Klasifikasi Pemberian Beasiswa Berprestasi Menggunakan Perbandingan Tiga Algoritma

Nanda Tri Haryati[1], Edi Surya Negara[1,*], Tri Basuki Kurniawan[1]

[1] Program Pascasarjana, Program Studi Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia

Email: [1] nandatriharyati99@gmail.com [1,*] e.s.negara@binadarma.ac.id, [1] tribasukikurniawan@binadarma.ac.id

Abstrak− Program Indonesia Pintar (PIP) merupakan program kerjasama dari tiga kementrian yaitu Kementrian Pendidikan dan Kebudayaan (Kemendikbud), Kementrian Sosial (Kemensos) dan Kementrian Agama (Kemenag). PIP melalui Kartu Indonesia Pintar (KIP) adalah pemberian bantuan tunai pendidikan kepada anak usia sekolah yang berasal dari keluarga miskin, rentan miskin, pemilik Kartu Keluarga Sejahtera (KKS), Program Keluarga Harapan (PKH), yatim piatu, penyandang disabilitas dan korban bencana alam/musibah. PIP merupakan bagian dari penyempurnaan program Bantuan Siswa Miskin (BSM) dengan tujuan untuk menjamin agar seluruh anak usia sekolah dari keluarga kurang mampu terdaftar sebagai penerima bantuan sampai anak lulus jenjang pendidikan menengah, baik melalui jalur pendidikan formal maupun non formal. Bantuan yang akan di terima yaitu berupa dana dengan besaran yang telah ditentukan sesuai tingkatan pendidikan. Banyaknya jumlah siswa yang mengajukan permohonan untuk menerima beasiswa dan kriteria penilaian yang banyak pula maka tidak semua siswa yang mengajukan permohonan beasiswa dapat dikabulkan. Permasalahan yang biasanya dihadapi oleh sekolah sehubungan dengan penentuan beasiswa adalah tidak adanya sistem yang dapat membantu untuk melakukan penyeleksian atau penentuan penerima beasiswa secara otomatis sesuai dengan kriteria-kriteria yang telah ditentukan. Untuk menyelesaikan hal ini, salah satu solusi yang akan digunakan adalah dengan mempelajari pola dari data siswa yang menerima beasiswa dan data siswa yang tidak menerima beasiswa. Dari data-data tersebut dapat dibentuk satu model yang dapat digunakan untuk memprediksi siswa siswa mana yang kemungkinan akan menerima beasiswa pada tahun berikutnya. Dalam penelitian ini dilakukan perbandingan dari 3 (tiga) algoritma klasifikasi untuk membantu menentukan prediksi siswa penerima beasiswa, sehingga pihak sekolah dapat dengan mudah menentukan data siswa mana yang akan diajukan. Algoritma yang dipilih untuk dibandingkan yaitu algoritma *Naïve Bayes*, *Random Fores*t dan *Support Vector Machine* dengan tujuan untuk mengetahui algoritma mana yang paling baik dalam hal tingkat akurasi dan berdasarkan literatur yang ada, tiga algoritma ini adalah algoritma yang umum digunakan dalam penelitian sejenis. Dari ke-3 algorithma tersebut, algoritma *Random Forest* memberikan nilai akurasi yang paling tinggi, yaitu 75%, diikuti oleh algoritma *Support Vector Machine* sebesar 59% dan nilai akurasi terendah yaitu algoritma *Naïve Bayes* sebesar 55%. Berikutnya, dengan menggunakan algoritma *Random Forest*, algoritma yang menghasilkan model dengan akurasi paling tinggi, dilanjutkan proses pemilihan fitur (*feature selection*) untuk melihat fitur mana yang paling memberikan pengaruh kepada label keputusan dan fitur mana yang tidak, yang selanjutkan dapat diabaikan atau tidak ikut dalam proses pembentukan model. Hasil yang diperoleh, menunjukkan fitur yang paling berpengaruh yaitu fitur penghasilan, sedangkan fitur mtk dan bhs_inggris dapat diabaikan, karena tidak memberikan pengaruh yang besar kepada label keputusan. Setelah itu, dilakukan pembandingan hasil model sebelum dilakukan pemilihan fitur dengan model setelah dilakukan pemilihan fitur dalam hal tingkat akurasi dan kecepatan proses menggunakan algoritma *Random Forest*. Hasil perbandingan menunjukan hasil peningkatan akurasi dan penuruan waktu proses yang cukup signifikan.

Kata Kunci: klasifikasi, beasiswa, program indonesia pintar (PIP), perbandingan algoritma

Abstract− The Smart Indonesian Program (SIP) is a collaborative program of three ministries, namely the Ministry of Education and Culture, the Ministry of Social Affairs and the Ministry of Religion. SIP through the Smart Indonesia Card (KIP) is the provision of educational cash assistance to school-age children who come from low-income families, vulnerable to poverty, owners of Prosperous Family Cards, Family Hope Programs, orphans, persons with disabilities and victims of natural disasters. SIP is part of the improvement of the Poor Student Assistance program to ensure that all school-age children from underprivileged families are registered as beneficiaries until the children graduate from secondary education, both through formal and non-formal education. The assistance that will be received is in the form of funds with a predetermined amount according to the level of education. The many students who apply to receive scholarships and the many assessment criteria mean that not all students who apply for scholarships can be granted. The problem that is usually faced by schools in connection with the determination of scholarships is that no system can help automatically select or determine scholarship recipients according to predetermined criteria. To solve this, one of the solutions that will be used is to study patterns from the data of students who receive scholarships and those who do not. From these data, a model can be formed that can be used to predict which students are likely to receive scholarships in the following year. In this study, 3 (three) classification algorithms were compared to help determine scholarship recipients' predictions, namely the Naïve Bayes algorithm, Random Forest and Support Vector Machine. Of the 3 algorithms, the Random Forest algorithm provides the highest accuracy value, 75%, followed by the Support Vector Machine algorithm at 59%, and the lowest accuracy value is the Naïve Bayes algorithm at 55%. Next, using the Random Forest algorithm, a feature selection process is carried out to see which features have the most influence on the decision label and which features do not, which can then be ignored or not involved in the model formation process. The results show that the most noteworthy feature is the Income feature, while the Math and English features can be ignored because they do not significantly influence the decision label.

Keywords: classfication, shcolarship, smart indonesian program (SIP), comparison of algorithms

1. PENDAHULUAN

Secara umum beasiswa merupakan bantuan keuangan yang diberikan kepada perorangan dengan tujuan untuk dipergunakan agar dapat melanjutkan pendidikan. Beasiswa ini dapat diberikan oleh lembaga pemerintah, perusahaan, ataupun yayasan. Adapun kategori pemberian beasiswa bisa dibagi menjadi dua, yaitu yang pertama pemberian beasiswa secara gratis dan yang kedua pemberian dengan ikatan pekerjaan atau ikatan dinas. Beasiswa tersebut diberikan kepada yang berhak menerima, berdasarkan klasifikasi, kualitas dan kompetensi penerimanya.

