

8276 | Profit each tiers



Total Spending € 4,923.87
Savings € 407.52
Foregone Savings € 167.75



Proposal Data Science



DataMinds

Personalized Learning Pathways

Presented By
Team DataMinds

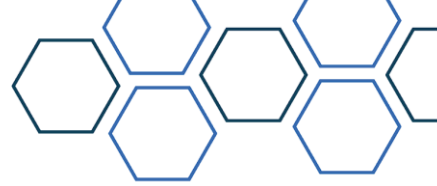
1. Maurino Audrian P.
2. Aurelio Ramadhan

Presented To

ICONIC IT 2024
Universitas Siliwangi



UNIVERSITAS ISLAM BANDUNG
POLITEKNIK ASTRA
2024



KATA PENGANTAR

Bismillahirrahmanirrahim,

Assalamualaikum Warahmatullahi Wabarakatuh

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya, kami dapat menyelesaikan proposal Data Science ICONIC IT 2024 Universitas Siliwangi. Kami telah berupaya menciptakan sebuah inovasi yang diharapkan dapat memberikan kontribusi positif dalam perkembangan teknologi di bidang pendidikan.

Dengan semangat yang membara dan antusiasme yang tinggi, tim DataMinds dari Universitas Islam Bandung dan Politeknik Astra dengan rendah hati menyampaikan kata pengantar ini dalam rangka mengikuti ICONIC IT 2024 Universitas Siliwangi pada cabang lomba Data Science. Kami memahami bahwa Data Science bukan sekadar ilmu data, tetapi juga seni dalam pengolahan data. Oleh karena itu, kami menggunakan kombinasi antara kemampuan teknis dan kreativitas, serta berkomitmen untuk menghadirkan karya-karya yang tidak hanya solutif tetapi juga estetis, sesuai dengan kebutuhan, keinginan, dan kenyamanan pengguna.

Sebagai perwakilan dari Universitas Islam Bandung dan Politeknik Astra, kami merasa bangga dapat berpartisipasi dalam ajang bergengsi ini. Kami juga ingin menyampaikan apresiasi yang tulus kepada para juri yang telah bersedia meluangkan waktu dan tenaga untuk menilai karya-karya kami dengan cermat.

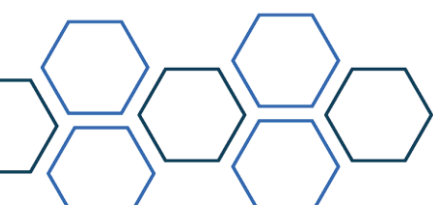
Kami yakin bahwa ICONIC IT 2024 Universitas Siliwangi akan menjadi ajang yang memunculkan inovasi-inovasi terbaru dalam dunia IT. Akhir kata, kami ingin mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan, bantuan, dan motivasi dalam perjalanan kami menuju ICONIC IT 2024 Universitas Siliwangi ini.

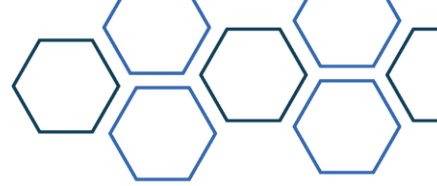
Wassalamualaikum warahmatullahi wabarakatuh

Bandung, 30 Agustus 2024

Penulis,

Tim DataMinds





DAFTAR ISI

KATA PENGANTAR	ii
DAFTAR ISI.....	iii
ABSTRAK	4
PENDAHULUAN	4
LATAR BELAKANG	4
TUJUAN	4
BATASAN MASALAH	5
MANFAAT	5
METODOLOGI.....	5
JENIS DAN SUMBER DATA	5
VARIABEL PENELITIAN	5
PERANGKAT PENELITIAN	6
METODE ANALISIS DATA.....	7
FLOWCHART.....	7
HASIL DAN PEMBAHASAN.....	8
PRE-PROCESSING DATA	8
PEMBAGIAN DATA TRAINING DAN DATA TESTING	11
IMPLEMENTASI RANDOM FOREST PADA PYTHON.....	11
RANCANGAN SISTEM	15
MEMBANGUN APLIKASI WEB MENGGUNAKAN STREAMLIT	15
KESIMPULAN.....	18
DAFTAR PUSTAKA	18
LAMPIRAN	19





ABSTRAK

Pada era digital yang semakin maju, terdapat kebutuhan mendesak untuk meningkatkan kompetensi sumber daya manusia (SDM) dalam menyesuaikan diri dengan perkembangan teknologi, terutama di bidang pendidikan. Penelitian ini bertujuan untuk menganalisis dan memprediksi jalur pembelajaran yang optimal bagi siswa dengan menggunakan pendekatan Data Science serta algoritma *machine learning*, khususnya *random forest*. Melalui pemanfaatan data sekunder yang terkait dengan performa akademik mahasiswa, penulis ini tidak hanya berupaya mengurangi kesalahan pemilihan jalur pembelajaran, tetapi juga menyediakan rekomendasi yang lebih personal melalui pengembangan aplikasi web berbasis Streamlit. Hasil penelitian ini diharapkan dapat memberikan kontribusi positif dalam mengoptimalkan kurikulum pendidikan serta meningkatkan kualitas dan efektivitas proses pembelajaran di Indonesia.

Kata Kunci: Pendidikan, Machine Learning, Random Forest, Jalur Pembelajaran, Aplikasi Web.

PENDAHULUAN

LATAR BELAKANG

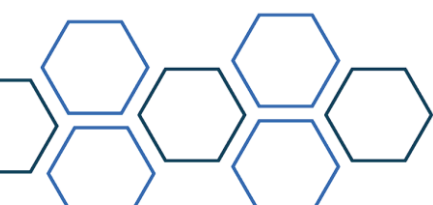
Pada era digital saat ini, manusia harus mampu menyesuaikan kompetensi, keahlian, dan wawasan mereka dengan perkembangan dan pertumbuhan di ranah digital (Purnama, 2023). Terutama dengan kehadiran revolusi industri 4.0 hingga menuju 5.0 yang membawa berbagai teknologi mutakhir. Namun menurut (Syafuruddin et al., 2022) mengemukakan bahwa transformasi ini tidak akan berjalan mulus jika sumber daya manusia (SDM) yang mengendalikan teknologi tersebut belum memadai atau tidak memenuhi standar kualitas yang dibutuhkan. SDM menjadi kunci penting karena pada dasarnya manusia yang mengendalikan dan menggunakan teknologi canggih tersebut, bukan sebaliknya (Kusumaryoko, 2021). Kualitas SDM menjadi kunci krusial pada era revolusi industri saat ini, karena kemajuan teknologi digital tumbuh dengan sangat cepat, terutama dengan munculnya teknologi seperti Data Science, Front End, dan Back End (Erwin et al., 2023). Akibat perubahan ini, terjadi transformasi digital yang mempengaruhi cara, proses, dan fungsi SDM sendiri yang harus dapat beradaptasi dengan penggunaan alat dan teknologi canggih ini dalam segala aspek, termasuk pengambilan keputusan, pemecahan masalah, dan pekerjaan-pekerjaan lainnya.

Berdasarkan temuan serta laporan (Jatmika & Widiarini, 2023), menyatakan bahwa Indonesia diperkirakan membutuhkan sekitar 9 juta individu berbakat di bidang digital hingga tahun 2030 untuk menyongsong era Industri 4.0. Dalam konteks pendidikan, Artificial Intelligence (AI), Machine Learning, ataupun Data Science dapat memberikan kontribusi yang beragam, mulai dari perbaikan dalam proses pembelajaran hingga personalisasi pengalaman belajar. Penelitian ini akan membahas mengenai dampak kecerdasan buatan bagi pendidikan (University, 2023).

Pendidikan merupakan salah satu sektor yang sangat penting dalam pembangunan suatu negara. Dengan hadirnya kecerdasan buatan, pendidikan dapat mengalami perubahan revolusioner dalam hal penyampaian materi, evaluasi, dan pengembangan kurikulum (Wirawan, 2019). Salah satu dampak utama kecerdasan buatan bagi pendidikan adalah adanya personalisasi pembelajaran. Dengan memanfaatkan AI, pendidik dapat merancang pengalaman belajar yang sesuai dengan kebutuhan individual setiap siswa. Hal ini akan meningkatkan efektivitas pembelajaran dan memungkinkan siswa untuk belajar secara lebih efisien (Sugihartono, 2020).

