Nama : Nadia Nur Oktaviani Sukma

Nim : 202010370311320

Pra – UTS Analisis Big Data

Prediction of Final Student Score Results

Dikumpulkan minggu ini hari JUMAT jam 23.59 di google classroom:

1. Create dataset (integration of multiple dataset based on your knowledge about OULAD) —> ALL OF DATA AMOUNT, not recommendation using sample.

2. Create Summary Data and Exploratory Data Analysis and describe clearly.

3. Find at least A REFERENCE INTERNATIONAL JOURNAL about modelling with machine learning/ deep learning

4. Create machine learning model/ deep learning model from your dataset about oulad.

5. Describe clearly step by step in your modelling (preprocessing, what model that you should use, evaluation)

6. Give conclusion about model and recommendation.

7. Collect in PDF. This PDF must include:

   1. Explanation about integration of dataset. Why and what data that you use from oulad

   2. Your dataset sample (head or tail dataset)

   3. Explanation about modelling (preprocessing until evaluation)

   4. Conclusion and recommendation

   5. Source code

   6. JOURNAL REFERENCE that you used (link journal and screenshot the title, author, and abstract)

1. **Penjelasan Integrasi Dataset**

Sebelum melelakukan integrasi atau penggabungan dataset ada satu pre processing yang saya lakukan yaitu penanganan missing value. Disini saya tidak menghilangkan missing valuenya, hanya saja mengisi nilai-nilai yang hilang dalam kolom ' DataFrame yang mengalami missing value dengan nilai mode dari kolom tersebut, dan perubahan ini diterapkan langsung ke DataFrame asli. Ini adalah teknik umum yang digunakan untuk mengatasi data yang hilang dalam analisis data. Setelah itu baru mulai untuk mengintegrasikan atau menggabungkan dataset.

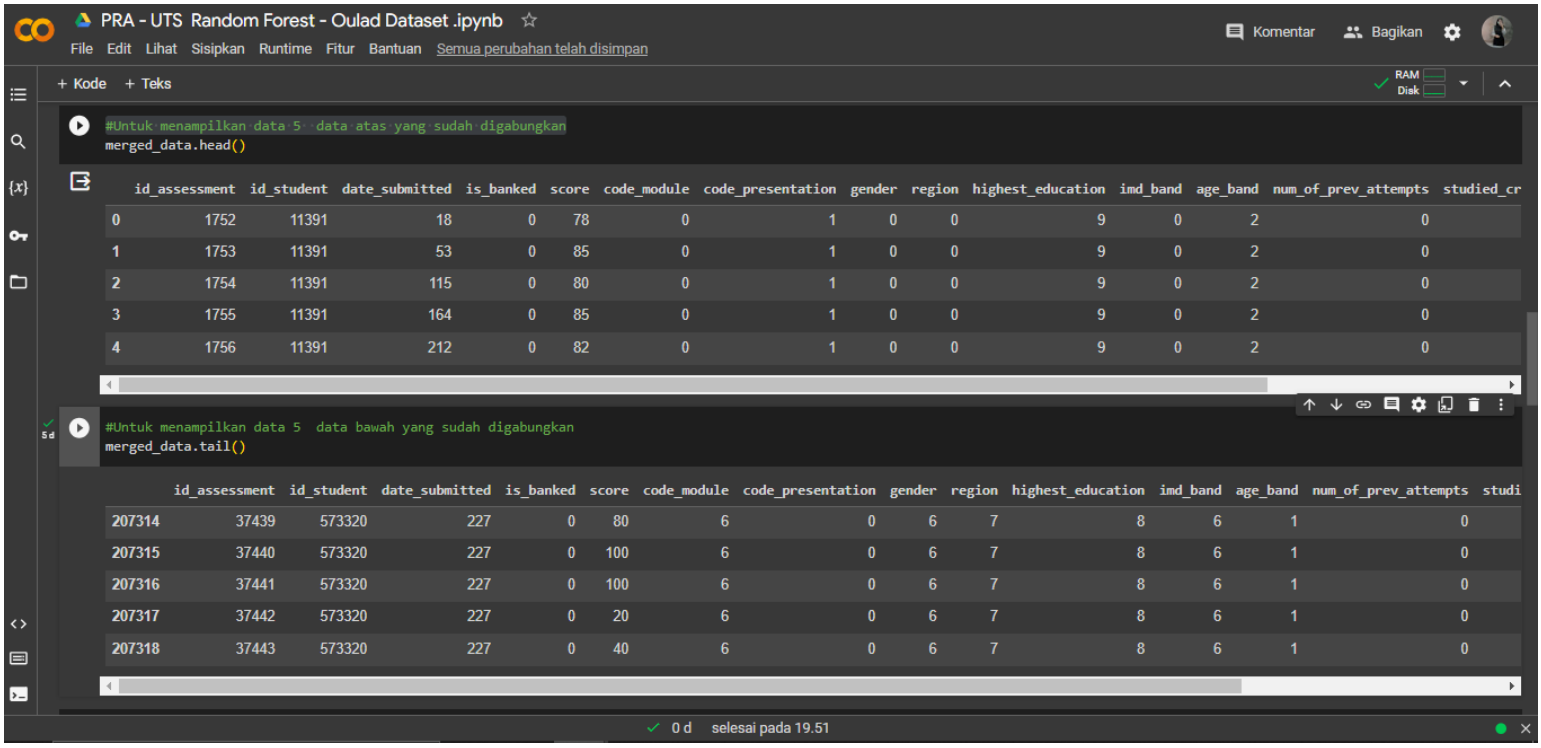
Integrasi dataset adalah proses menggabungkan dua atau lebih dataset yang berbeda menjadi satu dataset tunggal. Tujuan utama dari integrasi dataset adalah untuk menggabungkan informasi dari berbagai sumber data yang mungkin memiliki keterkaitan atau relevansi untuk analisis atau pemodelan yang lebih baik. Integrasi dataset antara dataset "Student Assessment" dan "Student Info" berarti menggabungkan informasi dari kedua dataset tersebut untuk mendapatkan pandangan yang lebih lengkap dan terinci tentang siswa dalam konteks lingkungan pendidikan. Dengan menggabungkan data dari "Student Assessment," yang berisi informasi tentang kinerja akademik siswa, seperti nilai ujian dan tugas, dengan data dari "Student Info," yang mungkin mencakup informasi demografis, seperti nama, alamat, tanggal lahir, dan jenis kelamin.

