EXPLORE
Jurnal Sistem Informasi dan Telematika
(Telekomunikasi, Multimedia dan Informatika)

Penerapan Seleksi Fitur *Analysis of Variance* Pada Algoritma *Random Forest Classifier* Dalam Klasifikasi Nilai Mahasiswa

Muhammad Fath Thoriq, Wawan Joko Pranoto, Faldi
Jurusan Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Muhammadiyah Kalimantan Timur
Samarinda, Indonesia
1911102441166@umkt.ac.id, wjp337@umkt.ac.id, fal146@umkt.ac.id

Abstract-In educational institutions, such as universities, it is important to pay attention to the performance of students so that they can complete their studies on time. However, there are still issues where some students are unable to finish their studies on time, and some even decide to drop out or become inactive as students. This is evidenced by the decline in student grades in the Indonesian Language course from the 2020 to 2021 cohorts at UMKT. To address this problem, a method is needed to measure students' performance in completing their studies. This study aims to identify the attributes that influence the decline in student grades in the Indonesian Language course, as well as improve the accuracy of the Random Forest Classifier algorithm using ANOVA feature selection. The data used in this study consists of UMKT student data who took the Indonesian Language course from the 2020/2021 to 2021/2022 academic years. The data was obtained from the academic administration department (BAA) of UMKT and the General Basic Course Unit (MKDU) of UMKT, with a total of 1028 data points. The data analysis process was conducted using the 5-Fold Cross Validation method. The results of the study indicate that attributes such as Progress, % Course completed, Assignment 1, and Assignment 2 have a significant influence on the decline in student grades in the Indonesian Language course. Furthermore, the use of ANOVA feature selection in the Random Forest Classifier algorithm improved its performance, with the accuracy increasing from 85.65% to 87.31%.

Keywords: Nilai Mahasiswa, Data Mining, Random Forest, 5-Fold Cross Validation, ANOVA.

Abstrak-Dalam lembaga pendidikan, seperti universitas, penting untuk memperhatikan kinerja mahasiswa agar mereka dapat menyelesaikan studi mereka tepat waktu. Namun, masih terdapat masalah di mana sebagian mahasiswa tidak mampu menyelesaikan studi mereka dengan tepat waktu, bahkan ada yang memutuskan untuk berhenti atau tidak lagi aktif sebagai mahasiswa. Hal ini diperlihatkan oleh penurunan nilai mahasiswa pada mata kuliah Bahasa Indonesia dari angkatan 2020 hingga 2021 di UMKT. Untuk mengatasi masalah ini, diperlukan metode yang dapat mengukur kinerja mahasiswa dalam menyelesaikan studi mereka. Penelitian ini bertujuan untuk mengetahui atribut-atribut yang berpengaruh terhadap penurunan nilai mahasiswa dalam mata kuliah Bahasa Indonesia, serta meningkatkan akurasi algoritma Random Forest Classifier dengan menggunakan seleksi fitur ANOVA. Data yang digunakan dalam penelitian ini adalah data mahasiswa UMKT yang mengambil mata kuliah Bahasa Indonesia pada periode 2020/2021 hingga 2021/2022. Data diperoleh dari bagian administrasi akademik (BAA) UMKT dan unit Mata Kuliah Dasar Umum (MKDU) UMKT, dengan jumlah data sebanyak 1028. Proses analisis data dilakukan menggunakan metode 5-Fold Cross Validation. Hasil penelitian menunjukkan bahwa atribut-atribut seperti Progress, % Course completed, Tugas 1, dan Tugas 2 memiliki pengaruh signifikan terhadap penurunan nilai mahasiswa dalam mata kuliah Bahasa Indonesia. Selain itu, penggunaan seleksi fitur ANOVA pada algoritma Random Forest Classifier mampu meningkatkan kinerja algoritma tersebut, dengan akurasi meningkat dari 85.65% menjadi 87.31%.

Kata Kunci: Nilai Mahasiswa, Data Mining, Random Forest, 5-Fold Cross Validation, ANOVA.

1. Pendahuluan

Saat ini pendidikan merupakan hal yang penting untuk setiap manusia karena dengan pendidikan, kita bisa mendapatkan ilmu yang berguna untuk menambah wawasan kita [1]. Selain itu pendidikan ini juga berkontribusi besar dalam menghasilkan lulusan yang berkompeten agar siap bersaing dalam dunia kerja. Pendidikan umumnya terdiri dari pendidikan usia dini, dasar, menengah, dan tinggi. Perkembangan teknologi

juga berkontribusi dalam dunia pendidikan salah satunya seperti pembelajaran online, pelaksanaan ujian nasional berbasis komputer, dan sistem informasi akademik.