TUJUAN

1. Mengidentifikasi dan menganalisis faktor-faktor yang mempengaruhi jalur pembelajaran siswa, sehingga dapat memberikan wawasan yang lebih mendalam mengenai faktor-faktor penentu jalur pembelajaran.
2. Mengurangi kesalahan pemilihan jalur pembelajaran dengan membangun model *machine learning* dengan performa yang optimal untuk memprediksi jalur pembelajaran yang sesuai untuk siswa.
3. Meningkatkan aksesibilitas prediksi jalur pembelajaran dengan mengembangkan aplikasi web berbasis Streamlit yang intuitif, memudahkan pengguna memasukkan data akademik, dan menampilkan hasil prediksi untuk mendukung pengambilan keputusan yang lebih baik.





BATASAN MASALAH

Penelitian ini memiliki beberapa batasan agar lebih spesifik dan terarah sehingga batasan masalah yang dimiliki yaitu:

1. Data yang digunakan adalah data sekunder dengan tema Pendidikan yang disediakan oleh ICONIC IT 2024 (<https://drive.google.com/drive/folders/1JM58y40nfn5UzUnr7hfIPLUpEfurxD7V>).
2. Variabel umum yang digunakan antara lain:
 - a. **Variabel independen atau atribut/fitur:**
 HOURS_DATASCIENCE, HOURS_BACKEND, HOURS_FRONTEND,
 NUM_COURSES_BEGINNER_DATASCIENCE,
 NUM_COURSES_BEGINNER_BACKEND,
 NUM_COURSES_BEGINNER_FRONTEND,
 NUM_COURSES_ADVANCED_DATASCIENC,
 NUM_COURSES_ADVANCED_BACKEND,
 NUM_COURSES_ADVANCED_FRONTEND, AVG_SCORE_DATASCIENCE,
 AVG_SCORE_BACKEND, dan AVG_SCORE_FRONTEND.
 - b. **Variabel dependen atau label/target:**
 PROFILE.

MANFAAT

Penelitian ini memberikan manfaat yang signifikan bagi berbagai pemangku kepentingan dalam pendidikan. Bagi institusi pendidikan, hasil penelitian dapat digunakan untuk mengoptimalkan kurikulum dan meningkatkan retensi peserta melalui rekomendasi yang lebih personal. Pengajar dapat memperoleh wawasan mendalam tentang cara terbaik menyampaikan materi, yang memungkinkan penyesuaian metode pengajaran untuk hasil yang lebih baik. Siswa dapat menikmati pengalaman belajar yang lebih efektif dan efisien, dengan rekomendasi pembelajaran yang disesuaikan dengan profil mereka, sehingga dapat mencapai skor dan pemahaman yang lebih tinggi. Secara lebih luas, penelitian ini berkontribusi pada peningkatan kualitas pendidikan di Indonesia, memberdayakan sumber daya manusia yang lebih kompeten, serta memperluas akses pendidikan berkualitas ke seluruh penjuru negeri.

METODOLOGI

JENIS DAN SUMBER DATA

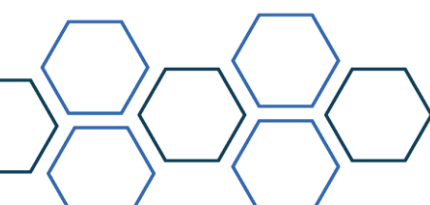
Penelitian ini menggunakan data sekunder yang disediakan oleh ICONIC IT 2024. Dataset yang digunakan dalam penelitian ini adalah dataset dengan tema Pendidikan (<https://drive.google.com/drive/folders/1JM58y40nfn5UzUnr7hfIPLUpEfurxD7V>). Dataset ini berisi informasi mengenai performa akademik mahasiswa serta berbagai faktor yang dapat mempengaruhi jalur pembelajaran mereka.

VARIABEL PENELITIAN

Penelitian ini melibatkan beberapa variabel. Berikut adalah Tabel 1 yang berisi variabel penelitian beserta definisi operasional masing-masing variabel.

Tabel 1. Definisi Operasional Variabel

No.	Variabel	Jenis Variabel	Definisi Operasional Variabel	Skala
1	HOURS_DATASCIENCE	Independen	Jumlah jam yang dihabiskan oleh pengguna untuk belajar atau mengikuti kursus terkait data science.	Numerik
2	HOURS_BACKEND	Independen	Jumlah jam yang dihabiskan oleh pengguna untuk belajar atau mengikuti kursus terkait pengembangan backend.	Numerik





3	HOURS_FRONTEND	Independen	Jumlah jam yang dihabiskan oleh pengguna untuk belajar atau mengikuti kursus terkait pengembangan frontend.	Numerik
4	NUM_COURSES_BEGINNER_DATASCIENCE	Independen	Jumlah kursus tingkat pemula yang diikuti pengguna dalam bidang data science.	Numerik
5	NUM_COURSES_BEGINNER_BACKEND	Independen	Jumlah kursus tingkat pemula yang diikuti pengguna dalam bidang pengembangan backend.	Numerik
6	NUM_COURSES_BEGINNER_FRONTEND	Independen	Jumlah kursus tingkat pemula yang diikuti pengguna dalam bidang pengembangan frontend.	Numerik
7	NUM_COURSES_ADVANCED_DATASCIENCE	Independen	Jumlah kursus tingkat lanjutan yang diikuti pengguna dalam bidang data science.	Numerik
8	NUM_COURSES_ADVANCED_BACKEND	Independen	Jumlah kursus tingkat lanjutan yang diikuti pengguna dalam bidang pengembangan backend.	Numerik
9	NUM_COURSES_ADVANCED_FRONTEND	Independen	Jumlah kursus tingkat lanjutan yang diikuti pengguna dalam bidang pengembangan frontend.	Numerik
10	AVG_SCORE_DATASCIENCE	Independen	Skor rata-rata yang diperoleh pengguna dalam kursus terkait data science.	Numerik
11	AVG_SCORE_BACKEND	Independen	Skor rata-rata yang diperoleh pengguna dalam kursus terkait pengembangan backend.	Numerik
12	AVG_SCORE_FRONTEND	Independen	Skor rata-rata yang diperoleh pengguna dalam kursus terkait pengembangan frontend.	Numerik
13	PROFILE	Dependen	Kategori atau profil pengguna berdasarkan pola belajar dan jumlah kursus yang diikuti. Bisa berupa beginner, advanced.	Kategorik

PERANGKAT PENELITIAN

Perangkat lunak yang digunakan peneliti dalam penelitian ini disajikan pada Tabel 2 dan spesifikasi perangkat keras disajikan pada Tabel 3.

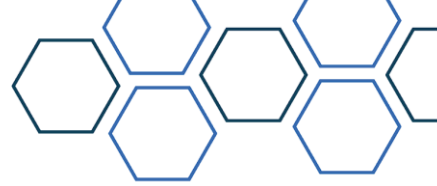
Tabel 2. Daftar Perangkat Lunak

No.	Perangkat Lunak	Kegunaan
1	Anaconda 24.1.2 dan Python 3.11.7	Bahasa Pemrograman untuk melakukan prediksi kelulusan mahasiswa
2	Visual Studio Code 1.87.0	Menulis dan meng- <i>edit</i> code untuk pembuatan aplikasi web

Tabel 3. Spesifikasi Perangkat Keras (Dell Inspiron 3505)

No.	Komponen	Spesifikasi
1	<i>Processor</i>	AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx 2.30 GHz





