JURNAL INFORMATIKA, Vol. 9 No. 2 Oktober 2022, Halaman 131~139
ISSN: 2355-6579 | E-ISSN**:** 2528-2247

Klasifikasi Pemilihan Jurusan Sekolah Menengah Kejuruan Menggunakan Gradient Boosting Classifier

Hadi Priyono[1], Retno Sari[2], Tati Mardiana[3]

[1,2] Fakultas Teknologi Informasi, Universitas Nusa Mandiri
Jl. Jatiwaringin No. 2, Cipinang Melayu, Makasar, Jakarta Timur

e-mail: [1]hadigaz@gmail.com, [2]retno.rnr@nusamandiri.ac.id, [3]tati.ttm@nusamandiri.ac.id

Informasi Artikel	Diterima: 27-03-2022	Direvisi: 15-09-2022	Disetujui: 19-09-2022

Abstrak

Pemilihan jurusan merupakan faktor penting bagi calon siswa yang akan melanjutkan pendidikan di SMK. Siswa cenderung mengikuti pilihan orang tua atau temannya tanpa mempertimbangkan kurikulum sesuai minat dan kemampuannya. Akibatnya banyak siswa yang mengalami kesulitan mengikuti pelajaran, dan prestasi belajarnya menurun. Model RIASEC merupakan salah satu metode pendeteksian minat yang digunakan untuk mengetahui tipe kepribadian siswa. Penelitian ini bertujuan membangun model untuk memprediksi pilihan jurusan di SMK Yadika 12 Depok. Penelitian ini membandingkan lima pengklasifikasi pada kumpulan data pemilihan jurusan di sekolah kejuruan. Proses selanjutnya melakukan tuning hyperparameter menggunakan GridsearchCV untuk mendapatkan parameter yang paling berpengaruh dari algoritma klasifikasi yang dipilih. Algoritma yang diimplementasikan adalah Multinomial Naive Bayes, Gaussian Naive Bayes, Bernoulli Naive Bayes, Gradient Boosting Classifier, Decision Tree Classifier, K Neighbors Classifier, dan Logistic Regression. Hasil pengujian menunjukkan bahwa Gradient Boosting Classifier dengan Hyperparameter Tuning menggunakan GridSearchCV memperoleh akurasi 72% dan class recall mencapai 76%.

Kata Kunci: jurusan, riasec, gradient boosting classifier

Abstract

The selection of the majors remains a crucial factor for prospective students who will pursue their education at SMK. However, students tend to follow the choices of their parents or friends. They are not considering the curriculum according to their interests and abilities. As a result, many students have difficulties following the lesson, and their academic achievement decreases. The RIASEC model is one of the interest detection methods used to determine the student's personality type. This study aims to develop a model to predict the choice of majors at SMK Yadika 12 Depok. We compared five classifiers on the major's selection data sets at vocational schools. In addition, we performed hyperparameter tuning using GridsearchCV to obtain the most influential parameters from the selected classification algorithm. The algorithms implemented are Multinomial Naive Bayes, Gaussian Naive Bayes, Bernoulli Naive Bayes, Gradient Boosting Classifier, Decision Tree Classifier, K Neighbors Classifier, and Logistic Regression. The test results show that the Gradient Boosting Classifier with Hyperparameter Tuning using GridSearchCV maintains an accuracy of 72% and class recall reaches 76%.

Keywords*:* majors, RIASEC, Gradient Boosting Classifier

1. Pendahuluan

Masyarakat saat ini lebih memilih sekolah menengah kejuruan untuk melanjutkan pendidikannya. Karena sekolah kejuruan membekali siswa dengan keterampilan hidup yang sesuai dengan standar industri, mereka akan siap bekerja setelah lulus (Wardani et al., 2018). Pada 2019, Kementerian Pendidikan dan Kebudayaan mencatat jumlah siswa SMK naik 5 juta. Tahun berikutnya, jumlah siswa SMK mencapai 5,2 juta. Data ini menunjukkan bahwa lulusan SMP memiliki minat yang signifikan terhadap sekolah kejuruan (Badan Pusat Statistik, 2020).

Pendidikan Kejuruan memiliki kekhususan atau ciri tertentu yang membedakan dengan subsistem pendidikan lainnya, yaitu ditandai dengan memiliki aspek- aspek lain yang berkaitan dengan perencanaan kurikulum yakni; tujuan peminatan, dasar kebenaran/ justifikasi, fokus, standar ketuntasan di sekolah, standar ketuntasan di luar sekolah, hubungan kerja sama dengan masyarakat, keterlibatan pemerintah pusat, kepekaan, logistik dan pengeluaran (Suyitno, 2020).

Pemerintah mempersiapkan sekolah kejuruan untuk mengatasi masalah pengangguran, meningkatkan taraf hidup dan mendorong masyarakat untuk bercita-cita untuk karir yang lebih menonjol. SMK saat ini memiliki 146 program keahlian yang sesuai dengan harapan masyarakat, pasar, dan dunia kerja saat ini (Kementerian Pendidikan dan Kebudayaan, 2018). Oleh karena itu, mahasiswa dapat memilih program keahlian sesuai dengan tujuan karir masa depan mereka.

