Penerapan Model Decision Tree pada Machine Learning untuk Memprediksi Calon Potensial Mahasiswa Baru

Simon Prananta Barus[1]

[1]Universitas Matana
Matana University Tower, Jl. CBD Barat Kav. 1. Gading Serpong Tangerang
simonbarus07@gmail.com[1]

ABSTRAK

Pada penelitian sebelumnya, "*Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University*" memiliki akurasi sebesar 0,73 atau 73%. Ini belum maksimal, akurasinya perlu ditingkatkan. Pada penelitian ini, untuk meningkatkan akurasi dengan penggunaan model yang berbeda, yaitu model Decision Tree. Alasan memilih Decision Tree adalah prediktor yang digunakan tidak banyak (empat prediktor) dan dapat digunakan untuk klasifikasi atau prediksi. Empat prediktor tersebut, yaitu frekuensi, lokasi, jurusan siswa di SMA/K, dan program studi yang diminati. Targetnya adalah status masuk dari calon mahasiswa. Metode penelitian yang dilakukan yaitu pengumpulan data, pra-pemrosesan, proses machine learning dengan model Decision Tree dan visualisasi hasil. Bahasa pemrograman yang digunakan adalah Python. Hasil dari Decision Tree ini memperlihatkan perubahan, melalui sepuluh kali eksekusi diperoleh rata – rata akurasi rasio data latih dan data uji, 70:30 sebesar 0,727 atau 72,7% (akurasi terendah 47% dan tertinggi 87%), untuk rasio 80:20 sebesar 0,73 atau 73% (akurasi terendah 60% dan tertinggi 90%). Dengan demikian, hasil dari model Decision Tree ini secara rata – rata belum melampaui hasil dari model Naïve Bayes Classifier. Penelitian lebih lanjut, meningkatkan jumlah dan variasi data, memperkecil selisih hasil setiap kali model dieksekusi, mencoba model lain, dan mengembangkan aplikasi siap pakai.

Kata kunci : decision tree, klasifikasi, prediksi, machine learning, pemrograman python

ABSTRACT

In a previous research, "Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University" has an accuracy of 0.73 or 73%. This is not maximized, accuracy needs to be improved. In this research, to increase accuracy by using a different model, i.e. the Decision Tree Model. The reason for choosing Decision Tree is that there are not many predictors used (four predictors) and can be used for classification or prediction. The four predictors are frequency, location, student major and the study program. The target is the entry status of prospective students. The research method used is data collection, pre-processing, machine learning process with Decision Tree model and visualization of results. The programming language used is Python. The results of this Decision Tree show a change, through ten executions the average accuracy of the ratio of training data and test data, 70:30 is 0.727 or 72.7% (lowest accuracy is 47% and highest is 87%), for a ratio of 80:20 by 0.73 or 73% (lowest accuracy 60% and highest 90%). Thus, the results of the Decision Tree model on average have not exceeded the results of the Naïve Bayes Classifier model. Further research, increasing the amount and variety of data, get a solution to reduce the difference in results every time the model is executed, use another model, and developing ready-to-use applications.

Keyword : decision tree, classification, prediction, machine learning, python programming

1. PENDAHULUAN

Menurut penelitian Gartner, tahun 2022 diperkirakan kecerdasan buatan (*artificial intelligence* (AI)) berpotensi bisnis senilai $3,9 triliun. (Gartner, 2018). Ini menjadikan kecerdasan buatan, termasuk juga di dalamnya *machine learning* ditempatkan pada posisi yang lebih strategis dalam bisnis. Pemasaran merupakan unit bisnis yang terdampak dari kemajuan kecerdasan buatan tersebut (Sterne, 2017). Penerapan kecerdasan buatan dan machine learning di dunia bisnis tentu tidak sembarangan. Salah satu hal yang paling penting dalam penerapan ini adalah akurasi. Semakin tinggi akurasinya, semakin dapat diandalkan. Hal ini sangat mendukung dalam pengambilan keputusan bisnis.

