Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Nov 2024, VII (4): 1800 – 1807 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

AI APPROACH TO PREDICT STUDENT PERFORMANCE (CASE STUDY: BATTUTA UNIVERSITY)

M. Rhifky Wayahdi[1], Fahmi Ruziq[2], Subhan Hafiz Nanda Ginting[3]
[1,2,3]Faculty of Technology, Battuta University
Email: [1]muhammadrhifkywayahdi@gmail.com, [2]fahmiruziq89@gmail.com, [3]subhanhafiz16@gmail.com

Abstract: *The research conducted uses an artificial intelligence (AI) approach to the process of predicting student performance at Battuta University. The artificial intelligence model used is the Random Forest model. The author used three different datasets with 300 decision trees for the training and testing process with the Random Forest model and conducted trials with three model variations. The first model (RF-1) showed a high accuracy of 90%, while the second (RF-2) and third (RF-3) models each obtained an accuracy of 89%. A confusion matrix and classification report (precision, recall, and f1-score) were used to evaluate the performance of the artificial intelligence models used. In the "pass" category, the three models performed well with a precision and recall of 90–95%. In the "distinction" category, the first (RF-1) and third (RF-3) models had better precision and recall than the second model (RF-2). While in the "fail" category, the second model (RF-2) shows slightly superior performance compared to other models. The results of this study show that the Random Forest model is able to produce quite high accuracy in predicting student performance, which is around 80–90%. Thus, the Random Forest model is a fairly effective method for predicting student performance. These results are expected to be used by the university to identify students who need early intervention and improve more effective learning strategies.*

Keywords: *Artificial intelligence, Predicting, Performance, Student, Random forest.*

Abstrak: Penelitian yang dilakukan menggunakan pendekatan kecerdasan buatan (AI) pada proses prediksi kinerja mahasiswa Universitas Battuta. Model kecerdasan buatan yang digunakan adalah model Random Forest. Penulis menggunakan tiga dataset berbeda dengan 300 pohon keputusan untuk proses pelatihan dan pengujian dengan model Random Forest dan melakukan uji coba dengan tiga variasi model. Model pertama (RF-1) menunjukkan akurasi yang tinggi yaitu sebesar 90%, sedangkan model kedua (RF-2) dan ketiga (RF-3) masing-masing memperoleh akurasi sebesar 89%. Matriks konfusi dan laporan klasifikasi (presisi, perolehan, dan skor f1) digunakan untuk mengevaluasi kinerja model kecerdasan buatan yang digunakan. Pada kategori "lulus", ketiga model memiliki performa yang baik dengan presisi dan perolehan 90–95%. Pada kategori "distinction", model pertama (RF-1) dan ketiga (RF-3) memiliki presisi dan recall yang lebih baik dibandingkan model kedua (RF-2). Sedangkan pada kategori "gagal", model kedua (RF-2) menunjukkan performa yang sedikit lebih unggul dibandingkan model lainnya. Hasil penelitian ini menunjukkan bahwa model Random Forest mampu menghasilkan akurasi yang cukup tinggi dalam memprediksi kinerja siswa, yaitu berkisar 80–90%. Dengan demikian, model Random Forest merupakan metode yang cukup efektif untuk memprediksi kinerja siswa. Hasil ini diharapkan dapat digunakan oleh universitas untuk mengidentifikasi mahasiswa yang memerlukan intervensi dini dan meningkatkan strategi pembelajaran yang lebih efektif.

Kata Kunci: Kecerdasan Buatan, Prediksi, Kinerja, Siswa, Random Forest.

INTRODUCTION

Educational institutions currently face the challenge of improving the quality of education and student performance in an increasingly competitive environment while utilizing student data for better decision-making (Akour et al., 2020). Prediction of student performance is a central issue in the world of higher education. With the increasingly complex learning environment and the continuing increase in the number of students, an educational institution is required to be able to identify students who have the potential to experience academic difficulties early on. Technological developments have changed the face of education. By utilizing educational data, we can predict student performance and improve the quality of learning (Aslam et al., 2021).

This allows for timely intervention that can improve student study success. Artificial intelligence (AI), with its ability to process data at scale and identify complex patterns, offers great potential in overcoming these challenges. Artificial intelligence (AI) has been applied in various fields, such as determining the shortest route (Wayahdi et al., 2021), detecting or predicting of cancer (Wayahdi & Ruziq, 2022), image classification (Wayahdi et al., 2020), determining the best campus (Wahyuni & Wayahdi, 2021), and including the field of education (Kar et al., 2022). Technologies such as IoT, AI, ML, DL, and Big Data have opened up opportunities to improve students learning experiences (Ojajuni et al., 2021). Learning Analytics (LA) is data science applied to the field of education (Rincón-Flores et al, 2020). AI has become an invaluable tool in many fields. AI is capable of analyzing data in depth, identifying patterns, and making accurate predictions (Deo et al., 2020).

