Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Rekomendasi Pilihan Program Studi Pada Mahasiswa Baru

Ade Mulyana[1], Yanto Hermawan[2], dan Nathania Juli Saputri[3]*

[123]Sistem Informasi, Fakultas Informatika dan Pariwisata, Institut Bisnis dan Informatika Kesatuan

** Penulis korespondensi : anto@ibik.ac.id*

ABSTRACT

The new student admission program is an essential annual agenda for universities to attract prospective students. Choosing the right study program is very important to match the abilities of prospective students, producing superior graduates. At the Kesatuan Business and Informatics Institute, new student admissions have significantly changed in recent years, reducing learning motivation. The problems that have emerged include a significant decrease in re-registration in the last five years, many student withdrawals, and students who change study programs after learning activities begin because they feel they do not match their interests and abilities. This indicates the need for a strategic approach to helping prospective students choose the right study program from the start of registration. Thus, a strategy is needed to recommend study programs that match the talents and potential of prospective students. The K-Means Clustering and Classification algorithms will be used to determine clusters from new student test data as a basis for recommending study programs. The results of K-Means Clustering show that 85 students were recommended to the Management study program, 84 students to the Tourism study program, 72 students to the Information Systems study program, 39 students to the Bioentrepreneurship study program, 39 students to the Accounting study program, and 38 students to the Information Technology study program. The results show accuracy using three classification models, namely Random Forest with an accuracy of 80.56%, KNN with an accuracy of 76.39%, and SVC with 52.78%. Thus, the Random Forest model was chosen for implementation in a web application using the Flask API because it provides the best accuracy in predicting study programs suitable for new students.

Article History

Received : 21-04-2025
Revised : 25-04-2025
Accepted : 28-04-2025

Keywords

Data Mining
Algoritma K-Means Clustering
Classification
Random Forest

ABSTRAK

Program penerimaan mahasiswa baru adalah agenda tahunan yang penting bagi perguruan tinggi untuk menarik calon mahasiswa. Pemilihan program studi yang tepat sangat penting agar sesuai dengan kemampuan calon mahasiswa, menghasilkan lulusan yang unggul. Di Institut Bisnis dan Informatika Kesatuan, penerimaan mahasiswa baru mengalami perubahan signifikan beberapa tahun terakhir, yang mengurangi motivasi belajar. Permasalahan yang muncul meliputi penurunan jumlah pendaftar ulang secara signifikan dalam lima tahun terakhir, tingginya angka pengunduran diri mahasiswa, serta adanya mahasiswa yang berpindah program studi setelah kegiatan belajar dimulai karena merasa tidak sesuai dengan minat dan kemampuan mereka. Hal ini menandakan perlunya pendekatan strategis dalam membantu calon mahasiswa memilih program studi yang sesuai sejak awal pendaftaran. Sehingga, diperlukan strategi untuk merekomendasikan program studi yang sesuai dengan bakat dan potensi calon mahasiswa. Dengan menggunakan algoritma K-Means Clustering dan Classification untuk menentukan cluster dari data tes mahasiswa baru sebagai dasar untuk merekomendasikan program studi. Hasil K-Means Clustering menunjukkan, 85 mahasiswa direkomendasikan ke prodi Manajemen, 84 mahasiswa ke prodi Pariwisata, 72 mahasiswa ke prodi Sistem Informasi, 39 mahasiswa ke prodi Biokewirausahaan, 39 mahasiswa ke prodi Akuntansi, dan 38 mahasiswa ke prodi Teknologi Informasi. Hasil menunjukan akurasi menggunakan tiga model klasifikasi yaitu Random Forest dengan akurasi 80.56%, KNN dengan akurasi 76.39%, dan *Support Vector Classifier (*SVC) 52.78%. Dengan demikian, model Random Forest dipilih untuk implementasi dalam aplikasi web menggunakan API Flask, karena memberikan akurasi terbaik dalam memprediksi program studi yang cocok untuk mahasiswa baru.

PENDAHULUAN

Penerimaan mahasiswa baru merupakan agenda tahunan yang dilaksanakan oleh perguruan tinggi untuk menjaring calon mahasiswa potensial dan menjaga keberlangsungan

institusi pendidikan. Institut Bisnis dan Informatika (IBI) Kesatuan Bogor secara rutin menyelenggarakan proses penerimaan ini setiap tahun dengan tujuan meningkatkan jumlah pendaftar serta kualitas calon mahasiswa [1]. Proses penerimaan tersebut meliputi beberapa tahapan, mulai dari pengambilan formulir, seleksi administrasi, tes masuk, hingga pendaftaran ulang [2]. Pemilihan program studi yang sesuai dengan minat dan kemampuan calon mahasiswa menjadi faktor penting untuk mendukung keberhasilan akademik dan mengurangi tingkat ketidakcocokan selama masa studi.

