INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

Pembuatan Aplikasi Data Mining Untuk Memperediksi Masa Studi Mahasiswa Menggunakan Algoritma *Naive Bayes*

Victor Tarigan

Teknik Informatika, Fakultas Teknik, Universitas Sam Ratulangi

Email : victortarigan@unsrat.ac.id

Abstract

Data mining is a series of processes to obtain additional value information that is not known manually from a database. Data mining also manages experience or even mistakes in the past to improve the quality of the analysis model, one of which is the learning ability of data mining techniques, namely classification. Class is a learning task that is a new object into one of the class labels or categories on the old object that has been previously defined. This classification uses one method of data mining algorithm that is Naive Bayes.Naive Bayes algorithm works based on a certain distance between two objects by setting the value of k. The value of k is a parameter to determine the distance between the new object to the old object.By using the data mining technique, the university can obtain student academic data, namely the Achievement Index (IP) to predict the student's study period.In this data mining application consists of data testing and data training with NIM input.PHP and the database used is MySQL.The results of this data mining application this system can predict the results of the classification of student study period based on GPA 4 first semester, the average value of high school time, and grades in high school.

Keywords: *Data Mining, Classification, Naive Bayes Algorithm, Student.*

I. Pendahuluan

Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuanpotensial dan berguna yang tersimpan di dalam database besar. Memiliki jumlah data yang sangat besar, misalnya data dosen, pegawai, sarana prasarana dalam Perguruan tinggi dapat melakukan analisa saat ini dituntut untuk memiliki kemampuan bersaing dengan memanfaatkan semua sumber daya yang dimiliki.

Sebagaimana yang telah dibentuk sebelumnya bahwa data yang disebutkan adalah data mahasiswa dalam data set misalnya data mahasiswa yang potensialdrop out. Pemahaman profil mahasiswa yang potensial drop out penting untuk diketahui Ukuran keberhasilan atau prestasi mahasiswa dapat dilihat dari Indeks Prestasi Kumulatif (IPK) mencerminkan seluruh nilai yang diperoleh mahasiswa sampai semester yang sedang berjalan, yang menunjukkan prestasi akademik

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

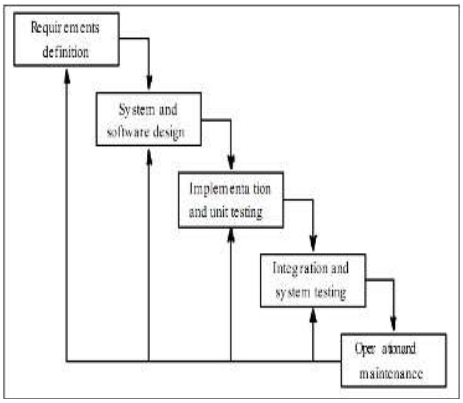
mahasiswa bersangkutan sampai semester tersebut. IPK diperoleh dengan cara menjumlahkan seluruh nilai mutu semua mata kuliah yang telah diambil dan membaginya dengan total sks (satuan kredit semester).

Dengan bantuan teknik data mining, seperti algoritma klasifikasi, yang memungkinkan untuk menemukan karakteristik-karakteristik dari prestasi mahasiswa dan menggunakan karakteristik mereka untuk memprediksi prestasi di masa depan. Penelitian ini sendiri dilakukan di universitas panca budi dimana data mahasiswa yang diperoleh dari sistem informasi Universitas Pancabudi lima tahun terakhir.

Dengan menggunakan salah satu teknik data mining yaitu metode klasifikasi menggunakan algoritma *Naive Bayes* dengan harapan dapat menemukan informasi tingkat kelulusan dan persentase kelulusan mahasiswa sehinga dapat digunakan oleh pihak Jurusan untuk mencari solusi atau kebijakan dalam proses evaluasi pembelajaran di jurusan tersebut.