Program Indonesia Pintar (PIP) [1] merupakan program kerjasama dari tiga kementrian yaitu Kementerian Pendidikan dan Kebudayaan (Kemendikbud), Kementerian Sosial (Kemensos) dan Kementrian Agama (Kemenag). PIP melalui Kartu Indonesia Pintar (KIP) [2] adalah pemberian bantuan tunai pendidikan kepada anak usia sekolah yang berasal dari keluarga miskin, rentan miskin, pemilik Kartu Keluarga Sejahtera (KKS) [3], Program Keluarga Harapan (PKH) [4], yatim piatu, penyandang disabilitas dan korban bencana alam/musibah. PIP merupakan bagian dari penyempurnaan program Bantuan Siswa Miskin (BSM) [5] dengan tujuan untuk menjamin agar seluruh anak usia sekolah dari keluarga kurang mampu terdaftar sebagai penerima bantuan sampai anak lulus jenjang pendidikan menengah, baik melalui jalur pendidikan formal maupun non formal. Bantuan yang akan diterima yaitu berupa dana dengan besaran yang telah ditentukan sesuai tingkatan pendidikan [6].

Melalui program ini pemerintah berupaya mencegah peserta didik dari kemungkinan putus sekolah dan diharapkan dapat menarik siswa putus sekolah agar kembali melanjutkan pendidikannya. Selain itu dengan adanya program ini dapat meringankan biaya personal pendidikan peserta didik baik biaya langsung maupun tidak langsung.

Banyaknya jumlah siswa yang mengajukan permohonan untuk menerima beasiswa dan kriteria penilaian yang banyak pula maka tidak semua siswa yang mengajukan permohonan beasiswa dapat dikabulkan. Permasalahan yang biasanya dihadapi oleh sekolah sehubungan dengan penentuan beasiswa adalah tidak adanya sistem yang dapat membantu untuk melakukan penyeleksian atau penentuan penerima beasiswa secara otomatis sesuai dengan kriteria-kriteria yang telah ditentukan.

Untuk menyelesaikan hal ini, salah satu solusi yang akan digunakan adalah dengan mempelajari pola dari data siswa yang menerima beasiswa dan data siswa yang tidak menerima beasiswa. Dari data-data tersebut dapat dibentuk satu model yang dapat digunakan untuk memprediksi siswa siswa mana yang kemungkinan akan menerima beasiswa pada tahun berikutnya. Banyak penelitian sudah dilakukan oleh peneliti lain, seperti yang dilakukan oleh Yuniar [7], meneliti tentang proses klasifikasi data pemberian beasiswa pada siswa bidik misi, dengan menggunakan algoritma naïve bayes. Hasil dari penelitian tersebut menunjukkan, algoritma naïve bayes dapat memprediksi dengan hasil yang baik, hampir menyerupai data yang sebenarnya, yaitu dengan akurasi mencapai 83,33%.

Rizal dan Shinta [8] melakukan penelitian dengan menggunakan data rata-rata nilai siswa yang lulus seleksi. Data yang ada, kemudian diolah dan diproses menjadi model dengan menggunakan algoritma naïve bayes. Selanjutnya dilakukan pengujian menggunakan *confusion matrix* dan kurva *Receiver Operating Characteristic* (ROC) untuk mengukur tingkat akurasi dari model yang dihasilkan. Sedangkan menurut Utamajaya [9], perlu dilakukan penambahan variabel data yang diproses. Hal ini dimaksudkan untuk dapat meningkatkan akurasi dari hasil prediksi model yang dibentuk. Pada penelitian ini, digunakan data siswa yang lulus dan tidak lulus seleksi beasiswa bidik misi. Dari kedua penelitian diatas, masing masing mendapatkan nilai akurasi yang sangat baik, yaitu 97.2% dan 96.67%.

Selain algoritma naïve bayes, terdapat beberapa algoritma lainnya yang juga mendapat perhatian dari peneliti lain, seperti penelitian yang dilakukan oleh Firdaus [10] menggunakan algoritma *Support Vector Machine* (SVM), seperti juga penelitian yang dilakukan oleh Damanik [11], selain itu juga menggunakan algoritma *Decision Tree*. Selain itu Hayati [12] dalam penelitiannya melakukan perbandingan antara algoritma *decision tree* dan algoritma *naïve bayes*.

Dalam penelitian ini dilakukan perbandingan dari 3 (tiga) algoritma klasifikasi untuk membantu menentukan prediksi siswa penerima beasiswa, yaitu algoritma *Naïve Bayes*, *Random Fores*t dan *Support Vector Machine* dan selanjutnya berdasarkan hasil yang memberikan nilai akurasi yang paling tinggi, akan dilakukan proses pemilihan fitur yang paling memberikan pengaruh terhadap keputusan.

2. METODE PENELITIAN

2.1 Desain Penelitian

Untuk memberikan gambaran yang jelas tentang tahapan-tahapan yang akan dilakukan dalam penelitian ini, maka disusun satu panduan tahapan penelitian di dalam satu desain penelitian, seperti diperlihatkan pada gambar 1, sebagai kerangka kerja penelitian.

Proses yang ditampilkan pada gambar 1 ini dikenal sebagai *pre-processing* dalam *Knowledge Discovery in Database* (KDD), yaitu salah satu metode yang umum digunakan dalam pemprosesan data mining. KDD adalah keseluruhan proses *non-trivial* untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang

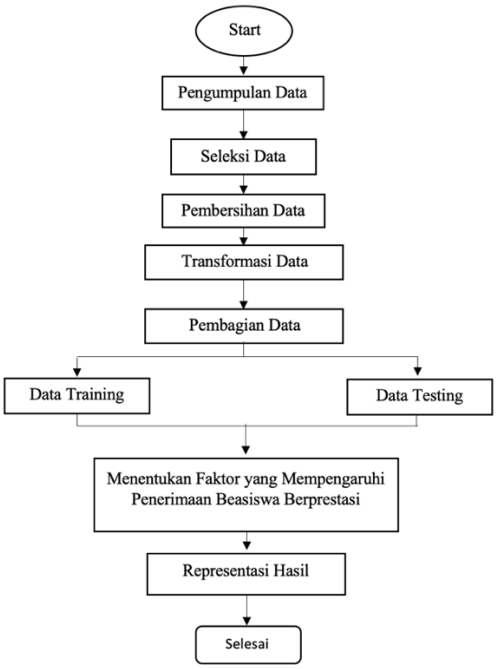
ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dari pola-pola sejumlah kumpulan data.

Pada gambar 1, tahapan pertama berupa pengumpulan data yang dalam hal ini mengumpulkan data langsung dari tempat dimana objek penelitian berada, yaitu berupa data siswa-siswa dimana penelitian ini dilakukan. Tahapan selanjutnya adalah seleksi data. Setelah data dikumpulkan, data tersebut selanjutnya dilakukan pemilihan. Tidak semua ada yang telah dikumpulkan sesuai atau memberikan informasi yang sesuai dengan tujuan dari penelitian ini. Untuk itu, hanya akan dipilih data-data yang akan memberikan pengaruh kepada hasil keputusan dari tiap tiap data, yaitu siswa yang mendapat beasiswa dan siswa yang tidak mendapat beasiswa.

Langkah ke-3, adalah proses pembersihan data. Pada proses ini akan dilakukan pembersihan dan pembuangan data yang tidak lengkap. Atau untuk data lain yang mungkin hilang atau terjadi kesalahan penulisan data, akan dilakukan proses penggantian data dengan data yang lebih sesuai atau tepat. Selanjutnya dilakukan proses transformasi data, bagi data-data yang memerlukan proses pengubahan data. Misalnya, jarak antara data-data numerik pada salah satu fitur terlalu jauh, maka bisa dilakukan proses normalisasi, sehingga semua data akan diubah ke dalam range antara 0 dan 1.