Dalam integrasi kali ini ada beberapa Langkah yang saya gunakan, diantaranya:

- Pemahaman Dataset, disini saya memahami kedua dataset dengan saya akan integrasikan, termasuk struktur data, atribut yang ada didalamnya, dan bagaimana data-data tersebut saling berhubungan. Dataset "Student Assessment" yang berisi informasi tentang penilaian akademik siswa, sedangkan dataset "Student Info" berisi data demografis dan profil siswa.
- Pencocokan Data, Mengidentifikasi atribut atau variabel yang dapat digunakan sebagai kunci pencocokan (seperti nomor identifikasi siswa) antara kedua dataset. Jadi dari kedua dataset harus terdapat atribut yang sama agar bisa menggabungkan kedua dataset.
- Pembersihan Data, Sebelum menggabungkan dataset, melakukan pembersihan data untuk mengatasi duplikasi, data yang hilang, atau lainnya yang mungkin muncul selama proses integrasi.

- Menggabungkan Data, Menggunakan atribut yang terdapat di kedua dataset untuk menggabungkan data dari kedua dataset sehingga memiliki satu dataset tunggal yang berisi informasi dari keduanya. Disini saya menggunakan python dengan modelling Machine Learning yaitu Random Forest

2. Sampel kumpulan data Anda (kumpulan data kepala atau ekor)



3. Penjelasan mengenai pemodelan (preprocessing hingga evaluasi)

Dalam proses pemodelan Prediction of Final Student Score Results ini, saya menggunakan Random Forest dari menggabungkan 2 dataset oulad yaitu "Student Assessment" dan "Student Info". Sebelum menggabungkan dataset, saya sudah melakukan satu preprocessing untuk masing – masing dataset seperti yang sudah dijelaskan di atas. Setelah menggabungkan ada ada beberapa hal yang saya lakukan dari pre preprocessing hingga evaluasi dan menampilkan akurasi, diantaranya:

a. Label encoding

Label encoding adalah Proses encoding ini mengubah nilai-nilai dalam kolom-kolom tersebut dari bentuk teks menjadi bentuk angka sehingga dapat digunakan dalam pemodelan atau analisis data. Disini saya mengubah semua atribut yang bertipe data kategori menjadi tipe data numerik, jika tidak data yang akan kita olah tidak bisa dilakukan karena masih ada tipe data yang tidak numerik.

b. Normalisasi data

Metode normalisasi yang saya gunakan adalah StandardScaler, dimana digunakan untuk mengubah distribusi nilai dari setiap fitur (kolom) dalam dataset agar memiliki rata-rata nol (0) dan deviasi standar satu (1). Tahapam normalisasi ini biasanya saya gunakan juga karena dapat meningkatkan suatu akurasi didalam beberapa model atau metode machine learning.

c. Membagi dataset menjadi fitur dan target

Dalam proses membagi dataset menjadi fitur (features) dan target (label) adalah salah satu langkah penting dalam pemodelan data. Proses ini dilakukan agar Anda dapat melatih model untuk memahami hubungan antara fitur-fitur tertentu dan label atau target yang akan diprediksi oleh model. Disini saya menggunakan “final_result” sebagai target dan fiturnya adalah semua atribut digunakan selain final_result, id_assessment, date_submitted, is_banked,gender.