Pemilihan jurusan menjadi faktor penting bagi calon siswa yang akan melanjutkan pendidikan di SMK. Pemilihan jurusan yang tepat akan mendorong siswa untuk menerima pelajaran dengan baik dan meningkatkan prestasi belajarnya. Sebaliknya, kesalahan dalam memilih jurusan menyebabkan siswa tidak bersemangat mengikuti pelajaran di kelas. Siswa yang sering tidak datang belajar, membuat kelas ribut, meninggalkan jam pelajaran, dan lain-lain menyebabkan prestasinya menurun (Rahmayu & Serli, 2018). Oleh karena itu, calon siswa SMK harus mempertimbangkan minat dan kemampuannya dalam memilih jurusan. Namun siswa cenderung mengikuti orang tua atau temannya (Fitri & Saraswati Sinta, 2021) tanpa mempertimbangkan kurikulum yang sesuai dengan minat dan kemampuannya (Rini et al., 2016). Akibatnya banyak siswa yang kesulitan mengikuti pelajaran, dan prestasi belajarnya menurun (Mahmudah & Lianawati, 2020).

Salah satu yang menentukan dalam memilih karir siswa terletak pada kepribadian siswa. Model RIASEC merupakan salah satu metode pendeteksian minat untuk menentukan tipe kepribadian seseorang. Siswa mengisi angket yang berisi ciri-ciri atau karakter dalam model RIASEC. Model ini mengelompokkan setiap sifat menurut minat dan bakatnya. Model RIASEC memetakan minat seseorang menjadi realistis, investigatif, artistik, sosial, giat, dan konvensional berdasarkan skor tertinggi dari kriteria yang dipilih. Tes hasil ini akan merekomendasikan jurusan pendidikan yang dapat dipilih siswa (Hidayat & Wahyuni, 2019). Orang-orang dengan tipe kepribadian yang sama dan mereka yang bekerja sama dalam pekerjaan yang mereka lakukan akan menciptakan dan memberikan lingkungan yang sesuai dengan tipe kepribadian (Hafsoh & Yusuf, 2021).

Penggunaan data mining telah meluas di berbagai domain. Data Mining adalah proses penggalian informasi dari kumpulan data menggunakan algoritme dan teknik dari domain statistik, pembelajaran mesin, dan sistem manajemen basis data dikenal sebagai penambangan data. Penambangan data adalah proses menganalisis kumpulan besar data yang disimpan dalam penyimpanan menggunakan teknik pengenalan pola seperti alat statistik dan matematika untuk mengidentifikasi tautan, pola, dan tren yang relevan (Sunge et al., 2019). Meskipun penggunaan data mining sudah meluas di berbagai domain, hanya sedikit peneliti yang memanfaatkannya untuk masalah pemilihan jurusan di SMK.

Prabowo dan Subiyanto (Prabowo & Subiyanto, 2017) meneliti pemilihan jurusan di SMK di 10 Batang Kabupaten Jawa Tengah. Penelitian ini menggunakan metode C.45 untuk mengklasifikasikan dataset pemilihan jurusan yang berjumlah 291 mahasiswa dengan tujuh atribut. Hasil akurasi sebesar 83,33%. Penelitian selanjutnya, Sumpena, Yuma, dan Nirat (Sumpena et al., 2018) melakukan penelitian jurusan di SMK PGRI Cibinong. Penelitian ini menggunakan metode C.45 untuk mengklasifikasikan dataset pemilihan jurusan yang berjumlah 30 mahasiswa dengan empat atribut. Hasil dikonfirmasi akurasi 90%. Selain itu, Fauziyah dan Nudin (Fauziyah & Nudin, 2021) melakukan penelitian jurusan SMKN 1 Pungging. Penelitian ini menggunakan metode Gradient Boosting Tree untuk mengklasifikasikan dataset pemilihan jurusan yang berjumlah 272 mahasiswa dengan lima atribut. Hasil mengkonfirmasi akurasi 96,34%.

Dalam penelitian ini mengusulkan model klasifikasi untuk memprediksi jurusan sekolah kejuruan sesuai dengan kemampuan dan minat siswa. Penelitian ini membandingkan lima pengklasifikasi pada kumpulan data pemilihan jurusan di sekolah kejuruan. Proses selanjutnya adalah melakukan *tuning hyperparameter* menggunakan *GridsearchCV* untuk mendapatkan parameter yang paling berpengaruh dari algoritma klasifikasi yang dipilih. Model yang diusulkan mencapai akurasi klasifikasi yang lebih baik.

2. Metode Penelitian

Jenis penelitian ini adalah penelitian eksperimen menggunakan desain pra-eksperimental dengan tipe desain One-shot Case Study. Dalam jenis penelitian eksperimen

ini hanya mempertimbangkan satu kelompok variabel terikat. Penelitian ini dilakukan selama tiga bulan, mulai Oktober hingga Desember 2021. Lokasi penelitian di SMK Yadika 12 dengan alamat Jalan Limo Raya No. 20 Depok.

