakukan perangkingan atribut pada dataset, metode ini banyak digunakan pada aplikasi kategorisasi teks, analilis data dan analisis data citra (Chormunge & Jena, 2016). Pengukuran nilai information gain didapatkan dari nilai entropy sebelum pemisahan dikurangi dengan nilai setelah pemisahan. Pengukuran ini akan digunakan sebagai tahap awal untuk menentukan atribut yang akan digunakan dan dibuang. Atribut yang memiliki kriteria atau bobot tertinggi nantinya akan digunakan untuk proses klasisfikasi (Bimantoro & Uyun, 2017).

### Symetrical uncertainly

|  |  |
| --- | --- |
| Sumber | hasil |
| (Piao et al., 2019 | The most often used feature selection method was filter-based such as mutual information,  Pearson correlation, chi-squared test, information gain, gain ratio and relief. Yu and Liu proposed  the fast correlation-based filter (FCBF) method to remove the irrelevant and redundant features.  The measurement of symmetrical uncertainty (SU) was defined to measure the redundancy |
| Saikhu et al., 2019) | In this paper, we propose a new integration strategy between Pearson Correlation and Symmetrical Uncertainty for relevant feature selection based on linear and non-linear relationships for multivariate time-series classification. |

Metode ini merupakan metode yang paling sering digunakan berbasis filter informasi, metode filter ini dengan berkolerasi dengan cepat untuk menghapus fitur yang tidak relevan dan berlebihan. Pengukuran metode ini digunakan untuk mengukur rudundansi (Piao et al., 2019). Dan merurut (Saikhu et al., 2019) metode Feature Selection yang digunakan untuk mengevaluasi hubungan non-linear antara variabel dan class.

### Gain Ratio

|  |  |
| --- | --- |
| Sumber | hasil |
| Priyadarsini et al., 2012) | In this paper high ranking features which play an important role in data mining algorithms results are selected using information gain feature selection method and these attributes, which can be linked to identify the underlying individual have been set as quasi identifiers of the dataset. High dimensional datasets like adult and census datasets have been used in this work. A hybrid perturbation method is proposed where both generalization and noise additive perturbation is combined for perturbing the values of the quasi identifiers. The utility of the privacy preserved datasets is evaluated by comparing their accuracy on the two functionalities of data mining namely classification and clustering .The maintenance of rank of attribute before and after perturbation is also evaluated. |

Pengembangan dari *information gain* disebut *gain ratio*, *gain ratio* merupakan modifikasi dari *information gain* yang mengurangi biasnya. *Gain ratio* mengambil angka dan ukuran dari cabang kedalam akun ketika memilih sebuah atribut, cara ini akan mengoreksi *information gain* dalam mengambil unsur informasi dari pecahan ke sebuah akun (Priyadarsini et al., 2012).

### Klasifikasi

|  |  |
| --- | --- |
| Sumber | hasil |
| (Jiawei Han, 2012). | Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example,we can build a classificationmodel to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large.Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis. We start off by introducing the main ideas of classification |

Klasifikasi adalah proses menemukan model (fungsi) yang menggambarkan dan membedakan kelas data atau kosep, untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Proses klasifikasi data, data latih akan dianalisa dengan algoritma klasifikasi. Disini label kelas adalah keputusan yang dipinjam, dan model yang dipelajari atau pengklasifikasi berbentuk aturan klasifikasi. Data uji digunakan untuk memperkirakan akurasi dari aturan klasifikasi. Jika akurasi bisa diterima, maka aturan dapat diterapkan pada klasifikasi data baru. Model yang didapatkan didasarkan pada analisis dari data latih (objek yang label kelasnya diketahui), terdapat banyak metode klasifikasi seperti *support vector machine*, *k-nearest-neighbor, naive bayes* dan lain-lain (Jiawei Han, 2012).