Namun, IBI Kesatuan menghadapi kendala dalam menentukan kesesuaian antara minat dan program studi pilihan calon mahasiswa. Ketidakcocokan tersebut sering kali berujung pada pengunduran diri atau perpindahan program studi setelah kegiatan akademik dimulai. Data internal Unit Marketing IBI Kesatuan mencatat bahwa pada tahun 2023 terdapat 20 mahasiswa yang mengundurkan diri dan 9 mahasiswa yang berpindah program studi, dengan dominasi kasus pada Program Studi Teknologi Informasi dan Sistem Informasi [3].

Fenomena serupa juga terjadi di berbagai perguruan tinggi di Indonesia, seperti yang diungkapkan oleh Wahyuni dan Susanto (2021) bahwa tingkat ketidakcocokan program studi menyumbang sekitar 25% dari total angka putus kuliah di perguruan tinggi swasta [4].

Untuk menghadapi tantangan tersebut, berbagai penelitian telah mengusulkan penerapan metode data mining, salah satunya adalah algoritma K-Means Clustering, guna mengelompokkan calon mahasiswa berdasarkan karakteristik tertentu. Setiawan dan Pratama (2020) berhasil menerapkan algoritma K-Means untuk mengelompokkan mahasiswa baru berdasarkan minat akademik, sehingga menghasilkan rekomendasi program studi yang lebih akurat [5]. Selanjutnya, penelitian oleh Dewi et al. (2021) menunjukkan bahwa K-Means Clustering dapat digunakan untuk mengidentifikasi kecenderungan pemilihan jurusan berdasarkan latar belakang pendidikan dan hasil tes masuk [6]. Pada tahun 2022, Putri dan Ramadhan menerapkan K-Means dalam sistem rekomendasi program studi berbasis minat dan nilai akademik calon mahasiswa, dengan hasil akurasi rekomendasi mencapai 82% [7]. Kemudian, penelitian terbaru oleh Fauziah dan Rachman (2023) mengembangkan model K-Means untuk memetakan preferensi calon mahasiswa terhadap program studi tertentu berdasarkan asal sekolah dan minat karier [8].

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan penggunaan algoritma K-Means Clustering untuk membantu panitia penerimaan mahasiswa baru di IBI Kesatuan dalam mengelompokkan calon mahasiswa berdasarkan karakteristik mereka. Tujuannya adalah untuk memberikan rekomendasi program studi yang lebih tepat, sehingga dapat meningkatkan kepuasan mahasiswa, mengurangi angka pengunduran diri, serta membantu pengembangan potensi akademik mahasiswa secara maksimal.

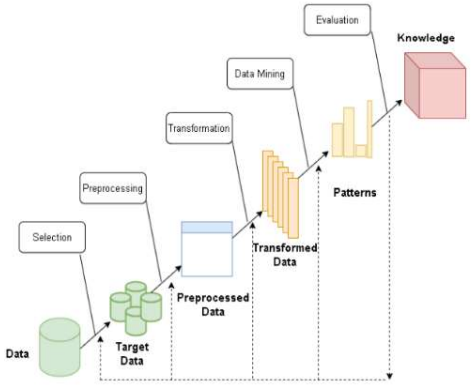
TINJAUAN PUSTAKA

Data Mining

Data adalah kumpulan fakta yang terekam atau entitas yang belum memiliki arti dan sering kali diabaikan [9]. Mining atau penambangan data merupakan tahap pencarian atau eksplorasi untuk menemukan pola tersembunyi dalam kumpulan data [10]. Dengan demikian, data mining dapat dijelaskan sebagai serangkaian langkah untuk menggali informasi berharga dari data yang tersedia, menghasilkan pengetahuan baru, pola, serta hubungan di antara data.

Data mining bertujuan untuk mengekstraksi esensi pengetahuan dari sekumpulan data sehingga dapat disajikan dalam bentuk struktur yang dapat dipahami manusia (ACM, 2019) [11]. Dalam proses penemuan pengetahuan dalam database (Knowledge Discovery in Database – KDD), terdapat beberapa tahap seperti pada gambar 1.

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i1.7624



Gambar 1. Proses Data Mining

1. Seleksi Data (Selection)
2. Pemilihan data (Preprocessing / Cleaning)
3. Transformasi (Transformation)
4. Data Mining
5. Interpretasi / Evaluasi (Interpretation / Evaluation)

Algoritma K-Means

Algoritma adalah serangkaian langkah sistematis yang dirancang untuk memecahkan suatu masalah dan dapat diimplementasikan ke dalam program komputer [12]. Salah satu metode pembelajaran tanpa pengawasan (unsupervised learning) yang paling populer adalah algoritma **K-Means** [13]. K-Means terkenal karena kemudahan implementasinya dan efisiensi dalam menangani data skala kecil hingga menengah.