2	RAM	8,00 GB
3	OS Type	Windows 11 Home Single Language
4	GPU	AMD Radeon Graphics

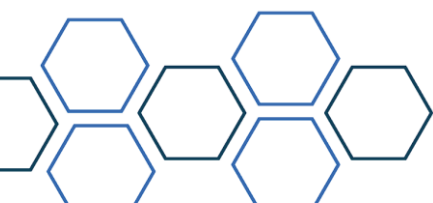
METODE ANALISIS DATA

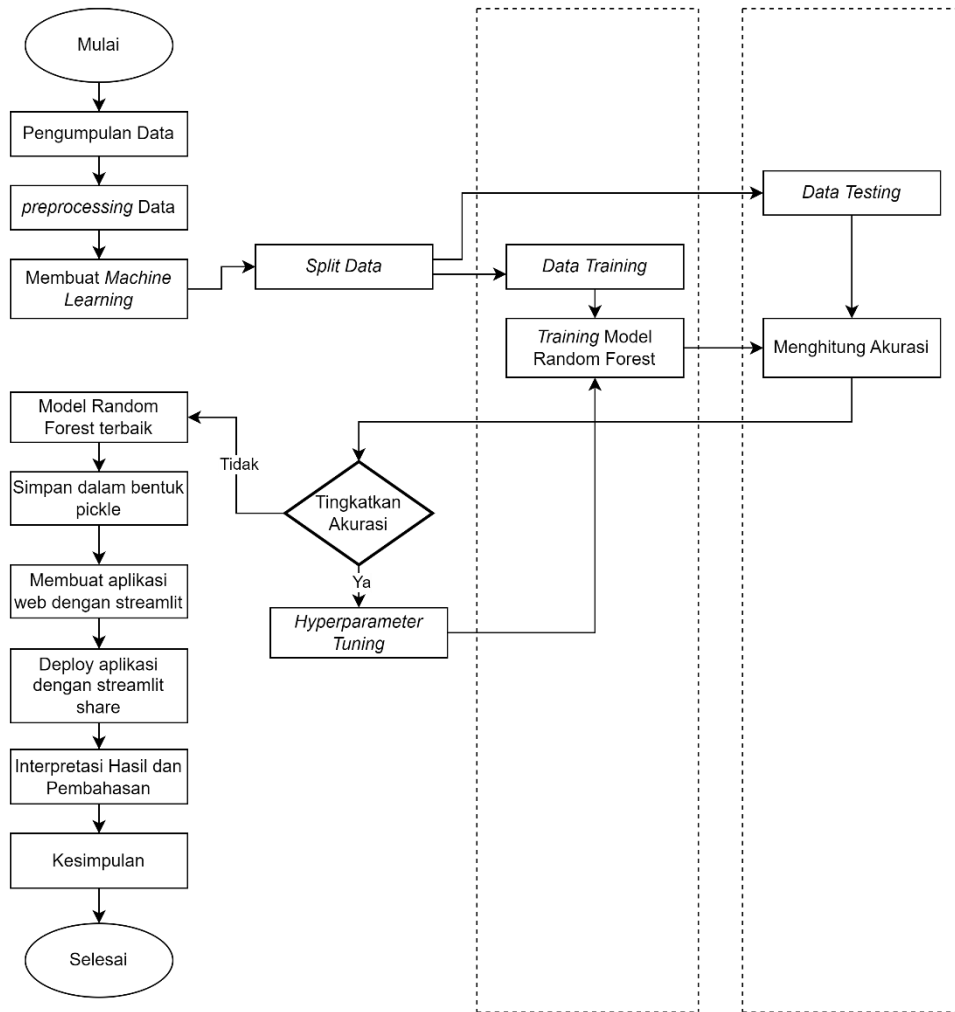
Dalam penelitian ini, metode analisis data melibatkan beberapa tahapan. Pertama, data sekunder dikumpulkan dari platform Kaggle dengan variabel seperti HOURS_DATASCIENCE, HOURS_BACKEND, HOURS_FRONTEND, NUM_COURSES_BEGINNER_DATASCIENCE, NUM_COURSES_BEGINNER_BACKEND, NUM_COURSES_BEGINNER_FRONTEND, NUM_COURSES_ADVANCED_DATASCIENC, NUM_COURSES_ADVANCED_BACKEND, NUM_COURSES_ADVANCED_FRONTEND, AVG_SCORE_DATASCIENCE, AVG_SCORE_BACKEND, AVG_SCORE_FRONTEND, dan PROFILE Data kemudian diproses untuk mengatasi *missing values* dan duplikasi, serta dibagi menjadi data pelatihan dan data pengujian.

Algoritma *random forest* digunakan untuk memilih variabel penting dan membangun model prediksi. Model ini dievaluasi dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Aplikasi web dikembangkan dengan Streamlit di Python untuk memudahkan penggunaan model prediksi. Penelitian ini bertujuan menghasilkan model prediksi jalur belajar siswa yang akurat dan aplikasi web yang berguna untuk meningkatkan kualitas pendidikan di Indonesia.

FLOWCHART

Tahapan yang dilakukan pada penelitian ini dapat digambarkan melalui diagram alir pada Gambar 1 berikut:





Gambar 1. Diagram Alir Penelitian

HASIL DAN PEMBAHASAN

PRE-PROCESSING DATA

Sebelum tahap pemodelan dilakukan *pre-processing* data. *Pre-processing* data dilakukan untuk memastikan bahwa data siap digunakan dalam model *machine learning*. Namun, Dataset pendidikan yang diunduh masih dalam bentuk file csv yang tidak beraturan (Gambar 2). Oleh karena itu, kami melakukan *cleaning data* dengan membuang variabel atau kolom yang tidak diperlukan, seperti Unnamed: 0, NAME, USER_ID.



Unnamed: 0		NAME	USER_ID	HOURS_DATASCIENCE	...	AVG_SCORE_DATASCIENCE	AVG_SCORE_BACKEND	AVG_SCORE_FRONTEND	PROFILE
0	28	Stormy Muto	58283940	7.0	...	84.0	74.0	NaN	beginner_front_end
1	81	Carlos Ferro	1357218	32.0	...	67.0	45.0	NaN	beginner_front_end
2	89	Robby Constantini	63212105	45.0	...	NaN	54.0	47.0	advanced_front_end
3	138	Paul Mckenney	23239851	36.0	...	NaN	71.0	89.0	beginner_data_science
4	143	Jean Webb	72234478	61.0	...	66.0	85.0	NaN	advanced_front_end
...
19995	20495	Rose Jurado	66754730	0.0	...	74.0	73.0	93.0	advanced_backend
19996	20496	Johnny Jones	6874888	0.0	...	50.0	83.0	94.0	advanced_front_end
19997	20497	Lawrence Givens	83752787	32.0	...	61.0	81.0	75.0	advanced_backend
19998	20498	Betty Diclaudio	45806698	0.0	...	64.0	68.0	68.0	advanced_front_end
19999	20499	Connie Harper	67068866	51.0	...	63.0	61.0	87.0	advanced_data_science

20000 rows × 16 columns

Gambar 2. Data Sebelum *Pre-processing*

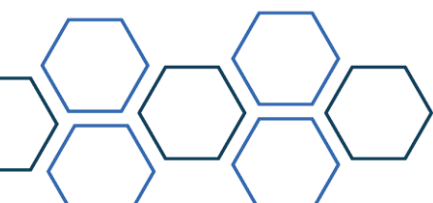
Penghapusan variabel Unnamed: 0, NAME, USER_ID dalam penelitian ini didasarkan pada beberapa alasan ilmiah yang kuat. Variabel Unnamed: 0, NAME, USER_ID dihapus karena tidak memiliki nilai prediktif yang relevan terhadap performa akademis dan untuk menghindari potensi masalah privasi dan etika data (Doshi & Chaturvedi, 2014; Helal et al., 2019). Dengan demikian, penghapusan variabel-variabel ini akan membantu meningkatkan akurasi model prediksi dan mengembangkan rencana tindakan akademik yang lebih efektif. Gambar 3 adalah data setelah penghapusan variabel-variabel yang tidak diperlukan

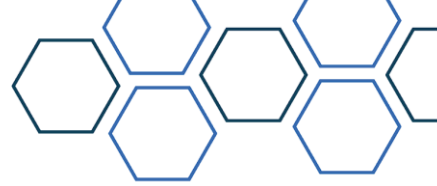
	HOURS_DATASCIENCE	HOURS_BACKEND	HOURS_FRONTEND	...	AVG_SCORE_BACKEND	AVG_SCORE_FRONTEND	PROFILE
0	7.0	39.0	29.0	...	74.0	NaN	beginner_front_end
1	32.0	0.0	44.0	...	45.0	NaN	beginner_front_end
2	45.0	0.0	59.0	...	54.0	47.0	advanced_front_end
3	36.0	19.0	28.0	...	71.0	89.0	beginner_data_science
4	61.0	78.0	38.0	...	85.0	NaN	advanced_front_end
...
19995	0.0	44.0	42.0	...	73.0	93.0	advanced_backend
19996	0.0	85.0	63.0	...	83.0	94.0	advanced_front_end
19997	32.0	50.0	22.0	...	81.0	75.0	advanced_backend
19998	0.0	96.0	69.0	...	68.0	68.0	advanced_front_end
19999	51.0	24.0	36.0	...	61.0	87.0	advanced_data_science

20000 rows × 13 columns

Gambar 3. Data setelah menghapus variabel-variabel yang tidak diperlukan

Setelah menghapus variabel yang tidak diperlukan, selanjutnya adalah mengatasi *missing values*. Data yang terdapat *missing value* akan mempengaruhi tingkat akurasi dalam model *machine learning*. Pengecekan *missing value* pada data dapat dilihat pada Gambar 4 berikut.





```
# tampilkan data yang memiliki missing value
data_cleaned.isna().sum()

HOURS_DATASCIENCE      14
HOURS_BACKEND           53
HOURS_FRONTEND          16
NUM_COURSES_BEGINNER_DATASCIENCE  26
NUM_COURSES_BEGINNER_BACKEND    18
NUM_COURSES_BEGINNER_FRONTEND   39
NUM_COURSES_ADVANCED_DATASCIENCE  2
NUM_COURSES_ADVANCED_BACKEND    8
NUM_COURSES_ADVANCED_FRONTEND   37
AVG_SCORE_DATASCIENCE      220
AVG_SCORE_BACKEND          84
AVG_SCORE_FRONTEND        168
PROFILE                   0
dtype: int64
```

