BAB 2

TINJAUAN PUSTAKA

2.1 Kinerja Mahasiswa

Menurut (Naomi, 2008) dalam buku Indra dkk (2021) Kinerja akademik merupakan hasil akhir yang dicapai oleh peserta didik sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan.

2.2 Data Mining

Menurut (Indah Werdianingsih, 2020) Data mining merupakan bidang ilmu yang digunakan untuk menangani masalah pengambilan informasi dari sebuah sumber data yang menkombinasikan antara mechine learning dan Teknik statistic, visualisasi data dan pengenalan pola.

Menurut (Suntoro, 2019) Data Mining adlah proses ekstraksi suatu data yang sebelumnya tidak diketahui, bersifat implisit dan dianggap tidak berguna menjadi informasi atau pengetahuan atau pola dari data yang jumlahnya besar.

Menurut Kusnawi (2007) dalam buku Sri Rahayu Ginantra dkk (2021) Data mining adalah aktivitas untuk mencari, memahami, dan pengambilan informasi dengan berbagai metode. Adapun jenis-jenis algoritma *data mining* sebagai berikut:

1. Klasifikasi (*Classification*)

   Diterapkan ke data baru untuk mengelompokkan tipe objek. Klasifikasi termasuk dalam model yang diawasi. Dalam masalah klasifikasi, kami memiliki sampel data dan memprediksi beberapa kelas yang ada berdasarkan sampel yang ada. Hanya satu dari sekian banyak atribut yang disebut atribut target, sedangkan atribut lainnya disebut atribut predator. Klasifikasi ini juga biasa digunakan dalam pemodelan bisnis dll. Misalnya, untuk mengidentifikasi penyakit tertentu berdasarkan klasifikasi atau mengidentifikasi pelanggan berdasarkan model pembayaran.

2. Klastering (*Clustering*)

   Berbeda dengan klasifikasi, pengelompokan adalah model yang tidak

diawasi. Mengelompokkan data grup dengan label yang tidak diketahui. Cluster yang diatur dalam hierarki menentukan klasifikasi data. Penerapan metode clustering yang tepat akan menghasilkan cluster yang berkualitas. Pengelompokan ditandai dengan sentroid atribut histogram dan model pengelompokan pohon hierarkis.

3. Regresi (*Regression*)

   Merupakan suatu fungsi yang digunakan untuk memodelkan data untuk meminimalkan hasil kesalahan prediksi. Umumnya regresi dilakukan dengan data yang bersifat time series.

4. *Association Rule*

   Merupakan permodelan kebergantungan. Fungsi asosiasi ini biasanya kita kenal dengan istilah "market basket analysis" yang merupakan fungsi untuk menemukan relasi atau korelasi antara himpunan item – item. Aturan asosiasi diartikan pada basket data yang digunakan untuk keperluan promosi, desain katalog untuk meningkatkan penjualan. Contoh penerapan asosiasi adalah ketika customer membeli pamper maka ada kemungkinan membeli bir.

5. *Anomaly Detection*

   Mengidentifikasi data yang tidak umum. Bisa berupa outlier, perubahan deviasi/bias yang penting dan perlu investigasi lebih lanjut.

6. *Summarization*

   Menyediakan representasi data yang lebih sederhana meliputi pelaporan, visualisasi data yang dipergunakan untuk menunjang informasi dan penguatan keputusan.

2.2.1. Metode Klasifikasi

Menurut (Indah Werdianingsih, 2020). Klasifikasi adalah penambangan data yang menetapkan item dalam koleksi ke kelas minat tertentu. Klasifikasi dimulai dengan pengumpulan data, dimana data tersebut telah diketahui kategorinya. Misalnya, klasifikasi yang digunakan untuk mengidentifikasi peringkat kredit, berupa pemohon kredit dengan risiko kredit rendah, sedang, dan tinggi. Atribut

yang digunakan bisa berupa data pemohon, antara lain riwayat kredit pemohon, riwayat pekerjaan pemohon, kepemilikan atau sewa rumah pemohon, tahun tinggal pemohon, jumlah uang dan investasi pemohon, dll. Peringkat Kredit akan menjadi target/kategori, data per pemohon.