K-Means merupakan algoritma clustering non-hierarki yang membagi data ke dalam dua atau lebih kelompok berdasarkan kemiripan karakteristik [14]. Algoritma ini bekerja dengan mengelompokkan data ke dalam beberapa cluster sehingga data yang memiliki kesamaan karakteristik ditempatkan dalam satu kelompok, sedangkan data yang berbeda dikelompokkan ke kelompok lain.

Clustering

Clustering, atau klasterisasi, adalah metode untuk menemukan dan mengelompokkan data berdasarkan kemiripan karakteristik [15]. Teknik ini merupakan salah satu pendekatan dalam data mining yang tidak memerlukan pengawasan (unsupervised), artinya proses pengelompokan dilakukan tanpa adanya label atau target output.

Terdapat dua pendekatan utama dalam clustering, yaitu:

1. **Hierarchical Clustering**: Mengelompokkan data dengan membangun struktur berbentuk pohon (dendrogram). Proses ini dimulai dengan menggabungkan objek yang paling mirip, kemudian secara bertahap menggabungkan kelompok berdasarkan tingkat kemiripan, hingga akhirnya seluruh objek bergabung menjadi satu kelompok besar.
2. **Non-Hierarchical Clustering**: Seperti algoritma K-Means, yang langsung membagi data menjadi sejumlah cluster berdasarkan inisialisasi pusat cluster (centroid).

Pendekatan-pendekatan tersebut membantu dalam memahami struktur data serta menemukan pola tersembunyi yang sebelumnya tidak diketahui.

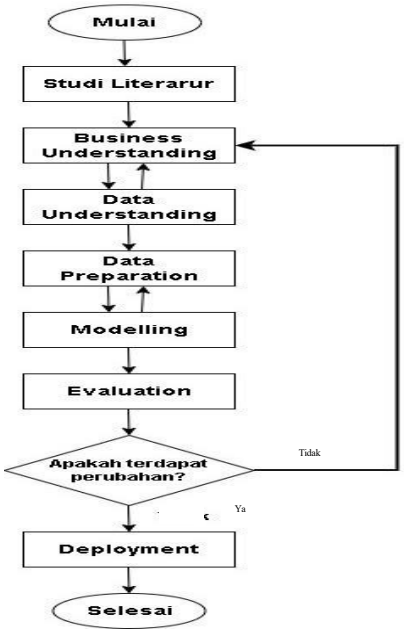
METODE

Penelitian ini dilaksanakan dalam trentang waktu 11 bulan, dari Agustus 2023 hingga Juli 2024. Subjek penelitian ini adalah calon mahasiswa baru di Institut Bisnis dan Informatika Kesatuan.

Data yang dikumpulkan mencakup informasi akademis dan hasil tes psikometri tahun 2023. Data ini digunakan untuk mengenalisis dan mengelompokkan calon mahasiswa berdasarkan kriteris tertentu, guna memberikan rekomendasi program studi yang sesuai.

Objek penelitian ini adalah data penerimaan mahasiswa baru di Institut Bisnis dan Informatika Kesatuan, yang mencakup data tes psikometri tahun 2023 yang diperoleh dari Unit Marketing Institut Bisnis dan Informatik Kesatuan. Data ini digunakan untuk menentukan program studi yang tepat bagi calon mahasiswa baru melalui penerapan metode data mining dengan algoritma *K-Means Clustering* dan beberapa model klasifikasi.

Analisis data dilakukan menggunakan metode *K-Means Clustering* dan *Classification* menggunakan bahasa pemrograman Python untuk mengelompokkan data mahasiswa ke beberapa cluster dan pengolahan data tersebut di uji untuk dilakukan prediksi menggunakan beberapa model evaluasi yaitu Random Forest, K- Nearest Neighbors, dan *Support Vector Classification*. Pada penelitian ini penulis ingin mendapatkan hasil yang sebaik mungkin, sehingga penulis berusaha untuk mengumpulkan data yang sesuai dengan kebutuhan. Berikutnya tergambar langkah-langkah metode atau teknik yang digunakan dalam penelitian ini sebagai gambar 2.



Gambar 2. Tahapan Prosedur Kerja

1. Studi Literatur

Penulis melakukan pengkajian literatur untuk mengidentifikasi informasi latar belakang yang relevan dengan permasalahan penelitian. Dalam tahap ini penulis membaca jurnal terkait untuk analisa terhadap penelitian yang sudah ada sebagai bahan pembanding dan memperhatikan referensi terkait penelitian sebelumnya yang telah dilakukan dalam aspek kelebihan dan kekurangan. Serta Mempelajari teori-teori tentang Data Mining, Clustering, K-Means, Python, Google Colab, API Flask, dan penelitian terdahulu.

