Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 53-57

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

PREDIKSI KINERJA DARI SISWA KETIKA MENJALANI UJIAN DENGAN MENGGUNAKAN KNN, *LOGISTIC REGRESSION*, DAN *DECISION TREE*

Enrico Liman
Program Studi Teknik Informatika, FakultasTeknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *enrico.535200003@stu.untar.ac.id*

ABSTRAK

Dalam paper ini, diangkat sebuah topik yang membahas kinerja atau performa dari murid- murid pada institusi pendidikan ketika menjalani ujian. Tujuan dalam pembuatan paper ini membantu merancang mekanisme efektif untuk meningkatkan kinerja hasil akademik dari siswa. Nilai ini terdiri dari gender, *race/ethinicity, lunch, test preparation course, math score, reading score, writing score*. Algoritma adalah sekumpulan instruksi atau langkah-langkah yang dituliskan secara sistematis dan digunakan untuk menyelesaikan masalah / persoalan logika dan matematika dengan bantuan komputer. Algoritma yang digunakan dalam paper yang dibuat adalah *K-Nearest Neighbors* (KNN), *Logistic Regression*, dan *Decision Tree*. Modal Evaluasi dalam paper ini menggunakan *Classification Report, Confusion Matrix*, dan *Cross Validation*.

Kata kunci: *K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree.*

ABSTRACT

In this paper, a topic is raised that discusses the performance of students in educational institutions when taking exams. The aim of making this paper is to help design effective mechanisms to improve students' academic performance. This value consists of gender, race/ethnicity, lunch, test preparation course, math score, reading score, writing score. An algorithm is a set of instructions or steps written systematically and used to solve logical and mathematical problems/issues with the help of a computer. The algorithms used in the paper are K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree. Evaluation capital in this paper uses Classification Report, Confusion Matrix, and Cross Validation.

Keywords: *K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree*.

1. PENDAHULUAN

Topik *machine learning* yang digunakan dalam pembuatan *paper* prediksi kinerja dari siswa ketika menjalani ujian ini adalah pembelajaran terawasi *(supervised learning)*, dimana algoritma akan mendapatkan pelatihan menggunakan *dataset* yang sudah diberikan label atau identitas. Masing-masing masukan *(input)* memiliki keluaran *(output)* yang sesuai, yang bertujuan untuk mempelajari pengenalan pola atau relasi antara *input* dan *output* sampai algoritma mampu melakukan prediksi yang akurat [1].

Topik ini dipilih dengan tujuan untuk membantu merancang mekanisme efektif untuk meningkatkan kinerja hasil akademik dari siswa di sekolah dan untuk mengidentifikasi faktor paling signifikan yang mempengaruhi kinerja siswa. Hal ini juga menentukan bagaimana performa model terpengaruh ketika model dijalankan pada data yang hanya menyertakan fitur paling penting. Dengan adanya paper ini diharapkan dapat membantu di bidang akademik khususnya di sekolah dengan memperbaiki kinerja dari siswa yang mengikuti ujian [2].

2. METODE PENELITIAN

Dalam pembelajaran mesin (machine learning), dataset dapat diartikan sebagai kumpulan data yang digunakan untuk melatih dan menguji model dari sebuah machine learning. Dataset yang

digunakan dalam pembelajaran mesin ini didapatkan dari kagle.com yang merupakan sebuah situs web (website) yang menyediakan berbagai dataset yang berhubungan dengan pembelajaran mesin (machine learning). Dataset ini terdiri dari 1000 baris dan 9 kolom, nilai yang terdapat di dalam dataset ini antara lain *gender, race/ethnicity, lunch, test preparation course, math score, reading score, writing score.* Pada Tabel 1 berikut ini dideskripsikan masing-masing dari kolom tersebut [3].