Dalam sebuah lembaga pendidikan seperti universitas, kinerja mahasiswa perlu diperhatikan agar mahasiswa tersebut dapat menyelesaikan studinya dengan tepat waktu, walaupun begitu, masih terdapat mahasiswa yang tidak dapat menyelesaikan studinya secara tepat waktu dan bahkan ada yang berhenti dari studinya atau tidak aktif menjadi mahasiswa lagi [2]. Hal ini dibuktikan dengan menurunnya nilai mahasiswa dari angkatan 2020 hingga 2021 pada Mata Kuliah Dasar Umum (MKDU) yaitu Bahasa Indonesia di UMKT. Nilai rata-rata mahasiswa mengalami penurunan pada angkatan tahun 2017 yang awalnya 75,84% lalu pada angkatan 2018 menurun menjadi 74,59%, setelah itu mengalami peningkatan di angkatan 2019 sebesar 78,02% dan pada angkatan 2020 meningkat lagi menjadi 82,46%, lalu menurun drastis pada angkatan 2021 yaitu sebesar 71.58%. Untuk mengatasi masalah tersebut dibutuhkannya cara yang dapat mengukur kinerja mahasiswa dalam menyelesaikan studinya.

Data mining merupakan proses memperoleh data tersembunyi dari suatu dataset yang besar. Teknik *data mining* digunakan secara luas dalam berbagai bidang. Dengan teknik *data mining,* kita dapat memprediksi, mengklasifikasi, memfilter, dan mengelompokkan data [3]. Beberapa teknik *data mining* yang sering digunakan adalah asosasi, klasifikasi, klastering, dan prediksi [4]. Penelitian ini menggunakan teknik klasifikasi yang merupakan suatu teknik pembelajaran untuk memprediksi dari beberapa atribut yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksi dari objek kelasnya yang tidak diketahui [5]. Salah satu teknik *data mining* yang digunakan dalam klasifikasi adalah algoritma *Random Forest.*

Dalam Penelitian ini, peneliti menggunakan algoritma *Random Forest* dengan seleksi fitur *ANOVA* karena mengambil saran dari penelitian sebelumnya yang menggunakan Algoritma *Random Forest Classifier* dengan *Correlation Based Feature Selection* (*CFS*). Dalam penelitian tersebut peneliti terdahulu menyarankan menggunakan seleksi fitur yang lain seperti *Chi-Square*, *Information Gain*, *ANOVA*, dan lain-lain agar diharapkan mendapatkan hasil akurasi yang lebih baik lagi [6]. Tujuan dari penelitian ini untuk mengetahui atribut-atribut yang memiliki pengaruh dalam menurunnya nilai mahasiswa pada mata kuliah Bahasa Indonesia dan meningkatkan akurasi dari algoritma *Random Forest Classifier* dengan menggunakan seleksi fitur *Anova.*

2. Metodologi

A. Objek Penelitian

Penelitian ini menjadikan nilai mahasiswa pada mata kuliah dasar umum (MKDU) Bahasa Indonesia di UMKT dari tahun angkatan 2020/2021 hingga 2021/2022 sebagai objek penelitian. Alasan peneliti memilih mata kuliah Bahasa Indonesia karena mata kuliah tersebut mengalami penurunan nilai yang cukup signifikan. Penelitian ini dilakukan di Universitas Muhammadiyah Kalimantan Timur Jl. Ir. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur.

B. Metode Pengumpulan Data

Terdapat 2 teknik pengumpulan data yang digunakan dalam penelitian yaitu:

1. Observasi

Teknik ini melakukan pengamatan langsung pada struktur data yang akan diambil dari riwayat perkuliahan di MKDU dan BAA UMKT. Data mahasiswa yang akan diambil dari MKDU dan BAA adalah data mata kuliah Bahasa Indonesia pada tahun angkatan 2020 sampai 2021. Pada tahap ini menghasilkan kumpulan dataset riwayat perkuliahan mahasiswa UMKT pada mata kuliah Bahasa Indonesia dari MKDU dan Unit BAA UMKT.

2. Studi Dokumen

Studi dokumen melakukan pengumpulan data pada dokumen-dokumen yang berkaitan dengan penelitian yang sedang dilaksanakan. Hal ini dilakukan dengan mempelajari jurnal, buku, dan referensi lainnya dalam mendukung penelitian ini.

