Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Prediction of student academic performance through Machine Learning

[1]Edgardo Martin Figueroa Donayre, [2]Richard Dante Ramirez Ormeño, [3]Abel Angel Sullon Macalupu, [4]Jhon Richard Huanca Suaquita, [5]Mia Lucia Guillen Guevara, [6]Rogger Humpiri Flores, Milton Edward Humpiri Flores*

[1] https://orcid.org/0000-0001-7891-3334
em.figueroa@unaj.edu.pe – Universidad Nacional de Juliaca
[2] https://orcid.org/0000-0003-0804-6989
rdramirezo@unaj.edu.pe – Universidad Nacional de Juliaca
[3]https://orcid.org/0000-0003-4142-7230
angeli@upeu.edu.pe – Universidad Peruana Unión
[4] https://orcid.org/0000-0001-6683-8859
jr.huanca@unaj.edu.pe – Universidad Nacional de Juliaca
[5]https://orcid.org/0000-0001-8641-0833
mguillen@unaj.edu.pe - Universidad Nacional de Juliaca
[6]https://orcid.org/0000-0003-4760-6467
rhumpiri@unap.edu.pe - Universidad Nacional del Altiplano
*https://orcid.org/0009-0006-9189-0994
milton.humpiri@upeu.edu.pe – Universidad Peruana Unión

* Correspondence: milton.humpiri@upeu.edu.pe

Abstract: The objective of this research study was to evaluate two machine learning techniques, including Decision Trees and Random Forests, in order to predict students' academic performance. The results indicated that the Random Forests algorithm with hyperparameters of 40 trees and a maximum depth of 20 levels, a higher accuracy of 0.77 is achieved. In addition, the value of the AUROC indicator in the ROC curve of the Random Forests algorithm is greater than the threshold of 0.7 reaching an acceptable estimate compared to the Decision Tree algorithm, whose value is 0. 69 thus demonstrating that the Random Forests algorithm was the most accurate in predicting academic performance, this finding is relevant for educational institutions in general, as it can help to define follow-up and support policies for those students at academic risk. In addition, the potential application of these machine learning techniques in the different careers of the Faculty of Industrial Process Engineering of the National University of Juliaca was estimated.
Keywords: Learning, Maching Learning, prediction, academic performance.

1. Introduction

As the amount of stored data grows, the need arises to examine and unravel valuable information from voluminous and complex datasets. Extracting useful and hidden knowledge from these datasets has become crucial in various fields in this competitive world, such as data mining, also known as knowledge discovery in databases (KDD), allows revealing hidden patterns and obtaining valuable and non-trivial information from the large amount of stored data (Alwarthan et al., 2022).

Student academic performance is the most critical indication of educational progress in any country. Essentially, student academic performance is influenced by gender, age, teaching staff, and student learning. Predicting students' academic success has gained much interest in education. In other words, student performance refers to the extent to which students achieve immediate and long-term learning goals plazo (Kumar Yadav & Saurabh, 2012). An excellent academic record is an essential factor for a high-quality university based on its rankings. As a result, its rankings improve when an institution has a strong academic record and achievements. From a student's perspective, maintaining outstanding academic performance increases the chances of getting a job, as excellent academic performance is one of the main aspects evaluated by employers (Shahiri et al., 2015).

The incorporation of information technology (IT) in education can generate superior educational outcomes for institutions. A clear example is the impact of artificial intelligence (AI) on the learning process, which offers a wide range of applications. AI-based technologies in education have gained popularity for their ability to engage students and improve both the quality of education and traditional teaching methods, highlighting their ability to collect large volumes of student data (Academic System). Learning assessment is a crucial component to measure its efficiency (educational process). It is desirable that most students complete their study plans within the established deadlines, but there are cases in which certain factors prevent some students from keeping pace and end up in a situation of academic risk, with the consequent failure to meet personal, family and social expectations.

The incursion of technology in education to maximize the learning experience and related aspects, has given rise to the establishment of Technology Enhanced Learning (TEL) and within this, various concepts related to analytics in the field of education such as Educational Data Mining (EDM), Academic Analytics (AA), and Learning Analytics (LA) have been defined (Ayesha et al., 2010). Concepts that go hand in hand with machine learning. These concepts convert educational data into useful information that allows to take previous actions or to promote teaching and learning.