Tabel 1. Tabel Contoh Dataset

Name	Description
gender	jenis kelamin dari siswa
race/ecthnicity	jenis ras atau etnik dari siswa
lunch	jenis makan siang dari siswa *(standard, free.reduced)*
test preparation course	kursus persiapan ujian *(completed, none)*
math score	nilai pelajaran matematika dari siswa
reading score	nilai pelajaran membaca dari siswa
writing score	nilai pelajaran menulis dari siswa

3.1 Pra-pemrosesan Data

Pra-pemrosesan data adalah proses mengubah data mentah menjadi bentuk yang lebih mudah dipahami. Proses ini dibutuhkan untuk memperbaiki kesalahan pada data mentah yang seringkali tidak lengkap dan formatnya tidak teratur. Tahapan pra-pemrosesan data yang dilakukan antara lain [4][5]:

1. *Data Cleaning*
   Langkah pertama yang harus dilakukan dalam pra-pemrosesan data adalah pembersihan data. Dimana data awal yang diperoleh harus diseleksi kembali. Selanjutnya, hapus atau hilangkan data yang tidak lengkap, tidak relevan, dan tidak akurat. Dengan melakukan langkah ini Anda akan terhindar dari kesalahpahaman saat menganalisis data.
2. *Data Transformation*
   Langkah selanjutnya yang dilakukan adalah transformasi data. Seperti dijelaskan di atas, data berasal dari berbagai sumber dan mungkin dalam format berbeda. Kita perlu menyeimbangkan semua data yang dikumpulkan sehingga kita dapat menyederhanakan proses analisis data. Di dalam transformasi data ada beberapa langkah yang digunakan dalam pembelajaran yang dilakukan yaitu standarisasi yang merupakan suatu proses yang memodifikasi data asli supaya rata-rata (*mean*) dari data tersebut menjadi 0 dan deviasi standarnya menjadi satu.
3. *Data Splitting*
   Langkah selanjutnya yang dilakukan adalah pemisahan data. Membagi data menjadi set pelatihan, set validasi, dan set pengujian untuk mengukur performa model. Pada dasarnya pembagian data dibagi menjadi dua bagian, yaitu data latih dan data uji. Data pelatihan digunakan untuk melatih dan mengembangkan model. Kumpulan data pelatihan sering kali digunakan untuk memperkirakan parameter yang berbeda atau untuk membandingkan performa model yang berbeda. Data pengujian digunakan setelah pelatihan selesai. Data pelatihan dan pengujian dibandingkan untuk memeriksa apakah model akhir yang digunakan memiliki kinerja yang benar.

3.2 Algoritma Klasifikasi

Metode atau algoritma klasifikasi yang akan digunakan antara lain [6][7][8]:

1. *K-Nearest Neighbors* (KNN)
   K-Nearest Neighbors (KNN) Merupakan sebuah algoritma untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pemelajaran digambarkan ke ruang berdimensi banyak dengan tiap-tiap dimensi mewakili tiap ciri/fitur dari data.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

2. *Logistic Regression*
   Logistic Regression merupakan salah satu algoritma dalam *machine learning* yang digunakan untuk klasifikasi data biner .Algoritma ini menggunakan pendekatan regresi linier untuk memodelkan hubungan antara variabel input dan output, dengan menghasilkan nilai probabilitas yang berkisar antara 0 dan 1.Dalam hal ini, output klasifikasi dilakukan berdasarkan probabilitas tersebut.
3. *Decision Tree*
   Decision Tree merupakan algoritma yang memungkinkan untuk memprediksi nilai output berdasarkan serangkaian kondisi atau atribut. Teknik ini banyak digunakan dalam berbagai aplikasi seperti kesehatan, keuangan, pemasaran, manufaktur, dan sumber daya manusia.