Populasi dalam penelitian ini adalah siswa SMK Yadika 12 Depok dengan karakteristik yang masih aktif pada Tahun Pelajaran 2021/2022. Pada tahun ajaran 2021/2022 terdapat 365 siswa aktif di SMK Yadika 12 Depok. Penelitian ini menggunakan rumus Slovin untuk menentukan jumlah sampel.

$$n = \frac{N}{1+Ne^2} \quad (1)$$

Where:
n: Ukuran sampel
N: Ukuran populasi
e: Batas toleransi error

Batas toleransi pada penelitian ini menetapkan nilai toleransi kesalahan (e) untuk mewakili 10%. Dengan jumlah populasi 365, penelitian ini menggunakan minimal 79 sampel data. Namun, penelitian ini menggunakan semua data yang dikumpulkan sebagai sampel untuk mendapatkan hasil penelitian yang lebih valid.

Eksperimen pada penelitian ini dengan memberikan perlakuan yang memungkinkan perubahan pada data sampel untuk mendapatkan model dengan akurasi tertinggi. Tahapan dalam penelitian ini adalah sebagai berikut:

1. Pengumpulan Data
   Tahap ini mengumpulkan data identitas siswa seperti nama, jenis kelamin, nilai ujian SMP, jurusan, dan minat siswa berdasarkan tes RIASEC. Berdasarkan sumber data yang digunakan, metode pengumpulan data dalam penelitian ini adalah sebagai berikut:
   a. Pengamatan
      Tahap ini melakukan observasi langsung terhadap kegiatan pembelajaran di SMK Yadika 12 Depok.
   b. Wawancara
      Tahap ini melakukan wawancara dengan pimpinan, guru, dan siswa SMK Yadika 12 Depok untuk mendapatkan informasi lebih mendalam mengenai proses seleksi jurusan.
   c. Tinjauan Literatur
      Tahap ini mengumpulkan dan mempelajari literatur terkait masalah penentuan jurusan di SMK. Sumber literatur yang dikumpulkan adalah buku teks, karya ilmiah, dan website pendukung.
   d. Survei
      Tahap ini melakukan penyebaran kuesioner kepada siswa di SMK Yadika 12 Depok. Siswa mengisi kuesioner yang berisi data nilai ujian sekolah menengah pertama dan memilih karakter yang sesuai dengan model RIASEC.
2. Pra-Pemrosesan Data
   Tahap ini membersihkan kumpulan data untuk menghilangkan inkonsistensi, data yang tidak lengkap, atau informasi yang berlebihan.
3. Analisis Data
   Tahap ini melakukan analisis data eksplorasi untuk memahami isi data yang digunakan, mulai dari distribusi, frekuensi, dan korelasi.
4. Pemodelan
   Tahapan ini memilih metode klasifikasi terbaik kemudian membangun model klasifikasi untuk merekomendasikan pilihan jurusan di SMK Yadika 12 Depok.
5. Evaluasi Model
   Tahap ini memvalidasi model yang terbentuk. Evaluasi model ini menggunakan akurasi, class recall, dan confusion matrix.

Sumber data merupakan faktor yang sangat penting dalam suatu penelitian karena akan menentukan kualitas hasil penelitian. Oleh karena itu, sumber data menjadi pertimbangan dalam menentukan metode pengumpulan data. Sumber data yang digunakan pada penelitian ini sebagai berikut:

a. Data Primer
   Data primer merupakan data yang diperoleh dari responden secara langsung untuk menjawab pertanyaan penelitian. Data tersebut adalah kondisi kelas, motivasi siswa, nilai ujian sekolah, dan pemetaan minat siswa berdasarkan model RIASEC.
b. Data Sekunder
   Data sekunder biasanya hadir dalam berbagai bentuk terkait pemilihan jurusan di SMK dari sekolah atau institusi yang terkait dengan penelitian ini.

Instrumen yang digunakan untuk mengumpulkan data, sebagai berikut:

1. Lembar Observasi
   Lembar observasi digunakan untuk memperoleh informasi hasil observasi kondisi siswa di SMK Yadika 12 Depok selama kegiatan pembelajaran.
2. Lembar Wawancara
   Lembar wawancara digunakan untuk memperoleh informasi tentang pemilihan jurusan, motivasi belajar, dan hasil belajar siswa di SMK Yadika 12 Depok.
3. Kuesioner
   Kuesioner digunakan untuk memperoleh data nilai ujian sekolah tingkat SMP dan data minat siswa berdasarkan RIASEC Test.