C. Teknik Analisis Data

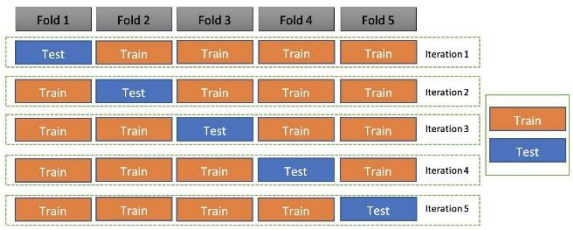
Pada penelitian ini teknik analisis data yang digunakan adalah CRISP-DM yang memiliki beberapa tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, dan *Evaluation* [7].

D. Metode

Dalam penelitian ini metode pembagian data yang digunakan adalah *Cross Validation* dengan skema *5-Fold*, lalu untuk pemodelan menggunakan algoritma *Random Forest* untuk mengetahui tingkat akurasi dari algoritma, setelah itu metode seleksi fitur *ANOVA* digunakan untuk mengetahui atribut yang berpengaruh dan meningkatkan akurasi dari algoritma *Random Forest*. Masing-masing metode dijelaskan sebagai berikut.

1. K-Fold Cross Validation

Cross Validation merupakan metode yang dapat digunakan untuk mengevaluasi kinerja dari suatu model atau algoritma yang dimana memisahkan data menjadi 2 subset yaitu data latih dan data uji [8]. Subset pembelajaran melatih model atau algoritma dan subset validasi, memvalidasinya. Lalu pemilihan jenis CV berdasarkan ukuran dari datasetnya. K-fold adalah sebuah metode yang memecah dataset menjadi dua bagian yaitu data latih dan data uji sebanyak k kelompok yang dimana jumlah data latih dan data uji pada tiap kelompok sama. Gambar proses pembagian data *K-Fold Cross Validation* dapat dilihat pada Gambar 1:



Gambar 1. K-Fold Cross Validation [9]

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i2.3187

2. Random Forest

Random Forest merupakan metode hasil pengembangan dari algoritma *Classification and Regression Tree* (CART) yang pada penerapannya menggunakan metode *bootstrap aggregating* (*bagging*) dan *random feature* selection [10]. *Random Forest* adalah salah satu algoritma *machine learning* yang merupakan pengembangan dari algoritma *Decision Tree*, *RF* bisa dilihat sebagai gabungan beberapa buah *decision tree* [11]. Rumus dari *RF* yang terdiri dari N *trees* dinyatakan sebagai berikut [12] yaitu:

$$l(y) = argmax_c \left(\sum_{n=1}^{N} I_{hn}(y) = c \right) \qquad (1)$$

Dimana variabel I adalah fungsi indikator dan *hn* merupakan *tree* ke-n dari *RF*. Metode *CART* yang digunakan untuk membangun pohon pada algoritma *Random Forest Classifier* menggunakan aturan *Gini Impurity* untuk menentukan pecahan dari pohon keputusan [10]. Perhitungan dimulai dengan penentuan nilai *Gini Index* untuk menentukan distribusi probabilitas atribut terhadap kelas target dan dilanjutkan pada perhitungan *Gini Impurity*. Berikut adalah rumus perhitungan *Gini* [13].

$$Gini = \sum_{i=1}^{n} P_{i^2} \qquad (2)$$

Dimana:

n = Merupakan jumlah kelas target

I = Merupakan kelas target

p = Merupakan rasio kelas target

Adapun perhitungan *Gini Impurity* adalah sebagai berikut:

$$Gini\ impurity = 1 - \sum_{i=1}^{n} P_{i^2} \qquad (3)$$

3. Analysis of Variance (ANOVA)

Analysis of Variance merupakan teknik standar untuk mengukur signifikansi statistik dari suatu set variabel independen dalam memprediksi variabel dependen [14]. *ANOVA* digunakan untuk memangkas fitur tanpa memengaruhi keakuratan prediktor [15]. Dalam penyelesaian menggunakan *ANOVA*, berikut langkah-langkahnya:

a. Semua fitur dipilih dari dataset.
b. Fungsi fitur target dari scikit-learn dihitung menggunakan *ANOVA F-Score* untuk setiap fitur. Dibawah ini merupakan rumus untuk menghitung *ANOVA*.

$$F = \frac{variance\ between\ groups}{variance\ within\ groups}$$

$$Variance\ between\ groups = \frac{\sum_{i}^{n} n_i (\bar{Y}_{i-} \bar{Y})^2}{(k-1)}$$

$$Variance\ within\ groups = \frac{\sum_{i=1}^{k} \sum_{j=1}^{ni} (Y_{ij-} \bar{Y}_i)^2}{(n-k)}$$

c. Hasil dari pengujian digunakan untuk melakukan pemilihan fitur yang memungkinkan pembuangan fitur yang tidak terkait dengan variabel target. Fitur yang memiliki pengaruh paling tinggi dengan varian terendah dipilih dalam eksperimen ini dan diuji dengan *SelectKBest(); K* mewakili jumlah fitur yang ada untuk dataset akhir.
d. Jumlah fitur dengan peringkat tertinggi digunakan untuk membuat berbagai subset fitur.