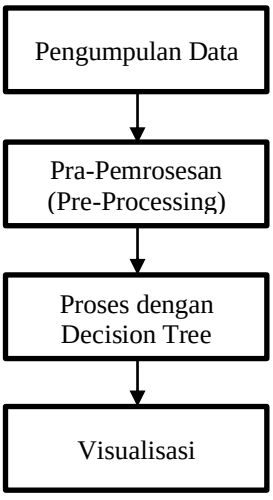
Pada penelitian sebelumnya, berjudul "*Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University", memperoleh akurasi sebesar* 0,73 atau 73% (Simon, 2021). Akurasi tersebut cukup tinggi, ada kemungkinan untuk ditingkatkan akurasinya. Cara yang dilakukan untuk meningkatkan akurasi tersebut dengan merubah model yang digunakan. Pada penelitian ini mencoba penerapan model Decision Tree.

Decision Tree merupakan bentuk pengambilan keputusan yang hirarkinya meniru pohon, memiliki akar, cabang dan daun. Ini digunakan juga sebagai teknik pembelajaran dalam kecerdasan buatan dengan istilah Decision Tree Learning (DTL). Menurut Suyanto, DTL adalah "*teknik pembelajaran mesin yang membangun representasi aturan klasifikasi berstruktur sekuensial hirarki dengan cara mempartisi himpunan data latih secara rekursif*" (Suyanto, 2018). Masalah klasifikasi dan regresi dapat diselesaikan dengan Decision Tree. Penelitian ini memilih Decision Tree dikarenakan prediktor yang digunakan tidak banyak (empat prediktor), yaitu frekuensi, JaBoDeTaBek, jurusan dan program studi dan dapat digunakan untuk klasifikasi atau prediksi. Penerapan Decision Tree dilakukan melalui pembuatan sebuah program. Melalui program berbasis Decision Tree ini diharapkan dapat mempermudah simulasi, kalkulasi dan visualisasi, mempercepat perolehan hasil, mengurangi kesalahan perhitungan, dibandingkan tanpa program.

Python merupakan bahasa pemrograman populer saat ini, khususnya untuk pengembangan machine learning, NLP dan neural network (Analytics Insight, 2021). Disamping itu, Python juga banyak digunakan dalam *data science*. Python memiliki banyak pustaka (*library*), seperti Numpy, Pandas, Sklearn, Matplotlib dan masih banyak lagi. Hal inilah yang menjadi alasan Python dipilih untuk memprediksi calon potensial mahasiswa baru dengan Decision Tree.

2. METODE PENELITIAN

Metode penelitian ini terdiri dari empat tahapan, yaitu pengumpulan data, pra-pemrosesan (*pre-processing*), proses machine learning dengan Decision Tree dan visualisasi hasil, Gambar 1.



Gambar 1. Metode penelitian

Data penelitian diperoleh dari hasil penelitian sebelumnya, dalam bentuk tabel dengan format *comma separated values* (csv). Pada berkas (*file*) tersebut terdiri dari empat kolom untuk prediktor (*predictor*) dan sebuah target. Kolom sebagai prediktor adalah frekuensi, JaBoDeTaBek, jurusan dan program studi. Kolom sebagai target adalah status masuk. Format dataset tersebut dapat dilihat pada Tabel 1. Pengkodean (*encoding*) untuk program studi dapat dilihat pada Tabel 2.

Tabel 1. Format dataset

Frekuensi	JaBoDeTaBek	Jurusan	Program Studi	Status Masuk

Tabel 2. Pengkodean program studi

Kode	Nama Program Studi
1	Arsitektur
2	Akuntansi
3	Desain Komunikasi Visual
4	Fisika
5	Hospitality dan Pariwisata
6	Manajemen
7	Sistem Informasi
8	Sistem Komputer
9	Statistika
10	Teknik Informatika

Pra-pemrosesan dilakukan melalui program dengan menggunakan bahasa pemrograman Python. Pustaka (*library*) yang dipakai adalah Numpy, Pandas dan Sklearn (Scikit-Learn). Pengkodean dilakukan untuk variabel prediktor frekuensi, JaBoDeTaBek, jurusan dan program studi. Data rangkap (duplikasi data) dihilangkan.

Pada tahapan proses dilakukan dengan menerapkan Decision Tree. Perbandingan data latih (*train*) dan data uji (*test*) dalam proses adalah 70:30 dan 80:20. Pada Python, dipakai pustaka Sklearn yang menerapkan algoritma Classification And Regression Tree (CART) untuk melatih model Decision Tree (Géron, 2019). Kriteria (*criterion*) yang dipakai adalah entropy (ukuran seberapa acak suatu kelompok data).

