PREDIKSI INTENTION TO ENROLL OLEH SISWA-SISWI SMA DALAM PEMILIHAN PERGURUAN TINGGI SWASTA DI INDONESIA DENGAN MENGGUNAKAN NAIVE BAYES

Hendra Achmadi

Universitas Pelita Harapan, Banten, Indonesia

Hendra.achmadi@uph.edu

ABSTRAK

Jumlah calon mahasiswa yang mendaftar di perguruan tinggi, terutama di perguruan tinggi swasta, merupakan masalah serius. Penurunan jumlah calon mahasiswa yang terjadi selama masa pandemi COVID-19 dari tahun 2018-2019 hingga tahun 2022 juga menjadi masalah serius bagi perguruan tinggi swasta di Indonesia. Oleh karena itu, penelitian ini berfokus pada mencari karakteristik utama siswa SMA dalam memilih perguruan tinggi swasta di Jakarta dan sekitarnya. Metode penelitian yang digunakan adalah data mining, dengan menggunakan data primer yang diperoleh dari kuesioner yang disebar kepada siswa SMA kelas 11 dan 12 di wilayah tersebut, dengan total 438 responden, yang kemudian melalui proses pembersihan data, menghasilkan 295 responden.Dengan menggunakan metode Naise Bias dan Supervisor Learning akan didapat prediksi pengambilan keputusan siswa-siswi SMA Dengan menggunakan algoritma naïve bayes maka didapat akurasi 94.9 persen dengan melakukan simulasi terhadap tiga universitas besar yang akan diprediksi yang dipilih yaitu Binus dengan kode 1 dan UPH dengan kode 2 dan Prasmul dengan kode 3. Kemudian dengan menggunakan Naïve Bayes algoritma , maka machine learning akan melakukan prediksi dari universitas yang akan di pilih, dan hasilnya adalah 94,9 persen akurasi yang didapatkan

Kata Kunci: Prediksi Intention to Enroll , Data Mining, Naise Bayes

2. LATAR BELAKANG

Dalam tiga tahun terakhir, sektor pendidikan di Indonesia mengalami perkembangan yang signifikan. Terbukti dengan meningkatnya jumlah lembaga pendidikan sebanyak 42,55% dari tahun 2017 ke 2018. Di Provinsi Banten, terjadi peningkatan sebesar 38,65% di periode yang sama. Peningkatan ini terjadi karena adanya penambahan jenis lembaga pendidikan, seperti politeknik, sehingga memberikan lebih banyak opsi bagi lulusan SMA. Oleh karena itu, lembaga pendidikan tinggi harus bersaing dalam menarik minat mahasiswa baru dari lulusan SMA. Hal ini diharapkan berlangsung hingga tahun 2022.

Persaingan antara institusi pendidikan tinggi akan mendorong pengembangan strategi pemasaran yang lebih efektif untuk menarik minat siswa SMA dan meyakinkan mereka untuk memilih universitas atau politeknik sebagai pilihan mereka setelah lulus. Di Indonesia, ada dua jenis institusi pendidikan tinggi, yaitu universitas dan politeknik, sehingga lulusan SMA memiliki banyak opsi dalam melanjutkan pendidikan.

Bagi perguruan tinggi, jumlah mahasiswa baru sangat penting untuk mempertahankan eksistensinya di industri pendidikan tinggi. Calon mahasiswa di perguruan tinggi kebanyakan adalah siswa-siswi SMA berusia 17 dan 18 tahun, yang termasuk dalam Generasi Z. Menurut penelitian Kusumaningtyas et al. (2020), Generasi Z memiliki kemampuan literasi teknologi yang baik. Selain itu, gaya hidup Generasi Z memiliki kekhasan dalam pengambilan keputusan, salah satunya adalah dengan melakukan online window shopping, seperti yang disebutkan dalam penelitian Santoso Geovani dan Anna (Santoso & Triwijayati, 2018).

Oleh karena itu dibutuhkan satu cara untuk dapat memprediksi keputusan pembelian dari para siswa-siswi SMA berdasarkan data historis dengan menanyakan nama universitas mana yang pertama kali teringat oleh siswa siswi SMA maka dapat dilakukan prediksi universitas mana yang akan dipilih nantinya dengan menggunakan algoritma naïve bayes.