3. Hasil dan Pembahasan

1. Business Understanding (Pemahaman Bisnis)

   Penelitian ini dilakukan untuk mengatasi masalah nilai mahasiswa pada mata kuliah Bahasa Indonesia yang mengalami penurunan dari 82.46% pada angkatan 2020 menjadi 71.58% pada angkatan 2021. Pada penelitian ini dilakukan pendekatan dengan data mining menggunakan algoritma Random Forest dan seleksi fitur ANOVA untuk mengetahui atribut yang berpengaruh dalam menurunnya nilai mahasiswa dan meningkatkan akurasi dari algoritma.

2. Data Understanding (Pemahaman Data)

Penelitian ini menggunakan data nilai mahasiswa yang diperoleh dari BAA UMKT dan MKDU. Mata kuliah yang digunakan sebagai data penelitian adalah Bahasa Indonesia pada tahun akademik 2020/2021 dan 2021/2022 semester genap dari seluruh program studi. Jumlah data yang diperoleh dari BAA UMKT dan MKDU sebanyak 1028 mahasiswa. Data nilai mahasiswa yang terdapat pada bagian administrasi akademik UMKT memiliki 6 atribut yang terdiri dari NIM, Nama, Jenis Kelamin, Nilai Akhir, Bobot, dan Simbol. Data yang di dapat dari BAA dapat dilihat pada tabel 4.1. Data mahasiswa yang terdapat pada MKDU memiliki 16 atribut yaitu, *Profile name, Learner name*, *Learner email*, *Enrolment ID*, *Institution Membership ID*, *Enrolment date*, *Completion date*, *Time spent on course, Progress*, *% Course completed*, *Certificate ID*, *Comments*, Tugas 1, Tugas 2, Tugas 3, dan UTS. Data yang di dapat dari MKDU dapat dilihat pada tabel 4.2.

Tabel 1 Atribut Data BAA UMKT

No	Atribut	Keterangan
1	NIM	Nomor Induk Mahasiswa
2	Nama	Nama Mahasiswa

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i2.3187

3	Jenis Kelamin	Jenis Kelamin Mahasiswa
4	Nilai Akhir	Nilai Akhir Mahasiswa
5	Bobot	Bobot Nilai Akhir
6	Simbol	Simbol Nilai Akhir

Tabel 2 Atribut Data MKDU

No	Atribut	Keterangan
1	*Profile name*	Id mahasiswa pada sistem *OpenLearning*
2	Learner name	Nama Mahasiswa
3	Learner email	Email Mahasiswa
4	Enrolment ID	Id pendaftaran *OpenLearning*
5	Institution Membership ID	Id anggota institusi
6	Enrolment Date	Tanggal daftar
7	Completion Date	Tanggal menyelesaikan mata kuliah
8	Time spent on course	Lama waktu mahasiswa berada di mata kuliah
9	Progress	Persentase kemajuan mahasiswa
10	% Course Completed	Persentase kemajuan mahasiswa menyelesaikan mata kuliah
11	Certificate ID	Id sertifikat
12	Comments	Banyaknya komentar mahasiswa selama perkuliahan
13	Tugas 1	Nilai Tugas 1
14	Tugas 2	Nilai Tugas 2
15	Tugas 3	Nilai Tugas 3
16	UTS	Nilai Ujian Tengah Semester

3. Data Preparation (Persiapan Data)

a. Seleksi dan Integrasi Data

Proses seleksi dan integrasi dilakukan dengan menyeleksi atribut yang tidak dibutuhkan. Dari data yang diperoleh dari BAA UMKT dan bagian MKDU, hanya beberapa atribut yang akan digunakan. Pada data BAA UMKT atribut yang dihapus adalah NIM, Nama, Nilai Akhir, dan Bobot. Sedangkan pada data bagian MKDU atribut yang dihapus adalah *Profile name*, *Learner name*, *Learner email*, *Enrolment ID*, *Institution Membership ID*, *Enrolment date*, *Completion date*, *Certificate ID*. Setelah atribut yang tidak dibutuhkan dihapus, selanjutnya menggabungkan data atribut yang tersisa dari BAA dan MKDU. Berikut adalah hasil dari seleksi dan integrasi.