d. Pemisahan data train dan test

Pemisahan data menjadi train dan test digunakan untuk membagi dataset menjadi data pelatihan (training data) dan data uji (testing data) menggunakan fungsi train_test_split dari Scikit-Learn. Pemisahan ini penting untuk mengukur kinerja model dan menghindari overfitting. Data pelatihan digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji kinerja model dan mengukur sejauh mana model dapat melakukan prediksi yang akurat pada data yang belum pernah dilihat.

e. Melatih model Klasifikasi yaitu Random forest

Dalam analisis kali ini saya menggunakan metode random forest. Setelah membuat objek model dengan konfigurasi tersebut, yaitu melatih model menggunakan data pelatihan yang telah dibagi sebelumnya. Dini saya akan menggunakan metode.fit() untuk melakukan

pelatihan. Dengan melatih model menggunakan data pelatihan, dapat memberikan model informasi yang diperlukan untuk memahami pola dalam data dan dapat digunakan untuk melakukan prediksi pada data uji atau data baru yang belum pernah dilihat. Model Random Forest adalah model ensemble yang terdiri dari beberapa pohon keputusan yang bekerja bersama untuk membuat prediksi yang kuat, terutama dalam tugas klasifikasi.

Setelah itu melakukan prediksi menggunakan model klasifikasi yang telah Anda latih sebelumnya (dalam hal ini, model Random Forest) pada data uji. Setelah Anda memiliki hasil prediksi (y_pred), Anda dapat melanjutkan dengan berbagai metrik evaluasi untuk mengukur kinerja model, seperti akurasi, presisi, recall, F1-score, dan lainnya.

f. Evaluasi Model

Terakhir masuk ketahap evaluasi model. Evaluasi model Random Forest, seperti halnya evaluasi model klasifikasi atau regresi lainnya, dimana proses penting untuk mengukur sejauh mana model Anda berhasil memahami pola dalam data dan menghasilkan prediksi yang akurat. Evaluasi model membantu Anda memahami sejauh mana model Anda efektif dalam tugas klasifikasi atau regresi yang Anda lakukan. Dalam prediksi kali ini dengan menggunakan gabungan dari 2 dataset yaitu student assessment dan student info dan menggunakan metode Random Forent menghasilkan akurasi sebesar 0.87.

4. Kesimpulan dan rekomendasi

Dalam proses analisis data "Prediction of Final Student Score Results" dengan menggunakan dua dataset, yaitu "Student Assessment" dan "Student Info," beberapa langkah penting dilakukan. Sebelum menggabungkan dataset, penanganan missing value dilakukan dengan mengisi nilai-nilai yang hilang dengan nilai mode dari kolom yang bersangkutan. Ini adalah teknik umum untuk mengatasi data yang hilang. Setelah itu, dilakukan integrasi dataset dengan menggabungkan informasi dari kedua dataset tersebut. Langkah-langkah yang diambil dalam integrasi termasuk pemahaman dataset, pencocokan data, pembersihan data, dan penggabungan data menggunakan Random Forest sebagai metode pemodelan. Model Random Forest digunakan untuk pelatihan dan prediksi, dan hasil prediksi dievaluasi menggunakan berbagai metrik, dengan akurasi mencapai 0.87.

Mungkin untuk yang ingin melakukan analisis dataset menggunakan data Oulad , agar mendapat akurasi yang tinggi dapat mengganti model atau mengganti dataset yang digabungkan.

5. **Kode sumber**

   https://colab.research.google.com/drive/1xqRB_s9onALLjUHZoQtuvY7UadLWSEtp?usp=sharing

6. **Referensi jurnal**

   https://www.mdpi.com/2071-1050/14/22/14795

sustainability

MDPI

Article

Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing

Khurram Jawad [1,*], Muhammad Arif Shah [2] and Muhammad Tahir [1,*]

[1] College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
[2] Department of IT & Computer Science, Pak-Austria Fachhochshule Institute of Applied Sciences & Technology, Haripur 22650, Pakistan
* Correspondence: k.allo@seu.edu.sa (K.J.); m.tahir@seu.edu.sa (M.T.)

Abstract: Virtual learning environment (VLE) is vital in the current age and is being extensively used around the world for knowledge sharing. VLE is helping the distance-learning process, however, it is a challenge to keep students engaged all the time as compared to face-to-face lectures. Students do not participate actively in academic activities, which affects their learning curves. This study proposes the solution of analyzing students' engagement and predicting their academic performance using a random forest classifier in conjunction with the SMOTE data-balancing technique. The Open University Learning Analytics Dataset (OULAD) was used in the study to simulate the teaching–learning environment. Data from six different time periods was noted to create students' profiles comprised of assessments scores and engagements. This helped to identify early weak points and preempted the students performance for improvement through profiling. The proposed methodology demonstrated 5% enhanced performance with SMOTE data balancing as opposed to without using it. Similarly, the AUC under the ROC curve is 0.96, which shows the significance of the proposed model.

Keywords: student academic performance; virtual learning environment; random forest; SMOTE

Citation: Jawad, K.; Shah, M.A.; Tahir, M. Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing. *Sustainability*