Berikutnya adalah proses pemisahan data (*spiliting data*) menjadi dua bagian, yaitu data training yang akan digunakan untuk melakukan pelatihan model dan data testing, yang digunakan untuk melakukan pengukuran akurasi dari model yang terbentuk dari hasil proses training. Setelah didapatkan hasil akurasi dari model yang terbentuk dengan melakukan perbandingan dari 3 algoritma yang digunakan dalam penelitian ini, selanjutnya dipilih satu algoritma yang memberikan hasil akurasi paling tinggi untuk dilakukan proses *feature selection*, atau pemilihan fitur yang paling memberikan kontribusi yang paling besar kepada label dari tiap-tiap data yang digunakan. Dengan mengetahui informasi ini, dalam proses pembentukan model berikutnya dapat hanya dipilih fitur yang memberikan kontribusi yang besar saja untuk mempercepat proses pembentukan model dan diharap dapat memberikan nilai akurasi yang lebih tinggi.

Hasil dari proses, selanjutnya dianalisis dan ditampilkan dalam bentuk grafik (proses visualisasi), sehingga lebih mudah untuk difahami oleh orang awam.



Gambar 1. Desain Penelitian

2.2 Jenis Data

Di dalam penelitian ini digunakan penelitian dengan metode eksperimen dengan menggunakan data siswa yang mengajukan beasiswa PIP di "SMK Nurul Iman Palembang" tahun angkatan 2018, 2019, 2020, 2021 dan 2022.

Terdapat 21 fitur data data asal yang diperoleh dari pihak SMK Nurul Iman Palembang, seperti dapat dilihat pada tabel 1, yang terdiri dari data 6 jurusan, AK, AP, TBSM, TKJ, TKJ1 dan TKJ2, untuk semua tingkatan. Jumlah keseluruhan data yang dikumpulkan berjumlah 2434 data

Tabel 1. Daftar Fitur data yang dikumpulkan

No	Nama Fitur	Keterangan	No	Nama Fitur	Keterangan
1	no_induk	No Induk Siswa	12	pekerjaan	Pekerjaan Orang Tua
2	nisn	NIS Nasional	13	penghasilan	Penghasilan Orang Tua
3	nama_siswa	Nama Siswa	14	alamat_ortu	Alamat Orang Tua
4	tahun_ajaran	Data Tahun Ajaran	15	pai	Nilai Pend. Agama Islam
5	alamat	Alamat Siswa	16	pkn	Nilai Pend. Kewarganegaraan
6	tempat_lahir	Tempat Lahir	17	bhs_indo	Nilai Bahasa Indonesia
7	ttl	Tanggal Lahir Siswa	18	mtk	Nilai Matematika
8	jenis_kelamin	Jenis Kelamin Siswa	19	bhs_inggris	Nilai Bahasa Inggris
9	asal_sekolah	Asal SMP	20	kelas	Kelompok Kelas
10	nama_ayah	Nama Ayah dari Siswa	21	label	Keputusan Dapat/Tidak Dapat
11	nama_ibu	Nama Ibu dari Siswa			Beasiswa di semester, tsb

Dari fitur yang ada pada tabel 1, tidak semua akan diikutkan di dalam proses pembentukan model. Terlebih dahulu akan dilakukan analisa dan pra-pemprosesan data, seperti dijelaskan sebelumnya.

2.3 Teknik Pengumpulan Data

Dalam penelitian ini metode pengumpulan data yang digunakan adalah pengumpulan data primer, berupa data siswa yang lulus seleksi beasiswa dan yang tidak lulus seleksi beasiswa. Sedangkan data pendukung lainnya didapat dari buku, jurnal dan publikasi lainnya.

Data siswa ini yang nanti akan digunakan dalam proses menentukan calon penerima beasiswa. Di dalam data tersebut dapat diketahui status siswa yang lulus seleksi dan tidak lulus seleksi.

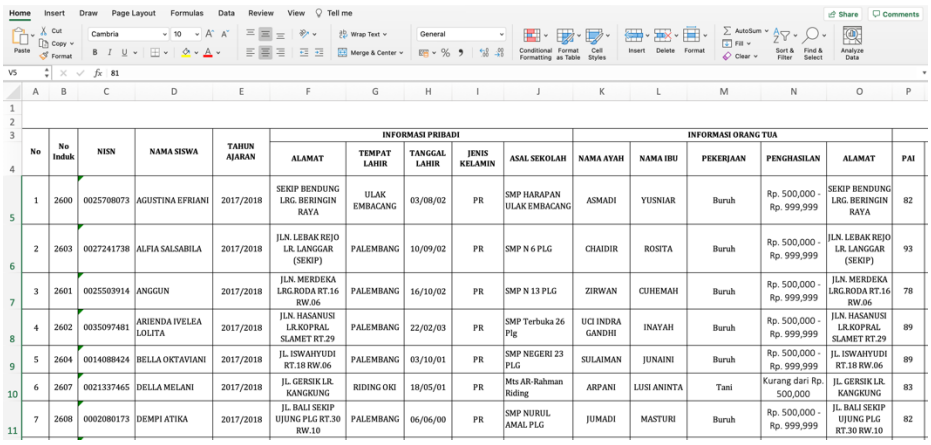
2.4 Teknik Analisis Data

Teknik analisis data yang digunakan dalam penelitian ini menggunakan metode kualitatif dengan pemrosesan data terhadap angka atau numerik dan nominal. Data yang akan diproses adalah data siswa SMK dengan nilai nilai mata pelajaran dari siswa yang lulus seleksi dan yang tidak lulus seleksi. Metode yang digunakan adalah metode klasifikasi, dengan membandingkan tiga (3) algoritma, antara lain: *Naïve Bayes*, *Random Forest* and *Support Vector Machine*. Nilai akurasi dari tiga (3) algoritma yang akan digunakan sebagai perbandingan dan penilian dan selanjutnya dilakukan analisis.

3. HASIL DAN PEMBAHASAN

3.1 Proses Pengumpulan Data

Hasil dari proses pengumpulan data yang dilakukan, berupa file excel dengan format tampilan seperti diperlihatkan pada gambar 2, berikut ini.