2.2.2. Algoritma Naïve Bayes

Menurut (Suntoro et al., 2018) Algoritma Naïve Bayes adalah salah satu algortima klasifikasi berdasarkan teorema Bayesian pada satistika. Algoritma Naïve Bayes dapat digunakan untuk memprediksi probabilitas keanggotan suatu kelas.

Menurut (Primartha, 2021) Metode Naïve Bayes atau Naïve Bayes Classifier berasal dari Bayes Theorem (Teorema Bayes) yang ditemukan oleh Thomas Bayes pada tahun 1770 yang merupakan penerapan teori probabiltas yaitu bagaimana sebuah peluang terjadi.

Menurut Dios Kurniawan (2021) Algortima Naïve Bayes adala metode yang menggunakan *probability* (Peluang) untuk membuat model prediksi klasifikasi. Dengan memanfaatkan data tentang kejadian di masa lalu, model bisa membuat perkiraan apa yang akan terjadi di masa depan. Metode probability suatu kejadian, dan bisa berubah bila informasi pendukung tambahan yang disediakan.

Rumus Algortima Naïve Bayes dapat kita jabarkan sebagai berikut :

$$P\ (H|X)\ = \frac{P(X|H)P(H)}{P(X)} \qquad (2.1)$$

X	=	Data yang belum diketahui (Data Testing)
H	=	Merupakan Hipotesis data X yang kelasnya lebih spesifik
P(H)	=	Probabilitas hipotesis H atau yang biasa disebut sebagai prior probability
P(X)	=	Probabilitas hipotesis X atau yang biasa disebut sebagai Prediction Prior
P(X|H)	=	Probilitas hipotesis X berdasarkan kondisi H atau biasa disebut sebagai Likelihood

Rumus Standar Deviasi dapat kita jabarkan sebagai berikut :

$$s = \sqrt{\frac{\sum_i^n (x_i - x)^2}{n-1}}$$

$$s = \sqrt{\frac{n\sum_{i=1}^n x_i^2 - (\sum_{=1}^n x_i)^2}{n(n-1)}} \quad (2.2)$$

s = Simpang Baku Populasi

N = Jumlah Populasi

X_i = Setiap Nilai dari Populasi

X = Rata – Rata Populasi

2.2.3. Data Preprocessing

Data preprocessing adalah tahapan awal dari *data mining* untuk menghasilkan analisis yang lebih akurat dalam pemakaian teknik-teknik *machine learning* (Santoso & Umam, 2018).

Menurut (Henny, 2021) Data preprocessing adalah salah satu teknik data pertama untuk mengubah data mentah menjadi informasi yang lebih efisien dan berguna. Format raw data dari berbagai sumber sering error, missing value dan tidak konsisten. Oleh karena itu perbaikan format hasil data mining akurat dan presisi.

Preprocessing melibatkan validasi dan imputasi data, dimana validasi ini bertujuan untuk menilai tingkat kelengkapan dan akurasi data. Sedangkan imputasi data bertujuan untuk mengkoreksi kesalahan dan memasukkan missing value (nilai hilang), melalui program business process automation (BPA).

Berikut 4 tahapan atau langkah-langkah Data Preprocessing yang digunakan, tergantung jenis pada kumpulan data menurut (Henny, 2021) :

1. Data Cleaning

   Langkah pertama adalah melakukan data cleaning. Data yang baru saja dikumpulkan kemungkinan tidak relevan dan banyak bagian yang hilang, sehingga dibutuhkan proses pembersihan. Dalam tahapan ini, data akan dibersihkan melalui beberapa proses seperti missing value dan noise.

2. Data Integration

   Data integration merupakan tahapan lanjutan dari data cleansing yang bertujuan untuk menghaluskan data. Pada tahap ini, data dengan representasi berbeda akan disatukan, dan konflik di dalamnya akan diselesaikan.

3. Data Transformation

   Tahapan ini digunakan untuk mengubah data menjadi bentuk yang sesuai dalam proses data mining. Pada tahap ini akan dinormalisasikan, dimana normalisasi ini adalah proses menskalakan nilai data dalam rentang tertentu untuk memastikan bahwa tidak ada data yang berlebihan.