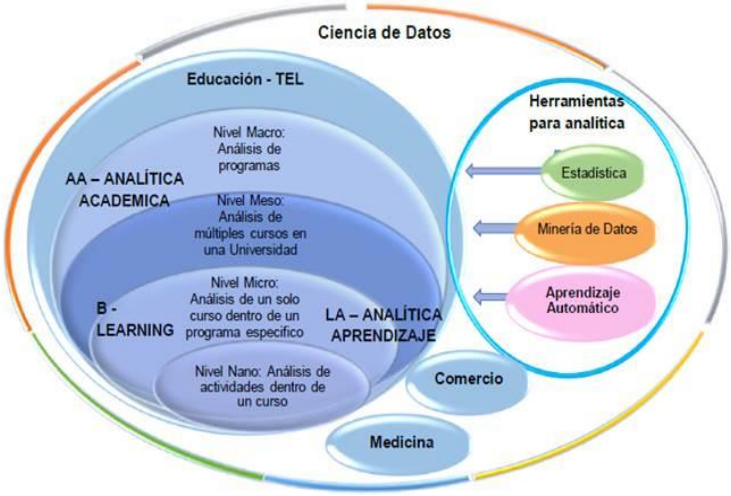

Fig 1: Data Science

Predicting and analyzing student performance is critical to help educators recognize students' weaknesses and help them improve their grades. Likewise, students can improve their learning activities and administrators can improve their operations (Hasan et al., 2020). Timely prediction of student performance enables educators to identify underperforming individuals and intervene early in the learning process to apply necessary interventions. ML is a novel approach with numerous applications that can make predictions on data (Kushwaha et al., 2020). ML techniques in educational data mining aim to model and detect meaningful hidden patterns and usable information from educational contexts (Salah Hashim et al., n.d.). Furthermore, in the academic field, ML approaches are applied to large datasets to represent a wide range of student characteristics as data points. These strategies can benefit various fields by achieving several goals, including extracting patterns, predicting behavior, or identifying trends (Gray & Perkins, 2019), allowing educators to offer the most effective methods for learning and tracking and monitoring student progress.

Academic Analytics is used by institutions to monitor the progress of institutional objectives, efficiency at the end of careers, impact of dissemination, etc. The reason lies in the fact that educational institutions are accountable to state agencies that evaluate the efficiency of education, so they are applying this type of analytics that has allowed them to change the way they make decisions (García-Tinisaray et al., 2018). Finally, Learning Analytics helps to personalize the student experience, predict dropout and dropout rates, and the design of actions to improve learning (Lonn et al., 2015). Academic analytics contains learning analytics.

In educational systems, both at the basic and higher levels, learning assessment plays a key role in measuring the effectiveness of the educational process. It is important that most students are able to complete their curricula

within the established deadlines. However, at times, various factors may hinder the progress of some students, placing them in a situation of academic risk. This leads to a failure to meet personal, family and social expectations that had been established.

1.1 Other background information

There are different definitions of academic performance. Some, such as Garbanzo Vargas (2007) conceptualize academic performance as the sum of different and complex factors that act on the learner. Touron (1985) states that academic performance is a result of learning, aroused by the pedagogical intervention of the teacher and produced in the student; Rojas Torres (2013)conceptualizes that it is a series of factors that revolve around the final results of the effort made by the student; and (García-Tinisaray et al., 2018), considers it as the main indicator of success or failure of the student, for such reason it has been considered as one of the important aspects when analyzing results on the teaching-learning process.

2. Academic performance

Another aspect is how to measure it. Page et al. (1990) state that it is the arithmetical result of passing a subject, contrary to what Vidal García (1999) thinks, since, according to him, grades are a measure of the results of teaching, but not strictly of its quality. Other authors express that objective tests are the adequate means to determine it because the answers are short and precise, without the subjective influence of the teacher. A third way is according to the number of subjects passed, since the number of subjects passed per year is a more adequate indicator of student performance than the average (Abu Bakar et al., 2023). Finally, Rodríguez Ayán & Ruiz Díaz (2011) state that it should be measured according to the number of credits accumulated, since this allows a comparison to be made between the credits accumulated by the student during a certain time of study and the credits that, according to the study plan, should have been accumulated in the programmed time.

Given this problem, the question arises as to whether it is possible to reliably predict the academic performance of students, especially those at academic risk? It is also relevant to determine which are the most determinant variables in the academic performance of students? The results to these questions would allow educational institutions to design policies and strategies to prevent the academic risk of students, such as curricula, teacher profile and especially teaching methodology.