No	No Induk	NISN	NAMA SISWA	TAHUN AJARAN	INFORMASI PRIBADI					INFORMASI ORANG TUA					
					ALAMAT	TEMPAT LAHIR	TANGGAL LAHIR	JENIS KELAMIN	ASAL SEKOLAH	NAMA AYAH	NAMA IBU	PEKERJAAN	PENGHASILAN	ALAMAT	PAI
1	2600	0025708073	AGUSTINA EFRIANI	2017/2018	SEKIP BENDUNG LRG. BERINGIN RAYA	ULAK EMBACANG	03/08/02	PR	SMP HARAPAN ULAK EMBACANG	ASMADI	YUSNIAR	Buruh	Rp. 500,000 - Rp. 999,999	SEKIP BENDUNG LRG. BERINGIN RAYA	82
2	2603	0027241738	ALFIA SALSABILA	2017/2018	JLN. LEBAK REJO LR. LANGGAR (SEKIP)	PALEMBANG	10/09/02	PR	SMP N 6 PLG	CHAIDIR	ROSITA	Buruh	Rp. 500,000 - Rp. 999,999	JLN. LEBAK REJO LR. LANGGAR (SEKIP)	93
3	2601	0025503914	ANGGUN	2017/2018	JLN. MERDEKA LRG.RODA RT.16 RW.06	PALEMBANG	16/10/02	PR	SMP N 13 PLG	ZIRWAN	CUHEMAH	Buruh	Rp. 500,000 - Rp. 999,999	JLN. MERDEKA LRG.RODA RT.16 RW.06	78
4	2602	0035097481	ARIENDA IVELEA LOLITA	2017/2018	JLN. HASANUSI LR.KOPRAL SLAMET RT.29	PALEMBANG	22/02/03	PR	SMP Terbuka 26 Plg	UCI INDRA GANDHI	INAYAH	Buruh	Rp. 500,000 - Rp. 999,999	JLN. HASANUSI LR.KOPRAL SLAMET RT.29	89
5	2604	0014088424	BELLA OKTAVIANI	2017/2018	JL. ISWAHYUDI RT.18 RW.06	PALEMBANG	03/10/01	PR	SMP NEGERI 23 PLG	SULAIMAN	JUNAINI	Buruh	Rp. 500,000 - Rp. 999,999	JL. ISWAHYUDI RT.18 RW.06	89
6	2607	0021337465	DELLA MELANI	2017/2018	JL. GERSIK LR. KANGKUNG	RIDING OKI	18/05/01	PR	Mts AR-Rahman Riding	ARPANI	LUSI ANINTA	Tani	Kurang dari Rp. 500,000	JL. GERSIK LR. KANGKUNG	83
7	2608	0002080173	DEMPI ATIKA	2017/2018	JL. BALI SEKIP UJUNG PLG RT.30 RW.10	PALEMBANG	06/06/00	PR	SMP NURUL AMAL PLG	JUMADI	MASTURI	Buruh	Rp. 500,000 - Rp. 999,999	JL. BALI SEKIP UJUNG PLG RT.30 RW.10	82

Gambar 2. Format data yang diperoleh dari tempat penelitian

Data yang diperoleh dari tempat penelitian, terdiri dari beberapa bagian data, yaitu informasi tentang data siswa, seperti no induk, NISN, nama siswa dan tahun ajaran. Berikutnya bagian data pribadi, seperti alamat, tempat tanggal lahir, jenis kelamin, jenis kelamin dan asal sekolah.

Pada bagian data orang tua, terdapat data nama ayah, nama ibu, pekerjaan, penghasilan orang tua dan alamat orang tua. Sedangkan bagian akhir adalah informasi mengenai nilai 5 mata pelajaran untuk semester 1 dan semester 2 pada tahun ajaran itu. Untuk 5 mata pelajaran yang digunakan dalam penelitian ini adalah nilai pelajaran Pendidikan Agama Islam, Pendidikan Kewarganegaraan, Bahasa Indonesia, Matematika dan Bahasa Inggeris.

Untuk memudahkan dalam pemrosesan data, dilakukan sedikit perubahan format pada data. Data perlu dipisah untuk setiap semesternya. Sehingga untuk semester 1 dan semester 2, akan digunakan data yang sama untuk bagian informasi data siswa, data pribadi siswa dan data orang tua.

Hasil dari pengumpulan data, diperoleh data sebanyak 2486 baris dan 22 kolom, seperti diperlihatkan pada gambar 3 dan 4, berikut ini.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2486 entries, 0 to 2485
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   no              2486 non-null   int64
 1   no_induk        2486 non-null   int64
 2   nisn            2454 non-null   float64
 3   nama_siswa      2486 non-null   object
 4   tahun_ajaran    2486 non-null   object
 5   alamat          2486 non-null   object
 6   tempat_lahir    2486 non-null   object
 7   ttl             2486 non-null   object
 8   jenis_kelamin   2486 non-null   object
 9   asal_sekolah    2486 non-null   object
 10  nama_ayah       2486 non-null   object
 11  nama_ibu        2486 non-null   object
 12  pekerjaan       2486 non-null   object
 13  penghasilan     2438 non-null   object
 14  alamat_ortu     2356 non-null   object
 15  pai             2486 non-null   int64
 16  pkn             2486 non-null   int64
 17  bhs_indo        2486 non-null   int64
 18  mtk             2486 non-null   int64
 19  bhs_inggris     2486 non-null   int64
 20  label           2481 non-null   object
 21  kelas           2486 non-null   object
dtypes: float64(1), int64(7), object(14)
memory usage: 427.4+ KB
None
```

Gambar 3. Informasi fitur yang ada

	no	no_induk	nisn	nama_siswa	tahun_ajaran	alamat	tempat_lahir	ttl	jenis_kelamin	asal_sekolah
0	1	2600	25708073.0	AGUSTINA EFRIANI	2017/2018	SEKIP BENDUNG LRG. BERINGIN RAYA	ULAK EMBACANG	03/08/02	PR	SMP HARAPAN ULAK EMBACANG
1	2	2603	27241738.0	ALFIA SALSABILA	2017/2018	JLN. LEBAK REJO LR. LANGGAR (SEKIP)	PALEMBANG	10/09/02	PR	SMP N 6 PLG
...	...	...	...	...	...	...	...	...	...	...
2484	2485	2889	69919894.0	NATASYA PERMATA WIJAYA	2021/2022	JL. SINTRAMAN JAYA B.28 RT.32 RW.09	PALEMBANG	13/11/2006	PR	SMP NEGERI 10 PLG
2485	2486	2914	69538577.0	HABIB PRAYOGA	2021/2022	JL. SIDOMULYO 1 RT.12 RW.04 NO.38 PLAJU	SEKAYU	09/08/2006	LK	Mts. NURUL AMAL ULAKEMBACANG

2486 rows × 22 columns

Gambar 4. Tampilan data dan jumlah data yang ada

Pada gambar 3, ditampilkan semu fitur yang telah diperoleh. Sedangkan pada gambar 4, detail dari data pada tiap-tiap baris diperlihatkan. Pada bagian bawah kanan gambar 4, terdapat informasi jumlah baris dan kolom dari data yang kita peroleh, yaitu sejumlah 2486 baris dengan 22 kolum data.

3.2 Seleksi Data

Pada tahapan seleksi data, dilakukan analisis terhadap data-data yang tersedia. Dari hasil analisis, dipilih beberapa fitur yang dianggap akan memberikan pengaruh kepada keputusan. Adapun fitur yang dipilih tersebut, seperti diperlihatkan pada tabel 2, berikut ini. Sedangkan fitur yang lain dianggap tidak mempunyai korelasi terhadap kemungkinan seorang siswa akan mendapatkan beasiswa ataupun tidak.

Tabel 2. Daftar Fitur data yang dipilih untuk diproses selanjutnya

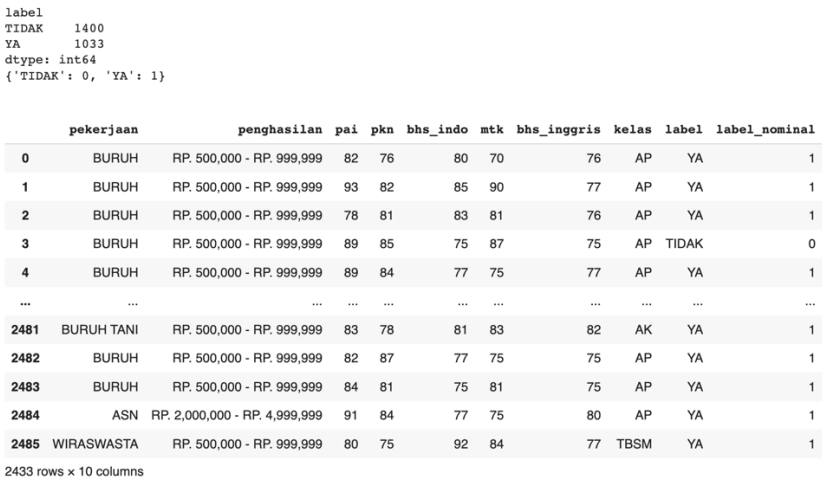
No	Nama Fitur	Keterangan
1	pekerjaan	Mempunyai korelasi terhadap kelompok ekonomi keluarga
2	penghasilan	Menentukan kelompok ekonomi keluarga
3	pai	Nilai dari mata pelajaran pokok
4	pkn	Nilai dari mata pelajaran pokok
5	bhs_indo	Nilai dari mata pelajaran pokok
6	mtk	Nilai dari mata pelajaran pokok
7	bhs_inggris	Nilai dari mata pelajaran pokok
8	kelas	Mempunyai korelasi terhadap kelompok nilai dari kelas yang sama
9	label	Keputusan menerima beasiswa ataupun tidak

Table 2 memperlihatkan daftar fitur yang akan diproses selanjutnya dengan masing-masing alasan mengapa fitur tersebut dipilih.

3.3 Pembersihan Data

Selanjutnya akan dilakukan proses pembersihan data. Dalam proses ini termasuk dilakukannya perbaikan data, atau format penulisan yang tidak tepat ataupun tidak konsisten. Untuk itu, perlu dilakukan proses analisis data dengan bantuan visualisasi kelompok item nilai yang ada pada tiap tiap kolom data.

Proses dimulai dengan kolom data yang paling penting, yaitu kolom label. Pertama, dilakukan pengelompokan data berdasarkan nilai yang ada di dalam kolom label. Dengan menggunakan pemrograman Python, kita kelompokkan data berdasarkan nilainya dan selanjutnya kita tampilkan dalam sebuah grafik jenis batang, seperti diperlihatkan pada gambar 5 dan 6 berikut ini. Selanjutnya nilai 'TIDAK' kita ubah menjadi nilai numerik 0 dan nilai 'YA', kita ubah menjadi nilai numerik 1.

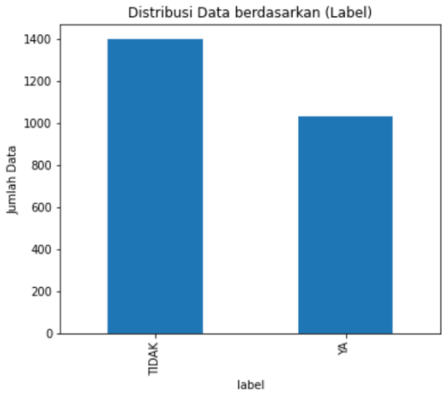