4. Data Reduction

   Memilah kumpulan data dengan volume besar akan memakan waktu yang cukup lama. Oleh karena itu, perlu adanya proses data reduction untuk membatasi kumpulan data, guna meningkatkan efisiensi penyimpanan, sekaligus mengurangi biaya dan menghemat waktu.

2.2.4. Metode Cross-Validation

Menurut (Daqiqil, 2021) Cross-validation adalah metode statistic yang dapat mengevaluasi kinerja dari suatu model atau algoritma pemisahan data yang dibagi menjadi dua subset, yaitu data pelatihan dan data uji, model dilatih oleh subset pelatihan dan divalidasi oleh subset, kemudian pemilihan jedis cross-validation didasrkan pada ukuran kumpulan data. Adapun jenis-jenis *cross-validation* sebagai berikut (Daqiqil, 2021) :

1. LOOCV (*leave one out cross-validation*)

   Pada LOOCV, data dibagi menjadi dua bagian, yaitu data uji dan data latih, namun pada data latih hanya menggunakan 1 data. Sehingga jika kita memiliki n data, maka kita memiliki sejumlah n-1 data latih dan dan uji.

2. *K-Fold cross-validation*

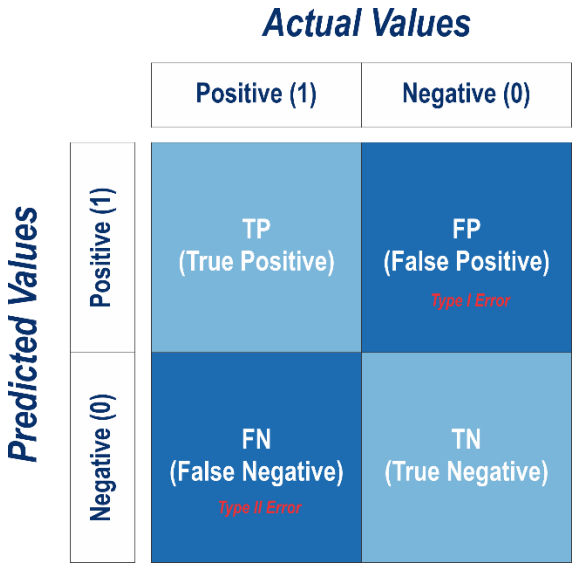
   K-fold adalah sebuah metode yang memecah dataset menjadi dua bagian yaitu data latih dan data uji sebanyak k kelompok di kelompok dimana jumlah data latih dan data uji pada tiap – tiap kelompok sama.

3. *Stratified cross-validation*

Stratified cross-validation mirip dengan *K-Fold*, namun perbedaannya dalam pembagian data tiap literasinya memperhatikan kondisi dataset seperti distribusi kelas, rata – rata dan varian, sehingga pembagian distribusi kelas lebih merata.

2.2.5. Confusion Matrix

Menurut (Mustika, 2021) Confusion matrix atau biasa juga dinamakan error matrix adalah sebuah metode yang dipergunakan untuk menyampaikan informasi perbandingan dari hasil penjabaran yang dilakukan oleh model yang digunakan dengan hasil klasifikasi sebenarnya. Agar lebih mudah dalam penjelasanya, confusion matrix biasanya digambarkan dalam berbentuk table matriks yang menjelaskan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya telah diketahui. Berikut ilustrasi dari table Confusion Matrix.



Gambar 2. 1 Confusion Matrix

TP (True Positive) merupakan jumlah data point berlabel yes yang nilainya diidentifikasi benar.

TP (True Negative) merupakan jumlah data point berlabel no yang nilainya diidentifikasi salah.

FP (False Positive) merupakan jumlah data point berlabel yes yang nilai sebenarnya diidentifikasi salah.

FN (False Negative) merupakan jumlah data point berlabel no yang nilai sebenarnya teridentifikasi benar.