Gambar 4. Missing Value pada setiap variabel data

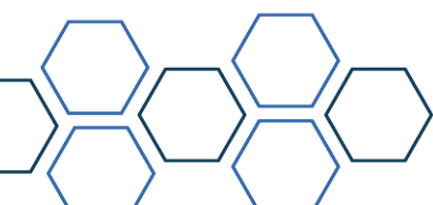
Berdasarkan Gambar 4 diketahui bahwa terdapat *missing value* pada beberapa variabel, maka dilakukan imputasi dengan nilai median. Hal itu dilakukan karena menurut tinjauan literatur oleh (Hasan et al., 2021) menyoroti bahwa imputasi menggunakan median berperan penting dalam meningkatkan performa model *machine learning* saat menangani nilai yang hilang pada data pendidikan. Meskipun ada metode yang lebih canggih seperti deep learning, median tetap menjadi metode yang andal dalam berbagai situasi. Setelah melakukan pengecekan *missing value* selanjutnya melakukan pengecekan terhadap data duplikat.

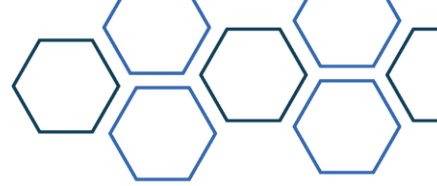
```
# cek data duplikat
data_cleaned.duplicated().sum()

0
```

Gambar 5. Pengecekan data duplikat

Berdasarkan Gambar 5 diketahui bahwa tidak terdapat data yang duplikat. Setelah melakukan pengecekan *missing value* dan data duplikat, Proses selanjutnya adalah melakukan transformasi, yaitu normalisasi pada variabel numerik dan *encoding* pada variabel kategorikal. Hal ini dilakukan agar algoritma *random forest* dapat melakukan komputasi terhadap data. Normalisasi dilakukan dengan fungsi *StandardScaler*, sedangkan *encoding* dilakukan dengan fungsi *LabelEncoder* yang terdapat pada module *sklearn.preprocessing*. Gambar 6 adalah hasil transformasi data pada variabel numerik dan kategorikal.





	HOURS_DATASCIENCE	HOURS_BACKEND	HOURS_FRONTEND	...	AVG_SCORE_BACKEND	AVG_SCORE_FRONTEND	PROFILE
0	-1.365333	-0.203549	-0.385618	...	0.476737	0.059916	5
1	-0.248584	-1.947946	0.343187	...	-1.567757	0.059916	5
2	0.332126	-1.947946	1.071992	...	-0.933259	-1.399848	2
3	-0.069904	-1.098112	-0.434205	...	0.265238	1.519681	4
4	1.046845	1.540848	0.051665	...	1.252235	0.059916	2
...
19995	-1.678022	0.020092	0.246013	...	0.406237	1.797731	0
19996	-1.678022	1.853945	1.266341	...	1.111235	1.867244	2
19997	-0.248584	0.288461	-0.725727	...	0.970236	0.546505	0
19998	-1.678022	2.345955	1.557863	...	0.053738	0.059916	2
19999	0.600145	-0.874471	-0.045509	...	-0.439760	1.380656	1

20000 rows × 13 columns

Gambar 6. Hasil Transformasi Data

PEMBAGIAN DATA TRAINING DAN DATA TESTING

Sebelum melakukan pemodelan dengan menggunakan klasifikasi *random forest*, langkah pertama yang harus dilakukan adalah membagi data menjadi dua bagian: data *training* dan data *testing*. Langkah ini bertujuan untuk mengukur kinerja model dengan mengevaluasi kesalahan prediksi yang mungkin terjadi. Data *training* digunakan untuk melatih algoritma dan membentuk model, sedangkan data *testing* digunakan untuk menguji keakuratan model yang telah dibuat. Jika performa yang dihasilkan oleh model tersebut tinggi, maka model tersebut dapat diandalkan untuk melakukan prediksi pada data baru. Data *training* dan data *testing* dibagi dengan proporsi 80% untuk data *training* dan 20% untuk data *testing* dari total dataset.

Tabel 4. Proporsi Data Training dan Data Testing

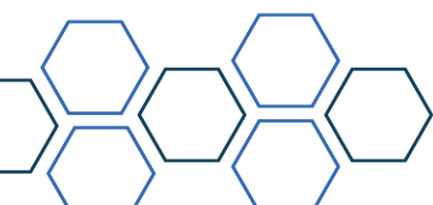
Keterangan	Data Training	Data Testing	Total
Proporsi	80%	20%	100%
Jumlah	16000	4000	20000

Berdasarkan pada Tabel 4 diketahui bahwa 20000 dataset yang ada, pembagian data untuk data *training* sebanyak 16000 data, sedangkan untuk data *testing* ada sebanyak 4000 data. Pembagian data *training* dan data *testing* pada dataset dilakukan secara *random* dengan bantuan *software* Python.

IMPLEMENTASI RANDOM FOREST PADA PYTHON

Setelah membagi data menjadi data *training* dan data *testing*, langkah selanjutnya adalah melakukan analisis klasifikasi menggunakan *random forest* pada data sampel *training* yang telah ditentukan. Variabel dependen/target dalam penelitian ini adalah PROFILE (jalur pembelajaran siswa) yang akan diprediksi, sementara variabel HOURS_DATASCIENCE, HOURS_BACKEND, HOURS_FRONTEND, NUM_COURSES_BEGINNER_DATASCIENCE, NUM_COURSES_BEGINNER_BACKEND, NUM_COURSES_BEGINNER_FRONTEND, NUM_COURSES_ADVANCED_DATASCIENCE, NUM_COURSES_ADVANCED_BACKEND, NUM_COURSES_ADVANCED_FRONTEND, AVG_SCORE_DATASCIENCE, AVG_SCORE_BACKEND, AVG_SCORE_FRONTEND berperan sebagai variabel independen/fitur.

Langkah pertama dalam membangun model klasifikasi *random forest* adalah melakukan *hyperparameter tuning* menggunakan fungsi *gridsearchCV* pada *scikit-learn* dan menggunakan





parameter *default* dari *random forest*. Parameter yang digunakan dalam penelitian ini adalah hasil modifikasi dari parameter penelitian sebelumnya (Asif et al., 2023; Hu & Sokolova, 2020; Mellal, 2022). Tabel 5 adalah hasil *hyperparameter tuning* yang diperoleh melalui proses *GridSearchCV* dengan menggunakan *10-fold cross validation* untuk mengevaluasi kinerja model *random forest* dengan sepuluh kali pengulangan dalam proses *GridSearchCV* untuk setiap parameter dan Tabel 6 adalah hasil parameter *default* yang diperoleh algoritma *random forest*.

Tabel 5. Hasil *Hyperparameter Tuning* Algoritma *Random Forest*

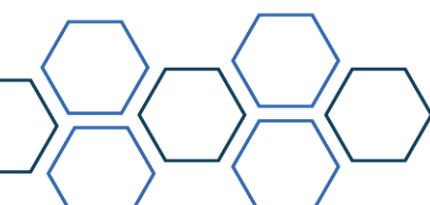
Parameter	Grid Search Values	Parameter Terbaik	Akurasi dengan Hyperparameter Tuning: 85%
<i>n_estimators</i>	100, 200, 500	500	
<i>max_depth</i>	4, 5, 6, 7, 8	8	
<i>min_samples_leaf</i>	1, 2, 4	2	
<i>max_features</i>	<i>sqrt</i> , log 2	<i>sqrt</i>	
<i>min_samples_split</i>	2, 5, 10	2	