2. *Business Understanding*

Penulis memahami bagaimana data mining berkaitan dengan bisnsi, Langkah ini melibatkan penentuan tujuan untuk data mining dan penyususnan rencana penelitian yang sesuai, serta memberikan memberikan rekomendasi program studi berdasarkan potensi mahasiswa baru.

3. *Data Understanding*

Proses pemahaman data dimulai dengan mengumpukan informasi awal, mengumpulkan, memeriksa, dan memahami data mahasiswa baru seberapa baik data tersebut.

4. *Data Preparation*

Penulis harus memilih tabel atau entitas yang sesuai dengan alat yang akan digunakan untuk pemodelan. Selanjutnya, membersihkan data dengan cara menghapus informasi yang tidak relevan, menangani data yang hilang, dan menghilangkan duplikasi agar data menjadi lebih tepat dan siap untuk dianalisis.

5. Modelling

Penulis akan memilih cara data mining yang akan digunakan, yaitu algoritma data mining dan nilai parameter terbaik. Beberapa teknik data mining memiliki persyaratan khusus yang harus dipenuhi sebelum penulis bisa melanjutkan ke tahap persiapan data. Dalam penelitian ini, pemodelan data dilakuakan dengan menggunakan metode K-Means Clustering dan Model Evaluation Classification dengan bantuan library dalam bahasa pemrograman Python.

6. Evaluation

Pada tahap ini, penting untuk mengevaluasi model untuk melihat apakah model tersebut memenuhi tujuan utama, yaitu dengan cara menggunakan model evaluasi untuk melihat apakah hasil prediksi sesuai atau tidak.

7. Deployment

Penulis melakukan deployment dalam bentuk visualisasi data menggunakan API Fask ataupun prediksi yang telah ditentukan untuk memudahkan dalam penggunaan model Random Forest.

HASIL DAN PEMBAHASAN

1. *Business Understanding*

Business Understanding adalah tahap awal dalam proses CRISP-DM untuk memahami konteks bisnis atau masalah yang akan diselesaikan menggunakan teknik *data mining*. Dalam konteks penerapan *data mining* untuk penentuan program studi pada penerimaan mahasiswa baru di Institut Bisnis dan Informatika Kesatuan (IBIK), tahap ini melibatkan pemahaman mendalam tentang proses penerimaan mahasiswa baru, kebutuhan institusi pendidikan, dan tujuan akhir yang ingin dicapai.

Dalam konteks penentuan program studi menggunakan *algoritma K- Means Clustering,* tahap ini juga mencakup pemahaman tentang penerapan algoritma tersebut untuk mengelompokkan calon mahasiswa berdasarkan kesamaan karakteristik, memungkinkan institusi pendidikan untuk menentukan program studi yang paling sesuai untuk setiap kelompok mahasiswa. Pemahaman mendalam tentang algoritma ini membantu institusi merekomendasikan keputusan yang tepat dan efektif dalam proses penerimaan mahasiswa baru.

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2024.v5i1.7624

2. Data Understanding

Data understanding adalah fase penting dalam proses *data mining* untuk memahami karakteristik dan kualitas data yang dimiliki. Fase ini melibatkan pengumpulan data awal dan deskripsi data. Selain itu, dilakukan juga *pre- processing data* untuk membersihkan dan mempersiapkan data agar siap digunakan dalam proses *data mining*, seperti penghapusan missing values, normalisasi data numerik, serta deteksi dan penanganan *outlier.* Dalam konteks penentuan program studi bagi penerimaan mahasiswa baru, pemahaman data mencakup identifikasi atribut relevan, serta persiapan dan normalisasi datauntuk analisis yang efektif. Ini memastikan analisis berjalan lancar dan memberikan hasil akurat.

Pengumpulan data awal *(collect initial data*) melibatkan pengumpulan informasi dari berbagai sumber seperti Institut Bisnis dan Informatika Kesatuan, observasi, dan studi literatur, dengan data tes masuk mahasiswa baru tahun 2023 sebanyak 357 data mahasiswa.

Describe data adalah proses analisis dan penggambaran data yang telah dikumpulkan, termasuk pemahaman tentang karakteristik data, seperti tipe data dan jumlah data, data mahasiswa baru tahun 2023 memiliki 28 kolom dan 337 baris, termasuk nilai tes psikologi untuk kemampuan kognitif, motivasi, pemecahan masalah, dan kecenderungan minat bakat. Deskripsi ini membantu institusi pendidikan memahami data yang digunakan untuk analisis penentuan program studi.