The objective of the work was to evaluate the feasibility of using Maching Learning techniques, specifically Decision Trees and Random Forests, to predict students' academic performance, using information from previous academic results, as well as a group of demographic and social variables that are not normally recorded by educational institutions, but that can be collected through questionnaires.

In the development of the research, the two techniques used were described, presenting the evaluations and results; as well as, the accuracy and convenience of each of them and their possible application at the National University of Juliaca.

3. Materials and Methods

What is Maching Learning?

Machine learning, known in Spanish as aprendizaje automático or aprendizaje de máquina, was born as an ambitious idea of AI in the 1960s. To be more precise, it was a subdiscipline of AI, a product of computer science and neuroscience (Adext, 2017).

Machine Learning is a form of AI that allows a system to learn from data instead of learning through explicit programming. However, machine learning is not a simple process. As the algorithm ingests training data, it is possible to produce more accurate data-driven models. A machine learning model is the information output that is generated when you train your machine learning algorithm with data. After training, providing a model with an input will give you an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a forecast based on the data that trained the model (IBM, 2022).

Pacheco (2017)states that, "The future is not just around the corner, it is already happening." At least that is what Philip Kotler, professor and one of the world's most influential marketing gurus, states with conviction.

3.1 Iterative Learning

Machine Learning allows models to be trained on data sets before being implemented. Some Machine Learning models are online and continuous. This iterative process of online modeling leads to an improvement in the types of associations made between data elements. Due to their complexity and size, these patterns and associations could easily have been missed by human observation. After a model has been trained, it can be used in real time to learn from the data. Improvements in accuracy are the result of the training process and automation that are part of Machine Learning (IBM, 2022).

3.2 Maching Learning Techniques Used

3.2.1 Decision Tree

A decision tree is a nonparametric supervised learning algorithm, which is used for both classification and regression tasks. It has a hierarchical tree structure, consisting of a root node, branches, internal nodes and leaf nodes; a decision tree starts with a root node, which has no incoming branches. Outgoing branches from the root node feed internal nodes, also known as decision nodes. Depending on the available features, both types of nodes perform evaluations to form homogeneous subsets, which are indicated by leaf nodes or terminal nodes. Leaf nodes represent all possible outcomes within the data set (IBM, 2020).

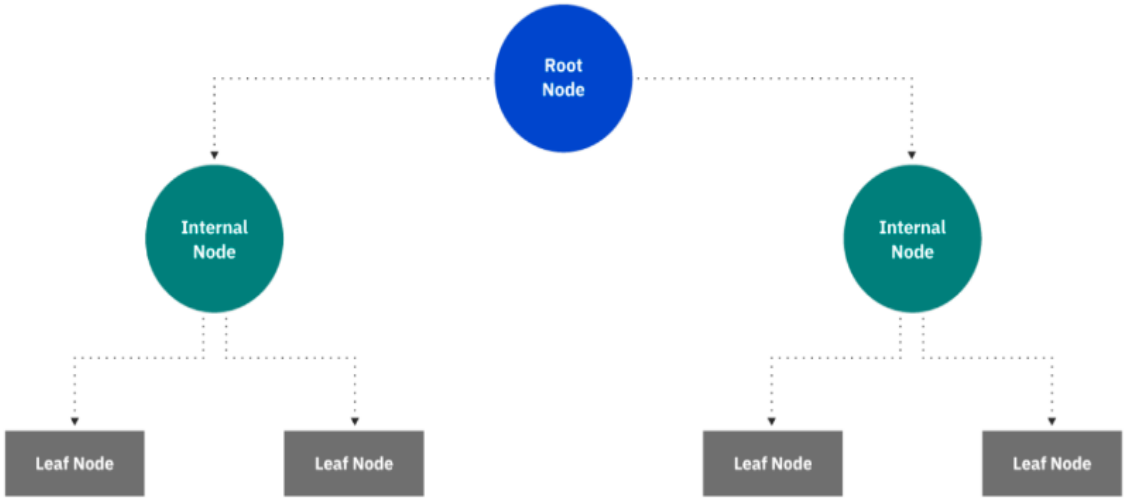

Fig 2: Decision tree

Decision Trees are a machine learning technique that employs a hierarchical structure similar to an inverted tree to recursively divide a data set into smaller groups. The root node of the tree represents the entire data set, while the terminal nodes contain the values of the target variable or classification sought. During the splitting process, attributes are selected that allow a homogeneous separation of the dataset into smaller groups (Charbuty & Abdulazeez, 2021).