Setelah pengumpulan dan pengolahan data, peneliti melakukan analisis data untuk menyimpulkan hasil penelitian. Analisis data dalam penelitian ini menggunakan bahasa pemrograman Python. Pustaka Python dasar yang digunakan untuk analisis data adalah Pandas. Metode analisis data dalam penelitian ini adalah sebagai berikut:

1. Analisis Deskriptif
   Analisis deskriptif adalah analisis untuk memperoleh karakteristik dataset jurusan di SMK Yadika 12 Depok seperti nilai mean, median, sum, Variance, Standard error, mean standard error, modus, range, minimum, maksimum, skewness, dan kurtosis.
2. Analisis Korelasi
   Analisis korelasi adalah analisis untuk mencari hubungan antar variabel numerik. Koefisien korelasi berkisar dari -1 hingga 1. Koefisien 1 mewakili korelasi positif yang terisi, koefisien -1 mewakili korelasi negatif, dan koefisien 0 mewakili koneksi non-linier.

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Penelitian ini mengumpulkan data primer dengan menyebarkan survei kepada siswa SMK Yadika 12 Depok melalui Google Form yang berisikan atribut identitas (ID), jenis kelamin (JK), nilai US bahasa Indonesia(N1), nilai US matematika (N2), nilai US IPA(N3), nilai US IPS(N4), nilai bahasa inggris (N5), jurusan dan minat siswa berdasarkan tes RIASEC. Tabel 1 menyajikan sampel data kuesioner pemilihan jurusan di SMK Yadika 12 Depok.

Tabel 1. Sampel Data Kuesioner

ID	JK	N1	N2	N3	N4	N5	Minat	Jurusan
S1	L	82	79	79	80	76	Realistic	TKJ
S2	P	80	82	83	84	82	Social	TKJ
S3	L	80	75	76	76	76	Realistic	TKJ
S4	L	87	88	85	87	85	Investigative	TKJ
S5	L	88	85	88	86	87	Realistic	TKJ
S6	P	87	88	88	89	85	Social	TMM
S7	L	86	68	65	64	61	Realistic	TMM
S8	L	70	70	85	75	70	Social	TKJ
S9	L	80	85	80	80	80	Artistic	TMM
S10	L	78	77	77	80	76	Investigative	TMM

Sumber : (Priyono et al., 2022)

3.2. Pra-Pemrosesan Data

Setelah data terkumpul kemudian data akan diolah menggunakan Bahasa pemrograman Pyton. Python merupakan Bahasa pemrograman yang menggabungkan kemampuan, kapabilitas, dan tingkat fleksibilitas yang tinggi. Sintaks pemrogramannya sederhana, dan memiliki perpustakaan standar yang luas dan lengkap (Fitri et al., 2017).

Pada tahap pra pemrosesan data akan dilakukan proses pembersihan data untuk menangani data yang kosong, duplikasi, menangani data outlier dan merubah data kategorikal menjadi numerik. Sebelum melakukan pra-pemrosesan data maka dilakukan persiapan data.

Adapun proses membersihkan data ke dalam Bahasa pemrograman Python, sebagai berikut :

1. Menangani data yang kosong (*handling missing value*)
   Pada tahap ini dilakukan pemeriksaan apakah dataset mengandung nilai kosong. Jika ditemukan nilai kosong makan akan dilakukan imputasi. Imputasi adalah imputasi adalah proses pengisian atau penggantian nilai-nilai yang hilang (*missing values*) pada sekumpulan data (*dataset*) dengan nilai-nilai yang mungkin berdasarkan informasi yang didapatkan pada dataset tersebut. Gambar 1. menunjukkan hasil pemeriksaan nilai yang kosong pada setiap kolom.

```
In [6]: #memeriksa nilai yang hilang di kolom
        print("Nilai yang Hilang Per Fitur:", df_riasec.isnull().sum())

Nilai yang Hilang Per Fitur: ID             0
JENIS KELAMIN    0
BINDO            0
MTK              0
IPA              1
IPS              2
BING             0
Minat            0
JURUSAN          0
dtype: int64
```

Sumber : (Priyono et al., 2022)

Gambar 1. Pemeriksaan *Missing Value*

Berdasarkan gambar 1 terlihat ada 1 data pada kolom IPA dan 2 data pada kolom IPS yang memiliki nilai kosong. Selanjutnya dilakukan imputasi untuk menangani nilai yang kosong. Nilai yang kosong akan diimputasi dengan nilai rata-rata dari semua nilai (nilai mean). Gambar 2 dan Gambar 3 menunjukkan proses imputasi pada kolom IPA dan IPS.

```
In [7]: #Missing value di kolom IPS akan diisi dengan rata-rata dari semua nilai (menggunakan mean)
        # Langkah 1
        nilai_ipa = df_riasec['IPA'].mean()
        # Langkah 2
        df_riasec['IPA'] = df_riasec['IPA'].fillna(nilai_ipa)
        # Langkah 3
        print(df_riasec.isnull().sum())

ID               0
JENIS KELAMIN    0
BINDO            0
MTK              0
IPA              0
IPS              2
BING             0
Minat            0
JURUSAN          0
dtype: int64
```

Sumber : (Priyono et al., 2022)

