



2. Teknik Analisis Data

© Hak Cipta Milik IBI KIKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Dalam penelitian ini, peneliti menggunakan teknik analisis yang dilakukan berdasarkan metode CRISP-DM (*Cross Industry Standard Process For Data Mining*) yaitu sebuah metode *data mining* yang masih digunakan saat ini. Menurut Chantal D.Larose (2015:7), CRISP-DM membagi proses *data mining* ke dalam enam fase yaitu :

Hak Cipta Dilindungi Undang-Undang

a. **Business Understanding**

Fase pertama yang berfokus pada pemahaman objektif proyek beserta kebutuhan dari perspektif bisnis, kemudian pengetahuan tersebut dikonversi menjadi *data mining*.

b. **Data Understanding**

Fase kedua yang berawal dengan pengumpulan data serta pemahaman data untuk mengidentifikasi masalah kualitas data untuk mengetahui *insight* dan membentuk hipotesis untuk informasi tersembunyi.

c. **Data Preparation**

Fase ketiga yang membentuk *dataset* final yang akan digunakan dalam alat modelling dari data mentah, tugas persiapan data akan dilakukan secara berulang-ulang seperti transformasi dan pembersihan data untuk alat *modelling*.

d. **Modelling**

Fase keempat yang menentukan teknik *modelling* yang akan diaplikasikan beserta *parameter* yang akan dikalibrasikan pada nilai optimal.

e. **Evaluation**

Fase kelima yang terdiri dari model yang sudah dibuat untuk dilakukan perspektif analisis data, sebelum diimplementasikan memerlukan evaluasi model dan

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.

2. Dilarang menggunakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



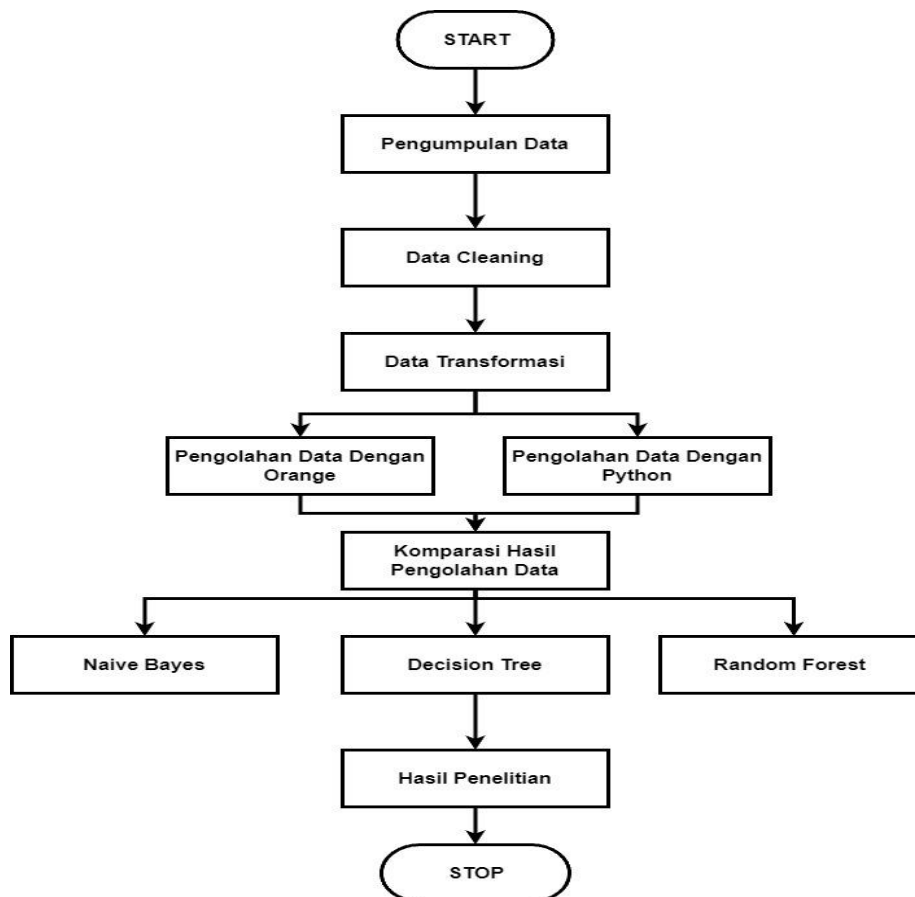
meninjau tahapan yang dieksekusi untuk membentuk model tersebut sesuai dengan tujuan bisnis.

Deployment

Fase keenam yang merupakan tahap implementasi dari sebuah model yang sudah dibuat yang berguna untuk meningkatkan pengetahuan tentang data.