```
label
TIDAK    1400
YA       1033
dtype: int64
{'TIDAK': 0, 'YA': 1}
```

	pekerjaan	penghasilan	pai	pkn	bhs_indo	mtk	bhs_inggris	kelas	label	label_nominal
0	BURUH	RP. 500,000 - RP. 999,999	82	76	80	70	76	AP	YA	1
1	BURUH	RP. 500,000 - RP. 999,999	93	82	85	90	77	AP	YA	1
2	BURUH	RP. 500,000 - RP. 999,999	78	81	83	81	76	AP	YA	1
3	BURUH	RP. 500,000 - RP. 999,999	89	85	75	87	75	AP	TIDAK	0
4	BURUH	RP. 500,000 - RP. 999,999	89	84	77	75	77	AP	YA	1
...	...	...	...	...	...	...	...	...	...	...
2481	BURUH TANI	RP. 500,000 - RP. 999,999	83	78	81	83	82	AK	YA	1
2482	BURUH	RP. 500,000 - RP. 999,999	82	87	77	75	75	AP	YA	1
2483	BURUH	RP. 500,000 - RP. 999,999	84	81	75	81	75	AP	YA	1
2484	ASN	RP. 2,000,000 - RP. 4,999,999	91	84	77	75	80	AP	YA	1
2485	WIRASWASTA	RP. 500,000 - RP. 999,999	80	75	92	84	77	TBSM	YA	1

2433 rows × 10 columns

Gambar 5. Distribusi nilai 'TIDAK' dan 'YA' dan hasil perubahan ke nilai nominal



Gambar 6. Distribusi jumlah data dengan nilai label 'TIDAK' dan 'YA'

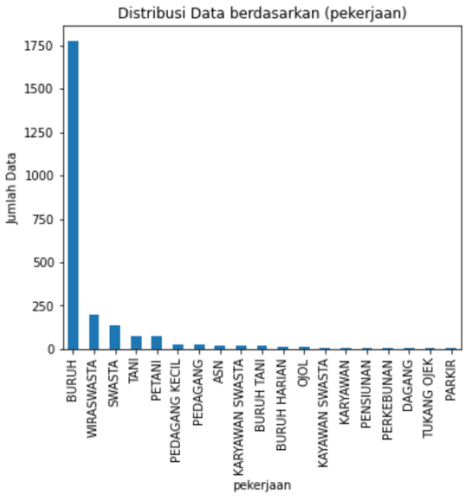
Pada gambar 5, diperlihatkan ada 1400 data yang mempunyai label 'TIDAK' dan 1033 data yang mempunyai label 'YA'. Disini dapat disimpulkan, terdapat lebih banyak data dengan label 'TIDAK' berbanding dengan data dengan label 'YA'. Dalam penelitian ini tidak membahas tentang kondisi inbalanced data dan dapat diusulkan untuk dilakukan dalam penelitian berikutnya.

Pada tahapan akhir, data label yang semula berjenis nominal atau teks, kita ubah menjadi nilai numerik. Hal ini diperlukan untuk menjadikan seluruh data nantinya bernilai numerik, karena nilai dari mata pelajaran, sudah dalam bentuk numerik semua.

Berikutnya kita proses untuk data pada kolom pekerjaan, penghasilan dan kelas, seperti diperlihatkan pada gambar 7 sampai dengan gambar 9, berikut ini.

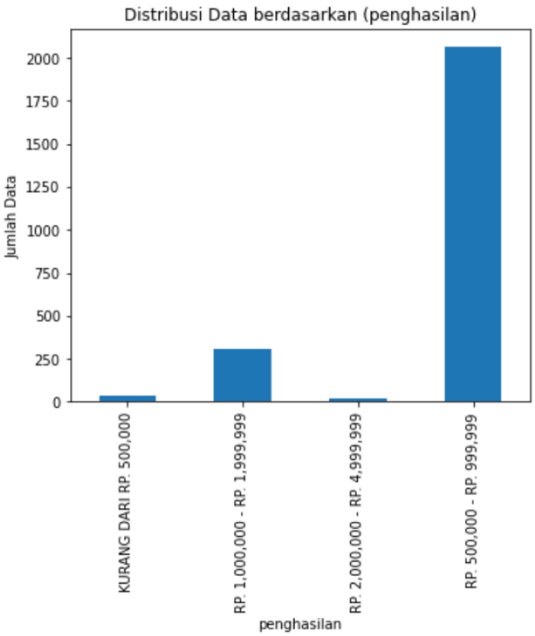
```
pekerjaan
BURUH               1775
WIRASWASTA           200
SWASTA               134
TANI                  72
PETANI                72
PEDAGANG KECIL        28
PEDAGANG              26
ASN                   20
KARYAWAN SWASTA       18
BURUH TANI            18
BURUH HARIAN          14
OJOL                  12
KAYAWAN SWASTA         8
KARYAWAN               6
PENSIUNAN              6
PERKEBUNAN             6
DAGANG                 6
TUKANG OJEK            6
PARKIR                 6
dtype: int64
```