Gambar 2. Menangani Nilai Kosong Pada Kolom IPA

```
In [8]: #Missing value di kolom IPS akan diisi dengan rata-rata dari semua nilai (menggunakan mean)
        # Langkah 1
        nilai_ips = df_riasec['IPS'].mean()
        # Langkah 2
        df_riasec['IPS'] = df_riasec['IPS'].fillna(nilai_ips)
        # Langkah 3
        print(df_riasec.isnull().sum())

ID               0
JENIS KELAMIN    0
BINDO            0
MTK              0
IPA              0
IPS              0
BING             0
Minat            0
JURUSAN          0
dtype: int64
```

Sumber : (Priyono et al., 2022)

Gambar 3. Menangani Nilai Kosong Pada Kolom IPS

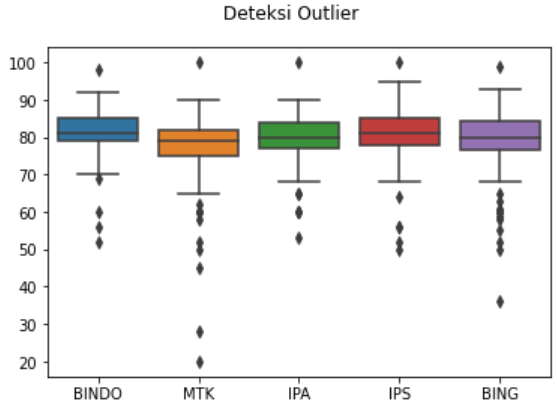
2. Menangani Data Duplikasi
   Pada tahap ini dilakukan pemeriksaan apakah dataset mengandung data yang duplikasi. Jika ada data yang duplikasi maka dilakukan reduksi terhadap data duplikasi tersebut. Pada gambar 4 menunjukkan tidak ada duplikasi data sehingga tidak melakukan proses reduksi data duplikasi.

```
In [10]: #mengecek apakah ada duplikat data?
         print("Jumlah Data Duplikat")
         df_riasec.duplicated().sum()

Jumlah Data Duplikat

Out[10]: 0
```

Sumber : (Priyono et al., 2022)

Gambar 4. Menangani Data Duplikasi

3. Menangani Data Outlier
   Outlier atau pencilan merupakan data yang nilainya jauh berbeda dengan nilai dari data yang lain pada suatu kelompok. Untuk melakukan deteksi dari outlier, dapat dilakukan dengan menggunakan boxplot. Pada gambar 5 menunjukkan data tidak ada yang outlier.



Sumber : (Priyono et al., 2022)

Gambar 5. Bloxpot Outlier Dataset Pemilihan Jurusan SMK

4. Feature Encoding
   Dataset pemilihan jurusan terdiri dari data kategorikal dan data numerik. Dalam membangun sebuah model *Machine Learning*, seluruh datanya harus berbentuk numerikal. Karena itu, data yang berupa kategorikal tersebut harus ditransformasikan terlebih dahulu menjadi data numerikal sehingga model *Machine Learning* dapat memproses data tersebut. Gambar 6 menjelaskan proses transformasi data kategorikal menjadi data numerik (integer) menggunakan encoding ordinal.

```
In [12]:
# 3. Label Ordinal
#Kadat encoding mengubah setiap nilai dalam kolom menjadi angka yang berurutan
#Membuat Dictionary dan Labelling

ljk = {
    'L':1,
    'P':2}

lminat = {
    'Realistic':1,
    'Investigative':2,
    'Artistic':3,
    'Social':4,
    'Enterprising':5,
    'Conventional':6}

ljrs = {
    'TKJ':1,
    'TMM':2}

In [13]:
df_riasec['JENIS KELAMIN'].replace(ljk, inplace=True)
df_riasec['Minat'].replace(lminat, inplace=True)
df_riasec['JURUSAN'].replace(ljrs, inplace=True)
```

Sumber : (Priyono et al., 2022)

Gambar 6. Proses Transformasi Data Kategorikal Menjadi Data Numerik

3.3. Analisis Data

Analis data eksplorasi (*Exploratory Data Analysis*) dilakukan untuk memahami isi dan komponen penyusun data. Hasil analisis ekspolorasi data, sebagai berikut :

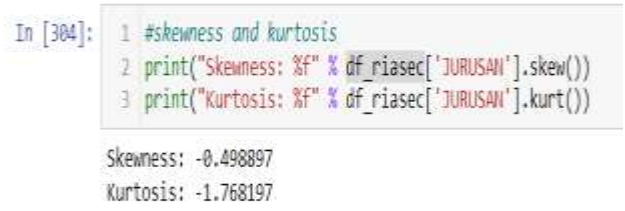
1. Analisis Eksplorasi Data Menggunakan Statistik Deskriptif
   Library Pandas menyediakan fungsi-fungsi statistika yang dapat diterapkan pada suatu DataFrame. Untuk mendapatkan ringkasan statistik secara cepat, DataFrame menyediakan fungsi describe() yang akan menghasilkan sebuah DataFrame baru yang berisi ringkasan statistik dari DataFrame yang padanya describe() diterapkan. Hasilnya dapat dilihat pada Gambar 7.