Tabel 3 Hasil Seleksi dan Integrasi Data

No	Jenis Kelamin	Time spent on course	Progress	% Course completed	Comments	Tugas 1	Tugas 2	Tugas 3	UTS	Simbol
1	Perempuan	10 Hrs 43 Mins	99.37	99.37	61	88	80	80	77	B
2	Laki-laki	10 Hrs 45 Mins	100	100	139	82	80	82	80	A
3	Laki-laki	4 Hrs 57 Mins	100	100	38	84	80	80	77	A
:	:	:	:	:	:	:	:	:	:	:
1028	Perempuan	10 Hrs 57 Mins	96.86	96.86	14	80	80	80	90	AB

b. Pembersihan Data

Pembersihan data dilakukan dengan mencari nilai kosong yang ada pada dataset. Setelah dicari, terdapat satu nilai kosong pada kolom atribut Tugas 3. Untuk mengatasi hal tersebut dilakukan dengan mengisi nilai yang kosong dengan nilai rata-rata pada atribut Tugas 3, sehingga nilai tersebut dapat digunakan sebagaimana mestinya.

c. Transformasi Data

Pada tahap ini, data yang telah di seleksi dan integrasi sebelumnya dilakukan proses pengubahan format, struktur, dan nilai pada beberapa atribut yang akan digunakan sebelum masuk pada tahap pemodelan. Atribut yang akan di transformasi pada data ini adalah Jenis Kelamin, *Time spent on course*, dan Simbol. Atribut jenis kelamin memiliki 2 nilai dalam *record*-nya yaitu, Laki-laki dan Perempuan. Nilai tersebut akan diubah ke dalam bentuk numerik yang mana nilai "laki-laki" diubah menjadi "1" dan "perempuan" diubah menjadi "2".

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i2.3187

Tabel 4 Transformasi Atribut Jenis Kelamin

No	Jenis Kelamin sebelum di transformasi	Jenis Kelamin setelah di transformasi
1	Perempuan	2
2	Laki-laki	1
3	Laki-laki	1
:	:	:
1028	Perempuan	2

Pada atribut *Time spent on course*, data mahasiswa yang berada pada mata kuliah Bahasa Indonesia memiliki dua jenis tipe data dalam *record*-nya, yaitu *integer* dan *string*. Hal ini dapat menyebabkan pemodelan tidak dapat memproses data yang akan di klasifikasi. Untuk mengatasi hal tersebut, nilai atribut ini diubah menjadi hitungan menit dari yang sebelumnya berupa akumulasi dari lamanya mahasiswa berada di *course*.

Tabel 5 Transformasi Atribut Time spent on course

No	Time spent on course sebelum di transformasi	Time spent on course setelah di transformasi
1	10 Hrs 43 Mins	643
2	10 Hrs 45 Mins	645
3	4 Hrs 57 Mins	297
:	:	:
1028	10 Hrs 57 Mins	657

Selanjutnya atribut simbol di transformasi ke dalam nilai "LULUS" dan "TIDAK LULUS" dari yang sebelumnya berupa huruf A, AB, B, BC, C, D, dan E. Hal ini dilakukan karena atribut simbol dijadikan sebagai kelas target. Dalam penilaian MKDU, syarat untuk lulus suatu matkul adalah memiliki nilai minimal B. Dibawah dari B maka tidak lulus dan harus mengulang, maka dari itu dalam transformasi atribut simbol, nilai A, AB, B di transformasi menjadi LULUS dan nilai BC, C, D, E di transformasi menjadi TIDAK LULUS.

Tabel 6 Transformasi Atribut Simbol Tahap 1

Simbol sebelum di transformasi	Simbol setelah di transformasi
A	LULUS
AB	LULUS
B	LULUS
BC	TIDAK LULUS
C	TIDAK LULUS
D	TIDAK LULUS
E	TIDAK LULUS
T	TIDAK LULUS

Setelah itu data di transformasi kembali menjadi 1 dan 0, yang mana LULUS diubah menjadi 1 dan TIDAK LULUS diubah menjadi 0. Hal ini dilakukan agar data dapat diolah di dalam *python*.