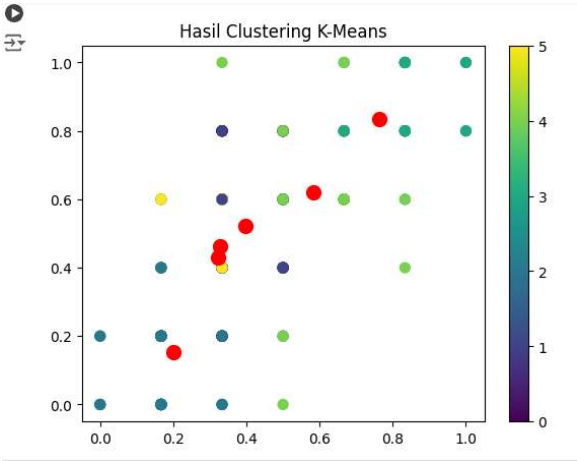
3. Data Preparation

Tahap data *preparation atau pre-processing* data dilakukan menggunakan *Google Colabs*. Langkah awal pada proses ini dengan mengaktifkan beberapa *library* yang diperlukan atau mengimpor semua pustaka yang diperlukan untuk analisis danta dan pembuatan model. Gambar 3 merupakan *library* yang diguankan dalam penelitian ini.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, f1_score, roc_auc_score
from google.colab import files
import plotly.express as px
```

Gambar 3. *Library* yang digunakan

Pada tahap ini, penulis melakukan penghapusan kolom yang tidak digunakan dalam pengolahan data yaitu kolom (Nama, Tanggal Lahir, Taraf Kcerdasan, Skor SEP, dan Level SEP) karena kolom tersebut dianggap tidak relevan, pengecekan data yang memiliki nilai NaN, membuat salinan dari *dataframe*, mengubah data mahasiswa ke dalam bentuk array, normalisasi data array, melakukan clustering pada data menggunakan *K-Means,* melihat informasi titik *centroid* dari tiap *cluster,* menampilkan *cluster* yang sudah ditambahkan pada data mahasiswa, dan menampilkan data dalam bentuk scatter plot untuk melihat titik *centroid,* seperti pada gambar dibawah ini :



Gambar 4. Scattere Plot K-Means Clustering

Penulis kemudian menempatkan program studi berdasarkan nilai tertinggi dari setiap indikator yang telah ditentukan ke dalam klaster yang sesuai. Hasil dari penempatan tersebut adalah sebagai berikut: Teknologi Informasi diletakkan pada klaster 0, Sistem Informasi pada klaster 1, Pariwisata pada klaster 2, Manajemen pada klaster 3, Akuntansi pada klaster 4, dan Biokewirausahaan pada klaster 5. Selanjutnya, program studi tersebut di mapping ke dalam data mahasiswa baru berdasarkan *cluster* yang sesuai.

Sehingga, dari penentuan cluster tersebut menghasilkan 85 mahasiswa yang direkomendasikan masuk ke program studi Manajemen, 84 mahasiswa direkomendasikan masuk ke dalam program studi Pariwisata, 72 mahasiswa direkomendasikan masuk ke dalam program studi Sistem Informasi, 39 mahasiswa direkomendasikan masuk ke dalam prgram studi Biokewirausahaan, 39 mahasiswa direkomendasikan masuk ke dalam program studi Akuntansi, dan 38 mahasiswa direkomendasikan masuk ke dalam program studi Teknologi Informasi.

Tahap selanjutnya adalah melakukan drop column, yang bertujuan untuk mempersiapkan data pada proses klasifikasi. Dengan cara ini, data yang sudah di drop atau variabel ‘X’ hanya akan berisi fitur-fitur yang akan digunakan untuk memprediksi program studi mahasiswa, sedangkan ‘y’ akan berisi label program studi yang akan diprediksi. Tahap terakhir dari data preparation adalah melakukan split data latih dan data uji digunakan untuk membagi data menjadi dua subset: data latih (training data) dan data uji (testing data) dengan rasio 80:20. Langkah ini penting untuk mengevaluasi kinerja model dengan baik. Dengan cara ini, data telah berhasil dibagi menjadi data latih (X_train, y_train) dan data uji (X_test, y_test), dan siap digunakan untuk melatih dan menguji model klasifikasi.

4. Modelling

Pada tahap ini bertujuan untuk melatih model klasifikasi menggunakan *algoritma Random Forest, algoritma K-Nighbors Classifier, dan Support Vector Classifier*, penulis mengguankan 3 model kalsifikasi untuk dilakukan perbandingan manakah model klasifikasi yang lebih cocok dari 3 model tersebut. Berikut merupakan source code pada tahap modelling menggunakan tiga model klasifikasi

a. Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
```

```
[59] # Pelatihan model klasifikasi (Random Forest)
     rf_model = RandomForestClassifier(random_state=123)
     rf_model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=123)
```

```
# Prediksi program studi untuk data uji
y_pred = rf_model.predict(X_test)
rf_model_probs = rf_model.predict_proba(X_test)
```