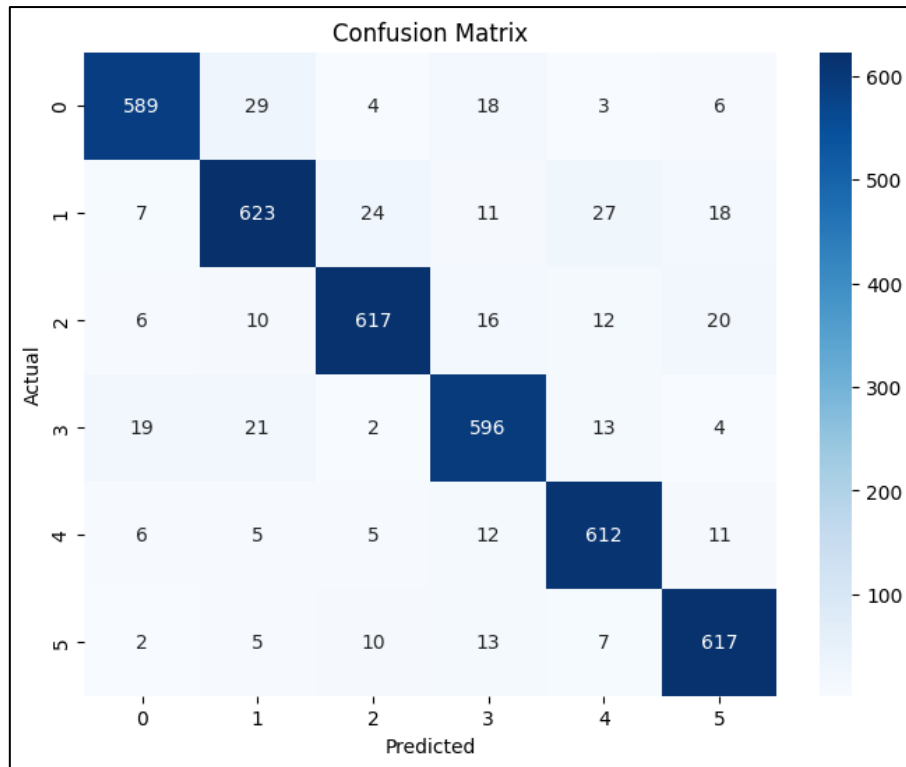
Tabel 6. Hasil Parameter *Default* Algoritma *Random Forest*

Parameter	Parameter Default	Akurasi dengan Parameter Default: 91%
<i>n_estimators</i>	100	
<i>max_depth</i>	<i>None</i>	
<i>min_samples_leaf</i>	1	
<i>max_features</i>	<i>sqrt</i>	
<i>min_samples_split</i>	2	

Berdasarkan Tabel 5 dan Tabel 6, Hasil menunjukkan bahwa model *random forest* dengan menggunakan parameter *default* mendapatkan hasil yang lebih baik dari hyperparameter tuning berdasarkan metrik akurasi. Menurut Penelitian dilakukan oleh (Muhamad Malik Matin, 2023) menemukan bahwa setelah melakukan *hyperparameter tuning* menggunakan *GridSearchCV*, akurasi model *random forest* justru menurun dibandingkan dengan penggunaan parameter *default*. Penurunan ini dikaitkan dengan kompleksitas model yang meningkat dan kemungkinan overfitting pada data pelatihan, sehingga performa pada data uji menurun. Oleh karena itu, penulis menggunakan parameter *default* sebagai dasar untuk membangun model *random forest* yang lebih optimal.

Setelah model diperoleh, langkah selanjutnya adalah evaluasi model. Langkah ini bertujuan untuk mengidentifikasi keakuratan model. Ukuran yang digunakan untuk mengevaluasi hasil prediksi model meliputi nilai *accuracy*, *precision*, *recall*, dan *F1-score* dengan menggunakan *confusion matrix*.





Gambar 7. Confusion Matrix Random Forest

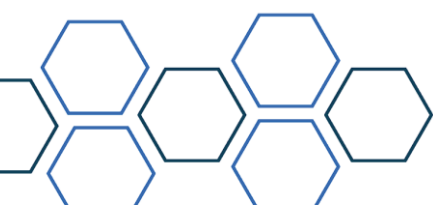
Berdasarkan Gambar 7 diketahui bahwa kelas 0 = advanced_backend, kelas 1 = advanced_data_science, kelas 2 = advanced_front_end, kelas 3 = beginner_backend, kelas 4 = beginner_data_science, dan kelas 5 = beginner_front_end. Kemudian data yang digunakan pada proses uji data sebanyak 4000 sampel. Untuk menilai performa model, penulis menggunakan empat metrik evaluasi. Tabel 7 adalah hasil metrik evaluasi model *random forest*.

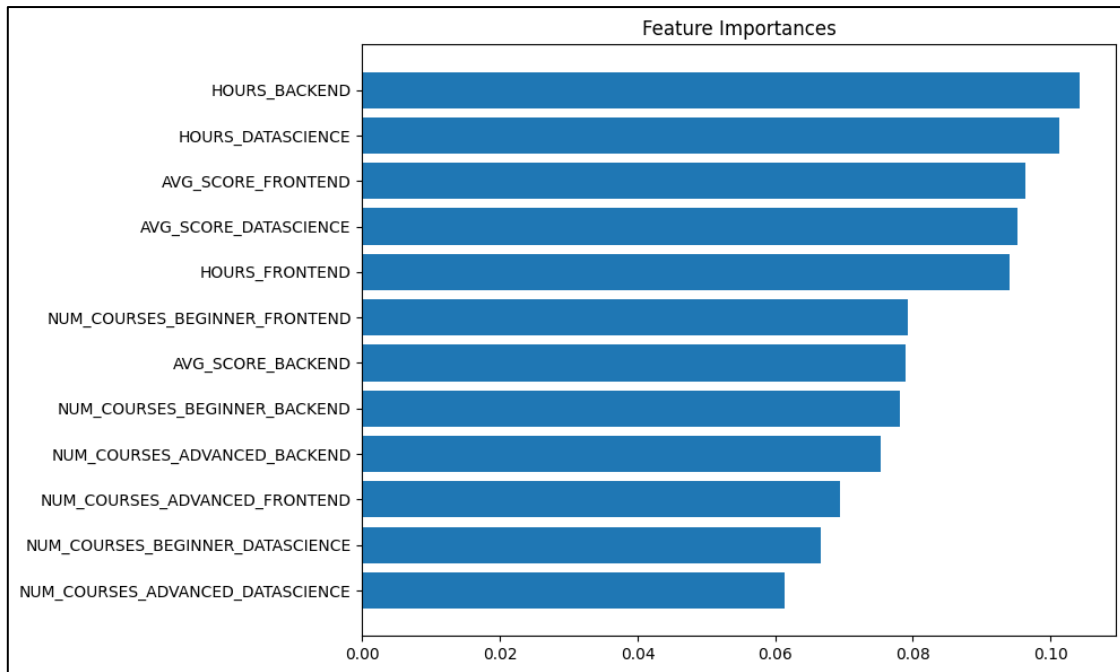
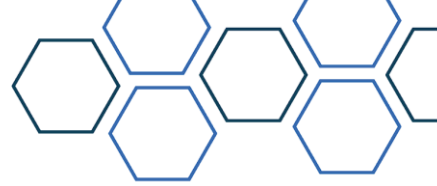
Tabel 7. Hasil Metrik Evaluasi Random Forest

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
91%	91%	91%	91%

Berdasarkan Gambar 7 dan Table 7, hasil *confusion matrix* dan metrik evaluasi menunjukkan bahwa model *random forest* yang digunakan memiliki kinerja yang sangat baik dalam mengklasifikasikan data uji. Dengan akurasi, presisi, *recall*, dan *F1-score* masing-masing sebesar 91%, model ini menunjukkan kemampuan yang konsisten dalam memprediksi kelas-kelas yang berbeda, yaitu kelas advanced backend, advanced data science, advanced frontend, beginner backend, beginner data science, dan beginner frontend. Setiap kelas memiliki tingkat kesalahan yang rendah, seperti yang terlihat dari distribusi diagonal pada *confusion matrix*, yang menandakan bahwa sebagian besar prediksi model berada pada kelas yang benar sesuai dengan kelas aktual. Hal ini mengindikasikan bahwa model *random forest* mampu menangani data dengan baik dalam tugas klasifikasi multi-kelas ini.

Selanjutnya adalah mengukur kepentingan variabel independent/fitur terhadap profil jalur belajar siswa berdasarkan nilai *feature importances*.