3.3 Metode Evaluasi

Tahap metode evaluasi dalam pembelajaran mesin adalah proses mengukur seberapa baik kinerja model Anda dalam memprediksi atau mengklasifikasikan data baru. Ada berbagai metode evaluasi yang digunakan untuk mengukur kinerja model, bergantung pada jenis masalah dan tujuan yang Anda pikirkan. Terdapat beberapa metode evaluasi yang digunakan dalam pembelajaran mesin ini yaitu [9][10]:

1. *Classification Report*
   Classification report adalah sebuah laporan atau ringkasan statistik yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam pembelajaran mesin. Laporan ini menyajikan berbagai metrik evaluasi yang membantu Anda memahami sejauh mana model klasifikasi Anda efektif dalam melakukan prediksi kelas. Classification report umumnya digunakan dalam masalah klasifikasi, di mana model mencoba mengklasifikasikan data menjadi beberapa kategori.
2. *Confusion Matrix*
   Confusion Matrix (Matriks Konfusi) adalah alat yang digunakan untuk mengukur kinerja model klasifikasi dalam pembelajaran mesin. Matriks konfusi adalah tabel berbentuk matriks yang membandingkan prediksi model dengan nilai aktual dari data yang diuji. Matriks ini berguna untuk memahami sejauh mana model Anda benar dalam memprediksi kelas tertentu dan di mana model tersebut mengalami kesalahan. Ada empat nilai yang dihasilkan di dalam tabel confusion matrix, di antaranya True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN).
3. *Cross Validation*
   Cross Validation adalah metode validasi model yang membagi data dengan cara yang kreatif untuk mendapatkan perkiraan kinerja model yang lebih baik dan meminimalkan kesalahan ketika memvalidasi model.

Skema eksperimen machine learning ini dimulai dengan *Data Understanding* untuk meringkas data dan mengidentifikasi potensi masalah. Tahap ini dilanjutkan dengan *Data Preparation*, di mana masalah data diperbaiki dan variabel *derived* dibuat untuk memastikan kesesuaian data dengan algoritma. Selanjutnya, tahap *Modelling* melibatkan pemilihan dan penerapan teknik serta algoritma *machine learning* pada data. Terakhir, tahap *Evaluation* berfokus pada interpretasi hasil yang diperoleh dari proses pemodelan sebelumnya untuk menilai kinerja model.

3. HASIL DAN PEMBAHASAN

Setelah dilakukan pembelajaran mesin model yang menggunakan algoritma klasifikasi *K-Nearest Neighbors (KNN), Logistic Regression*, dan *Decision Tree* dengan menggunakan dataset di atas, didapatkan hasil dan pembahasan sebagai berikut. Nilai yang terdapat di dalam dataset ini antara lain *gender, race/ethnicity, lunch, test preparation course, math score, reading score, writing score.* Untuk akurasi dari algoritma klasifikasi *K-Nearest Neighbors* (KNN), *Logistic Regression*, dan *Decision Tree* didapatkan hasil seperti pada Gambar 1 dibawah ini dimana didapatkan akurasi 1.00.

```
Akurasi KNN: 1.00
Akurasi Decision Tree: 1.00
Akurasi Logistic Regression: 1.00
```