Hussain et al. (2021) developed a regression model to analyze student academic performance using a deep learning approach. Machine learning plays an important role in predicting students' academic performance and helping them achieve higher grades. Predicting student academic success is important for identifying students at risk of failure and providing appropriate remediation (Nabil et al., 2021). The research area related to predicting student performance is multidimensional and can be explored and analyzed through various perspectives, including early prediction of dropout and withdrawal in ongoing studies, analyzing intrinsic factors that impact their performance, and applying statistical techniques to measure student performance (Waheed et al., 2020).

This study aims to explore the application of AI approaches in predicting student academic performance using data from Battuta University as a case study. By utilizing machine learning algorithms, this research seeks to build a model that is able to predict accurately and reliably. The author chose the random forest model in this research. Random Forest does not require any assumptions about data distribution. Random Forest is generally used in two different classes, namely regression and classification (Gholizadeh et al., 2020; Li et al., 2020). The random forest model was proven to be accurate in predicting decreased quality of life in thyroid cancer patients three months after thyroidectomy. These findings can be applied clinically to optimize health care (Liu et al., 2022). In their research, Palimkar et al. (2022) used a random forest algorithm to diagnose diabetes in patients with good accuracy. In their research, Hu et al. (2021) developed a real-time disturbance predictor using Random Forest for high-density disturbances and used it in the EAST tokamak plasma control system (PCS) for the first time.

Battuta University was chosen as the research object because it has a fairly complete and diverse database. Apart from that, this university has also demonstrated its commitment to implementing information technology in the learning process. Thus, the data

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Nov 2024, VII (4): 1800 – 1807 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

obtained from Battuta University is expected to provide a comprehensive picture of student performance and the factors that influence it. The hope is that the results of this research can be used to improve the quality of learning by providing more personalized recommendations for each student at Battuta University.

METHOD

1. Feature Engineering

a. Assessments

It's interesting to include the performance on each assessment as a component in the final model since it serves as a decent gauge of the students' understanding of the material and determines their final evaluation score. However, it is not practical to develop a feature for every evaluation because there are numerous courses, each with a unique structure. To incorporate the evaluations, we will create two features: The ultimate grade determined by adding the weight and score of each evaluation is one of them.

The other is a pass rate, which is based on the idea that a student needs to receive a score of at least 40% to pass an assessment. It determines the proportion of assessments that the student passed with success. Because final exams have a different status and participation in the final evaluation than previous assessments, we will also separate them from the other assessments.

```
#Final exam scores
stud_exams=pd.merge(studAss,exams,how="inner",on=["id_assessment"])
stud_exams["exam_score"]=stud_exams["score"]
stud_exams.drop(["id_assessment","date_submitted","is_banked",
"score","assessment_type","date","weight"],axis=1,inplace=True)
stud_exams.head()
```

Table 1. Final exam scores

	id_student	code_module	code_presentation	exam_score
0	558914	CCC	2014B	32
1	559706	CCC	2014B	78
2	559770	CCC	2014B	54
3	56011	CCC	2014B	64
	4			
...	...	...	...	...

b. Virtual Learning Environment

The student interaction feed with the content that is available for reference for the entire duration of the term is contained in the datasets related to the Virtual Learning Environment (VLE). We may deduce a student's level of engagement with their subjects, whether they studied them thoroughly, and how they applied the material from this data.

```
#General average per student per module
avg_per_student=avg_per_site.groupby(["id_student","code_module",
"code_presentation"]).mean()[["date","sum_click"]].reset_index()
avg_per_student.head()
```

Table 2. General average per student per module

	id_student	code_module	code_presentation	date	sum_click
0	6516	AAA	2014J	105.292	5.816
1	8462	DDD	2013J	38.794	1.734
2	8462	DDD	2014J	10	3
3	11391	AAA	2013J	111.739	4.231
...	...	...	...	...	...

c. Student Info

Although there are many details about the students in the student_info table, the following are pertinent to our analysis: the number of times the student has attempted to complete the module and the student's final score.

```
#Compiling all relevant tables
df_2=pd.merge(studInfo,assessment_info,how="inner",on=
["id_student","code_module","code_presentation"])
final_df=pd.merge(df_2,avg_per_student,how="inner", on=
["id_student","code_module","code_presentation"])
final_df.drop(["id_student","code_module","code_presentation"],
axis=1,inplace=True)
final_df.head()
```