### Support Vector Machine

|  |  |
| --- | --- |
| Sumber | hasil |
| (Nugroho, 2007) | Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai state of the art dalam pattern recognition, dan dewasa ini merupakan salah satu tema yang berkembang dengan pesat. |
| (Octaviani et al., 2014) | SVM memiliki prinsip dasar linier *classifier* yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada problem non-linier dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* (*hyperplane*) yang dapat memaksimalkan jarak (margin) antara kelas data. |

*Support Vector Machine* (SVM) diperkenalkan pertama kali oleh Vapnik di tahun 1992 sebagai suatu rangkaian dengan konsep yang unggul pada bidang pengenalan pola. Metode ini masih terbilang muda, walapun begitu kemampuan evaluasinya dalam berbagai hal menempatkannya menjadi salah satu tema yang berkembang pesat (Nugroho, 2007). *Support Vector Machine* (SVM) sendiri memiliki prinsip dasar linier clasifier, dimana klasifikasi yang secara linear dapat dipisahkan. Tetapi *Support Vector Machine* (SVM) telah dikembangkan agar dapat beroperasi pada permasalahan *non-linear* dengan cara memasukkan konsep kernel (Octaviani et al., 2014).

### K-Nearest Neighbor

|  |  |
| --- | --- |
| Sumber | hasil |
| (Bode, 2017). | Algoritma *K-Nearest Neighbor* (KNN) adalah salah satu metode yang menerapkan algoritma *supervised* yang mana algoritma tersebut terbagi atas dua jenis yaitu *supervised learning* dengan *unsupervised learning.* Algoritma *supervised learning* bertujuan untuk mendapatkan pola baru sedangkan *unsupervised learning* untuk mendapatkan pola dalam sebuah data Ketepatan algoritma KNN ditentukan oleh ada dan tidak adanya data yang tidak relevan, atau jika bobot fitur tersebut setara dengan relevansinya terhadap klasifikasi . Algoritma KNN adalah salah satu metode yang digunakan untuk analisis klasifikasi, namun beberapa dekade terakhir metode KNN juga digunakan untuk prediksi. |
| (Yustanti, 2012) | Ketepatan algoritma KNN sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Algoritma KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar. Sedangkan, kelemahan KNN adalah KNN perlu menentukan nilai dari parameter K (jumlah dari tetangga terdekat), training berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik, dan biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap query instance pada keseluruhan training sample. |

Algoritma *K-Nearest Neighbor* (KNN) merupakan salah satu metode yang menerapkan *supervised learning*, dengan kata lain algoritma ini bertujuan untuk menemukan pola baru. Ketepatan akurasi algoritma ini ditentukan oleh ada atau tidaknya data yang tidak relevan. Algoritma ini biasanya digunakan untuk melakukan proses analisis klasifikasi, tetapi belakangan ini metode KNN dapat juga diguanakan untuk prediksi (Bode, 2017). Perhitungan metode ini dilakukan berdasarkan data pembelajaran data yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran akan ditampilkan ke ruang berdimensi banyak, dimana masing masing dimensi menampilkan fitur dari *dataset*. Ruang tersebut akan dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Ruang akan ditandai sebuah titik kelas c, dimana kelas c merupakan klasifikasi yang paling bnyak ditemukan pada buah tetangga terdekat (Yustanti, 2012).

### Naive Bayes

|  |  |
| --- | --- |
| Sumber | hasil |
| (Bustami, 2014) | Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. |
| (Saleh, 2015) | Persamaan di atas merupakan model dari teorema *Naive Bayes* yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss* :  𝑃(𝑋𝑖=𝑥𝑖|𝑌= 𝑦𝑗)= 1√2𝜋𝜎𝑖𝑗𝑒−(𝑥𝑖−𝜇𝑖𝑗)22𝜎2𝑖𝑗 |

*Naive Bayes classifier* (NBC) merupakan salah satu algoritma pada teknik klasifikasi yang ditemukan oleh ilmuan inggris Thomas Bayes. *Naive bayes* merupakan salah satu metode pembelajaran mesin dengan perhitungan probabilitas dan statistik yang memprediksi peluang dimasa depan dengan menggunakan pengalaman dimasa sebelumnya atau yang lebih dikenal dengan teorema *bayes*. Dikombinasikan dengan *naive* dimana bahwa semua atribut independen atau tidak ada ciri tertentu dari sebuah kelas yang berhubungan dengan ciri kelas lainnya (Bustami, 2014).

Dalam penelitian (Saleh, 2015)klasifikasi menggunakan data numerik menggunakan rumus *densitas gauss* :

|  |  |
| --- | --- |
|  | (2.5) |

Keterangan :

P : Peluang

Xi : Atribut ke i

xi : Nilai atribut ke i

Y : Kelas yang dicari

yi : Sub kelas Y yang dicari

µ : *mean* (rata – rata dari seluruh atribut)

σ : *Deviasi* standar (menyatakan varian dari seluruh atribut)