Evaluasi Kinerja Datamining Pada Dataset Pendaftaran Mahasiswa Baru Dengan Class Yang Tidak Seimbang

Herianto[1], Nur Syamsiyah[2], Adam Arif B[3], Yahya[4]

[1,3]Program Studi Teknologi Informasi Universitas Darma Persada
[2,4]Program Studi Sistem Informasi Universitas Darma Persada

Jl. Taman Malaka Selatan Pondok Kelapa Jakarta Timur 13450

E-mail : heri.unsada@gmail.com[1], nurs.syamsiyah@gmail.com[2], ariadam@gmail.com[3], yahya.unsada@gmail.com [4]

ABSTRAK

Topik penelitian ini di bidang EDM (*Educational Datamining*) yang bertujuan memanfaatkan datamining dalam memperoleh informasi yang lebih bernilai dari database akademik Perguruan Tinggi. Penggunaan data pendaftaran mahasiswa baru karena karakteristik datanya yang umumnya tidak seimbang (*imbalanced class*) sehingga dapat digunakan untuk menguji dan membandingkan kinerja model pembelajaran *machine learning*. Perkakas yang digunakan adalah Jupyter sebagai editor, Python sebagai bahasa pemrogramannya dan Library sklearn sebagai modul terpopular di bidang *machine learning*. Metodologi penelitian mengacu kepada CRISP-DM sebagai metodologi yang bersifat terbuka. Percobaan dilakukan menggunakan 7491 data, 5 kolom sebagai fitur dan 1 kolom sebagai target. Kelas bernilai 1 sebanyak 6197 (82,5%) dan yang bernilai 0 sebanyak 1312 (17,5%). Kemudian dibangun model klasifikasi dengan pembelajaran berbeda yaitu : SVM (*Support Vector Machine*), *Logistic Regression*, *Decision Tree*, *Naive Bayes*, dan *K-Nearest Neighbors* (K-NN). Dari hasil percobaan diperoleh rata-rata akurasi semua model sebesar 0,81 dan nilai rata-rata F Score 0,46. Nilai Akurasi tertinggi 0,82, dan akurasi terendah 0,81. Nilai F Score tertinggi 0,51 dan nilai F Score terendah adalah 0,44. Hasil ini kembali mengungkap bahwa pada kasus dengan komposi data target yang tidak seimbang (*Imbalanced Classes*) memungkinkan untuk menghasilkan akurasi yang baik tetapi tidak menjamin nilai F Score baik.

Kata kunci : EDM, Imbalanced Classes, Sklearn, Klasifikasi, Akurasi, F Score

ABSTRACT

The topic of this research is in the field of EDM (Educational Datamining), utilizing datamining in obtaining more valuable information from university academic databases. The use of new student registration database because the characteristics of the data are generally imbalanced class and suitable to be used to test and compare the performance of machine learning models. The tools used are Jupyter as an editor, Python as a programming language and the Sklearn Library as

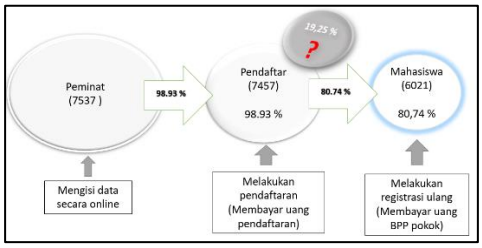
a popular module in the field of machine learning. This research methodology refers to CRISP-DM as an open methodology. The experiment was carried out using 7491 datasets, 5 columns as features and 1 column as targets, class 1 was 6197 (82.5%) and 0 was 1312 (17.5%). Classification modeling was carried out with several lessons: Support Vector Machine (SVM), Logistic Regression, Decision Tree, Naive Bayes, and K-Nearest Neighbors (K-NN). From the evaluation, it was found that the average accuracy of the model was 0.81 and the average value of the F Score was 0.46. The highest accuracy value is 0.82 and the lowest accuracy value is 0.81, while the highest F Score value is 0.51 and the lowest F Score value is 0.44. The data above shows that the composition of the unbalanced target data makes it possible to produce good accuracy but does not guarantee a good F Score value.

Keywords : EDM, Imbalanced Classes, Sklearn, Classification, Accuracy, F Score

1. PENDAHULUAN

Jumlah pendaftaran mahasiswa baru pada Perguruan Tinggi swasta umumnya lebih banyak yang diterima dibandingkan dengan yang tidak diterima. Status calon mahasiswa (pendaftar) yang diterima dan yang tidak diterima tersebut merupakan data yang tidak seimbang (*imbalanced data*). Pada model *machine learning* dengan pembelajaran yang terbimbing (*supervised learning*) atribut status calon mahasiswa baru tersebut umumnya dijadikan sebagai atribut target atau class.