3. Penerapan Algoritma

Pada penelitian ini akan dibahas bagaimana penggunaan komparasi algoritma untuk menentukan masa studi mahasiswa dapat digambarkan dalam diagram sebagai berikut :



Gambar 3.1
Komparasi Algoritma Klasifikasi

Hak Cipta Dilindungi Undang-Undang

© Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Institut Bisnis dan Informatika Kwik Kian Gie

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



Berikut ini adalah penjelasan diagram dari gambar 3.1 yaitu :

- a. **Pengumpulan Data** yang dilakukan pada masa studi mahasiswa Sistem Informasi dan Teknik Informatika pada tahun 2009 sampai dengan 2022 dengan jumlah data 532 *record*.
- b. **Data *cleaning*** untuk membersihkan data kotor menjadi data berkualitas untuk dilakukan tahap *data mining*.
- c. **Data transformasi** untuk mengelompokan data yang diperlukan saja dalam tahapan *data mining*.
- d. **Pengolahan data dengan *Orange*** yaitu data diolah dalam aplikasi *Machine Learning* berbasis *Orange*.
- e. **Pengolahan Data Dengan *Python*** yaitu data diolah menggunakan bahasa pemrograman *Python*.
- f. **Komparasi Hasil Pengolahan Data** dilakukan untuk komparasi hasil dari pengolahan data menggunakan *Machine Learning* berbasis *Orange* dan bahasa pemrograman *Python*.
- g. **Komparasi Algoritma *Naive Bayes*, *Decision Tree* dan *Random Forest*** dari pengolahan data menggunakan *Machine Learning* berbasis *Orange* dan bahasa pemrograman *Python*.
- h. **Hasil penelitian** merupakan hasil dari komparasi pengolahan data menggunakan *Orange* dan *Python* yang sudah disimpulkan.

© Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Hak Cipta Dilindungi Undang-Undang

Institut Bisnis dan Informatika Kwik Kian Gie

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.

2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

4. Teknik Pengukuran Data

Peneliti menggunakan *Confusion Matrix* Dan Kurva *ROC* sebagai pengukuran data klasifikasi berdasarkan pengaruh dari tiap variabel yang digunakan.



a. Confusion Matrix

Menurut Ibnu Daqiqil (2021:101) definisi *Confusion Matrix* yaitu :

“*Confusion Matrik adalah matrik yang berukuran $N \times N$ dimana N adalah jumlah kelas yang diprediksi.*”

Confusion matrix menyajikan ringkasan hasil prediksi yang dihasilkan oleh *Machine Learning* dengan membandingkan antara hasil prediksi dan hasil yang diharapkan.

	Actual = Yes	Actual = No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

Gambar 3.2

Contoh *Confusion Matrix* Dengan Dua Kelas Yaitu *Yes* Dan *No*

Sumber : Ibnu Daqiqil ID (2021:102)

Pada gambar 3.2, *confusion matrix* ada 4 kolom menurut Ibnu Daqiqil ID (2021:102) yaitu :

- (1) TP (*True Positif*) berisi jumlah data *points* diberi label Yes
- (2) TN (*True Negatif*) berisi jumlah data *points* diberi label No
- (3) FP (*False Positif*) berisi jumlah data *points* diberi label Yes tetapi nilai sebenarnya bernilai No. *Error* ini diberi label *Error Type 1*.
- (4) FN (*False Negatif*) berisi jumlah data *points* diberi label No tetapi nilai sebenarnya adalah Yes. *Error* ini diberi label *Error Type 2*.

Hak Cipta Dilindungi Undang-Undang

© Hak cipta milik IBIKKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Institut Bisnis dan Informatika Kwik Kian Gie

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.

2. Dilarang menggunakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

Berdasarkan nilai-nilai *Confusion matrix*, ada perhitungan yang perlu dilakukan

menurut Ibnu Daqiqil ID (2021:103) yaitu :

- (1) **Akurasi** : Perbandingan jumlah item yang diprediksi benar dengan total seluruh prediksi yang dilakukan.

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

Gambar 3.3
Perhitungan Akurasi

Sumber : Ibnu Daqiqil ID (2021:103)

- (2) **Recall** : Perbandingan jumlah item yang relevan diidentifikasi benar dengan seluruh item yang benar.

$$Recall = \frac{TP}{TP + FN}$$

Gambar 3.4
Perhitungan Recall

Sumber : Ibnu Daqiqil ID (2021:103)

- (3) **Specificity** : Perbandingan jumlah item yang diidentifikasi sebagai negatif dengan yang benar-benar negatif.

$$Specificity = \frac{TN}{TN + FP}$$

Gambar 3.5
Perhitungan Specificity

Sumber : Ibnu Daqiqil ID (2021:103)

- (4) **Precision** : Perbandingan jumlah item yang diidentifikasi sebagai positif secara benar terhadap jumlah item yang diidentifikasi positif.





$$Precision = \frac{TP}{TP + FP}$$

Gambar 3.6

Perhitungan Precision

Sumber : Ibnu Daqiqil ID (2021:104)

(5) **False Positive Rate (FPR)** : Perbandingan jumlah item yang salah diidentifikasi sebagai positif dengan jumlah item yang dibenarkan tetapi negatif, ini juga disebut sebagai Type 1 Error.