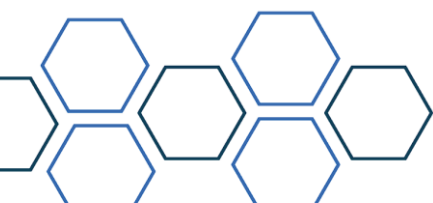


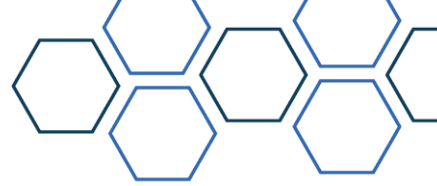
Gambar 8. Feature Importances

Berdasarkan hasil *feature importances* pada Gambar 8, dapat disimpulkan bahwa jumlah jam belajar di bidang Backend dan Data Science (**HOURS_BACKEND** dan **HOURS_DATASCIENCE**) merupakan faktor yang paling dominan dalam memprediksi profil pengguna, menunjukkan betapa pentingnya durasi pembelajaran di kedua bidang tersebut. Skor rata-rata dalam kursus Frontend dan Data Science (**AVG_SCORE_FRONTEND** dan **AVG_SCORE_DATASCIENCE**) juga memberikan kontribusi signifikan, menandakan bahwa performa akademis di bidang ini berpengaruh besar terhadap profil pengguna. Meskipun jam belajar di Frontend (**HOURS_FRONTEND**) juga relevan, pengaruhnya sedikit di bawah dua bidang lainnya. Jumlah kursus yang diambil, baik di tingkat pemula maupun lanjutan, masih berperan, namun dengan tingkat kepentingan yang lebih rendah, terutama pada Data Science lanjutan. Hal ini mengindikasikan bahwa fokus utama dalam memahami profil pengguna sebaiknya diberikan pada jam belajar dan performa akademis di bidang Backend dan Data Science. Nilai *feature importances* dari masing-masing variabel dapat dilihat pada Tabel 8 berikut.

Tabel 8. Feature Importances

No.	Variabel	Features Importance
1	HOURS_BACKEND	0.104287
2	HOURS_DATASCIENCE	0.101225
3	AVG_SCORE_FRONTEND	0.096338
4	AVG_SCORE_DATASCIENCE	0.095130
5	HOURS_FRONTEND	0.094014
6	NUM_COURSES_BEGINNER_FRONTEND	0.079294
7	AVG_SCORE_BACKEND	0.078881
8	NUM_COURSES_BEGINNER_BACKEND	0.078094
9	NUM_COURSES_ADVANCED_BACKEND	0.075301
10	NUM_COURSES_ADVANCED_FRONTEND	0.069414
11	NUM_COURSES_BEGINNER_DATASCIENCE	0.066651
12	NUM_COURSES_ADVANCED_DATASCIENCE	0.061371

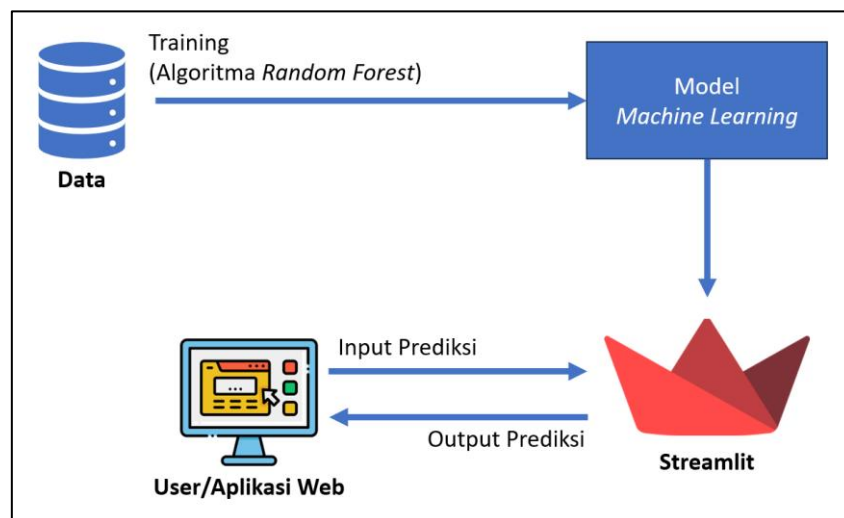




Setelah analisis selesai dilakukan, model yang telah diperoleh kemudian disimpan untuk pengembangan aplikasi web, pada Python disimpan dengan perintah *pickle*. Model ini juga akan tersimpan dan memudahkan untuk memprediksi suatu data karena peneliti hanya memanggil model tanpa melakukan analisis yang sebelumnya dilakukan. Model ini akan digunakan dalam aplikasi Streamlit.

RANCANGAN SISTEM

Rancangan sistem dalam penelitian ini bertujuan untuk mengembangkan aplikasi web yang dapat memprediksi jalur pembelajaran siswa menggunakan algoritma *random forest* dan platform Streamlit. Sistem ini terdiri dari beberapa komponen utama: dataset profil jalur pembelajaran siswa, model *machine learning*, dan aplikasi web Streamlit. Arsitektur sistem ditampilkan pada Gambar 9.



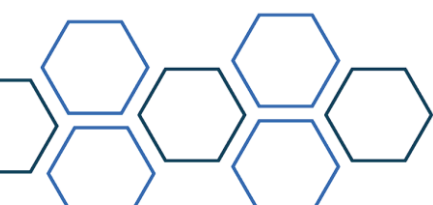
Gambar 9. Arsitektur Sistem

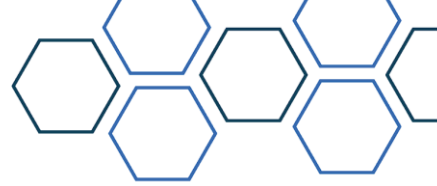
Setelah model dilatih, hasilnya disimpan dalam format *.pkl* untuk digunakan dalam aplikasi web. Aplikasi web dikembangkan menggunakan Streamlit, yang menyediakan antarmuka pengguna interaktif untuk memasukkan data mahasiswa dan mendapatkan prediksi kelulusan. Aplikasi ini memiliki fitur input data, tombol prediksi, hasil prediksi, dan probabilitas prediksi. *Deployment* aplikasi dilakukan melalui Streamlit Sharing, yang memungkinkan aplikasi diakses secara online melalui URL publik. Dengan rancangan sistem ini, aplikasi web yang dikembangkan dapat membantu pengguna memprediksi jalur pembelajaran yang sesuai untuk siswa.

MEMBANGUN APLIKASI WEB MENGGUNAKAN STREAMLIT

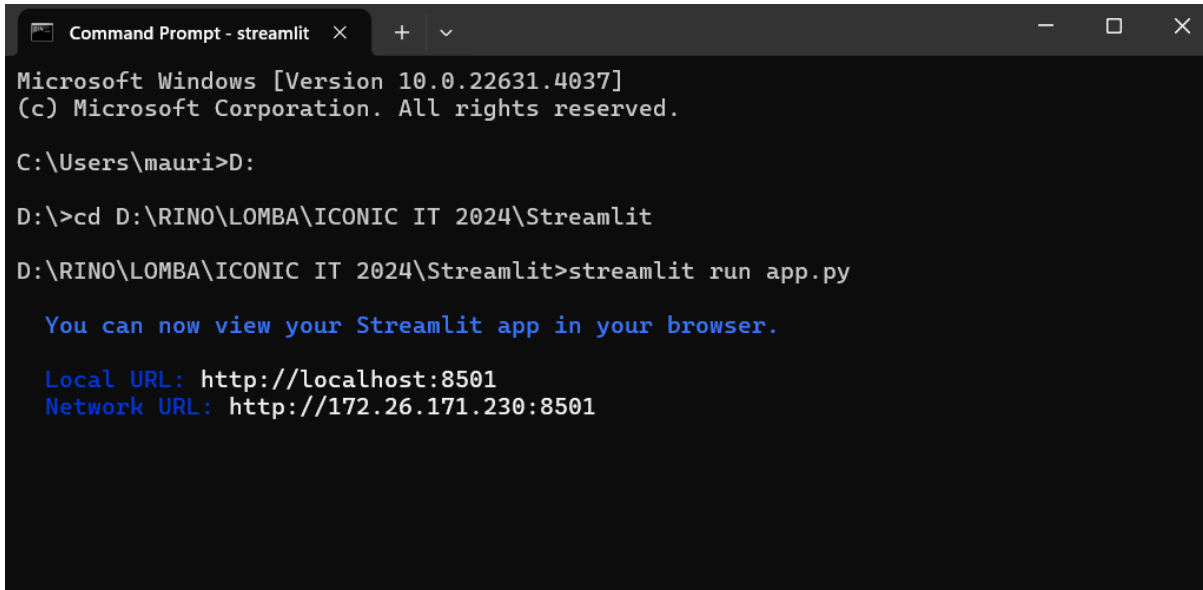
Setelah tahapan pembentukan model, model terbaik akan digunakan menjadi *prototype* aplikasi yang dapat memprediksi jalur pembelajaran yang tepat untuk siswa beserta probabilitas prediksinya. Prediksi ini dihasilkan menggunakan model *random forest* yang dipengaruhi oleh beberapa variabel independent/fitur seperti jam belajar, jumlah kursus, dan skor rata-rata di Data Science, Backend, dan Frontend. Berikutnya adalah membuat file aplikasi Streamlit *app.py*. Di dalam file ini, memuat model yang telah dilatih, judul, deskripsi aplikasi, dan form input dari pengguna. Komponen yang digunakan adalah *st.number_input* untuk mendapatkan input jam belajar, jumlah kursus, dan skor rata-rata di Data Science, Backend, dan Frontend.