Gambar 5. Model klasifikasi Random Forest

b. K-Neighbors Classifier

```
[63] from sklearn.neighbors import KNeighborsClassifier
```

```
# Inisialisasi model KNN
knn_model = KNeighborsClassifier()
# Latih model KNN
knn_model.fit(X_train, y_train)
```

```
KNeighborsClassifier
KNeighborsClassifier()
```

```
# Prediksi untuk data uji
y_pred = knn_model.predict(X_test)
y_probs = knn_model.predict_proba(X_test)
```

Gambar 6. Model klasifikasi KNN

c. Support Vector Classifier

```
41] from sklearn.svm import SVC
```

```
# Inisialisasi model SVC dengan probabilitas
svc_model = SVC(probability=True)
# Latih model SVC
svc_model.fit(X_train, y_train)
```

```
SVC
SVC(probability=True)
```

```
# Prediksi untuk data uji
y_pred = svc_model.predict(X_test)
y_probs = svc_model.predict_proba(X_test)
```

Gambar 7. Model klasifikasi SVC

Hasil dari proses K-Means Clustering dalam penelitian ini menghasilkan enam kelompok rekomendasi program studi bagi calon mahasiswa baru. Sebanyak 85 mahasiswa direkomendasikan untuk memilih program studi. Tabel 1merupakan daftar Program Studi dan jumlah mahasiswanya.

Tabel 1. Program Studi dan Jumlah Mahasiswa

No.	Program Studi	Jumlah Mahasiswa
1	Manajemen	85
2	Pariwisata	84
3	Sistem Informasi	72
4	Biokewirausahaan	39
5	Akuntansi	39
6	Teknologi Informasi	38
Total		**357**

Proses ini dilakukan dengan tujuan untuk membantu calon mahasiswa memilih program studi yang paling sesuai dengan potensi dan kemampuan mereka berdasarkan hasil tes masuk. Selanjutnya, hasil clustering ini diimplementasikan dalam bentuk aplikasi web menggunakan Flask API. Aplikasi ini berfungsi memberikan rekomendasi program studi secara otomatis kepada calon mahasiswa. Dalam tahap implementasi, model machine learning yang dipilih untuk deployment adalah Random Forest, karena dinilai memberikan hasil prediksi yang paling optimal. Dengan adanya sistem ini, diharapkan proses seleksi dan penempatan calon mahasiswa baru di program studi yang tepat dapat berjalan lebih efektif dan efisien.

Dengan menggunakan 3 model klasifikasi, penulis mendapatkan prediksi program studi untuk data uji serta probabilitas prediksi untuk setiap kelas. Hal ini memungkinkan untuk melakukan evaluasi lebih lanjut terhadap kinerja model, seperti menghitung akurasi prediksi atau menggunakan metrik evaluasi lainnya.

Adapun Hasil dari ke 3 pengujian terhadap tiga model klasifikasi untuk menentukan model terbaik dalam merekomendasikan pilihan program studi kepada mahasiswa baru. Model pertama, Random Forest, menghasilkan akurasi sebesar 80,56%, yang merupakan nilai akurasi tertinggi di antara ketiga model yang diuji. Model kedua, yaitu K-Nearest Neighbor (KNN), menunjukkan akurasi sebesar 76,39%, sedangkan model ketiga, Support Vector Classifier (SVC), menghasilkan akurasi paling rendah yaitu sebesar 52,78%. Berdasarkan hasil pengujian ini, model Random Forest dipilih untuk diimplementasikan dalam aplikasi web menggunakan API Flask, karena memberikan tingkat akurasi terbaik dalam memprediksi program studi yang sesuai dengan potensi mahasiswa baru.

5. Evaluation

Evaluation (evaluasi) adalah proses untuk mengevaluasi kinerja model yang telah dibuat. Tujuan dari evaluasi adalah untuk memahami seberapa baik model tersebut dapat melakukan tugas-tugas tertentu yang telah ditetapkan. *Source code* tahap *evaluation* merupakan sambungan untuk melihat *output* pada tahap *modelling,* berikut merupakan hasil evlauasi menggunakan tiga model klasifikasi seperti gambar berikut:

a. Evaluasi *Random Forest Calassifier*

```
# Evaluasi model dengan accuracy
accuracy = accuracy_score(y_test, y_pred)
```

```
[62] # Cetak hasil evaluasi
     # format akurasi dengan 4 angka di belakang koma
     print("Evaluasi Model:")
     print(f"Accuracy: {accuracy:.4f}")
```

```
Evaluasi Model:
Accuracy: 0.8056
```

Gambar 8. Evaluasi Random Forest

b. Evaluasi *K-Neighbors Classifier*

```
[67] # Evaluasi model
     accuracy = accuracy_score(y_test, y_pred)
```

```
# Cetak hasil evaluasi
print("Evaluasi Model KNN:")
print(f"Accuracy: {accuracy:.4f}")
```

```
Evaluasi Model KNN:
Accuracy: 0.7639
```