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity$$

Gambar 3.7

Perhitungan False Positive Rate

Sumber : Ibnu Daqiqil ID (2021:104)

(6) **False Negative Rate (FNR)** : Perbandingan jumlah item yang salah identifikasi sebagai negatif dengan jumlah item yang dibenarkan positif

$$FNR = \frac{FN}{FN + TP}$$

Gambar 3.8

Perhitungan False Negative Rate

Sumber : Ibnu Daqiqil ID (2021:104)

b. ROC (Receiver Operating Characteristic Curve) dan AUC (Area Under The ROC Curve)

Menurut Ibnu Daqiqil ID (2021:105) definisi ROC dan AUC yaitu :

“Kurva AUC - ROC merupakan pengukuran performa untuk masalah klasifikasi pada berbagai pengaturan threshold. ROC adalah kurva probabilitas dan AUC mewakili derajat atau ukuran keterpisahan.”

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



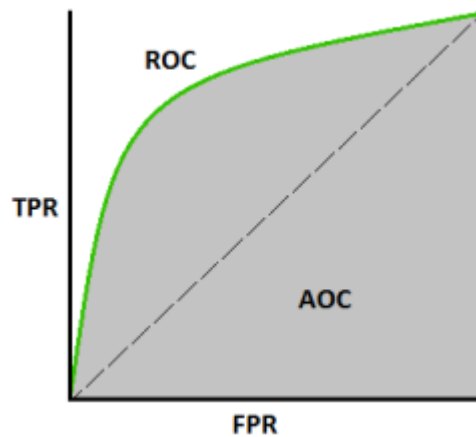
Semakin tinggi AUC, semakin baik model memprediksi true sebagai true dan false sebagai false.

(a)

Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Hak Cipta Dilindungi Undang-Undang

Kurva KOP diplotkan dengan TPR terhadap FPR dimana TPR berada pada sumbu Y dan FPR pada sumbu x berdasarkan gambar berikut.



Gambar 3.9

Contoh Grafik ROC

Sumber : Ibnu Daqiqil ID (2021:106)

Apabila *AUC* mendekati angka 1 berarti memiliki model kelas variabel yang terpisahkan dengan baik, tetapi apabila mendekati angka 0 berarti memiliki variabel yang terpisahkan terburuk.

Berikut ini adalah tabel nilai *AUC* beserta penjelasan dari setiap nilai tersebut :

Tabel 3.1
Penjelasan Nilai *AUC*

<i>AUC</i>	Keterangan
0.9 – 1	Klasifikasi yang sangat baik
0.8 – 0.9	Klasifikasi yang baik
0.7 – 0.8	Klasifikasi yang sedikit baik
0.6 – 0.7	Klasifikasi yang kurang baik
<0.6	Klasifikasi gagal

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



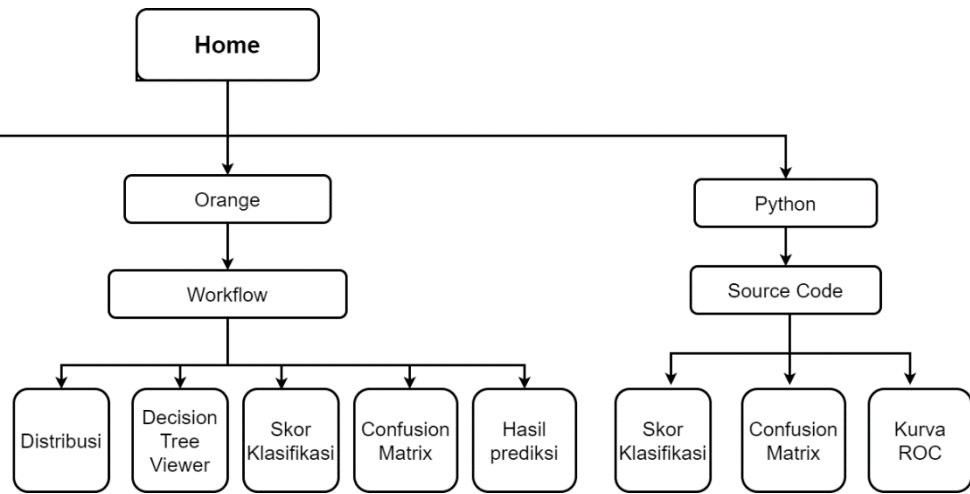
C. Teknik Perancangan *Graphic User Interface* (GUI)

C Hak Cipta Milik Institut Bisnis dan Informatika Kwik Kian Gie

Pada perancangan *Graphic User Interface* (GUI) menggunakan platform *Google Sites* untuk menyajikan hasil proses *data mining* berikut tampilan data dalam bentuk *dashboard*. Detail rancangan *gui* tersebut meliputi *Home*, *Dataset*, *workflow* dari *Orange*, hasil skor klasifikasi dan *confusion matrix*, *source code* dari *Python*, hasil skor klasifikasi, *confusion matrix* dan Kurva *ROC*.

Hak Cipta Dilindungi Undang-Undang

Data Mahasiswa
2009-2021



Gambar 3.10
Hirarki Rancangan *GUI*

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang menggunakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.