```
{'BURUH': 0,
 'WIRASWASTA': 1,
 'SWASTA': 2,
 'TANI': 3,
 'PETANI': 4,
 'PEDAGANG KECIL': 5,
 'PEDAGANG': 6,
 'ASN': 7,
 'KARYAWAN SWASTA': 8,
 'BURUH TANI': 9,
 'BURUH HARIAN': 10,
 'OJOL': 11,
 'KAYAWAN SWASTA': 12,
 'KARYAWAN': 13,
 'PENSIUNAN': 14,
 'PERKEBUNAN': 15,
 'DAGANG': 16,
 'TUKANG OJEK': 17,
 'PARKIR': 18}

Jumlah jenis pekerjaan 19
```



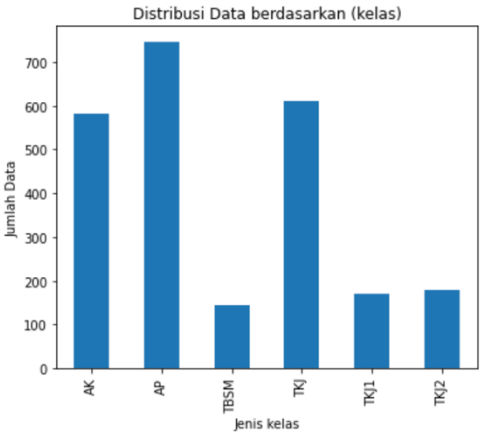
Gambar 7. Distribusi nilai yang ada di dalam kolom pekerjaan dan konversi ke bentuk numerik

```
penghasilan
KURANG DARI RP. 500,000            36
RP. 1,000,000 - RP. 1,999,999     310
RP. 2,000,000 - RP. 4,999,999      20
RP. 500,000 - RP. 999,999        2067
dtype: int64

{'KURANG DARI RP. 500,000': 0,
 'RP. 1,000,000 - RP. 1,999,999': 1,
 'RP. 2,000,000 - RP. 4,999,999': 2,
 'RP. 500,000 - RP. 999,999': 3}

Jumlah jenis penghasilan 4
```



Gambar 8. Distribusi nilai yang ada di dalam kolom penghasilan dan konversi ke bentuk numerik

```
kelas
AK      582
AP      745
TBSM    144
TKJ     612
TKJ1    170
TKJ2    180
dtype: int64

{'AK': 0, 'AP': 1, 'TBSM': 2, 'TKJ': 3, 'TKJ1': 4, 'TKJ2': 5}

Jumlah jenis kelas 6
```



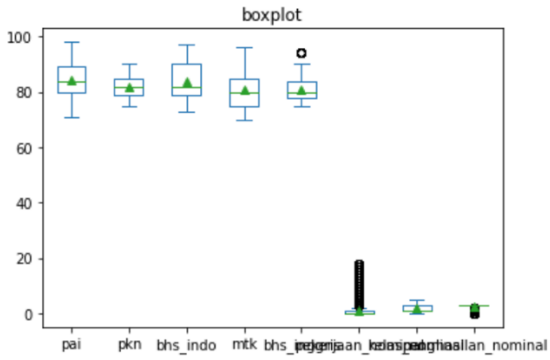
Gambar 9. Distribusi nilai yang ada di dalam kolom kelas dan konversi ke bentuk numerik

Pada gambar 7, dapat dilihat jumlah data untuk nilai buruh yang paling tinggi yaitu 1775 sedangkan jumlah nilai terbanyak kedua yaitu wiraswasta, sebanyak 200 atau hanya 11,27%, sangat berbeda jauh dengan nilai lain pada kolom pekerjaan. Hal ini akan mempengaruhi bagaimana model yang akan dibentuk nantinya belajar tentang data data yang ada. Berikutnya pada gambar 8, terlihat untuk penghasilan antara Rp 500.000 – Rp. 999.999 yang mempunyai data paling banyak, yaitu sebanyak 2067 data. Sedangkan nilai terbanyak keduanya adalah nilai Rp.1.000.000 – Rp. 1.999.999 yaitu sebanyak 310 data, atau hanya 14,51%.

Pada gambar 9, perbandingan jumlah data untuk masing-masing nilai terbagi pada dua kelompok, yaitu kelompok data yang jumlahnya cukup besar yaitu untuk nilai AK, AP dan TKj, dan kelompok yang jumlah jumlah nya kecil, yaitu TBSM, TKj1 dan TKj2. Perbandingan antara jumlah terbanyak dari kelompok pertama terhadap jumlah terbanyak dari kelompok kedua adalah sebesar 24,16%.

3.4 Transformasi Data

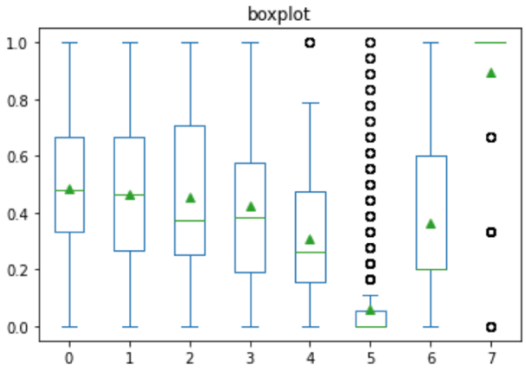
Pada tahapan ini, kita mula-mula perlu melihat apakah jangkauan data untuk tiap-tiap kolom sudah seimbang. Untuk itu kita perlu menampilkan data dalam grafik boxplot, seperti diperlihatkan pada gambar 10.



Gambar 10. BoxPlot data

Pada gambar 10, terlihat perbedaan yang sangat jauh antara nilai-nilai yang ada pada kolom mata pelajaran berbanding dengan nilai-nilai yang ada pada kolom pekerjaan, penghasilan dan kelas. Untuk itu perlu dilakukan proses transformasi data ke nilai normalisasinya. Hal ini dapat diselesaikan dengan mengubah nilai-nilai tersebut menggunakan fungsi MinMaxScalar yang disediakan oleh *library sklearn*.

Setelah dilakukan proses perubahan data, maka hasil yang diperoleh, seperti diperlihatkan pada gambar 11, berikut ini.



Gambar 11. BoxPlot data setelah dilakukan normalisasi proses

Pada gambar 11, semua kolom sudah mempunyai cakupan nilai yang sama. Artinya pada tahapan ini, data kita sudah siap untuk diproses.

3.5 Pembagian Data

Pada tahapan ini, data akan kita bagi menjadi dua bagian, yaitu data training sebanyak 70% dan data testing sebanyak 30%, atau dengan pembagian sebanyak 1703 data untuk data training dan 730 data untuk data testing, seperti diperlihatkan pada gambar 12.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=64, stratify=y)
print(len(X_train), len(X_test), len(y_train), len(y_test))
print()
```

Gambar 12. Tampilan code dalam Python untuk membagi data

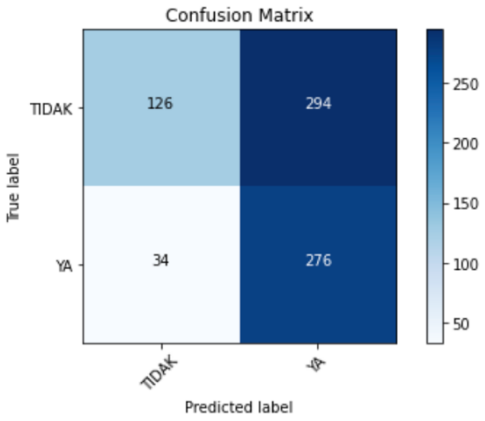
Pada gambar 12, kita menggunakan fungsi *train_test_split* pada *library sklearn* untuk membagi data menjadi dua bagian, data training dan data testing. Selanjutnya data training akan kita gunakan untuk melatih model yang akan kita buat.