Tabel 7 Transformasi Atribut Simbol Tahap 2

Simbol sebelum di transformasi	Simbol setelah di transformasi
LULUS	1
LULUS	1
LULUS	1
TIDAK LULUS	0
TIDAK LULUS	0
TIDAK LULUS	0
TIDAK LULUS	0
TIDAK LULUS	0

d. Reduksi Data

Reduksi data dilakukan untuk menyeimbangkan jumlah data terhadap kelas target. Setelah data melewati tahap seleksi, integrasi, dan transformasi kelas target memiliki data yang tidak seimbang. Cara melakukan reduksi data adalah dengan menggunakan teknik *undersampling* yaitu mengurangi jumlah data mayoritas agar seimbang dengan data minoritas. Data minoritas memiliki 59 data sedangkan data mayoritas memiliki 969 data, maka dari itu jumlah data mayoritas diseimbangkan dengan data minoritas menjadi 59. Setelah melalui proses reduksi data, jumlah data yang siap digunakan sebanyak 118 data.

Tabel 8 Hasil Akhir dari Data Preparation

No	Jenis Kelamin	Time spent on course	Progress	% Course completed	Comments	Tugas 1	Tugas 2	Tugas 3	UTS	Simbol
1	1	490	100	100	23	90	85	80	46	1
2	1	338	100	100	24	64	73	73	65	1
3	2	4080	100	100	24	76	78	78	61	1
:	:	:	:	:	:	:	:	:	:	:
118	1	236	62,89	62,89	10	78	67	78	85	0

4. Modeling (Pemodelan)

a. Implementasi Random Forest pada Python

Proses *import* dataset dimulai dengan mengimport *library pandas* dan menginisialisasikannya sebagai variabel "pd". Selanjutnya, variabel "data" dibuat untuk menampung dataset menggunakan fungsi bantuan dari *pandas* untuk membaca dataset dalam format csv.

```
import pandas as pd
data = pd.read_csv('balanced_bindo.csv')
```

Code 1 Import Dataset

Tahapan pengujian pemodelan akan melibatkan pembagian data menggunakan 5-fold *Cross validation* lalu memisahkan atribut fitur dengan atribut target yaitu Simbol. Proses pembagian data dilakukan menggunakan fungsi KFold dari library scikit-learn. Selama proses pembagian data, dilakukan pengacakan data menggunakan hyperparameter random_state. Pemodelan *Random Forest* dimulai dengan mengimport library *RandomForestClassifier* untuk membuat modelnya, mengimport library *metrics, accuracy_score, confusion_matrix* untuk melihat akurasi dari algoritma dan menampilkan tabel *confusion matrix,* dan terakhir menjalankan program untuk mengetahui akurasi yang di dapatkan.

```
from sklearn import metrics
from sklearn.model_selection import KFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,
confusion_matrix

# Pemisahan atribut dan label
X = data.drop('Simbol', axis=1)  # Data atribut
y = data['Simbol']               # Data label

# Membuat objek/insialisasi Random Forest Classifier
rf_classifier =
RandomForestClassifier(n_estimators=30)

# Inisialisasi 5-fold cross-validation
kfold = KFold(n_splits=5, shuffle=True,
random_state=42)

fold = 1
total_accuracy = 0
num_folds = 0
total_cm = None

for train_index, test_index in kfold.split(X):
    print(f"Fold {fold}:")
    fold += 1

    # Split data into train and test set
    X_train, X_test = X.iloc[train_index],
X.iloc[test_index]
    y_train, y_test = y.iloc[train_index],
y.iloc[test_index]

    # Train Random Forest model
    rf_classifier.fit(X_train, y_train)

    y_pred = rf_classifier.predict(X_test)

    # Evaluate model on test set
    accuracy = metrics.accuracy_score(y_test,
y_pred)
    total_accuracy += accuracy
    num_folds += 1
    print(f"Akurasi: {accuracy}\n")

    # Calculate confusion matrix
    cm = confusion_matrix(y_test, y_pred)
    if total_cm is None:
        total_cm = cm
    else:
        total_cm += cm

average_accuracy = total_accuracy / num_folds
print(f"Rata-rata akurasi: {average_accuracy}")

# Menghitung dan mencetak confusion matrix
print("Confusion Matrix:")
print(total_cm)
```

Code 2 Pemodelan Random Forest

Hasil yang didapat yaitu berupa akurasi tiap *fold*, rata-rata akurasi, dan *confusion matrix*. Disini dapat diperhatikan rata-rata dari akurasi yang keluar yaitu 85.65%, hasilnya mendekati dengan perhitungan *confusion matrix* yaitu 85.59%. Selanjutnya membandingkan akurasi setelah menggunakan *ANOVA*.