Menurut (Pratiwi et al., 2021) Confusion matrix dapat dihitung dengan menggunakan perhitungan sebagai berikut :

1. Akurasi

   Nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan persamaan atau rumus sebagai berikut :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\% \quad (2.3)$$

2. *Sensitivity*

   Recall menunjukkan beberapa persen data kategori positif yang terklasifikasi dengan benar oleh sistem. Rumus yang digunakan adalah sebagai berikut :

$$Sensitivity = \frac{TP}{TP + FN} X\ 100\% \quad (2.4)$$

3. *Precision*

   Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasi secara benar dibagi dengan total data yang diklasifikasi positif, Presisi dapat diperoleh dengan rumus :

$$Presisi = \frac{TP}{TP + FP} X\ 100\% \quad (2.5)$$

4. *Error*

   Error adalah kasus yang diidentifikasi salah dalam sejumlah data, sehingga dapat dilihat seberapa besar tingkat kesalahan pada sistem yang digunakan. Berikut rumus daripada error atau kasus yang diidentifikasi salah tersebut.

$$Error = \frac{FN}{TP} X\ 100\% \quad (2.6)$$

2.2.6. CRISP-DM (Cross-Industry Standard Process for Data Mining)

Menurut (Huber et al., 2019) CRISP-DM adalah sebuah medel atau sebuah tahapan dalam mengelola dan menyempurnakan sebuah data mining. Ada beberapa point yang ada dalam CRISP-DM yakni sebagai berikut :

1. *Business Understanding*

   Tahapan awal dari metodologi CRISP-DM adalah tahapan business understanding berisi tentang menentukan tujuan bisnis, menilai situasi saat ini dan menetapkan tujuan dilakukan data mining. Tahapan ini sangat penting, namun sering diabaikan ketika seseorang terjun ke dunia data mining.

2. *Data Understanding*

   Tahapan kedua adalah kegiatan persiapan, mengevaluasi persyaratan data, dan termasuk pengumpulan data. Pada tahapan ini, data yang berhasil dikumpulkan kemudian dideskripsikan bagian mana yang atribut, kelas, dan tipe data.

3. Data Preparation

   Tahapan ketiga setelah data dikumpulkan, data-data tersebut perlu di identifikasi, dipilih, dibersihkan, kemudian dibangun ke dalam bentuk atau format yang diinginkan. Data preparation , disebut juga dengan data pre-processing.

4. *Modeling*

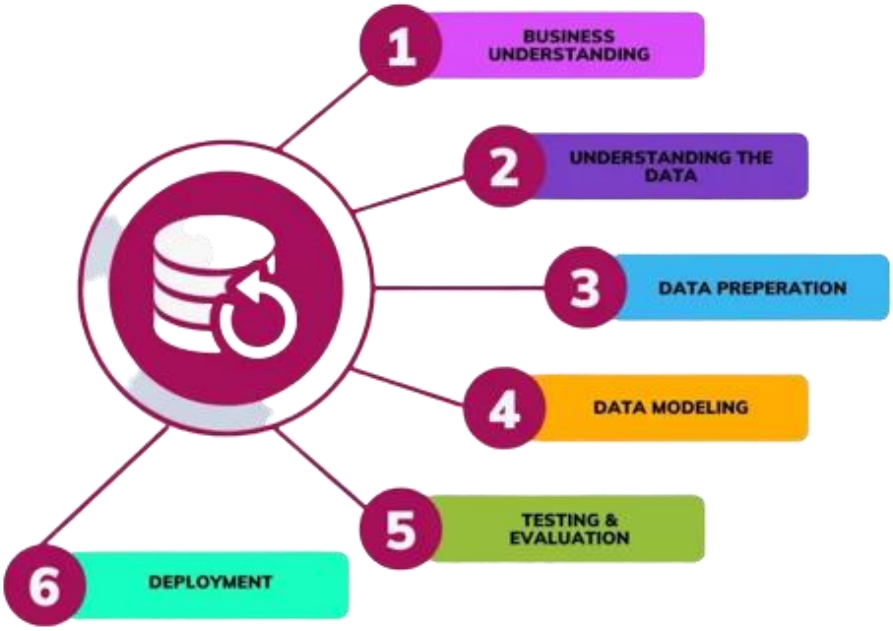
   Modeling adalah aplikasi dari algoritma untuk mencari, mengidentifikasi, dan menampilkan pola. Pemilihan algoritma berdasarkan tipe data karena dari tipe data kita bisa mengetahui apakah data tersebut akan diestimasi, prediksi, klasifikasi, clustering, atau melihat hubungan asosiatif.