Gambar berikut memperlihatkan jumlah calon mahasiswa pada setiap proses pendaftaran mahasiswa di salah satu perguruan tinggi swasta di Jakarta :



Gambar 1. Jumlah Pendaftar dan Mahasiswa

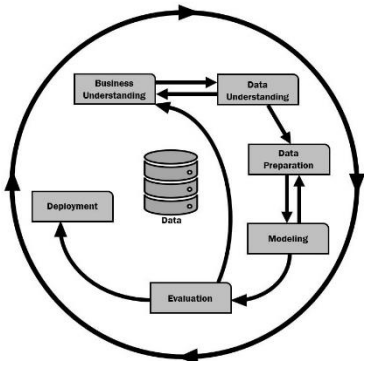
Data di atas memperlihatkan sebuah kasus dimana perbandingan pendaftaran mahasiswa baru yang diterima sebesar 80,74% dibandingkan dengan jumlah mahasiswa yang tidak diterima dengan alasan tertentu sebesar 19,25%

Penelitian di bidang EDM (*Educational Datamining*) seringkali menggunakan database akademik Perguruan Tinggi termasuk data calon mahasiswa baru bagian pendaftaran (*admission*) ini.

Sebelum dataset pendaftaran mahasiswa baru digunakan lebih jauh dalam menghasilkan pengetahuan (informasi yang lebih bernilai), maka perlu diteliti sejauhmana ketidakseimbangan dataset tersebut mempengaruhi kinerja algoritma *machine learning* sebagai proses datamining.

2. METODOLOGI

Tahapan penelitian ini mengacu ke metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) seperti berikut :

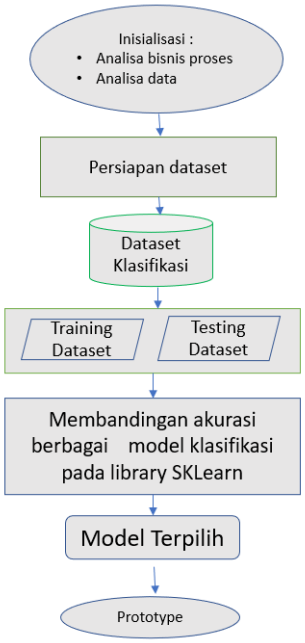


Gambar 2 Metodologi CRISP-DM

Metodologi CRISP-DM merupakan standar proses datamining yang bersifat terbuka, terdiri dari 6 tahap : (1) Business Understanding, (2) Data Understanding, (3)

Data Preparation, (4) Modelling, (5) Evaluation, dan (6) Deployment.

Lebih spesifik penelitian kasus ini menggunakan tahapan seperti berikut :



Gambar 3. Metodologi Berdasarkan Kasus

Seperti terlihat pada diagram alir di atas, proses utama yang dilakukan pada penelitian ini adalah tahap pemilihan data dari database akademik untuk memperoleh dataset yang paling sesuai yang digunakan oleh model klasifikasi.

Perkakas (*Tool*) yang digunakan adalah Library sklearn pada python. Output yang dihasilkan berupa prototype model klasifikasi dan hasil evaluasinya.

3. LANDASAN TEORI

3.1 EDM (Educational Datamining)

Penelitian di bidang EDM (*Educational Datamining*) merupakan bidang datamining dengan dataset yang bersumber dari database akademik terutama data perguruan tinggi. Penelitian ini muncul sebagai bidang baru yang semakin popular.(Saleh et al., 2021). Selanjutnya (Bai et al., 2021) menyebut bidang ini sebagai *educational big data* yang dapat dimanfaatkan untuk menghasilkan berbagai aplikasi di dunia Pendidikan baik di bagian administrasi, inovasi mengajar dan bentuk riset lainnya.

Sampai saat ini belum ditemui paper yang secara khusus meneliti dataset dari database pendaftaran mahasiswa baru (*admission*) khususnya membahas perbedaan jumlah pendaftar dengan mahasiswa yang memunculkan fenomena tidak seimbang (*imbalanced class dataset*).