$$Accuracy = \frac{49 + 52}{49 + 52 + 10 + 7} = \frac{101}{118} = 0.8559322 = 85.59\%$$

b. Implementasi Random Forest dengan ANOVA pada Python

Tahap pertama adalah memanggil fungsi "SelectKBest dan f_classif" dari *library scikit-learn* untuk menggunakan seleksi fitur *ANOVA*, lalu menjalankan program untuk mengetahui atribut yang berpengaruh dalam dataset.

```
from sklearn.feature_selection import SelectKBest,
f_classif

# Membuat objek SelectKBest dengan ANOVA dan k=4
selector = SelectKBest(score_func=f_classif, k=4)

# Melakukan seleksi fitur dengan ANOVA
X_selected = selector.fit_transform(X, y)

# Mendapatkan masker fitur terpilih
feature_mask = selector.get_support()

# Mendapatkan daftar fitur terpilih
selected_features = X.columns[feature_mask]

# Menampilkan daftar fitur terpilih
print("Fitur terpilih:")
for feature in selected_features:
    print(feature)
```

Code 3 Seleksi Fitur

Setelah dijalankan, hasilnya menyatakan bahwa atribut "*Progress*", *"% Course completed"*, "Tugas 1", dan "Tugas 2" merupakan atribut yang berpengaruh dalam dataset, selanjutnya menghapus atribut yang kurang berpengaruh dalam dataset tersebut yaitu, "Jenis kelamin", "*Time spent on course*", *"Comments"*, "Tugas 3", dan "UTS".

```
data = data.drop(['Jenis Kelamin','Time spent on
course', 'Comments', 'Tugas 3', 'UTS'], axis=1)
```

Code 4 Menghapus atribut yang kurang berpengaruh

Setelah atribut tersebut dihapus, selanjutnya menjalankan kembali model algoritma *Random Forest* sebelumnya untuk mengetahui apakah akurasi telah meningkat atau tidak. Setelah dijalankan, hasilnya menyatakan terdapat peningkatan akurasi pada algoritma

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i2.3187

Random Forest dengan akurasi rata-rata sebesar 87.31%, hasilnya mendekati dengan perhitungan *confusion matrix* yaitu 87.28%.

$$Accuracy = \frac{50+53}{50+53+9+6} = \frac{103}{118} = 0.87288135 = 87.28\%$$

Tabel 9 Hasil Akurasi tiap fold

Fold	Akurasi (Sebelum Seleksi Fitur)	Akurasi (Setelah Seleksi Fitur)	Status
1	91%	87%	Turun
2	83%	83%	Tetap
3	75%	87%	Naik
4	91%	91%	Tetap
5	86%	86%	Tetap

Tabel diatas menampilkan hasil akurasi yang ada pada tiap fold sebelum dan sesudah seleksi fitu;, pada fold 1 mengalami penurunan dari 91% ke 87%, lalu pada fold 2, fold 4, dan fold 5 tidak terjadi perubahan akurasi, dan terakhir pada fold 3 mengalami peningkatan dari 75% ke 87%.

Table 10 Hasil Akurasi rata-rata dari fold

Akurasi rata-rata (Sebelum Seleksi Fitur)	Akurasi rata-rata (Setelah Seleksi Fitur)	Status
85%	87%	Naik

Tabel diatas menunjukkan hasil rata-rata akurasi pada seluruh fold sebelum dan sesudah seleksi fitur, terdapat peningkatan akurasi 2% setelah seleksi fitur.

5. Evaluation (Evaluasi)

Dari pemodelan *random forest* yang dilakukan dengan pembagian data *5-fold cross validation*, didapatkan hasil akurasi rata-rata dari seluruh *fold* pada algoritma *random forest* sebesar 85.65%, dari hasil akurasi tersebut masih bisa ditingkatkan lagi menggunakan seleksi fitur *ANOVA*. Selanjutnya menggunakan seleksi fitur *ANOVA* untuk mengetahui atribut yang berpengaruh dan meningkatkan akurasi dari algoritma *Random Forest*. Hasilnya atribut yang berpengaruh adalah *Progress, % Course completed,* Tugas 1, dan Tugas 2. Selain itu akurasi rata-rata meningkat menjadi 87.31%, mengalami peningkatan sekitar 2%. Hal ini membuktikan bahwa dengan seleksi fitur *ANOVA* dapat mengetahui atribut yang berpengaruh pada dataset dan meningkatkan akurasi dari algoritma *Random Forest.*