Table 3. Compiling all relevant tables

	num_of_prev_...	final_result	weighted_grade	pass_rate	exam_score	…
0	0	Distinctio	89.65	1	94	…

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Nov 2024, VII (4): 1800 – 1807 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

		n				
1	0	Pass	84.58	1	76	…
2	0	Pass	51.44	0.625	66	…
3	0	Pass	75.13	1	50	…
…	…	…	…	…	…	…

2. Exploratory Data Analysis

We are unable to include the objective feature in a correlation matrix since it is categorical; nevertheless, we can observe a propensity for correlation between the grading features (weighted_grade, pass_rate, and exam_score). The correlation matrix can be seen in Figure 1.

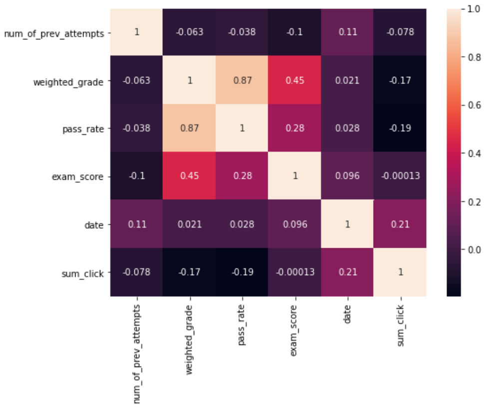

Figure 1. The correlation matrix

The link between the variables in the dataset is described by the correlation matrix. The correlation between two variables is represented by each box in the matrix. Each box's colour and number represent the correlation's strength and direction.

```
# Plot final result
plt.figure(figsize=(8,6))
sns.countplot(data=final_df, x="final_result")
```

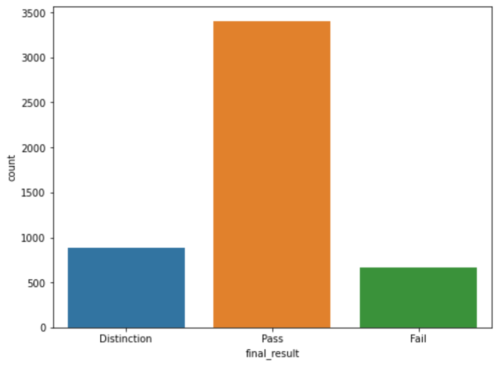

Figure 2. Final result chart

Although the 'Pass' category has much higher numbers than other categories, we need to pay close attention to the model performance metrics. This is especially true for underrepresented cases, such as the 'Distinction' and 'Fail' categories. A deeper analysis of these cases can provide valuable insight into the factors influencing model performance and help us identify areas for improvement.

```
# Paired plots
sns.pairplot(final_df)
```

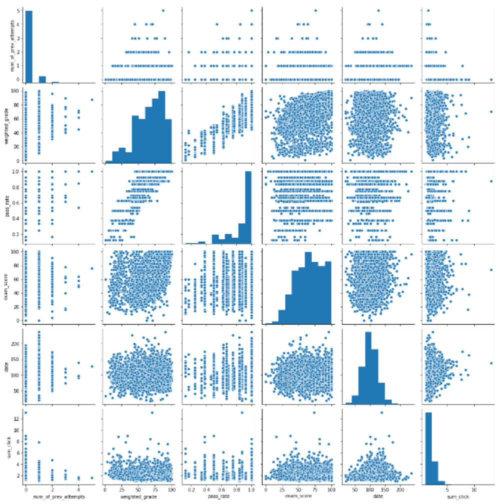

Figure 3. Paired plots

Two outliers can be seen in the pair plot (Figure 3): one has an average click count that is far higher than average, while the other has just one instance of a certain number of prior tries. These cases will be eliminated to preserve the consistency of our data as much as feasible.