Gambar 9. Evaluasi KNN

c. Evaluasi *Support Vector Classification*

```
[92] # Evaluasi model
     accuracy = accuracy_score(y_test, y_pred)
```

```
# Cetak hasil evaluasi
print("Evaluasi Model SVC:")
print(f"Accuracy: {accuracy:.4f}")
```

```
Evaluasi Model SVC:
Accuracy: 0.5278
```

Gambar 10. Evaluasi SVC

Dengan demikian, berdasarkan hasil evaluasi di atas, dapat disimpulkan bahwa *Random Forest Classifier* adalah model terbaik untuk dataset ini, dengan akurasi tertinggi sebesar 80,56%. *K-Neighbors Classifier* memberikan kinerja yang baik, tetapi sedikit lebih rendah dibandingkan dengan Random forest, dengan akurasi sebesar 76.39%. *Support Vector Classifier* menunjuukkan kinerja yang paling rendah, dengan akurasi sebesar 52.78%, sehingga mungkin kurang sesuai untuk dataset ini. Dalam kasus ini, *Random Forest Classifier* terbukti menjadi pilihan terbaik untuk tahap *deployment*. Oleh karena itu, pada tahap selanjutnya, model Random Forest akan dikembangkan dan diimplementasikan pada API.

6. Deployment

Pada tahap *deployment* ini, penulis menggunakan modul pickle untuk mengubah objek ke format *byte system* agar dapat disimpan ke dalam file atau ditransfer ke dalam jaringan API. API (*Application Programming Interface*) adalah sekumpulan kode pemrograman yang memungkinkan transmisi data natara satu produk perangkat lunak dengan produk perangkat lunak lain, sehingga keduanya dapat lancar berkomunikasi satu sama lain. Berikut merupakan proses yang dilakukan pada tahap *deployment* :

a. Save model random forest menggunakan pickle

```
import pickle

rf_model_filename = 'rf_model.pkl'
with open(rf_model_filename, 'wb') as file:
    pickle.dump(rf_model, file)

print(f"Random Forest Model saved to {rf_model_filename}")
```

```
Random Forest Model saved to rf_model.pkl
```

Gambar 11. Save model Random Forest

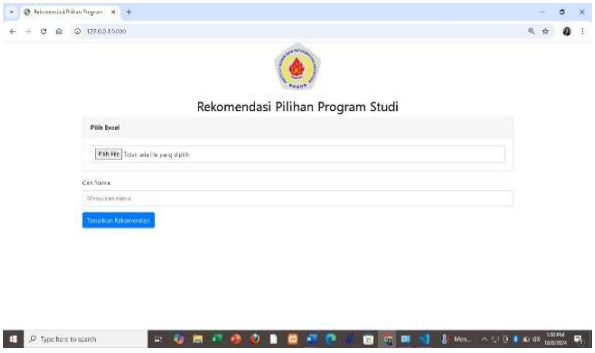
Kode diatas menunjukkan bagaimana cara menyimpan model Random Forest yang telah dilatih ke dalam sebah file dengan menggunanakan 'pickle'. Langkah-langkah tersebut mencakup menentukan nama file, membuka file dalam mode penulisan biner menggunakna 'pickle.dump()' untuk menyimpan model, dan mencetak pesan konfirmasi setelah model disimpan.

b. Menggunakan HTML untuk tampilan Web

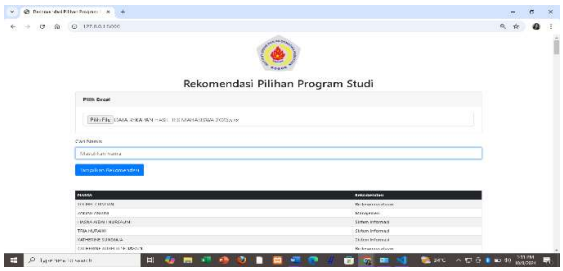
Template ini menggunakan *Bootstrap* untuk tata letak dan *sytling,* dan menerima *input* dari pengguna untuk kemudian diproses oleh aplikasi Flask untuk menghasilkan rekomendasi program studi. Template HTLM ini membentuk antarmuka pengguna untuk memasukkan data yang diperlukan untuk prediksi program studi bagi mahasiswa baru. Data tersebut kemudian dikirim ke *server* Flask untuk diproses, dan hasil rekomendasi program studi ditampilkan kembali kepada pengguna. Integrasi dengan Bootstrap memastikan bahwa *form* dan elemen lainnya memiliki tampilan yang *responsif* dan modern.