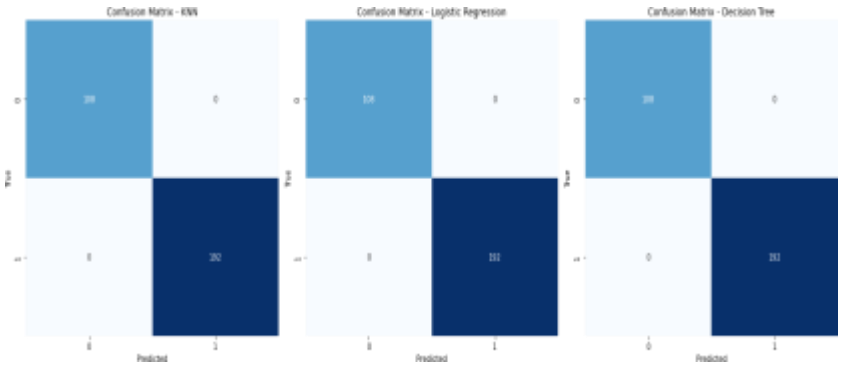
Gambar 1. Hasil Akurasi Algoritma

Untuk metode evaluasi menggunakan *classification report* didapatkan hasil sebagai berikut yang dapat dilihat pada Gambar 2 dimana dari ketiga algoritma tersebut mendapatkan nilai *precision, recall, f1-score* dan *support* yang identik atau sama.

```
Classification Report for KNN:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        78
           1       1.00      1.00      1.00       122

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

Classification Report for Decision Tree:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        78
           1       1.00      1.00      1.00       122

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        78
           1       1.00      1.00      1.00       122

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200
```

Gambar 2. *Classification Report*

Untuk metode evaluasi menggunakan *confusion matrix*, didapatkan hasil sebagai berikut yang dapat dilihat pada Gambar 3 dibawah dimana dari ketiga algoritma tersebut mendapatkan nilai *true positive* (TP), *true negative* (TN), *false positive* (FP), *false negative* (FN) yang identik dari ketiga algoritma klasifikasi tersebut.



Gambar 3. *Confusion Matrix*

Kemudian yang terakhir yaitu metode evaluasi menggunakan *cross-validation* didapatkan hasil sebagai berikut yang dapat dilihat pada Gambar 4 dimana dari ketiga algoritma tersebut *cross-validation* untuk KNN mendapatkan hasil yang lebih rendah apabila dibandingkan dengan *cross-validation* untuk *Logistic Regression* dan *Decision Tree*.

```
Cross-Validation Scores for KNN:
[0.645 0.675 0.69  0.69  0.74 ]
Mean Accuracy: 0.69
Standard Deviation: 0.03


Cross-Validation Scores for Logistic Regression:
[1. 1. 1. 1. 1.]
Mean Accuracy: 1.00
Standard Deviation: 0.00


Cross-Validation Scores for Decision Tree:
[1. 1. 1. 1. 1.]
Mean Accuracy: 1.00
Standard Deviation: 0.00
```

Gambar 4. *Cross Validation*

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

4. KESIMPULAN

Berdasarkan pembelajaran mesin terhadap model yang telah dilakukan di atas didapatkan kesimpulan. Pembelajaran mesin yang telah dilakukan dengan menggunakan algoritma klasifikasi *K-Nearest Neighbors* (KNN), *Logistic Regression, dan Decision Tree* didapatkan nilai yang identik apabila menggunakan metode evaluasi *classification report* dan *confusion matrix*. Akan tetapi jika berdasarkan metode evaluasi *cross validation* , KNN mendapatkan hasil yang lebih rendah apabila dibandingkan dengan *cross-validation* untuk *Logistic Regression* dan *Decision Tree*.

DAFTAR PUSAKA

[1] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences, 10(3), 1042.

[2] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

[3] Zafari, M., Sadeghi-Niaraki, A., Choi, S. M., & Esmaeily, A. (2021). A practical model for the evaluation of high school student performance based on machine learning. Applied Sciences, 11(23), 11534.

[4] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.

[5] Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. Applied Sciences,11(22), 10907.

[6] Singh, R., & Pal, S. (2020). Application of machine LearningAlgorithms to predict students performance.International Journal of Advanced Science and Technology,29(5), 7249- 7261

[7] Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021, February). Prediction of students performance using machine learning. In IOP conference series: Materials science and engineering (Vol. 1055, No. 1, p. 012122). IOP Publishing.

[8] Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: a systematic review. *Applied Sciences*, *11*(21), 10007.

[9] Alhothali, A., Albsisi, M., Assalahi, H., & Aldosemani, T. (2022). Predicting student outcomes in online courses using machine learning techniques: A review. *Sustainability*, *14*(10), 6199.

[10] Enughwure, A. A., & Ogbise, M. E. (2020). Application of machine learning methods to predict student performance: a systematic literature review. Int. Res. J. Eng. Technol, 7(05), 3405- 3415.