Setelah mendapatkan input dari pengguna, tombol prediksi ditambahkan menggunakan *st.button*. Saat tombol ditekan, aplikasi akan menampilkan hasil prediksi dengan warna hijau untuk prediksi jalur pembelajaran yang paling sesuai dengan siswa beserta dengan probabilitas prediksinya.





Untuk menjalankan aplikasi Streamlit, gunakan perintah "streamlit run *app.py*" di terminal seperti yang terlihat pada Gambar 10. Gambar 11 adalah Aplikasi yang sudah dapat diakses secara lokal.



```
Command Prompt - streamlit
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mauri>D:

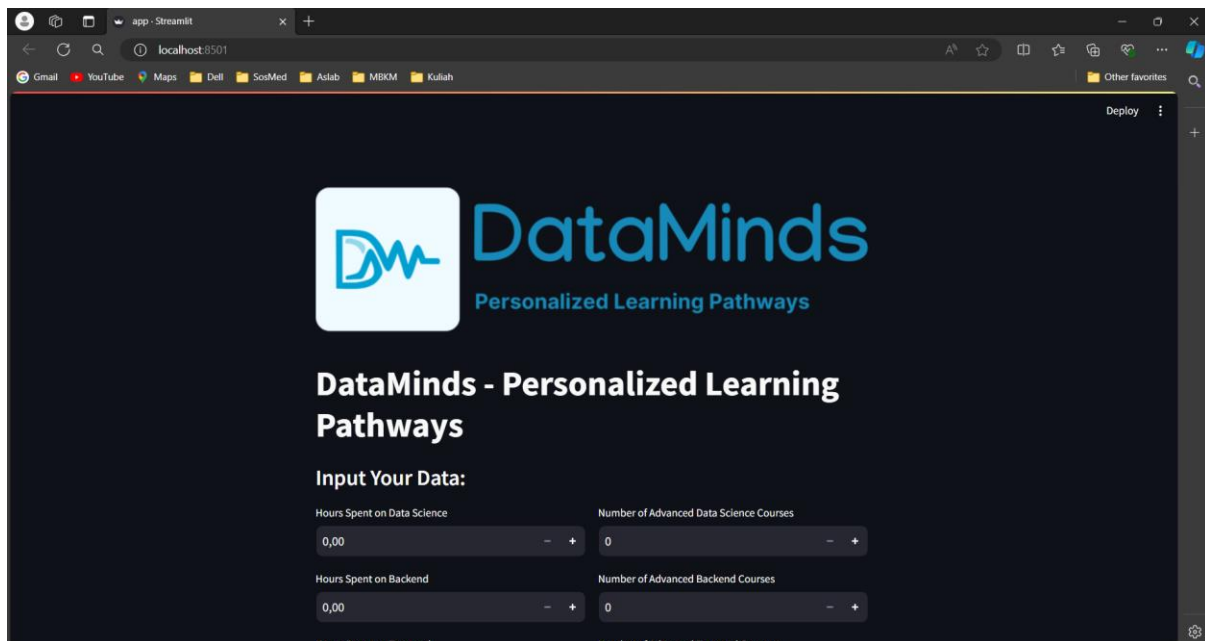
D:\>cd D:\RINO\LOMBA\ICONIC IT 2024\Streamlit

D:\RINO\LOMBA\ICONIC IT 2024\Streamlit>streamlit run app.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://172.26.171.230:8501
```

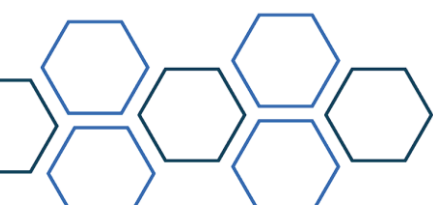
Gambar 10. Perintah untuk menjalankan aplikasi secara lokal

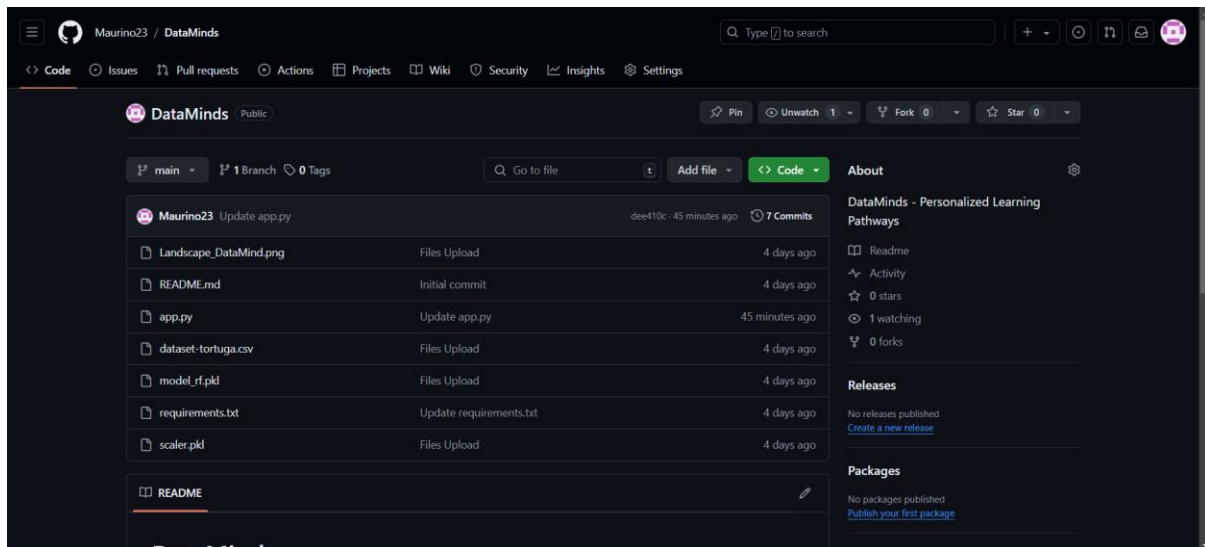
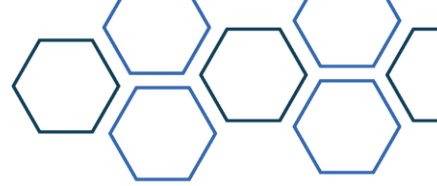


Gambar 11. Tampilan Aplikasi yang sudah berjalan secara lokal

Tahap terakhir adalah *deployment* agar aplikasi dapat diakses secara publik. Streamlit Sharing adalah platform hosting gratis yang disediakan oleh Streamlit untuk memudahkan *deployment* aplikasi web berbasis Streamlit. Dengan Streamlit Sharing, aplikasi dapat dengan cepat dibagikan secara publik tanpa perlu konfigurasi server yang kompleks.

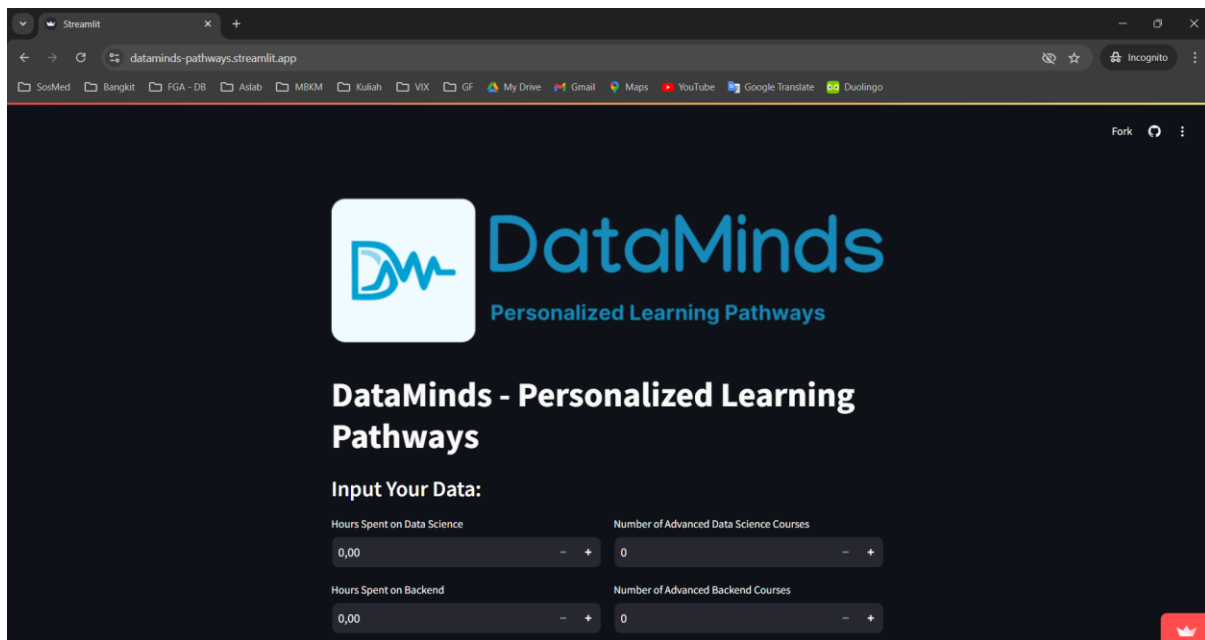
Untuk mendistribusikan aplikasi web prediksi kelulusan mahasiswa menggunakan Streamlit Sharing, berikut adalah beberapa langkah pentingnya. Pertama, buat *repository* baru di GitHub dan unggah semua file yang diperlukan, termasuk file aplikasi (*app.py*), model *machine learning* (*model_rf.pkl*), dan *requirements.txt* yang berisi daftar dependensi seperti Streamlit, pandas, dan scikit-learn. Gambar 12 adalah tampilan *repository* DataMinds Github yang sudah memuat file yang diperlukan (<https://github.com/Maurino23/DataMinds>).