4. Kesimpulan

Dalam penelitian ini dapat disimpulkan bahwa Data yang digunakan merupakan data mata kuliah Bahasa Indonesia tahun angkatan 2020/2021 dan 2021/2022 dengan atribut Jenis Kelamin, *Time spent on course, Progress, % Course completed, Comments,* Tugas 1, Tugas 2, Tugas 3, UTS, dan Simbol sebagai atribut target. Dengan menggunakan seleksi fitur *ANOVA* pada Algoritma *Random Forest Classifier*, diketahui atribut *Progress, % Course completed,* Tugas 1, dan Tugas 2 merupakan atribut yang berpengaruh dalam menurunnya nilai mahasiswa pada mata kuliah Bahasa Indonesia. Dalam pemodelannya, dilakukan pembagian data dengan menggunakan 5-fold Cross Validation. Hasilnya menunjukkan bahwa algoritma Random Forest Classifier mengalami peningkatan kinerja dalam mengklasifikasi nilai mahasiswa sebesar 2%. Sebelumnya, tingkat akurasi algoritma tersebut adalah 85.65%, namun setelah menggunakan seleksi fitur ANOVA, tingkat akurasinya meningkat menjadi 87.31%.

5. Saran

Untuk meningkatkan akurasi, bisa menambahkan atribut lain seperti kehadiran mahasiswa, dan nilai ujian akhir semester (UAS). Disarankan untuk penelitian selanjutnya menggunakan seleksi fitur yang lain terhadap algoritma *Random Forest* seperti *Correlation Pearson, Adaboost*, *Particle Swarm Optimization*, dan lainnya untuk meningkatkan akurasi.

6. Daftar Pustaka

[1] Makkawaru, M. (2019). Pentingnya Pendidikan Bagi Kehidupan dan Pendidikan Karakter dalam Dunia Pendidikan. *Jurnal Konsepsi*, *8*(3), 116–119.

[2] Gunawan Sudarsono, B., & Ulan Bani, A. (2020). Prediksi Mahasiswa Berpotensi Berhenti Kuliah Secara Sepihak Menggunakan Data Mining Algoritma C4.5. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *4*(2), 359–367.

[3] Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*(April), 100178. https://doi.org/10.1016/j.imu.2019.100178.

[4] Osman, A.S. (2019). Data mining techniques: Review. *International Journal of Data Science Research, 2*(1), 1-4.

[5] Ardiyansyah, Rahayuningsih, P. A., & Maulana, R. (2018). Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner. *Jurnal Khatulistiwa Informatika*, *VI*(1), 20-28.

[6] Priantama, Y., & Yoga Siswa, T. A. (2022). Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(2), 251. https://doi.org/10.26798/jiko.v6i2.651.

[7] Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403-408. Https://doi.org/10.1016/j.procir.2019.02.106.

[8] Daqiqil, I. (2021). *MACHINE LEARNING: Teori, Studi Kasus dan Implementasi Menggunakan Python*. UR PRESS. https://www.researchgate.net/publication/353338909_Machine_Learning_Teori_Studi_Kasus_dan_Implementasi_Menggunakan_Python.

[9] Gopal Krishna Ranjan. (2021, july 12) *Introduction to k-fold Cross Validation in Python*. SQLRelease. https://sqlrelease.com/introduction-to-k-fold-cross-validation-in-python.
[10] Breiman, L. (2001). Random Forest. Machine Learning, 45(1), 5-32. Springer.
[11] Primartha, R. (2021). *Algoritma Machine Learning*. Penerbit Informatika.
[12] Liparas, D., Hacohen-kerner, Y., & Moumtzidou, A. (2014). News Articles Classification Using Random Forests and Weighted Multimodal Features. *Springer International Publishing Switzerland*, 63–75.
[13] Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with Gini index. Advances in Mathematics: Scientific Journal, 9(10), 8237- 8247.
[14] Wenda, A. (2022). Support Vector Machine Untuk Pengenalan Bentuk Manusia Menggunakan Kumpulan Fitur Yang Dioptimalkan. *JST (Jurnal Sains Dan Teknologi)*, *11*(1), 77–84. https://doi.org/10.23887/jstundiksha.v11i1.4443.
[15] Chen, Z., Jiao, S., Zhao, D., Zou, Q., Xu, L., Zhang, L., & Su, X. (2022). The Characterization of Structure and Prediction for Aquaporin in Tumour Progression by Machine Learning. *Frontiers in Cell and Developmental Biology*, *10*(February), 1–11. https://doi.org/10.3389/fcell.2022.845622.