3.2 Data Mining

Datamining adalah rentetan proses untuk memperoleh nilai tambah pada bentuk yang sebelumnya tidak diketahui informasinya dari database (Pattiasina et al., 2020). Datamining juga merupakan bagian dari *data science* dan *Artificial Intelligence* (AI). *Machine learning* merupakan salah satu Teknik dari datamining.

3.3 Library Sklearn Pemrograman Python

Python merupakan Bahasa pemrograman yang paling popular saat ini terutama karena banyaknya jumlah library yang tersedia. Salah satu library yang banyak digunakan untuk penerapan *machine learnin*g adalah scikit-learn atau biasa juga disebut sklearn. *Scikit-learn* merupakan modul Python yang terintegrasi luas pada algoitma machine learning terkini yang digunakan untuk skala menengah baik pembelajaran supervised maupun unsupervised (Pedregosa et al., 2011).

Supervised Learning sendiri berarti model pembelajaran yang terbimbing oleh atribut target (class), dan Unsupervised Learning merupakan model pembelajaran yang tidak terbimbing oleh atribut target bahkan pada awalnya tidak memiliki atribut target.

3.4 Confusion Matrix, Akurasi dan F Score

Confusion Matrix adalah tabel yang sering digunakan untuk mendeskripsikan kinerja model klasifikasi berdasarkan nilai prediksi (*predicted*) dan nilai sebenarnya (*actual*) yang diketahui.

Tabel 1. Confusion Matrix

	Predicted-P	**Predicted-N**
Actual-P	TP	FN
Actual-N	FP	TN

Akurasi adalah jumlah nilai prediksi yang benar dibagi dengan total data yang diprediksi.

F-Score atau kadang disebut juga F-measure merupakan ukuran yang banyak digunakan saat menguji machine learning. F score merupakan rata-rata harmonik dari nilai Recall dan Precision (Powers, 2019).

4. HASIL DAN PEMBAHASAN

Dataset pendaftaran mahasiswa baru diperoleh dari database akademik salah satu universitas swasta di Jakarta, memiliki 56 variabel/atribut, seperti berikut :

```
idperiode, idgelombang,idpendaftar,
nim,namapendaftar,jk,idagama,agama,
idkotatinggal,kotatinggal,idpropinsiti
nggal, propinsitinggal, idgaji_ortu,
gajiortu, tmplahir, tgllahir,
idtransport, namatransport,
idpropinsiasalsekolah,
propinsiasalsekolah,idkotaasalsekolah,
kotaasalsekolah, idjenissekolah,
namajenissekolah, jurusanasalsekolah,
nem, kodeprodi, prodi,
idjalurpendaftaran,
namajalurpendaftaran, idsistemkuliah,
namasistemkuliah, jalurmendaftar,
bebastes, islulus, alihjenjang,
idpend_ayah, gaji_ayah, instansi_ayah,
idkerja_ayah, idpend_ibu,
idgaji_ibu, instansi_ibu, idkerja_ibu,
idmendapat_info, sumberinfo,
isdaftarulang, tgladministrasi,
tgldaftarulang,tglawalpendaftaran,
tglakhirpendaftaran, tglregistrasi,
gdb, namapromo_ipmb, sedang_promo,
namamasa_daftar_ptn, sedang_ptn
```

Selanjutnya dilakukan seleksi fitur menggunakan berbagai tool pada python sehingga tersisa 25 atribut/variabel terpenting berikut :

```
idperiode, idgelombang, jeniskelamin,
idagama, tgllahir,, idkotatinggal,
idkotaasalsekolah, idjenissekolah,
namatransport, nem, kodeprodi,
idjalurpendaftaran, idsistemkuliah,
jalurmendaftar, bebastes, alihjenjang,
idpend_ayah,    idpend_ibu,
idmendapat_info, tglregistrasi,
ekonomi_nasional,        sedang_promo,
sedang_ptn, islulus, gaji_ortu,
isdaftarulang
```

Dari 25 atribut di atas dilakukan lagi pemilihan variabel berdasarkan pemahaman di kasusnya (*domain knowledge*). Dari pemilihan atribut terkait permasalahan kasus dan kebutuhan model klasifikasi yang akan digunakan, maka ditentukan 6 variabel, yang terdiri dari 5 variabel sebagai atribut penentu/fitur dan 1 variabel sebagai target/class.