c. Tampilan Web

Tahap terakhir ini adalah bagian dari *deployment*, dimana aplikasi web sudah dapat diakses dan digunakan oleh pengguna akhir. Pengguna dapat mengisi data, mengirimkan data untuk diproses, dan melihat hasil rekomendasi program studi yang diberikan oleh model *machine learning*. Ini adalah aplikasi *machine learning* yang dapat diintegrasikan dengan antarmuka web untuk menyediakan layanan yang interaktif dan bermanfaat. Berikut merupakan tampilan web yang sudah dapat diakses :



Gambar 12. Halaman web aplikasi flask rekomendasi program studi



Gambar 13. Contoh aplikasi flask setelah data diinput

KESIMPULAN

Berdasarkan hasil penelitian mengenai penerapan data mining menggunakan algoritma K-Means Clustering untuk rekomendasi pilihan program studi pada mahasiswa baru di Institut Bisnis dan Informatika Kesatuan, diperoleh kesimpulan bahwa proses clustering berhasil mengelompokkan mahasiswa ke dalam enam program studi berdasarkan nilai tes masuk. Hasil clustering tersebut kemudian divalidasi menggunakan tiga model klasifikasi, yaitu Random Forest, K-Nearest Neighbor (KNN), dan Support Vector Classifier (SVC). Dari ketiga model tersebut, Random Forest menunjukkan kinerja terbaik dengan akurasi sebesar 80,56%, diikuti oleh KNN dengan akurasi 76,39%, dan SVC dengan akurasi terendah sebesar 52,78%. Dengan demikian, model Random Forest dipilih untuk diimplementasikan dalam aplikasi web berbasis API Flask sebagai sarana rekomendasi program studi kepada mahasiswa baru. Hasil ini membuktikan bahwa penerapan algoritma K-Means Clustering, yang dikombinasikan dengan model klasifikasi Random Forest, efektif dalam membantu proses rekomendasi program studi sesuai potensi mahasiswa berdasarkan nilai tes masuk mereka.

DAFTAR PUSTAKA

[1] Institut Bisnis dan Informatika Kesatuan. (2023). Panduan Penerimaan Mahasiswa Baru 2023/2024. Bogor: IBI Kesatuan.
[2] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2022). Buku Panduan SNPMB 2022. Jakarta: Kemendikbudristek.
[3] Data Internal Unit Marketing IBI Kesatuan. (2023). Laporan Penerimaan Mahasiswa Baru Tahun 2023.
[4] Wahyuni, S., & Susanto, A. (2021). "Analisis Faktor Dropout Mahasiswa di Perguruan Tinggi Swasta," Jurnal Pendidikan dan Pengajaran, 54(2), 198–207.
[5] Setiawan, R., & Pratama, D. (2020). "Penerapan Algoritma K-Means dalam Pengelompokan Minat Studi Mahasiswa Baru," Jurnal Teknologi Informasi dan Pendidikan, 13(1), 24–30.
[6] Dewi, F. A., Anwar, A., & Yuliana, R. (2021). "Clustering Data Calon Mahasiswa Berdasarkan Minat dan Hasil Tes Masuk Menggunakan K-Means," Jurnal Sistem Informasi, 17(2), 45–53.
[7] Putri, D. F., & Ramadhan, M. I. (2022). "Pengembangan Sistem Rekomendasi Program Studi Menggunakan K-Means Clustering," Jurnal Riset Teknologi Informasi dan Komputer, 8(1), 91–99.
[8] Fauziah, N., & Rachman, A. (2023). "Implementasi K-Means Clustering untuk Pemetaan Minat Program Studi Calon Mahasiswa Baru," Jurnal Teknologi dan Sistem Informasi, 9(1), 12–20.
[9] Pratama, D. (2019). Dasar-Dasar Data dan Informasi dalam Teknologi Informasi. Jakarta: Mitra Wacana Media.
[10] Hidayat, A., & Lestari, R. (2020). "Pengenalan Data Mining dan Aplikasinya di Dunia Industri," Jurnal Informatika dan Teknologi Informasi, 6(1), 12–20.

[11] Association for Computing Machinery (ACM). (2019). Standard Glossary of Data Mining Terms. ACM Press.
[12] Susanto, R. (2021). Algoritma dan Struktur Data untuk Pemula. Bandung: Informatika Bandung.
[13] Fitriani, E., & Mahendra, A. (2022). "Implementasi K-Means untuk Segmentasi Data Mahasiswa Baru," Jurnal Teknologi dan Komputasi, 10(2), 45–52.
[14] Purnama, R. Y. (2023). "Analisis Pengelompokan Data Menggunakan K-Means Clustering," Jurnal Riset Komputer dan Aplikasi (JRKA), 11(1), 88–95.
[15] Fauziah, N., & Ramadhani, A. (2024). Pengantar Data Mining: Teori dan Implementasi. Yogyakarta: Andi Publisher.