3. Modelling with Random Forest Method

The final result (Distinction, Pass, Fail) from a set of data is predicted using a classification model with Random Forest. Three separate feature sets (X1, X2, and X3) are used to train the model.

```
# Training 1
rf1=RandomForestClassifier(n_estimators=300)
rf1.fit(X1_train,y_train)
result_rf1=rf1.predict(X1_test)
print(confusion_matrix(y_test,result_rf1))
print("\n")
print(classification_report(y_test,result_rf1))
```

```
[[222   0  45]
 [  0 152  50]
 [ 33  22 961]]

              precision    recall  f1-score   support

 Distinction       0.87      0.83      0.85       267
        Fail       0.87      0.75      0.81       202
        Pass       0.91      0.95      0.93      1016

    accuracy                           0.90      1485
   macro avg       0.88      0.84      0.86      1485
weighted avg       0.90      0.90      0.90      1485
```

```
# Training 2
rf2=RandomForestClassifier(n_estimators=300)
rf2.fit(X2_train,y_train)
result_rf2=rf2.predict(X2_test)
print(confusion_matrix(y_test,result_rf2))
print("\n")
print(classification_report(y_test,result_rf2))
```

```
[[222   0  45]
 [  0 155  47]
 [ 45  21 950]]

              precision    recall  f1-score   support

 Distinction       0.83      0.83      0.83       267
        Fail       0.88      0.77      0.82       202
        Pass       0.91      0.94      0.92      1016

    accuracy                           0.89      1485
   macro avg       0.87      0.84      0.86      1485
weighted avg       0.89      0.89      0.89      1485
```

```
# Training 3
rf3=RandomForestClassifier(n_estimators=300)
rf3.fit(X3_train,y_train)
result_rf3=rf3.predict(X3_test)
print(confusion_matrix(y_test,result_rf3))
print("\n")
print(classification_report(y_test,result_rf3))
```

```
[[218   0  49]
 [  0 148  54]
 [ 32  25 959]]

              precision    recall  f1-score   support

 Distinction       0.87      0.82      0.84       267
        Fail       0.86      0.73      0.79       202
        Pass       0.90      0.94      0.92      1016

    accuracy                           0.89      1485
   macro avg       0.88      0.83      0.85      1485
weighted avg       0.89      0.89      0.89      1485
```

.

RESULT AND DISCUSSION

The author used 3 different datasets with 300 decision trees for the training and testing process with the Random Forest model, and carried out trials with variations of the 3 models. While the second and third models (RF-2) each achieve an accuracy of 89%, the first model (RF-1) demonstrates a high accuracy of 90%. The accuracy, recall, and f1-score of the classification reports as well as the confusion matrix are used to assess how well the artificial intelligence model is performing. The three models performed well in the "pass" category, with recall and precision ranging from 90 to 95 percent. The first (RF-1) and third (RF-3) models in the "distinction" category had greater precision and recall than the second model (RF-2). In contrast, the second model (RF-2) performed somewhat better than the other models in the "fail" category. The findings:

Overall accuracy: The overall accuracy of the three Random Forest models is comparable, ranging from 89 to 90%. This demonstrates how well the model predicts the outcome in the end.

Precision and Recall: For each category (Distinction, Pass, Fail), each model has a different precision and recall. The precision measure indicates the accuracy of "positive" forecasts, such as Distinction predictions. Recall is the frequency with which the model correctly classified every instance of a category (all students who ought to have received Distinction, for example).

Best model: It is hard to say which model is the best overall based on these results. In the Distinction category, Model 1 has the best precision, but in the Fail category, it has the lowest recall. Model 2, on the other hand, has marginally worse precision for Distinction but stronger recall for the Fail category.

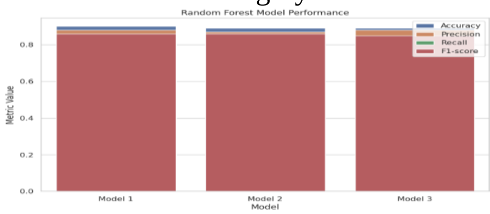

Figure 4. Random forest model performance

Based on the visualization in Figure 4, it can be concluded that the three Random Forest models have quite good performance in predicting the final result. However, if precision is the priority, then Model 1 may be more suitable. If the priority is recall, then the Model 3 may be a better fit.

CONLUSSION

When it comes to ultimate outcome prediction, the Random Forest model performs admirably. However, additional research based on categorization priorities could be necessary in order to choose the optimal model. The findings of this study demonstrate that the Random Forest model can predict student performance with an accuracy of 80–90%. As a result, the Random Forest model is a decent predictor of student achievement. It is hoped that the institution would be able to use these findings to better target more effective learning tactics and identify students who require early intervention.