	ID	JENIS KELAMIN	BINDO	MTK	IPA	IPS	BING	Minat	JURUSAN
count	208.000000	208.000000	208.000000	208.000000	208.000000	208.000000	208.000000	208.000000	208.000000
mean	104.500000	1.322115	81.312500	77.346577	79.726962	80.721154	79.586538	3.764423	1.620192
std	60.188992	0.468414	5.074158	8.729969	5.933058	6.280022	7.868029	1.652778	0.486510
min	1.000000	1.000000	52.000000	20.000000	53.000000	50.000000	35.000000	1.000000	1.000000
25%	52.750000	1.000000	79.000000	75.000000	77.000000	78.000000	76.750000	3.000000	1.000000
50%	104.500000	1.000000	81.000000	79.000000	80.000000	81.000000	80.000000	4.000000	2.000000
75%	156.250000	2.000000	85.000000	82.000000	84.000000	85.000000	84.250000	5.000000	2.000000
max	208.000000	2.000000	98.000000	100.000000	100.000000	100.000000	99.000000	6.000000	2.000000

Sumber : (Priyono et al., 2022)

Gambar 7. Ringkasan Statistik

Berdasarkan dari ringkasan statistik dari dataframe diketahui nilai rata-rata hasil ujian Sekolah berada pada rentang 77 sampai 81. Standar deviasi untuk setiap mata pelajaran yang dujikan paling besar adalah 7.86. Nilai terendah ada 3 mata pelajaran ujian sekolah yaitu : Matematika, IPA dan IPS. Terdapat dua ukuran yang biasa digunakan untuk menjelaskan bentuk distribusi data yakni ukuran kemencengan (skewness) dan ukuran keruncingan (kurtosis). Untuk mendapatkan nilai Skeweness dan Kurtosis, DataFrame menyediakan fungsi skew() dan kurt().



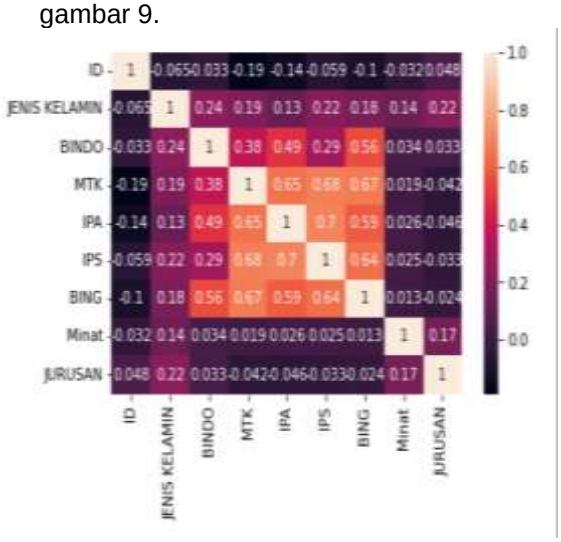
```
In [304]:
#skewness and kurtosis
print("Skewness: %f" % df_riasec['JURUSAN'].skew())
print("Kurtosis: %f" % df_riasec['JURUSAN'].kurt())

Skewness: -0.498897
Kurtosis: -1.768197
```

Sumber : (Priyono et al., 2022)

Gambar 8. Nilai Skewness dan Nilai Kurtosis Dataset Pemilihan Jurusan

Pada gambar 8 menunjukkan skewness yang bernilai -0.498897 berarti ekor distribusi berada di sebelah kiri, menunjukkan bahwa sebagian besar nilai berada di sisi kanan kurva. Sedangkan nilai Kurtosis bernilai -1.768197 berarti data tidak tersebar merata.

2. Analisis Korelasi
   Analisis korelasi menunjukkan hubungan antar dua atribut. Untuk mendapatkan nilai korelasi antar atribut, DataFrame menyediakan fungsi corr(). Visualisasi nilai korelasi antar atribut dapat dilihat pada gambar 9.



Sumber : (Priyono et al., 2022)

Gambar 9. Visualisasi Nilai Korelasi

Hasil analisis korelasi menunjukkan bahwa atribut ID dengan parameter lain kurang memiliki korelasi dengan semua parameter karena memiliki nilai negatif. Sedangkan, atribut jenis kelamin, bahasa indonesia, minat memiliki korelasi yang kuat dengan jurusan.