II. Metodologi Penelitian

Metode yang digunakan pada kasus ini adalah model *Waterfall*. Model ini mengusulkan sebuah pendekatan kepada perkembangan *software* yang sistematik yang mulai pada tingkat dan kemajuan sistem pada seluruh analisis, desain, kode, pengujian, dan pemeliharaan. Proses yang terdapat dalam model *Waterfall* dapat dilihat pada gambar berikut ini:



Gambar 1. Model Waterfall

Penjelasan mengenai tahapan-tahapan yang terdapat dalam gambar 1model *Waterfall* adalah sebagai berikut ini:

1. *System Engineering*
   Rekayasa perangkat lunak merupakan tahapan yang pertama kali dilakukan untuk merumuskan sistem yang akan dibangun. Hal ini bertujuan untuk memahami sistem yang akan dibangun.
2. *Analysis*
   Analisis dilakukan terhadap permasalahan yang dihadapi serta untuk menetapkan kebutuhan perangkat lunak dari aplikasi yang dibangun.
3. *Design*
   Tahap desain merupakan tahap penerjemahan dari data yang telah dianalisis ke dalam bentuk yang mudah dimengerti oleh pengguna.
4. *Coding*
   Coding merupakan tahap penerjemahan data yang telah dirancang ke dalam bahasa pemrograman tertentu.
5. *Testing*
   Tahap pengujian dilakukan terhadap perangkat lunak yang

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

telah dibangun. Proses pengujian berfokus pada logika internal perangkat lunak serta memastikan apakah hasil yang diinginkan tercapai atau tidak.

6. *Maintenance*
   Maintenance merupakan penanganan dari suatu perangkat lunak yang telah selesai dibangun sehingga dapat dilakukan perubahan-perubahan atau penambahan sesuai dengan permintaan pengguna.

III. Landasan Teori

Data Mining

Data Mining merupakan proses penggalian data dari database yang berukuran besar untuk menemukan informasi penting dan bermanfaat. Klasifikasi adalah salah satu teknik yang ada pada data mining. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. *Data mining* merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

Data mining juga merupakan proses untuk menemukan hubungan baru yang mempunyai arti, pola dan kebiasaan dengan memilah-milah sebagian besar data yang disimpan dalam media penyimpanan dengan menggunakan teknologi pengenalan pola seperti teknik statistic dan matematika. *Data mining* merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar.

Data mining mempunyai fungsi yang penting untuk membantu mendapatkan informasi yang berguna serta meningkatkan pengetahuan bagi pengguna. Pada dasarnya, *data mining* mempunyai empat fungsi dasar yaitu :

1. Fungsi Klasifikasi (*Classification*)
   Data mining dapat digunakan untuk mengelompokkan data-data yang jumlahnya besar menjadi data-data yang lebih kecil.
2. Fungsi Segmentasi (*Segmentation*)
   Data mining dapat digunakan untuk melakukan segmentasi (pembagian) terhadap data berdasarkan karakteristik tertentu.
3. Fungsi Asosiasi (*Association*)
   Data mining digunakan untuk mencari hubungan antara karakteristik tertentu.
4. Fungsi Pengurutan (*Sequencing*)
   Data mining digunakan untuk mengidentifikasikan perubahan pola yang telah terjadi dalam jangka waktu tertentu

Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia.Dalam klasifikasi ada dua pekerjaan utama yang dilakukan.Pertama, pembangunan model sebagai prototipe untuk disimpan sebagai memori.Kedua, penggunaan model untuk melakukan pengenalan/klasifikasi/prediksi pada

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

suatu objek lain, agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya.

Klasifikasi dapat didefinisikan sebagai perkerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang memetakan setiap set atribut (fitur) x ke satu dari sejumlah label kelas y yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori.

Model dalam klasifikasi mempunyai arti yang sama dengan kotak hitam, di mana ada suatu model yang menerima masukan, kemudian mampu melakukan pemikiran terhadap masukan tersebut, dan memberikan jawaban sebagai keluaran dari hasil pemikirannya.

Naïve Bayes

Naive Bayes Classifier merupakan sebuah proses klasifikasian probabilistik sederhana yang berdasarkan pada penerapan Teorema *Bayes* (atau Aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat, dengan kata lain, dalam *Naïve Bayes*, model yang digunakan adalah model fitur independen. Dalam Bayes (terutama Naïve Bayes),maksud independensi yang kuat pada fitur adalah sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Keuntungan dari klasifikasi adalah metode ini hanya membutuhkan sejumlah kecil data pelatihanuntuk memperkirakan parameter (sarana dan varians dari variabel) yang diperlukan untuk klasifikasi.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output,probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metodeini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalamproses pengklasifikasian.Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan. Persamaan dari teorema Bayes adalah.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu class spesifik

P(H|X): Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

P(H) : Probabilitas hipotesis H (prior probabilitas)

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

Adapun alur dari metode Naive Bayes adalah sebagai berikut :