3.6 Pembuatan Model dan Pengukuran Akurasi

Proses pembentukan model menggunakan 3 (tiga) algoritma, yaitu Naïve Bayes, Random Forest dan Support Vector Machine, yang masing-masing menggunakan fungsi *GaussianNB*, *RandomForestClassifier* dan *svm* dari *library sklearn*.

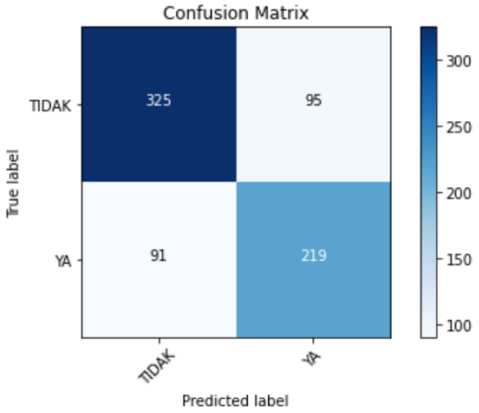
Setelah model terbentuk, selanjutnya dilakukan pengukuran akurasi dengan memprediksi data testing dan selanjutnya membandingkan hasil prediksi dengan label yang ada di data testing. Dengan menggunakan *confusion matrix*, akurasi dari model dapat kita ukur. Berikut ini, seperti diperlihatkan pada gambar 13 sampai dengan gambar 15, diperlihatkan hasil proses pengukuran akurasi kepada ketiga model dan selanjutnya pada tabel 3 diperlihatkan hasil perbandingan dari ketiga model.

```
******************************************************
Naive Bayes Gaussian Classifier
******************************************************

Confusion matrix, without normalization

              precision    recall  f1-score   support

           0       0.79      0.30      0.43       420
           1       0.48      0.89      0.63       310

    accuracy                           0.55       730
   macro avg       0.64      0.60      0.53       730
weighted avg       0.66      0.55      0.52       730


0.5506849315068493
```



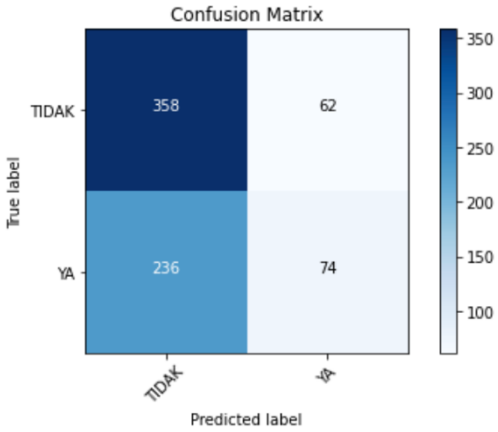
Gambar 13. Hasil pengukuran akurasi model dengan algoritma Naïve Bayes Gaussian

```
******************************************************
Random Forest Classifier
******************************************************

Confusion matrix, without normalization

              precision    recall  f1-score   support

           0       0.78      0.77      0.78       420
           1       0.70      0.71      0.70       310

    accuracy                           0.75       730
   macro avg       0.74      0.74      0.74       730
weighted avg       0.75      0.75      0.75       730


0.7452054794520548
```



Gambar 14. Hasil pengukuran akurasi model dengan algoritma Random Forest

```
******************************************************
Support Vector Machine Classifier
******************************************************

Confusion matrix, without normalization

              precision    recall  f1-score   support

           0       0.60      0.85      0.71       420
           1       0.54      0.24      0.33       310

    accuracy                           0.59       730
   macro avg       0.57      0.55      0.52       730
weighted avg       0.58      0.59      0.55       730


0.5917808219178082
```



Gambar 15. Hasil pengukuran akurasi model dengan algoritma Support Vector Machine

Tabel 3. Hasil pengukuran akurasi untuk setiap model

No	Algoritma	Akurasi (%)
1	Naïve Bayes	55,07
2	Random Forest	75,52
3	Support Vector Machine	59,18

Dari gambar 13 sampai dengan gambar 15, diperlihatkan hasil prediksi dari ketiga model, *Naïve Bayes*, *Random Forest* dan *Support Vector Machine*. Selanjutnya dari tabel 3 dapat disimpulkan bahwa algoritma *Random Forest* membentuk model yang tingkat akurasinya paling baik, yaitu mencapai 75,52%. Sedangkan akurasi tertinggi kedua didapat dari model yang dibentuk menggunakan algoritma *Support Vector Machine* dengan nilai akurasi 59,18% dan terakhir didapat dari model dengan algoritma *Naïve Bayes*, sebesar 55,07%.

Selanjutnya dengan menggunakan fungsi *cross_val_score* dari *library sklearn*, dilakukan percobaan dengan pengulangan sebanyak 10 kali lalu diambil nilai rata-ratanya. Hal ini dilakukan untuk melihat seberapa baik dan cukup kuat dan konsisten hasil akurasi yang diperoleh, sekiranya dilakukan percobaan lebih dari satu kali. Hasil dari ketiga algoritma dapat dilihat pada tabel 4.

Tabel 4. Hasil pengukuran akurasi menggunakan *cross validation* untuk setiap model

No	Algoritma	Hasil akurasi untuk 10 kali percobaan	Akurasi rata-rata – (Max) – (Min) (%)
1	Naïve Bayes	[0.53835616 0.5109589 0.51917808 0.49452055 0.50547945 0.53561644 0.53287671 0.53424658 0.53287671 0.5109589]	52,15 – (53,84) – (49,45)
2	Random Forest	[0.72328767 0.71780822 0.71369863 0.76438356 0.72328767 0.76438356 0.75890411 0.74109589 0.73835616 0.70958904]	73,55 – (76,44) – (70,96)
3	Support Vector Machine	[0.58767123 0.59178082 0.58356164 0.60410959 0.57671233 0.57534247 0.58356164 0.59589041 0.58219178 0.58630137]	58,67 – (60,41) – (57,53)

Dari tabel 4 dapat dilihat terdapat konsistensi hasil yang diperoleh dengan menggunakan satu kali percobaan saat dibandingkan dengan yang menggunakan 10 kali percobaan melalui *cross validation*. Dapat disimpulkan artinya, masing-masing algoritma cukup stabil dan tangguh untuk mendapatkan hasil yang lebih kurang sama. Dari hasil percobaan, algoritma *Random Forest* tetap memberikan nilai akurasi rata-rata yang paling tinggi, yaitu 73,55%. Walaupun lebih sedikit rendah dibandingkan hasil akurasi yang dilakukan pada percobaan sebelumnya (75,52%), tetapi terdapat satu kali percobaan dengan hasil melampaui nilai akurasi dari percobaan sebelumnya, yaitu dengan nilai 76,44% (nilai tertinggi yang diperoleh dalam 10 kali percobaan).