2. LANDASAN TEORI

2.1 Langkah Proses Data Mining

Proses Data Mining dilakukan dengan persiapan data dan dilanjutkan dengan data pemrosesan atau pembersihan data, di sini persiapan data dimulai untuk diproses lebih lanjut, untuk contoh apakah data memiliki jenis nomor atau faktor atau tanggal, dan kemudian data dalam data pembersihan juga dilakukan dengan menghilangkan karakter khusus, kemudian setelah itu dilakukan transformasi dilakukan yaitu mengubah data dari cleansing data menjadi data target yaitu proses selanjutnya adalah melakukan data mining atau model data berdasarkan metode yang cocok untuk data tersebut, dan yang terakhir adalah proses interpretasi pengetahuan yang diperoleh dari pengolahan data.(Jiawei, 2012)

2.2. Naïve Bayes

Menurut (Rish, 2000)Naïve bayes dapat digunakan untuk mengklasifikasikan keputusan yang akan dibentuk. Hal ini juga didukung oleh penelitian yang dilakukan oleh (Muzumdar et al., 2022) dimana Muzumdar melakukan pengklasifikasian Kesehatan mental dari para siswa dengan menggunakan machine learning, dan hal ini juga dilakukan oleh (Mansoor, 2022) yang membandingkan algoritma naïve bayes dengan k-Nearest Neigbors, dimana Naïve bayes lebih akurat dibandingkan dengan k-nearest neighbor yang hanya dapat memberikan perkiraan persentase kemiripan terhadap satu klasifikasi tertentu saja.

2.3 Supervisor Learning

Menurut (Duarte et al., 2019) teknik klasifikasi biasanya adalah program komputer yang belajar dari data input yang diberikan, dan menggunakan data pelatihan ini dengan tujuan untuk belajar mengklasifikasikan berdasarkan pola pengamatan pada data tersebut. Di sisi lain pembelajaran terawasi untuk regresi adalah seperangkat algoritma yang digunakan untuk memprediksi nilai kontinu.

Menurut (Charbuty & Abdulazeez, 2021), Algoritma DT adalah bagian dari keluarga algoritma pembelajaran yang diawasi, dan tujuan utamanya adalah untuk membangun model pelatihan yang dapat digunakan untuk memprediksi kelas atau nilai variabel target melalui aturan keputusan pembelajaran yang disimpulkan. dari data pelatihan.

Menurut (Müller & Guido, 2017) Pembelajaran yang diawasi adalah jenis pembelajaran mesin di mana algoritme belajar dari kumpulan data berlabel untuk membuat prediksi atau keputusan tentang data baru yang tidak terlihat. Dalam pembelajaran terawasi, algoritme dilatih pada sekumpulan data input dan data output yang sesuai, juga dikenal sebagai label. Algoritme belajar untuk memetakan data input ke data output dengan menggeneralisasi pola dalam data pelatihan. Pembelajaran terbimbing menjadi area untuk banyak aktivitas penelitian dalam pembelajaran mesin. Banyak dari teknik pembelajaran yang diawasi telah menemukan penerapannya dalam pemrosesan dan analisis berbagai data

3. METODOLOGI

Penelitian ini bersifat kuantitatif, yang pertama adalah untuk mengetahui gambaran dari setiap profil pelanggan yang tertunda yang diambil melalui kuesioner kepada 202 responden dengan menggunakan Google form, kemudian dilakukan pengolahan data dan pembersihan data dengan menggunakan metode data mining, sehingga dapat diketahui dari lima belas karakteristik atau fitur, dimana dari fitur tersebut penting untuk menentukan keputusan, dengan menggunakan algoritma random forest, akan menggunakan algoritma pohon keputusan. Untuk membuat perhitungan algoritma random forest menggunakan program python, dan untuk membuat algoritma pohon keputusan menggunakan python juga.

4. HASIL

Data Preperation

Data diambil data primer dari questioner yang dibagikan kepada para siswa siswa SMA kelas 11 dan 12 di daerah Jakarta dan sekitarnya dengan menggunakan google form, dan didapat 438 responden , dan kemudian dilakukan data cleansing dan tersisa 295 responden

	SMA	TIPESKS	SEX	GRADE	JURUSAN	DOMISILI	UANGSAKU	TRANSPORT	BIMBEL	PEKERJAAN	DIDIKPP	DIDIKIBU	SOCIALMEDIA	PRESENTASI	UNIV1
0	UPH College	SMS SWASTA	PRIA	XI	IPA	DKI Jakarta	1-3 Juta	Antar jemput dengan sopir	Ya	Wiraswasta	S2	S1	Ya	Ya	UPH
1	UPH College	SMS SWASTA	WANITA	XII	IPA	DKI Jakarta	1-3 Juta	Mobil	Ya	Dosen/Guru	S1	S1	Ya	Ya	UPH
2	UPH College	SMA NEGERI	WANITA	XII	IPS	DKI Jakarta	1-3 Juta	Antar jemput dengan sopir	Ya	Dosen/Guru	S1	S1	Ya	Ya	ITB
3	SMA 2	SMS SWASTA	WANITA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	1-3 Juta	Mobil	Tidak	Wiraswasta	S3	S1	Ya	Ya	UPH
4	SMA 2	SMA SWASTA	PRIA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Sepeda	Tidak	Wiraswasta	SMA	SMA	Ya	Ya	UPH
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
289	SMAI PB SOEDIRMAN 1 BEKASI	SMA NEGERI	PRIA	XI	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	Karyawan Swasta	SMA	SMA	Ya	Ya	UI
290	SMAK 5 Penabur	SMA SWASTA	WANITA	XII	IPA	Jawa	1-3 Juta	Mobil	Ya	Wiraswasta	S1	SMA	Ya	Ya	PRASMUL
291	SMAK 5 Penabur	SMA NEGERI	PRIA	XI	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	Karyawan Swasta	SD	sederajat (SMP)	Ya	Ya	UGM
292	SMAI PB SOEDIRMAN 1 BEKASI	SMA NEGERI	WANITA	XII	IPS	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	PNS	SMP	SMA	Ya	Ya	PTS
293	SMA Kolese Kanisius	SMA NEGERI	PRIA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Motor	Ya	Wiraswasta	S1	S1	Ya	Ya	IPB