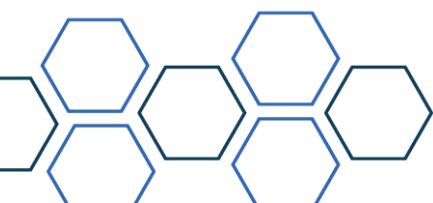


Gambar 12. Repository Github

Selanjutnya, daftarkan akun di Streamlit Sharing dan hubungkan dengan akun GitHub. Setelah itu, pada *dashboard* Streamlit Sharing, pilih opsi untuk membuat aplikasi baru, dan hubungkan dengan *repository* GitHub yang telah dibuat. Isi detail aplikasi seperti branch yang digunakan dan nama file utama (*app.py*). Klik tombol "Deploy" untuk memulai proses *deployment*. Streamlit Sharing akan otomatis menyiapkan lingkungan, menginstal dependensi, dan menjalankan aplikasi yang telah dibuat. Setelah proses ini selesai, URL publik akan didapatkan yang dapat dibagikan kepada pengguna untuk mengakses aplikasi (<https://dataminds-pathways.streamlit.app/>). Dengan menggunakan Streamlit Sharing, penulis dapat dengan mudah dan cepat melakukan *deployment* aplikasi web berbasis *machine learning* tanpa perlu melakukan konfigurasi server yang rumit. Gambar 13 adalah tampilan aplikasi web yang telah berhasil di-deploy.



Gambar 13. Tampilan aplikasi yang telah berhasil di-deploy





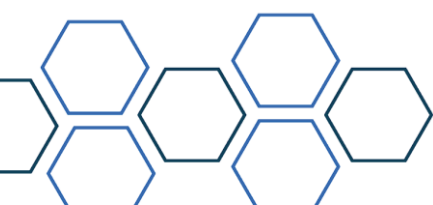
KESIMPULAN

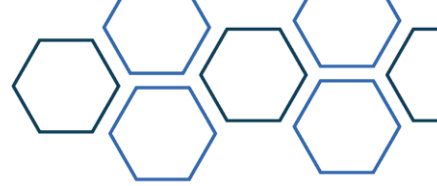
Secara keseluruhan, solusi DataMinds membantu meningkatkan hasil belajar siswa melalui rekomendasi jalur pembelajaran yang dipersonalisasi, mengoptimalkan pendidikan, dan mendukung pengambilan keputusan yang lebih baik. Berdasarkan hasil analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut:

1. Faktor-faktor yang secara signifikan mempengaruhi jalur pembelajaran yang akan diambil seorang siswa telah diidentifikasi berdasarkan nilai *feature importances*. Faktor utama adalah jumlah jam belajar di bidang Backend dan Data Science (**HOURS_BACKEND** dan **HOURS_DATASCIENCE**) merupakan faktor yang paling dominan dalam memprediksi profil pengguna.
2. Prediksi jalur pembelajaran siswa dapat dilakukan dengan algoritma *random forest* yang telah dioptimalkan melalui *hyperparameter tuning* dan *default* parameter. Model ini menghasilkan akurasi, presisi, *recall*, dan *F1-score* masing-masing sebesar 91%.
3. Aplikasi web untuk memprediksi jalur pembelajaran siswa telah dikembangkan dan dapat diakses melalui <https://dataminds-pathways.streamlit.app/>. Aplikasi ini memudahkan pengguna dengan menyediakan bagian input, tampilan data, tombol prediksi, hasil prediksi, dan probabilitas prediksi.

DAFTAR PUSTAKA

- Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms*, 16(6), 308. <https://doi.org/10.3390/a16060308>
- Doshi, M., & Chaturvedi, S. K. (2014). Correlation Based Feature Selection (CFS) Technique to Predict Student Performance. *International Journal of Computer Networks & Communications*, 6(3), 197–206. <https://doi.org/10.5121/ijcnc.2014.6315>
- Erwin, E., Chatra P, A., Pasaribu, A., Novel, N., Sepriano, Thaha, A., Adhicandra, I., Suardi, C., Nasir, A., & Syafaat, M. (2023). *Transformasi Digital*. PT. Sonpedia Publishing Indonesia.
- Hasan, Md. K., Alam, Md. A., Roy, S., Dutta, A., Jawad, Md. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, 27, 100799. <https://doi.org/10.1016/j.imu.2021.100799>
- Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics*, 7(3), 227–245. <https://doi.org/10.1007/s41060-018-0141-y>
- Hu, Y., & Sokolova, M. (2020). *Explainable Multi-class Classification of Medical Data*.
- Jatmika, A., & Widiarini. (2023, June 9). *Indonesia Butuh 9 Juta Talenta Digital pada 2030, Apa yang Perlu Dipersiapkan Pelaku Industri?* Money.Kompas.Com. <https://money.kompas.com/read/2023/06/09/150600726/indonesia-butuh-9-juta-talenta-digital-pada-2030-apa-yang-perlu-dipersiapkan>
- Kusumaryoko, P. (2021). *Manajemen Sumber Daya Manusia di Era Revolusi Industri 4.0*. Deepublish.
- Mellal, M. A. M. (2022). *Design and Control Advances in Robotics*. IGI Global.
- Muhamad Malik Matin, I. (2023). Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware. *MULTINETICS*, 9(1), 43–50. <https://doi.org/10.32722/multinetics.v9i1.5578>
- Purnama, Y. H. (2023). Strategi Pengembangan Eksistensi Karyawan di Era Digital Perspektif Teori Core Competence. *Journal of Management and Bussines (JOMB)*, 5(2), 882–895. <https://doi.org/10.31539/jomb.v5i2.6838>





- Sugihartono. (2020). Pendidikan Personalisasi dalam Era Kecerdasan Buatan: Kajian Implementasi di Indonesia. *Jurnal Pendidikan Dan Teknologi Informasi*, 7(1), 13–22.
- Syafruddin, S. E., Periansya, S. E., Farida, E. A., Tawaf, N., Palupi, F. H., Butarbutar, D., & Satriadi. (2022). *Manajemen Sumber Daya Manusia*. CV Rey Media Grafika.
- Wirawan, Y. E. (2019). Pendidikan di Era Kecerdasan Buatan: Konsep, Aplikasi, dan Tantangan. *Jurnal Pendidikan IPA Indonesia*, 8(4), 529–537.

LAMPIRAN

Link Video Presentasi dan Demo Aplikasi

<https://drive.google.com/file/d/1C5KPkSDLmMSaPMACYCA53zn3YclkZdbN/view?usp=sharing>

Link DataMinds Web App

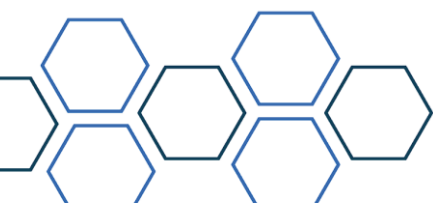
<https://dataminds-pathways.streamlit.app/>

Link Google Colab

<https://colab.research.google.com/drive/163YXgx4BJVYVXnij33TlsvOdWxY7uryv?usp=sharing>

Link Github

<https://github.com/Maurino23/DataMinds>



8276 | Profit each tiers



Total Spending € 4,923.87
Savings € 407.52
Foregone Savings € 167.75



DataMinds

DataMinds

Personalized Learning Pathways

"Emphasizes the creation of tailored educational recommendations based on your data."

(DataMinds, 2024)

DATA
SCIENCE