3.4. Pemodelan

Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Pada penelitian ini menggunakan enam atribut dengan kelas jurusan. Proposisi distribusi Kelas pada dataset pemilihan jurusan tidak seimbang. Kelas TKJ memiliki 79 data dan kelas TMM memiliki 129 data. Kelas yang tidak seimbang adalah masalah umum dalam klasifikasi pembelajaran mesin di mana terdapat rasio yang tidak proporsional di setiap kelas. Untuk menangani ketidakseimbangan data menggunakan Metode *Synthetic Minority Over-sampling Technique* (SMOTE). Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas (Sutoyo & Fadlurrahman, 2020). Gambar 10 menunjukkan distribusi kelas data yang telah seimbang dengan jumlah dataset menjadi 258 dengan komposisi 50% kelas TKJ dan 50% kelas TMM.

```
In [21]: from imblearn.over_sampling import SMOTE

         sm = SMOTE(random_state=42)

         X_sm, y_sm = sm.fit_resample(X, y)

         print(f'''Shape of X before SMOTE: {X.shape}
         Shape of X after SMOTE: {X_sm.shape}''')

         print('\nBalance of positive and negative classes (%):')
         y_sm.value_counts(normalize=True) * 100

Shape of X before SMOTE: (208, 7)
Shape of X after SMOTE: (258, 7)

Balance of positive and negative classes (%):

Out[21]: 1    50.0
         2    50.0
         Name: JURUSAN, dtype: float64
```

Sumber : (Priyono et al., 2022)

Gambar 10. Menangani Ketidakseimbangan Kelas Jurusan

Dalam membuat model *Machine Learning*, data dibagi menjadi data pelatihan dan data pengujian dengan komposisi 70 : 30. Data pelatihan dan data pengujian dipisahkan dengan menggunakan fungsi train_test_split(). Gambar 11 menunjukkan hasil pemisahan menjadi 180 data pelatihan dan 78 data pengujian.

```
In [344]: X_train, X_test, y_train, y_test = train_test_split(X_sm, y_sm, test_size=0.3, random_state=5)
          print(X_train.shape)
          print(y_train.shape)
          print(X_test.shape)
          print(y_test.shape)

(180, 7)
(180,)
(78, 7)
(78,)

Data Training = 180
Data Testing = 78
```

Sumber : (Priyono et al., 2022)

Gambar 11. Pemisahan Data pelatihan dan Data Pengujian

Penelitian ini melakukan klasifikasi menggunakan beberapa algoritma klasifikasi untuk menentukan algoritma klasifikasi terbaik. Algoritma yang digunakan adalah Multinomial Naïve Bayes, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Gradient Boosting Classifier, Decision Tree Classifier, K Neighbors Classifier, dan LogisticRegression. Berdasarkan tabel 2 terlihat bahwa algoritma *Gradient Boosting Classifier* memiliki nilai akurasi tertinggi sebesar 71.79%.

Tabel 2. Nilai Akurasi Algoritma Klasifikasi

No	Algoritma Klasifikasi	Akurasi	Log Loss
1	Multinomial Naïve Bayes	6.41%	0.677
2	Gaussian Naïve Bayes	70.51%	0.656
3	Bernoulli Naïve Bayes	48.72%	0.693
4	Gradient Boosting Classifier	71.79%	0.702
5	Decision Tree Classifier	61.54%	13.284
6	K Neighbors Classifier	66.67%	1.881
7	LogisticRegression	62.82%	0.673

Sumber : (Priyono et al., 2022)

Model klasifikasi ini masih dapat ditingkatkan nilai akurasinya dengan melakukan *Hyperparameter Tuning*. Hyperparameter Tuning merupakan proses mengidentifikasi parameter terbaik dari algoritma klasifikasi.

```
from sklearn.model_selection import GridSearchCV
svr = GradientBoostingClassifier()
parameters = {
    "n_estimators":[5,50,100,150],
    "max_depth":[1,3,5,7,9],
    "learning_rate":[0.01,0.1,1,10,100]
}

gbc = GridSearchCV(svr, parameters, cv=5,n_jobs=-1)
gbc.fit(X_train,y_train)
print("Parameter terbaik algoritma Gradient Boosting Classifier adalah ", gbc.best_params_)
```

```
Parameter terbaik algoritma Gradient Boosting Classifier adalah  {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 150}
```

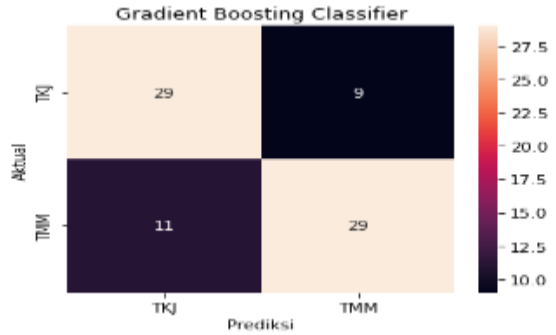
Sumber : (Priyono et al., 2022)

Gambar 12. *Hyperparameter Tuning*

Gambar 12 menunjukaan parameter terbaik dari algoritma *Gradient Boosting Classifier* berupa *learning rate* dengan nilai 1, maksimal kedalaman adalah 5 dan nilai estimatornya adalah 50. Pembangunan model klasifikasi algoritma algoritma *Gradient Boosting Classifier* menggunakan parameter *learning rate* = 1, max_depth = 5 dan n_estimator = 50.