Types of decision trees

- ID3: Ross Quinlan is credited with the development of ID3, which is short for "Iterative Dichotomiser 3". This algorithm leverages entropy and information gain as metrics for evaluating candidate splits.
- C4.5: This algorithm is considered a later iteration of ID3, which was also developed by Quinlan. It can use information gain or gain ratios to evaluate split points within decision trees.
- CART: The term CART is an abbreviation for "classification and regression trees" and was introduced by Leo Breiman. This algorithm generally uses the Gini impurity to identify the ideal attribute for division. The Gini impurity measures the frequency with which a randomly chosen attribute is misclassified. When evaluated using the Gini impurity, a lower value is more ideal.

3.2.2 Random Forest

The random forest algorithm is an extension of the bagging method, as it uses both bagging and feature randomization to create an uncorrelated forest of decision trees. Feature randomization, also known as feature clustering or "the random subspace method", generates a random subset of features, which ensures low correlation

between decision trees. This is a key difference between decision trees and random forests. While decision trees consider all possible feature splits, random forests only select a subset of those features (IBM, 2023).

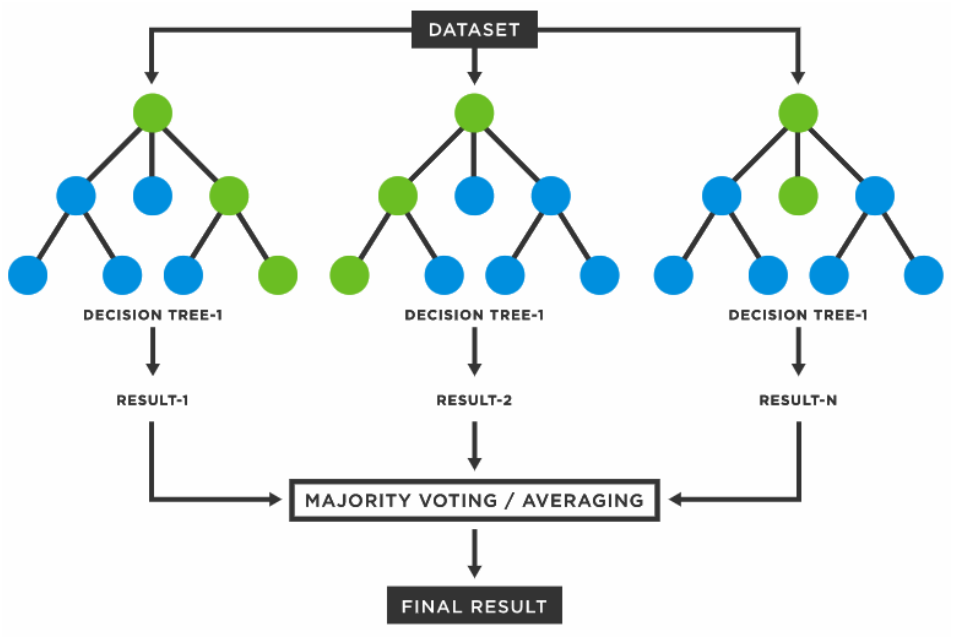

Fig 3: Random forest

Random Forests are supervised machine learning algorithms that use an ensemble learning technique, based on the use of multiple Decision Trees, constructed from a random selection of attributes from the total set of independent variables. Each tree in the forest evaluates a random sample with replacement from the data set (a process known as bootstrapping). Subsequently, the results of all classification trees are considered and, using the "wisdom of crowds" principle, the most frequent classifications of the set of trees are considered as the final solution (Breiman, 2004).

Data used: For the evaluation of the described techniques, the database corresponding to the Portuguese Language course, hosted in the Maching Learning Repository of the University of California, Irvine, will be used. This database is composed of 649 records, each representing one student, and 33 variables.