Variabel fitur dan variable target/class tersebut dengan deskripsinya adalah sebagai berikut :

Variabel penentu (fitur) :

- `ekonomi_nasional` : merupakan nilai income per kapita nasional yang dilakukan proses scalling sebanyak 5 bins sehingga bernilai : 1- Sangat rendah , 2-Rendah , 3-cukup, 4-Baik, 5-Sangat baik
- `sedang_ptn` : bernilai 0 dan 1, nilai 0 menyatakan pada saat tersebut sedang tidak ada pendaftaran PTN (Perguruan tinggi negeri) dan 1 menyatakan sdang ada pendaftaran PTN
- `sedang_promo` : merupakan variable yang menyimpan informasi apakah pada saat pendaftaran sedang ada promosi internal pada kampus setempat, dan jika ada jenis promosi apa. Untuk kasus ini 0-sedang tidak ada promos, 1-sedang ada promosi ULTAH kampus, 2-sedang ada promosi hardiknas, 3-sedang ada promosi proklamasi.
- `gaji_ortu` : merupakan nilai gaji orang tua yang telah dilakukan transformasi scalling dengan nilai : 1-di bawah 2 juta, 2-di bawah 4 juta,3-di bawah 6 juta, 4-di bawah 8 juta, 5-di atas 10 juta
- `idjenissekolah` : nilai asal sekolah yang nerupakan nilai kategori dengan nilai : 1-asal sma/ma negeri, 2-asal smk negeri,3-asal sma/ma swasta, 4-asal smk stasta, 5- asal lain

Variabel target :

- `Isdaftarulang` : nilainya 0 dan 1, 0 berarti mendaftar ulang, 1 berarti tidak mendaftar ulang, merupakan nilai yang diperoleh dari database, dan bukan proses labelling secara manual.

Selanjutnya menggunakan method info() pandas pada jupyter-python terhadap dataset di atas dihasilkan output berikut :

```
1 X.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7491 entries, 0 to 7490
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ekonomi_nasional  7491 non-null   int32
 1   sedang_ptn        7491 non-null   int64
 2   sedang_promo      7491 non-null   int64
 3   gaji_ortu         7491 non-null   int64
 4   idjenissekolah    7491 non-null   int64
```

Gambar. 4 Struktur dataset final

Dari gambar di atas diperoleh informasi utama dari dataset final :

- Jumlah record dataset = 7491
- Jumlah fitur : 5

Dari 7491 data di atas, lalu diproses menggunakan method : train_test_split(X, y, train_size = 0.7), untuk mendapatkan dataset training dan dataset testing. Hasilnya sebagai berikut :

- Dataset Training : 5243
- Dataset Testing : 2248

Jumlah di atas diperoleh dari parameter train_size=0.7 sehingga data training sebanyak 70% dari keseluruhan dataset.

Untuk atribut target dilakukan cek kesimbangan class 0 dan 1 dengan method value_counts() pada dataseries pandas, dan hasilnya seperti berikut :

Tabel 1. Perbandingan Class Atribut Target

Dataset	Class 0	Class 1	Ratio
Training	1839	4340	30:70
Testing	409	903	29:69

Dari table di atas tampak jelas bahwa ada ketidakseimbangan *(imbalanced data*) pada atribut target (isdaftarulang), artinya dari keseluruhan pendaftar ada sekitar 70% yang melakukan daftar ulang dan sisanya tidak melakukan daftar ulang.

Selanjutnya berdasarkan dataset training di atas dilakukan proses training menggunakan beberapa model klasifikasi terbimbing yang terdapat pada library sklearn python. Model yang dihasilkan langsung dievaluasi menggunakan modul metrics python dan hasilnya adalah seperti berikut :

Tabel 2. Hasil Evaluasi Awal Model Klasifikasi

Learning	Accuracy	F Score
SVM	0,81	0,44
L Regression	0,82	0,44
Decision Tree	0,81	0,49
Naive Bayes	0,82	0,45
K-NN	0,81	0,51

Dari data di atas terlihat bahwa akurasi dari model klasifikasi yang dicobakan terhadap imbalanced dataset tersebut hasilnya baik (di atas 0,8). Akurasi yang terbaik dicapai oleh Logistic Regression dan Naïve Bayes (0,82).

Sedangkan nilai F Score yang terbaik dicapai oleh K-NN (0,51) dan terendah SVM dan Logistic Regression yaitu 0,44.

Jika sekedar melihat nilai akurasi sebesar 0,8 maka model klasifikasi yang diperoleh sudah memenuhi syarat untuk masuk ke tahap deployment.