Visualisasi merupakan tahap terakhir pada metode penelitian ini. Pada tahap ini hasil atau output dari proses divisualisasikan dalam bentuk grafik sehingga mempermudah dalam pemahaman dan pengambilan keputusan.

3. HASIL DAN PEMBAHASAN

Penerapan model Decision Tree pada machine learning untuk memprediksi calon potensial mahasiswa baru telah dilakukan dalam sebuah program dengan Python sebagai bahasa pemrogramannya. Pada program tersebut dilakukan sepuluh kali eksekusi model Decision Tree tersebut dengan dua tingkat rasio data latih dan data uji yang berbeda, yaitu 70:30 dan 80:20. Hasil sepuluh kali eksekusi tersebut dapat dilihat pada Tabel 3.

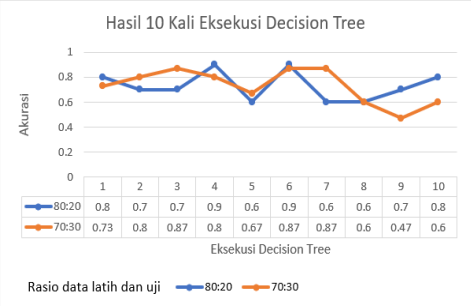
Tabel 3. Hasil proses Decision Tree

No	70:30	80:20
1	0,73	0,8
2	0,8	0,7
3	0,87	0,7
4	0,8	0,9
5	0,67	0,6
6	0,87	0,9
7	0,87	0,6
8	0,6	0,6
9	0,47	0,7
10	0,6	0,8
$\bar{x}$	**0,727**	**0,73**

Dari hasil Tabel 3 tersebut dapat dilihat rata – rata yang tidak jauh berbeda antara rasio 70:30 dan 80:20, yaitu 0,727 dan 0,73. Namun, jika mengamati tingkat akurasi teratas dan akurasi terbawahnya, terlihat rasio 70:30 memiliki akurasi terbawah sebesar 0,47 dan teratasnya sebesar 0,87, sedangkan rasio 80:20 memiliki akurasi terbawah sebesar 0,6 dan akurasi teratasnya sebesar 0,9.

Visualisasi hasil tersebut dapat dilihat pada Gambar 2. Dari visualisasi

tersebut terlihat eksekusi kedelapan memiliki kesamaan hasil, sedangkan pertama, kelima dan keenam masing – masing rasio tidak jauh berbeda. Adapun hasil yang jauh berbeda dapat dilihat pada eksekusi ketujuh dan kesembilan.



Gambar 2. Visualisasi hasil

4. KESIMPULAN

Pada penelitian sebelumnya, model Naïve Bayes Classifier dengan perolehan akurasi 0,73 atau 73%, tidak jauh berbeda dengan rata – rata yang diperoleh melalui model Decision Tree baik rasio 70:30 dan 80:20. Di sini terlihat bahwa model Decision Tree tidak stabil dimana data latih berubah, hasil atau outputnya dapat berubah. Walaupun akurasi pernah sampai 0,9 atau 90% (rasio 80:20), tapi pernah juga sampai 0,47 atau 47% (rasio 70:30), dengan rata – rata 0,727 atau 72,7% (rasio 70:30) dan 0,73 atau 73% (rasio 80:20). Dengan demikian, secara keseluruhan akurasi model Decision Tree belum stabil dan secara rata – rata belum melampaui model Naïve Bayes Classifier.

Penelitian lebih lanjut, meningkatkan jumlah dan variasi data, mendapatkan solusi untuk memperkecil selisih hasil setiap kali model tersebut dieksekusi, menggunakan model lain, dan mengembangkan aplikasi siap pakai.

DAFTAR PUSTAKA

Analytics Insight. (2021). *What Are The Best Programming Languages For Artificial Intelligence*, https://www.analyticsinsight.net/what-are-the-best-programming-languages-for-artificial-intelligence/

Barus, S.P. (2021). *Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University*. Journal of Physics: Conference Series 1842 (1), 012008

Gartner. (2018). *Gartner Says Global Artificial Intelligence Business Value to Reach $1.2 Trillion in 2018*. https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial-intelligence-business-value-to-reach-1-point-2-trillion-in-2018

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media

Sterne, J. (2017). *Artificial Intelligence for Marketing, Practical Applications*. John Wiley & Sons

Suyanto. (2018). *Machine Learning: Tingkat Dasar dan Lanjut*. Bandung: Informatika