1. Mulai
2. Baca data training
3. Tampilkan hasil prediksi

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

1. Hitung P(Ci) untuk setiap kelas
2. Hitung P(X|Ci) untuk setiap kriteria dan setiap kelas
3. Cari P(X|Ci) yang paling besar menjadi kesimpulan

Basis Data

Basis data (bahasa Inggris: *database*), atau sering pula dieja basisdata, adalah kumpulan informasi yang disimpan di dalam komputer secara sistematik sehingga dapat diperiksa menggunakan suatu program komputer untuk memperoleh informasi dari basis data tersebut. Perangkat lunak yang digunakan untuk mengelola dan memanggil kueri (*query*) basis data disebut sistem manajemen basis data (*database management system*, DBMS).

Sistem basis data dipelajari dalam ilmu informasi Istilah "basis data" berawal dari ilmu komputer. Meskipun kemudian artinya semakin luas, memasukkan hal-hal di luar bidang elektronika, artikel ini mengenai basis data komputer.Catatan yang mirip dengan basis data sebenarnya sudah ada sebelum revolusi industri yaitu dalam bentuk buku besar, kuitansi dan kumpulan data yang berhubungan dengan bisnis.

Konsep dasar dari basis data adalah kumpulan dari catatan-catatan, atau potongan dari pengetahuan. Sebuah basis data memiliki penjelasan terstruktur dari jenis fakta yang tersimpan di dalamnya: penjelasan ini disebut skema. Skema menggambarkan obyek yang diwakili suatu basis data, dan hubungan di antara obyek tersebut. Ada banyak cara untuk mengorganisasi skema, atau memodelkan struktur basis data: ini dikenal sebagai model basis data atau model data.

IV. Analisa dan Pembahasan

Analisis Sistem

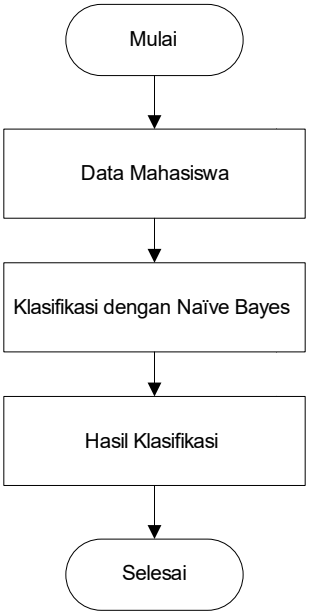
Universitas Sam Ratulangi merupakan perguruan tinggi Negeri di Manado.Universitas Sam Ratulangi memiliki beberapa program studi salah satunya adalah Program Studi Teknik Elektro.Program Studi ini termasuk kategori yang sangat susah untuk dapat lulus tepat waktu.Setiap tahun, Program Studi Teknik Elektro hanya menghasilkan beberapa mahasiswa yang lulus tepat 4 tahun atau 5 tahun.

Karena jumlah kelulusan tiap tahunnyahanya sedikit, maka penulis memanfaatkan data nilai IP mahasiswa di JurusanTeknik Komputer untuk menemukan informasi atau pengetahuan baru yangberguna dalam mengambil sebuah keputusan dan membantu dalam evaluasisistem pembelajaran di Jurusan Teknik Komputer. Informasi yang dibutuhkan adalah memprediksi masa studi mahasiswa dengan atribut IP dari semester satusampai semester empat, jurusan mahsiswa pada saat SMA/SMK, dan nilai ujian nasional pada saat SMA/SMK.

***Framework* Penelitian**

Adapun *framework* dari penelitian ini adalah sebagai berikut :

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)



Gambar 2. Framework Penelitian

Analisis Perhitungan Klasifikasi Masa Studi Mahasiswa Dengan Metode *Naive Bayes*

Untuk mencari klasifikasi masa studi mahasiswa Sistem Komputer diperlukan data training untuk menentukan masa studi mahasiswa seperti tabel 1 di bawah ini :