Tetapi dari nilai F score yang rata-rata hanya sekitar 0,5 merupakan nilai yang jauh dari ideal (1). Oleh karena nilai F Score merupakan turunan nilai nilai Precision, Recall, False-Negatip (FN) dan False-Positip (FP) maka nilai tersebut dapat dijadikan indikasi sebagai akibat dari dataset target yang tidak seimbang.

Lebih detail nilai precision dan recall masing-masing model klasifikasi seperti pada table 3 berikut :

Tabel 3. Nilai Recall dan Precision model klasifikasi

Learning	Recall	Precision
SVM	0,50	0,41
L Regression	0,50	0,41
Decision Tree	0,51	0,55
Naive Bayes	0,50	0,41
K-NN	0,51	0,56

Dari table di atas tampak bahwa nilai recall dan precison juga masih jauh dari nilai 1 yang merupakan gambaran dari nilai False-Positip (FP) dan False-Negatip (FN) yang masih cukup besar.

Selanjutnya besar nilai FP dan FN untuk semua model yang dibangun dapat diamati melalui evaluasi Confusion Matrix berikut :

Tabel 4. Evaluasi Confusion Matrix Model SVM

	Predicted-P	**Predicted-N**
Actual-P	TP=1860	FN=388
Actual-N	FP=0	TN=0

Tabel 5. Evaluasi Confusion Matrix Model L Regression

	Predicted-P	**Predicted-N**
Actual-P	TP=1860	FN=388
Actual-N	FP=0	TN=0

Tabel 6. Evaluasi Confusion Matrix Model Decision Tree

	Predicted-P	**Predicted-N**
Actual-P	TP=1800	FN=365
Actual-N	FP=60	TN=23

Tabel 7. Evaluasi Confusion Matrix Model Naïve Bayes

	Predicted-P	**Predicted-N**
Actual-P	TP=1860	FN=388
Actual-N	FP=0	TN=0

Tabel 8. Evaluasi Confusion Matrix Model K-NN

	Predicted-P	**Predicted-N**
Actual-P	TP=1804	FN=364
Actual-N	FP=54	TN=24

Dari table matrik confusion di atas semakin jelas pada bagian mana terjadi eror terhadap fenomena imbalanced class ini, yaitu pada nilai TN yang tidak tidak sepadan dengan nilai TP dan jumlah FP dan FN yang cukup besar. Hal ini jika diimplementasikan akan menghasilkan model klasifikasi yang tidak andal karena hasil prediksinya akan tidak seperti yang diharapkan

Akhirnya, sebelum dilakukan tahap deployment untuk model klasifikasi menggunakan data imbalaned class seperti kasus dataset pendaftaran mahasiswa baru ini, ini, maka masalah ketidakseimbangan *class* harus diatasi terlebih dahulu.

5. KESIMPULAN

- Dataset pendaftaran mahasiswa baru dengan atribut target (class) : isdaftarualng (nilai binernya : daftar ulang atau tidak), memiliki potensi besar berkarakteristik *imbalanced class dataset*
- Hasil evaluasi model machine learning pada data target yang tidak seimbang (*Imbalanced Classes*) untuk kasus Dataset pendaftaran mahasiswa baru di penelitian ini memungkinkan untuk menghasilkan akurasi yang baik tetapi tidak menjamin nilai F Score (recall dan precision) yang baik (nilai 1).
- Nilai F score yang kurang baik pada kasus ini (sekitar 0,5) menjadi indikasi sebagai akibat dari dataset dengan nilai class yang tidak seimbang tersebut.
- Dari Analisa table Confusion Matrix terlihat bahwa terdapat nilai yang kecil sekali untuk TN dan terdapat cukup besar nilai FP dan FN yang mengindikasikan bahwa model dengan imbalanced class ini belum layak untuk memasuki tahap deployment.

DAFTAR PUSTAKA

Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational Big Data: Predictions, Applications and Challenges. *Big Data Research*, *26*, 100270. doi: https://doi.org/10.1016/j.bdr.2021.100270

Pattiasina, T., & Rosiyadi, D. (2020). Comparison of Data Mining Classification Algorithm for Predicting the Performance of High School Students. *Jurnal Techno Nusa Mandiri*, *17*(1), 22–30. doi: 10.33480/techno.v17i1.1226

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Powers, D. M. W. (2019). What the F-measure doesn't measure …. *ArXiv*.

Saleh, M. A., Palaniappan, S., Ali, N., & Abdalla, A. (2021). *Education is An Overview of Data Mining and The Ability to Predict the Performance of Students*. *15*(1), 19–28.