5. *Evaluation*

   Tahapan kelima yang digunakan untuk membantu pengukuran evaluasi pada model adalah kita bisa mengukur model mana yang paling baik digunakan untuk proses data mining. Pada penerapan klasifikasi, pengukuran evaluasi yang banyak digunakan adalah akurasi, sensitivity, G- Mean, F-Measure, dan lain sebagainya.

6. *Deployment*

Deployment adalah tahapan akhir dalam CRISP-DM. Setelah model dievaluasi dan dipilih algoritma dengan hasil pengukuran terbaik, dilanjutkan ke tahapan deployment. Tahapan deployment digunakan untuk melakukan otomatisasi model atau pengembangan aplikasi, terintegrasi dengan sistem informasi manajemen atau operasional yang ada.



Gambar 2. 2 CRISP-DM

2.2.7. Optimasi Forwad Selection

Menurut (Astuti et al., 2018) Metode Sequential Forward Selection atau metode seleksi maju adalah algoritma pencarian paling sederhana. Forward Selection didasarkan pada model Regresi Linear. Forward Selection adalah salah satu teknik untuk mereduksi dimensi dataset untuk menghapus atribut-atribut yang tidak relevan. Metode Forward Selection merupakan model yang diawali dengan nol variable, untuk selanjutnya variable dimasukkan satu persatu sampai pada kriterianya terpenuhi.

Metode Forward Selection menerapkan prinsip model regresi linier dengan mereduksi atribut pada sebuah dataset. Proses pencarian, attribute dengan forward selection diawali dengan empty model, selanjutnya tiap variable

dimasukan hingga kriteria kombinasi model attribute terpenuhi dengan baik. Adapun tahapan dan rumus dari Seleksi Fitur Forward Selection yakni sebagai berikut (Samosir & Siagian, 2014) :

- Mulai dari 0 fitur (F=0).
- Menambahkan fitur satu demi satu pada setiap stepnya yang menghasilkan error terkecil atau akurasi terbesar.
- Berhenti jika tidak ada perubahan error atau perubahannya tidak signifikan.

2.3 Python.

Python adalahbahasa pemrograman dinamis, tingkat tinggi, dimana merupakan bahasa pemrograman interpreteryaitu bahasa yang mengkonversi source codemenjadi machine codesecara langsung ketika program dijalankan.Bahasa ini juga mendukung pendekatan pemrograman Berorientasi Objek untuk pengembangan aplikasidan mudah dipelajari serta menyediakan banyak struktur data tingkat tinggi.Python adalah bahasa skrip yang mudah dipelajari namun kuat dan serbaguna, yang membuatnya menarik untuk Pengembangan Aplikasi.Sintaks dan pengetikan Python sangat dinamis dengan sifat interpretasinya menjadikannya bahasa yang ideal untuk skrip dan pengembangan aplikasi yang cepat.

Python mendukung banyak pola pemrograman, termasuk gaya pemrograman berorientasi objek, imperatif, dan fungsional serta prosedural.Python tidak hanya dimaksudkan untuk bekerja di area tertentu, seperti pemrograman web. Itulah mengapa ini bahasa pemrograman ini dikenal sebagai Bahasamultiguna karena dapat digunakan untuk web, enterprise,CAD 3D, dll.Deklarasi variabel pada Bahasa pythontidak perlu menggunakan tipe datakarena ini diketik secara dinamis sehingga kita dapat menulis a=10 untuk menetapkan nilai bilangan bulat dalam variabel bilangan bulat. (Mathematics, 2018).

2.4 Scikit – Learn

Scikit – Learn adalah modul python yang mengintegrasikan berbagai algoritma pembelajaran mesin state-of-the-art untuk masalah yang diawasi dan tidak diawasi skala menengah. Paket ini berfokus pada membawa pembelajaran mesin

ke non-spesialis menggunakan bahasa tingkat tinggi tujuan umum. Penekanan diberikan pada kemudahan penggunaan, kinerja, dokumentasi, dan konsistensi API. Ini memiliki ketergantungan minimal dan didistribusikan dibawah lisensi BSD yang disederhanakan, mendorong penggunaannya baik dalam aturan akademis dan komesial. (Riadi Silitonga et al., 2019)