3.7 Penentuan Fitur yang Paling Berpengaruh

Tahapan akhir yang dilakukan dalam penelitian ini adalah pemilihan fitur. Hal ini dilakukan untuk menentukan fitur mana yang paling memberikan pengaruh dan paling tidak memberikan pengaruh kepada keputusan atau label. Informasi ini dapat digunakan untuk memilih fitur fitur yang memberikan pengharuh saja untuk diproses dalam pembentukan model. Selain akan mengurangi waktu proses karena fitur yang digunakan lebih sedikit, juga diharapkan dapat meningkatkan akurasi dari model. Proses ini dilakukan dengan menggunakan fungsi *SelectKBest* dan fungsi *chi2* dari *library sklearn*. Hasil dari percobaan yang dilakukan menunjukan bahwa dari 8 fitur yang digunakan, seperti diperlihatkan pada gambar 16, fitur yang paling memberikan pengaruh adalah fitur penghasilan, seperti diperlihatkan pada gambar 17. Sedangkan fitur yang paling tidak memberikan pengaruh adalah fitur matematika (mtk), seperti diperlihatkan pada gambar 18.

```
All features:
Index(['pai', 'pkn', 'bhs_indo', 'mtk', 'bhs_inggris', 'pekerjaan', 'kelas',
       'penghasilan'],
      dtype='object')
```

Gambar 16. Semua fitur yang digunakan dalam proses klasifikasi

```
7 [0 1 2 4 5 6 7]
**********************************************************************
Index(['pai', 'pkn', 'bhs_indo', 'bhs_inggris', 'pekerjaan', 'kelas',
       'penghasilan'],
      dtype='object')
Exclude Index(['mtk'], dtype='object')
**********************************************************************
```

Gambar 17. Fitur yang paling tidak berpengaruh, matematika (mtk)

```
1 [7]
**********************************************************************************************
Index(['penghasilan'], dtype='object')
Exclude Index(['pai', 'pkn', 'bhs_indo', 'mtk', 'bhs_inggris', 'pekerjaan', 'kelas'], dtype='object')
**********************************************************************************************
```

Gambar 18. Fitur yang paling berpengaruh, penghasilan

Dapat disimpulkan, fitur matematik yang paling tidak memberikan pengaruh, hal ini terlihat dari gambar 17, pada saat hanya dipilih 7 fitur dari 8 fitur yang ada, fitur matematika tidak terpilih. Sedangkan fitur penghasilan adalah fitur yang paling berpengaruh, terlihat pada gambar 18, dimana fitur penghasilan adalah fitur yang terpilih pada saat hanya satu (1) fitur yang diperlukan.

4. KESIMPULAN

Dalam penelitian ini telah dilakukan proses pemodelan untuk memprediksi data siswa yang mendapat beasiswa dan data yang tidak mendapat beasiswa. Mulai dari proses pengumpulan data telah dijelaskan proses yang perlu dilakukan sampai dengan didapatkan hasil akurasi bagi tiga algoritma yang digunakan dalam penelitian ini. Akurasi yang paling tinggi diperoleh oleh algoritma *Random Forest*.

Selanjutnya, dengan menggunakan metode *Random Forest*, telah dilakukan percobaan untuk mengetahui fitur mana yang paling memberikan pengaruh dan fitur mana yang paling tidak memberikan pengaruh kepada keputusan atau label. Dari hasil pemilihan fitur dapat dapat diketahui bahwa fitur penghasil adalah fitur yang paling memberikan pengaruh, sedangkan fitur matematika adalah fitur yang paling tidak memberikan pengaruh. Dengan informasi ini dapat dilakukan pembentukan model dengan lebih cepat karena tidak semua fitur harus diikutkan dalam proses dan dapat memberikan hasil akurasi yang lebih baik.

REFERENCES

[1] M. Pendidikan, dan Kebudayaan Republik, and Indonesia, “Permen-10-Tahun-2020 PIP,” 2020.

[2] N. E. Rohaeni and O. Saryono, “Implementasi Kebijakan Program Indonesia Pintar (PIP) Melalui Kartu Indonesia Pintar (KIP) dalam Upaya Pemerataan Pendidikan,” *J. Educ. Manag. Adm. Rev.*, vol. 2, no. 1, pp. 193–204, 2018.

[3] A. Anggleni, “Implementasi Kebijakan Program Kartu Keluarga Sejahtera (KKS) dalam Meningkatkan Kesejahteraan Masyarakat Miskin di Kelurahan Sekip Jaya Kecamatan Kemuning Kota Palembang,” *J. PPS UNISTI*, vol. 1, no. 1, pp. 24–39, 2018.

[4] D. Utomo, A. Hakim, and H. Ribawanto, “Pelaksanaan Program Keluarga Harapan dalam Meningkatkan Kualitas Hidup Rumah Tangga Miskin (Studi pada Unit Pelaksana Program Keluarga Harapan Kecamatan Purwoasri, Kabupaten Kediri),” *Jap*, vol. 2, no. 1, pp. 29–34, 2013.

[5] P. Suprastowo, "Contributions of Students Aid Program Towards Sustainability and Continuity of Students' Education," *J. Pendidik. dan Kebud.*, vol. 20, no. 2, pp. 149–172, 2014.

[6] L. N. Saraswati, “Implementasi Kebijakan Program Indonesia Pintar (Pip) Pada Jenjang Sekolah Dasar Di Kecamatan Sungai Pinang Kota Samarinda,” *Adm. Negara*, vol. 5, no. 4, pp. 6738–6749, 2017.

[7] Y. Farida and N. Ulinnuha, “Klasifikasi Mahasiswa Penerima Program Beasiswa Bidik Misi Menggunakan Naive Bayes,” *Syst. Inf. Syst. Informatics J.*, vol. 4, no. 1, pp. 17–22, 2018.

[8] R. A. Saputra and S. Ayuningtias, “Penerapan Algoritma Naive Bayes Untuk Penentuan Calon Penerima Beasiswa Pada Smk Pasim Plus Sukabumi,” *Swabumi*, vol. IV, no. 2, pp. 114–120, 2016.

[9] J. N. Utamajaya, A. Mentari, and S. Masnunah, “Penerapan Algoritma Naive Bayes Untuk Penentuan Calon Penerima Beasiswa PIP Pada SDN 023 Penajam,” *J. Sist. Inf.*, vol. 3, no. 1, pp. 11–17, 2019.

[10] M. I. Firdaus and M. G. L. Putra, “Seleksi Beasiswa Bidik Misi Uniska Mab Banjarmasin Hibah Lldikti Xi

Kalimantan Menggunakan Metode Svm Dan Topsis," *Al Ulum J. Sains Dan …*, pp. 1–6, 2020.

[11] W. A. Damanik and Prihandoko, "Analisis Penentuan Pemberian Beasiswa Berprestasi Menggunakan Metode Decision Tree dan SVM (Support Vector Machine)," *J. Tek. Dan Inform.*, vol. 6, pp. 2018–2020, 2019.

[12] I. Hayati, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Decision Tree C4 . 5 Dan Naive Bayes Di Universitas Jambi," p. 115032, 2021.