REFERENCES

Akour, M., Sghaier, H.A., & Qasem, O.A. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 387-393. DOI: https://doi.org/10.11591/ijeecs.v19.i1

Aslam, N., Khan, I.U., Alamri, L.H., & Almuslim, R.S. (2021). An Improved Early Student's Performance Prediction Using Deep Learning, iJET, 16(12), 108-122. DOI: https://doi.org/10.3991/ijet.v16i12.20699

Deo, R.C., Yaseen, Z. M., Al-Ansari, N. Nguyen-Huy, T., Langlands, T. A. M., & Galligan, L. (2020). Modern Artificial Intelligence Model Development for Undergraduate Student Performance Prediction: An Investigation on Engineering Mathematics Courses. In IEEE Access, 8(1), 136697-136724. DOI: https://doi.org/10.1109/ACCESS.2020.3010938

Gholizadeh, M., Jamei, M., Ahmadianfar, I. Pourrajab, R. (2020). Prediction of nanofluids viscosity using random forest (RF) approach. Chemometrics and Intelligent Laboratory Systems, 201, 104010, ISSN 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2020.104010

Hu, W.H., Rea, C., Yuan, Q.P., Erickson, K.G., Chen, D.L., Shen, B., Huang, Y., Xiao, J.Y., Chen, J.J., Duan, Y.M., Zhang, Y., Zhuang, H.D., Xu, J.C., Montes, K.J., Granetz, R.S., Zeng, L., Qian, J.P. Xiao, B.J., Li, J.G., & EAST Team. (2021). Real-time prediction of high-density EAST disruptions using random forest. Nuclear Fusion, 61(6), 066034. DOI: https://doi.org/10.1088/1741-4326/abf74d

Hussain, S., Gaftandzhieva, S., & Maniruzzaman, M. (2021). Regression analysis of student academic performance using deep learning. Educ Inf Technology, 26, 783–798. DOI: https://doi.org/10.1007/s10639-020-10241-0

Kar, A.K., Choudhary, S.K., Singh, V.K. (2022). How can artificial intelligence impact sustainability: A systematic literature review. Journal of Cleaner Production, 376, 134120, ISSN 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2022.134120

Li, M., Zhang, Y., Wallace, J., & Campbell, E. (2020). Estimating annual runoff in response to forest change: A statistical method based on random forest. Journal of Hydrology, 589, 125168, ISSN 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2020.125168

Liu, Y.H., Jin, J., & Liu, Y.J. (2022). Machine learning–based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Support Care Cancer, 30, 2507–2513. DOI: https://doi.org/10.1007/s00520-021-06657-0

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students Academic Performance Based on Courses Grades Using Deep Neural Networks. In IEEE Access, 9, 140731-140746. DOI: https://doi.org/10.1109/ACCESS.2021.3119596

Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., & Mbarika, V. (2021). Predicting Student Academic Performance Using Machine Learning. International Confrence on Computational Science and Its Applications-ICCSA, 12957, 481–491. DOI: https://doi.org/10.1007/978-3-030-87013-3_36

Palimkar, P., Shaw, R.N., & Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. Advanced Computing and Intelligent Technologies, Lecture Notes in Networks and Systems, 218, 219–244. DOI: https://doi.org/10.1007/978-981-16-2164-2_19

Rincón-Flores, E.G., López-Camacho, E., Mena, J., & López, O.O. (2020). Predicting academic performance with Artificial Intelligence (AI), a new tool for teachers and students. IEEE Global Engineering Education Conference (EDUCON), 1049-1054. DOI: https://doi.org/10.1109/EDUCON45650.2020.9125141

Waheed, H., Hassan, S., Aljohani, N.R. Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189, ISSN 0747-5632. DOI: https://doi.org/10.1016/j.chb.2019.106189

Wahyuni, D. & Wayahdi, M.R. (2021). Decision Support System Determining the Best Private Universities Using the Analytical Hierarchy Process Method (Case Study: LLDIKTI Area I North Sumatra). INFOKUM, 10(1), 244-252.

Wayahdi, M.R. & Ruziq, F. (2022). KNN and XGBoost Algorithms for Lung Cancer Prediction. Journal of Science Technology (JoSTec), 4(1), 179-186. DOI: https://doi.org/10.25008/ijadis.v2i1.1206

Wayahdi, M.R., Ginting, S.H.N., & Syahputra, D. Greedy, A-Star, and Dijkstra's Algorithms in Finding Shortest Path. (2021). International Journal of Advances in Data and Information Systems (IJADIS), 2(1), 45-52. DOI: https://doi.org/10.25008/ijadis.v2i1.1206

Wayahdi, M.R., Syahputra, D., & Ginting, S.H.N. (2020). Evaluation of the K-Nearest Neighbor Model with K-Fold Cross Validation on Image Classification. INFOKUM, 9(1), 1-6.

Wayahdi, M.R., Tulus, & Lydia, M.S.

(2020). Combination of k-means with naïve bayes classifier in the process of image classification. IOP Conference Series: Materials Science and Engineering, 725(1), 1-7. DOI: https://doi.org/10.1088/1757-899X/725/1/012126