Tabel 1. Klasifikasi Masa Studi Mahasiswa

No	Nama	IP Semester 1	IP Semester 2	IP Semester 3	IP Semester 4	Rata-rata Nilai Ujian Nasional	Jurusan SMA	Masa Studi
1	Andi	3.00	3.11	1.89	1.70	85	IPA	> 4 Tahun
2	Anita	2.72	2.32	1.89	1.60	78	IPS	> 4 Tahun
3	Anggi	3.28	3.37	3.32	2.00	87	IPA	<=4 Tahun
4	Farhan	3.61	3.16	2.89	3.11	82	IPA	<=4 Tahun
5	Fatin	3.50	3.42	3.27	2.84	89	IPS	<=4 Tahun
6	Sinka	2.83	2.63	2.56	2.25	78	IPA	> 4 tahun
7	Fajar	1.50	1.89	1.93	1.44	81	IPS	> 4 tahun
8	Widya	3.06	3.89	3.53	3.55	89	IPS	<=4 tahun
9	Nikita	3.11	3.21	2.89	3.22	88	IPA	<=4

Setelah mendapatkan data untuk dijadikan sebuah data training, kemudian akandibuatkan data test yang terliht pada tabel di bawah ini atau data baru dan ditentukan apakah data baru tersebut berada di level yang mana.

Tabel 2. Data Testing

No	Nama	IP Sem 1	IP Sem 2	IP Sem 3	IP Sem4	Rata-rata Nilai UAN	Jurusan SMA	Masa Studi
1	Agus	3,97	3.79	3.88	3.35	90	IPA	-

Sebelum mencari nilai masa studi untuk data testing, akan ditentukan dulu batas-batas nilai untuk IP semester 1-4 dan rata-rata nilai ujian nasional seperti yang ditunjukkan pada tabel di bawah ini.

No	Atribut	Batas
1	IP Semester 1	<= 2.8 2.81-3.2 >=3.21
2	IP Semester 2	<= 2.8 2.81-3.2 >=3.21
3	IP Semester 3	<= 2.8 2.81-3.2 >=3.21
4	IP Semester 4	<= 2.8 2.81-3.2 >=3.21
5	Rata-rataNilaiUjian Nasional	<= 80 81-86 >=87

Setelah melakuakn proses ambang nilai untuk kritera yang mempunyai tipe

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

nilai integer kemudian akan dilakukan proses klasifikasi dengan menggunakan metode naïve bayes berdasarkan data training yang ada seperti yang akan ditampilkan pada proses di bawah ini.

P(Ci)
P(Masa Studi > 4 tahun) = 5/10 = 0.5
P(Masa Studi <= 4 tahun) = 5/10 = 0.5
P(X|Ci)
Data Testing Untuk Semester 1 adalah : 3.97, jadi ambang batas yang diambil adalah >= 3.21

1. IP Semester 1
   P(IP Semester 1 = " >= 3.21" | Masa Studi = " > 4 tahun ") = 0 / 5 = 0
   P(IP Semester 1= " >= 3.21" | Masa Studi = " <= 4 tahun ") = 3 / 5 = 0.6
2. IP Semester 2
   Data Testing Untuk Semester 2adalah :3.79, jadi ambang batas yang diambil adalah >= 3.21
   P(IP Semester 2 = " >= 3.21" | Masa Studi = " > 4 tahun ") = 0 / 5 = 0
   P(IP Semester 2 = " >= 3.21" | Masa Studi = " <= 4 tahun ") = 4 / 5 = 0.8
3. IP Semester 3
   Data Testing Untuk Semester 3 adalah : 3.88, jadi ambang batas yang diambil adalah >= 3.21
   P(IP Semester 3 = " >= 3.21" | Masa Studi = " > 4 tahun ") = 0 / 5 = 0
   P(IP Semester 3 = " >= 3.21" | Masa Studi = " <= 4 tahun ") = 3 / 5 = 0.6
4. IP Semester 4
   Data Testing Untuk Semester 4 adalah : 3.35, jadi ambang batas yang diambil adalah >= 3.21
   P(IP Semester 4 = " >= 3.21" | Masa Studi = " > 4 tahun ") = 0 / 5 = 0
   P(IP Semester 4 = " >= 3.21" | Masa Studi = " <= 4 tahun ") = 2 / 5 = 0.4
5. Nilai Ujian
   Data Testing Untuk Rata-rata nilai ujian nasional adalah : 90, jadi ambang batas yang diambil adalah >= 87
   P(Nilai UAN = " >= 87" | Masa Studi = " > 4 tahun ") = 0 / 5 = 0
   P(Nilai UAN = " >= 87" | Masa Studi = " <= 4 tahun ") = 3 / 5 = 0.6
6. Penjurusan
   Data Testing Untuk Penjurusan adalah : IPA
   P(jurusan = "IPA" | Masa Studi = " > 4 tahun ") = 3 / 5 = 0.6
   P(jurusan = "IPA" | Masa Studi = " <= 4 tahun ") = 4 / 5 = 0.8
   P(X | Masa Studi > 4 tahun) = 0 * 0 * 0 * 0 * 0 * 0.6 = 0
   P(X | Masa Studi <= 4 tahun) = 0.6 * 0.8 * 0.6 * 0.4 * 0.6 * 0.8 =0.055296
   P(X|Ci)*P(Ci)
   P(Masa Studi > 4 tahun) * P(X | Masa Studi > 4 tahun) = 0 * 0.5 = 0
   P(Masa Studi <= 4 tahun) * P(X | Masa Studi <= 4 tahun) = 0.055296 * 0.5 = 0.0276