2.5 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian
1	(Syahputra et al., 2018)	Implementasi Data Mining Untuk Prediksi Mahasiswa Pengambil Mata Kuliah Dengan Algoritme Naive Bayes	Dari hasil pengujian yang dilakukan menggunakan dataset mahasiswa tahun 2016 semester ganjil dapat diketahui nilai akurasi dari proses klasifikasi menggunakan Naive Bayes cukup baik. Untuk mata kuliah sampel yaitu mata kuliah Manajemen Hubungan Pelanggan dan Jaringan Nirkabel, nilai Accuracy berdasarkan Confussion Matrix adalah sebesar 85,88%

2	(Tyas et al., 2020)	Implementasi Algoritma Naïve Bayes Dalam Penentuan Rating Buku	Model naïve bayes memiliki tingkat kesalahan yang sangat minimum jika dibandingkan dengan algoritma klasifikasi lainnya. Hasil penelitian menentukan bahwa hasil penentuan rating buku menggunakan metode naïve bayes memiliki accuracy 66.98%, precision 74.47% dan recall 62.47%
3	(Sainanda et al., 2020)	Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus Stmik Primakara)	Pada penelitian tersebut Algoritma Naïve Bayes Mendapatkan hasil dengan recall sebesar 80%, accuracy sebesar 80% dan precision sebesar 80%.
4	(Barus, 2021)	Implementation Of Naïve Bayes Classifier-Based Machine Learning To Predict And Classify New Students At Matana	Hasil implementasi algoritma naïve bayes pada penelitian tersebut mendapatkan hasil

		University	akurasi yakni 73%
5	(Devita et al., 2018)	Perbandingan Kinerja Metode Naive Bayes Dan K-Nearest Neighbor Untuk Klasifikasi Artikel Berbahasa Indonesia	Berdasarkan penelitian dapat diketahui bahwa kinerja dari metode Naive bayes dan memiliki tingkat akurasi sebesar 70% lebih baik dari metode K-Nearest Neighbor
6	(Ubaedi & Djaksana, 2022)	Nggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit	Hasil penelitian menunjukkan bahwa pohon keputusan yang dihasilkan dari algoritma C4.5 memiliki akurasi cukup baik sebesar 79,11% dan metode Feature Selection dan Stratified Sampling berhasil meningkatkan akurasi algoritma C4.5 sebesar 9,2% dalam memprediksi kelayakan kredit.
7	(Nurlia & Enri, 2021)	Penerapan Fitur Seleksi Forward Selection Untuk	Pengujian algoritma C4.5 menghasilkan akurasi sebesar 77.89%

		Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5	dan sedangkan algoritma C4.5 berbasis forward selection memperoleh akurasi sebesar 84,29%. Forward selection pada algoritma C.45 dinilai memiliki performa yang cukup baik ,karena meningkatkan akurasi sebesar 6,4%
8	(Astuti et al., 2018)	Algoritma Naive Bayes Dengan Fitur Seleksi Untuk Mengetahui Hubungan Variabel Nilai Dan Latar Belakang Pendidikan	dilihat hasil akurasi dari optimasi forward selection pada algoritma naïve bayes sebesar 78,08%. Artinya ada peningkatan performa dengan melakukan optimasi sebesar 13,31%.

Berdasarkan uraian penelitian yang telah dilaksanakan sebelumnya, Algoritma Naïve Bayes Classifer digunakan pada beberapa studi kasus yang ada, salah satu yakni pada bidang Pendidikan. Tidak hanya itu, seleksi fitur Forward Selection juga telah digunakan pada beberapa penelitian sebelumnya untuk melakukan peningkatan akurasi atau performa pada sebuah algoritma untuk mendapatkan hasil yang optimal seperti peningkatan akurasi menggunakan forward selection pada algoritma C4.5 . Namun pada penelitian ini penulis akan mengangkat permasalahan tentang Analisis Optimasi Forward Selection pada Klasifikasi Nilai Mahasiswa dalam perkuliahan hybrid mahasiswa di Universitas Muhammadiyah Kalimantan Timur.

Namun pada penelitian ini yang membedakan adalah Forwad Selection akan digunakan untuk melakukan peningkatan akurasi pada Algoritma Naïve Bayes dan indikator yang digunakan diperoleh dari Platform OpenLearning serta data dari Bagian MKDU. Sehingga penelitian ini memunculkan sebuah penelitian baru yang belum pernah dilakukan sebelumnya.