294 rows × 15 columns

Gambar 1. Data Sumber dari Data Primer

Encoding

Data yang sudah didapat kemudian dilakukan proses encoding , atau merubah dari tipe data kategorikal ke numberik, supaya dapat diproses selanjutnya

```
from sklearn.preprocessing import LabelEncoder
enc= LabelEncoder()
data['SMA']=enc.fit_transform(data['SMA'].values)
data['TIPESKS']=enc.fit_transform(data['TIPESKS'].values)
data['SEX']=enc.fit_transform(data['SEX'].values)
data['GRADE']=enc.fit_transform(data['GRADE'].values)
data['JURUSAN']=enc.fit_transform(data['JURUSAN'].values)
data['DOMISILI']=enc.fit_transform(data['DOMISILI'].values)
data['UANGSAKU']=enc.fit_transform(data['UANGSAKU'].values)
data['TRANSPORT']=enc.fit_transform(data['TRANSPORT'].values)
data['BIMBEL']=enc.fit_transform(data['BIMBEL'].values)
data['PEKERJAAN']=enc.fit_transform(data['PEKERJAAN'].values)
data['DIDIKPP']=enc.fit_transform(data['DIDIKPP'].values)
data['DIDIKIBU']=enc.fit_transform(data['DIDIKIBU'].values)
data['SOCIALMEDIA']=enc.fit_transform(data['SOCIALMEDIA'].values)
data['PRESENTASI']=enc.fit_transform(data['PRESENTASI'].values)
data['UNIV1']=enc.fit_transform(data['UNIV1'].values)
data['PILIH']=enc.fit_transform(data['PILIH'].values)

✓ 0.0s
```

Gambar 2. Encoding proses

Hasil dari proses encoding dapat di lihat di gambar 3. Hasil proses encoding

```
data
✓ 0.1s
```

	SMA	TIPESKS	SEX	GRADE	JURUSAN	DOMISILI	UANGSAKU	TRANSPORT	BIMBEL	PEKERJAAN	DIDIKPP	DIDIKIBU	SOCIALMEDIA	PRESENTASI	UNIV1	PILIH
0	55	3	0	1	3	0	0	4	2	32	6	4	1	1	42	2
1	55	3	2	2	3	0	0	20	2	7	5	4	1	1	42	2
2	55	1	2	2	4	0	0	4	2	7	5	4	1	1	10	0
3	8	3	2	2	3	4	0	20	1	32	7	4	1	1	42	2
4	8	2	0	2	3	4	2	35	1	32	10	7	1	1	42	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
289	32	1	0	1	3	4	2	26	1	13	10	7	1	1	31	0
290	36	2	2	2	3	2	0	20	2	32	5	7	1	1	20	3
291	36	1	0	1	3	4	2	26	1	13	8	13	1	1	30	0
292	32	1	2	2	4	4	2	26	1	16	12	7	1	1	21	0
293	16	1	0	2	3	4	2	21	2	32	5	4	1	1	9	0

294 rows × 16 columns

Gambar 3. Hasil Proses Encoding

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.tree import DecisionTreeClassifier
import sklearn.model_selection as ms
X=data[['SMA','TIPESKS','SEX','GRADE','JURUSAN','DOMISILI','UANGSAKU','TRANSPORT','BIMBEL','PEKERJAAN','DIDIKPP','DIDIKIBU','SOCIALMEDIA','PRESENTASI']]
X
y=data['PILIH']
y
```

[392] ✓ 0.1s

```
0      2
1      2
2      0
3      2
4      2
      ..
289    0
290    3
291    0
292    0
293    0
Name: PILIH, Length: 294, dtype: int64
```

```
from sklearn.naive_bayes import GaussianNB
X_train, X_test, y_train,y_test=ms.train_test_split(atr_data,cls_data,test_size=0.2)
tree_data=GaussianNB()
tree_data.fit(X_train,y_train)
print(tree_data.score(X_test,y_test))
```

5] ✓ 0.0s