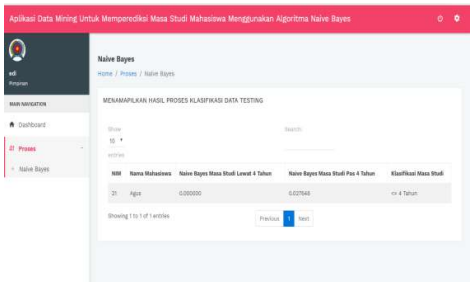
   Jadi untuk data testing yang ada, masa studi kuliah termasuk dalam klasifikasi <= 4 tahun

Hasil Penerapan Naive Bayes

Penerapan Algoritma Naïve Bayes untuk mengklasifikasi masa

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

studi mahasiswa sendiri akan diterapkan dengan menggunakan bahasa pemrograman PHP yang akan menampilkan proses klasifikasi dengan menggunakan metode naive bayes dimanakolom tabel yang ditampilkan adalah nim, nama mahasiswa, nilai probabilitas naive bayes untuk masa studi lewat 4 tahun, nilai probabilitas naive bayes untuk masa studi pas 4 tahun, dan klasifikasi masa studi berdasarkan data training yang ada. Penerapan algoritma Naïve Bayes dapat dilihat pada gambar.



Gambar 3. Hasil Penerapan Naive Bayes

V. Kesimpulan dan Saran

Kesimpulan

Berdasarkan permasalahan yang di pada Aplikasi Diagnosa Penyakitpada gigi dan mulut Serta Cara Penanganannya Berbasis Web dengan metode Forward Chaining, maka dapat diambilbeberapa kesimpulan adalah sebagaiberikut:

1. Aplikasi ini memudahkan masyarakat umum untuk mendiagnosa lebih dini jenis penyakit pada gigi dan mulut dimana saja sehingga penanganan lebih lanjut terhadap penyakit tersebut dapat dengan cepat dilakukan.
2. Aplikasi ini memudahkan masyarakat umum khususnyapara penderita penyakit pada gigi dan mulut untuk mengetahui penyebab, akibat dan gejala tanpa harus datang ke klinik atau rumah sakit.
3. Mempermudah masyarakat untuk mencari informasi yang lebih cepat, detail danakurat tentang jenis penyakit pada gigi dan mulut.
4. Aplikasi ini diharapkan dapat membantu para medis dalam melakukan pengolahan data penyakit.

Saran

Setelah mengevaluasi sistem secara menyeluruh, diharapkan sistem ini nantinya dapat dikembangkan lebih lanjut dengan saran sebagai berikut :

1. Perlu diadakan penambahan data untuk jenis penyakit gigi dan mulut beserta gejala sehingga informasi yang ditambahkan pada sistem akan semakin banyak.
2. Pengembangan program dan analisis data agar dapat lebih diperluas cakupannya sesuai dengan kebutuhan program.

VI. Daftar Pustaka

Alkhairi, P., & Windarto, A. P. (2019). Penerapan K-Means Cluster pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara.*Seminar Nasional Teknologi Komputer & Sains*, 762–767

Rosmini, R., Fadlil, A., & Sunardi, S. (2018). Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah. *It Journal*

INFORMATIKA
Manajemen Informatika
Universitas Labuhanbatu
Vol. 11 No.1 / Januari/2023
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

Research and Development, *3*(1), 22–31. https://doi.org/10.25299/itjrd.2018.vol3(1).1773.

A. W. Indra Purnama, Ragil Saputra, "Implementasi Data Mining Menggunakan Crisp-Dm Pada Sistem Informasi Eksekutif Dinas Kelautan Dan Perikanan Provinsi Jawa Tengah," *Annual Review of Information Science and Technology*, vol. 36, 2017.