Table 1: Number of Portuguese language students

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

4. Results

4.1 Decision Tree Algorithm

1. Installing Library

```
import pandas as pd
```

2. Upload the Database

```
d = pd.read_csv('/content/drive/MyDrive/student/student-por.csv', sep=';')
```

```
len(d)
```

3. Calculating whether the student passes or fails the subject (0: failed and 1: passed).

```
d['pass'] = d.apply(lambda row: 1 if (row['G1']+row['G2']+row['G3']) >= 35 else 0, axis=1)
d = d.drop(['G1', 'G2', 'G3'], axis=1)
```

4. Turning qualitative data into numerical data

```
d = pd.get_dummies(d, columns=['school','sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', 'guardian', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic'])
```

5. Training and testing

```
d_train = d[:500]
d_test = d[500:]
d_train_att = d_train.drop(['pass'], axis=1)
d_train_pass = d_train['pass']
d_test_att = d_test.drop(['pass'], axis=1)
d_test_pass = d_test['pass']
d_att = d.drop(['pass'], axis=1)
d_pass = d['pass']
```

6. Number of students selected

```
import numpy as np
print("Pasan: %d de %d (%.2f%%)" % (np.sum(d_pass), len(d_pass), 100*float(np.sum(d_pass)) / len(d_pass)))
```

7. Adjusting the decision tree

```
from sklearn import tree
t = tree.DecisionTreeClassifier(criterion="entropy", max_depth=5)
t = t.fit(d_train_att, d_train_pass)
```

8. Calculate and display the Accuracy as well as display the average score +/- 2 standard deviations (covering 95% of the scores).

```
t.score(d_test_att, d_test_pass)
from sklearn.model_selection import cross_val_score
scores = cross_val_score(t, d_att, d_pass, cv=5)
print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

9. Visualizing the decision tree

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
ax.errorbar(depth_acc[:,0], depth_acc[:,1], yerr=depth_acc[:,2])
plt.show()
```

10. Calculating and displaying the confusion matrix, with the maximum depth (2) having the highest level of accuracy

```
from sklearn import tree
t = tree.DecisionTreeClassifier(criterion="entropy", max_depth=2)
t = t.fit(d_train_att, d_train_pass)
from sklearn.metrics import confusion_matrix, roc_curve, roc_auc_score
#t = t.fit(d_train_att, d_train_pass)
y_pred = t.predict(d_train_att)
cm = confusion_matrix(d_train_pass, y_pred)
print("Confusion matrix")
print(cm)
```

11. Calculate and display the ROC curve

```
r_probs = [0 for _ in range(len(d_train_pass))]
nb_probs = t.predict_proba(d_train_att)
nb_probs = nb_probs[:, 1]
r_auc = roc_auc_score(d_train_pass, r_probs)
nb_auc = roc_auc_score(d_train_pass, nb_probs)
print("AUROC = %.3f" % (nb_auc))
nb_fpr, nb_tpr, _ = roc_curve(d_train_pass, nb_probs)
plt.plot(nb_fpr, nb_tpr, marker=".", label="AUROC = %.3f" % nb_auc)
plt.title("ROC plot")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.legend()
plt.show()
```

After verifying that the optimal depth to achieve the highest level of accuracy is 2 levels, we proceed to use this depth to run the algorithm, and the following confusion matrix is obtained as a result.

Table 2: *Confusion matrix - Decision trees*

		Prediction	
		0	1
Actual Value	0	101	151
	1	3	245

The model has been successful in 245 cases of passing the course and in 101 cases of not passing the course. However, 3 cases have been identified where the model misclassified passes as non-passes (false positives) and 151 cases where the model misclassified non-passes as passes (false negatives).
In addition, an accuracy (Accuracy) of the model has been calculated: 0.69 and the ROC curve can be observed.