3.5. Evaluasi Model

Tahap ini penyajian hasil dari proses pemodelan pada tahap sebelumnya. Tujuan evaluais Tujuan evaluasi model ini adalah menghasilkan model sesuai dengan tujuan yang ingin dicapai dalam tahap pertama. Gambar 13 menunjukkan *visualisasi confusion* matriks model klasifikasi data pemilihan jurusan. Hasil evaluasi model klasifikasi memiliki nilai akurasi sebesar 72% dan nilai sensivitas (*recall*) sebesar 76%. Visualisasi Confusion Matrix dapat dilihat pada gambar 13.



Sumber : (Priyono et al., 2022)

Gambar 13. Visualisasi Confusion Matrix Model Klasifikasi Dataset Pemililihan Jurusan SMK

4. Kesimpulan

Penelitian klasifikasi pemilihan jurusan sekolah menengah kejuruan menunjukkan algoritma Gradient Boosting Classifier memiliki nilai akurasi yang lebih baik dibandingkan dengan algoritma yang lain. Hasil pengujian menggunakan parameter terbaik dari Algoritma Gradient Boosting Classifier Gradient Boosting Classifier dengan Hyperparameter Tuning menggunakan GridSearchCV memperoleh akurasi 72% dan class recall mencapai 76%. Penelitian selanjutnya dapat melakukan penambahan parameter dan jumlah data sehingga menghasilkan akurasi yang lebih baik.

Referensi

Badan Pusat Statistik. (2020). *Potret Pendidikan Indonesia : statistik Pendidikan 2020*.

Fauziyah, E. N., & Nudin, S. R. (2021). *Sistem Pendukung Keputusan Penentuan Jurusan di SMKN 1 Pungging Menggunakan Gradient Boosting Tree*. *3*, 42–50.

Fitri, H. Y. F., & Saraswati Sinta. (2021). Pengaruh Self Determination Dan Prestasi Akademik Terhadap Kematangan Karier Siswa MA NU Nurul Huda. *G-COUNS: Jurnal Bimbingan Dan Konseling*, *5*(2), 247–257.

Fitri, R, K. R., Rahmansyah, A., & Darwin, W. (2017). Penggunaan Bahasa Pemrograman Python Sebagai Pusat Kendali Pada Robot 10-D. *5th Indonesian Symposium on Robotic Systems and Control*, 23–26.

Hafsoh, S., & Yusuf, A. M. (2021). *Knowing Student ' S Personality Types in the Determination of Career Selection According To Holland ' S Theory*. *1*(3), 88–98.

Hidayat, F. K., & Wahyuni, S. N. (2019). Pendeteksian Minat Dan Bakat Menggunakan Metode Riasec. *Indonesian Journal of Business Intelligence (IJUBI)*, *2*(1), 32. https://doi.org/10.21927/ijubi.v2i1.1023

Kementerian Pendidikan dan Kebudayaan. (2018). *Spektrum Keahlian Sekolah Menengah Kejuruan (SMK)/Madrasah Aliyah Kejuruan (MAK)*.

Mahmudah, S. N., & Lianawati, A. (2020). Bimbingan Kelompok Berbasis RIASEC Efektif Meningkatkan Kemantapan Pemilihan Karier Siswa Kelas XII SMA. *Jurnal Bimbingan Dan Konseling Teraputik*, *4*(2), 126–132. https://doi.org/10.26539/teraputik-42427

Prabowo, I. M., & Subiyanto. (2017). Sistem rekomendasi penjurusan sekolah menengah kejuruan dengan algoritma

c4.5. *Jurnal Kependidikan*, *1*(1), 139–149.

Priyono, H., Sari, R., & Mardiana, T. (2022). Skripsi Penerapan Algoritma Gradient Boosting Classifier Untuk Rekomendasi Pilihan. Jakarat : Universitas Nusa Mandiri.

Rahmayu, M., & Serli, R. K. (2018). Sistem Pendukung Keputusan Pemilihan Jurusan Pada SMK Putra Nusantara Jakarta Menggunakan Metode Analytical Hierarchy Process (AHP). *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *9*(1), 551–564.

Sumpena, Akbar, Y., & Nirat. (2018). Penerimaan Calon Siswabaru Dan Penentuan Penjurusan Dengan Algoritma C4.5 SMK Plus PGRI 1 Cibinong Selection and peer-review under responsibility of The 11th STIKOM CKI on SPOT. *CKI On SPOT*, *11*(2), 181–191.

Sunge, A. S., Fidiawan, H., Studi, P., Informatika, T., Tinggi, S., & Pelita, T. (2019). Data Mining, Penjualan Produk, Decision Tree, Algoritma C4.5 . . -Jurnal Teknologi Pelita Bangsa, 9, 97–103.

Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. JEPIN (Jurnal Edukasi Dan Penelitian Informatika), 6(3), 379–385.

Suyitno. (2020). *Pendidikan Vokasi Dan Kejuruan Strategi Dan Revitalisasi Abad 21* (M. Darmiati (ed.); 2020th ed.). K-Media.