```
0.9491525423728814
```

Gambar 4. Proses Supervisory learning

Proses selanjutnya adalah dengan menggunakan proses Supervisory learning, dimana data akan dibagi menjadi 2 yaitu file training dan file testing, kemudian baru dijalankan algoritma logistic regression.

Gambar 5. Intercept dan Slope dari Model Logistic Regression

Dari gambar 5 akan didapat intercept dan slope untuk membuat model dari Logistic Regression

No	Actual	Prediction	Hasil
108	**0**	**0**	**1**
167	**1**	**1**	**1**
267	**0**	**0**	**1**
22	**2**	**2**	**1**
10	**2**	**2**	**1**

193	**1**	**1**	**1**
118	**0**	**0**	**1**
218	**1**	**1**	**1**
66	**0**	**0**	**1**
171	**0**	**0**	**1**
262	**0**	**0**	**1**
84	**0**	**0**	**1**
53	**0**	**0**	**1**
176	**1**	**1**	**1**
242	**0**	**0**	**1**
51	**0**	**0**	**1**
63	**0**	**0**	**1**
279	**1**	**1**	**1**
91	**3**	**0**	**0**
147	**0**	**0**	**1**
14	**0**	**0**	**1**
287	**0**	**0**	**1**
205	**0**	**0**	**1**
25	**1**	**1**	**1**
73	**1**	**0**	**0**
128	**1**	**0**	**0**
27	**2**	**2**	**1**
282	**3**	**3**	**1**
23	**2**	**2**	**1**
137	**0**	**0**	**1**
101	**2**	**2**	**1**
221	**1**	**1**	**1**
269	**0**	**0**	**1**

129	**1**	**1**	**1**
133	**1**	**1**	**1**
105	**0**	**0**	**1**
261	**0**	**0**	**1**
272	**0**	**0**	**1**
94	**1**	**1**	**1**
200	**1**	**1**	**1**
100	**0**	**0**	**1**
263	**0**	**0**	**1**
55	**2**	**2**	**1**
115	**1**	**1**	**1**
52	**2**	**2**	**1**
240	**0**	**0**	**1**
164	**2**	**2**	**1**
225	**1**	**1**	**1**
146	**1**	**1**	**1**
86	**0**	**0**	**1**
40	**0**	**0**	**1**
180	**2**	**2**	**1**
248	**2**	**2**	**1**
106	**0**	**0**	**1**
79	**0**	**0**	**1**
145	**1**	**1**	**1**
223	**3**	**3**	**1**
64	**0**	**0**	**1**
207	**0**	**0**	**1**
			56
		Akurasi	**0,949153**

5. PEMBAHASASN

Dengan menggunakan metode naïve bayes maka akan didapat akurasi 94.9 persen dengan melakukan simulasi terhadap tiga universitas besar yang akan diprediksi yang dipilih yaitu Binus dengan kode 1 dan UPH dengan kode 2 dan Prasmul dengan kode 3. Kemudian dengan menggunakan Naïve Bayes algoritma , maka machine learning akan melakukan prediksi dari universitas yang akan di pilih, dan hasilnya adalah 94,9 persen akurasi yang didapatkan. Sehingga dengan menggunakan kata kunci universitas apa yang muncul pertama kali dalam pemikiran siswa siswa SMA secara historis dapat mempredik calon mahasiswa selanjutnya akan melakukan pemilihan perguruan tinggi.

6. KESIMPULAN

Dengan menggunakan algoritma naïve bayes maka didapat akurasi 94.9 persen dengan melakukan simulasi terhadap tiga universitas besar yang akan diprediksi yang dipilih yaitu Binus dengan kode 1 dan UPH dengan kode 2 dan Prasmul dengan kode 3. Kemudian dengan menggunakan Naïve Bayes algoritma , maka machine learning akan melakukan prediksi dari universitas yang akan di pilih, dan hasilnya adalah 94,9 persen akurasi yang didapatkan.

REFERENCES

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Duarte, V., Zuñiga-Jara, S., & Contreras, S. (2019). *Machine Learning and Marketing: a literature review*. https://ssrn.com/abstract=4006436

Jiawei. (2012). *Data Mining Third Edition*.

Mansoor, F. (2022). *Increasing Generalizability: Naïve Bayes Vs K-Nearest Neighbors*. https://doi.org/10.21203/rs.3.rs-1578985/v1

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python*.

Muzumdar, P., Prasad Basyal, G., & Vyas, P. (2022). An Empirical Comparison of Machine Learning Models for Student's Mental Health Illness Assessment. In *Asian Journal of Computer and Information Systems* (Vol. 10, Issue 1). www.ajouronline.com

Rish, I. (2000). *An empirical study of the naive Bayes classifier*.