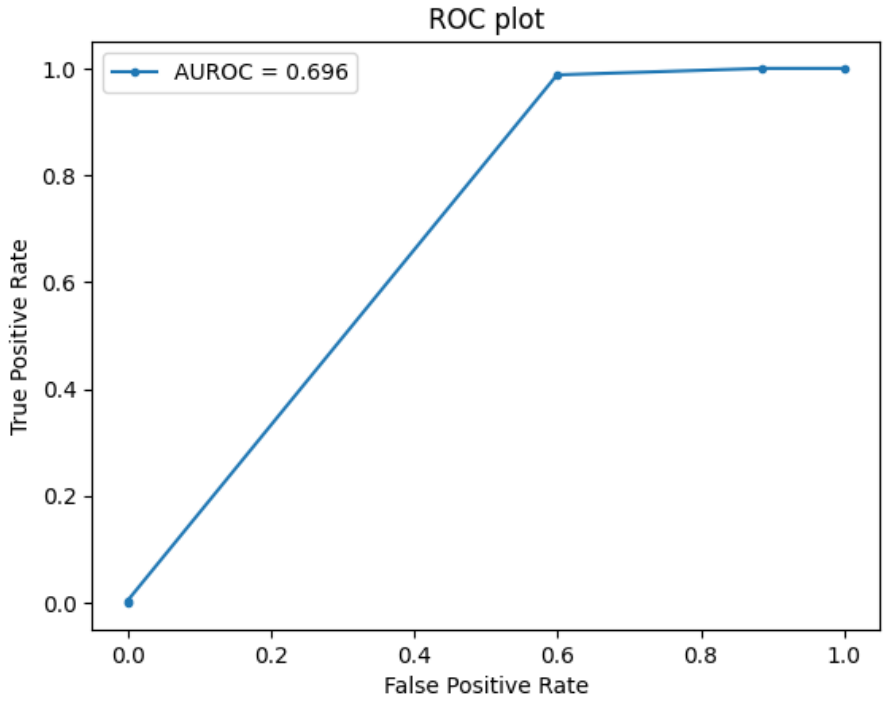

Fig 4: ROC curve - Decision trees

4.2 Random Forest Algorithm

1. Installing Library

```
import pandas as pd
```

2. Define the rf_bch function for:

a. Normalize the data.
b. Split the data into training and testing.
c. Create the random forest model.
d. Train with the training data.
e. Calculate and display the confusion matrix.
f. Calculate and display the ROC curve.
g. Measure the level of precisión

```
def rf_bch(d,i,j,nom):
 var_dep = d[nom]
 var_ind = d.drop(nom,axis= 1)
 var_ind_norm = StandardScaler().fit_transform(var_ind)
 var_ind_ent, var_ind_test, var_dep_ent, var_dep_test =
train_test_split(var_ind_norm,var_dep,test_size=0.3)
rf= RandomForestClassifier(n_estimators=10,max_depth=4)
 rf.fit(var_ind_ent,var_dep_ent)
 y_pred = rf.predict(var_ind_test)
 cm = confusion_matrix(var_dep_test, y_pred)
 print("Confusion matrix")
 print(cm)
 r_probs = [0 for _ in range(len(var_dep_test))]
```

```
nb_probs = rf.predict_proba(var_ind_test)
nb_probs = nb_probs[:, 1]
r_auc = roc_auc_score(var_dep_test, r_probs)
nb_auc = roc_auc_score(var_dep_test, nb_probs)
print("AUROC = %.3f" % (nb_auc))
nb_fpr, nb_tpr, _ = roc_curve(var_dep_test, nb_probs)
plt.plot(nb_fpr, nb_tpr, marker=".", label="AUROC = %.3f" % nb_auc)
plt.title("ROC plot")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.legend()
plt.show()
return rf.score(var_ind_test,var_dep_test)
```

3. Upload the Database

```
d = pd.read_csv('/content/drive/MyDrive/student/student-por.csv', sep=';')
len(d)
```

4. Turning qualitative data into numerical data

```
d = pd.get_dummies(d, columns=['school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', 'guardian', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic'])
```

5. Define the rf_arq function for:

a. Normalize the data.
b. Split the data into training and testing.
c. Create the random forest model.
d. Train with the training data.
e. Measure the level of accuracy.

```
def rf_arq(d,i,j,nom):
 var_dep = d[nom]
 var_ind = d.drop(nom,axis= 1)
 var_ind_norm = StandardScaler().fit_transform(var_ind)
 var_ind_ent, var_ind_test, var_dep_ent, var_dep_test =
train_test_split(var_ind_norm,var_dep,test_size=0.3)
rf= RandomForestClassifier(n_estimators=i,max_depth=j)
 rf.fit(var_ind_ent,var_dep_ent)
 return rf.score(var_ind_test,var_dep_test)
```

6. Call the rf_arq function

```
mayor = 0
for i in range(10, 101, 10):
    for j in range(10, 101, 10):
        A = rf_arq(d, i, j, 'pass')
        if A > mayor:
```

```
        mayor = A
        n1 = i
        n2 = j
print("Arquitectura: %dx%d, Accuracy: %0.4f" % (n1, n2, mayor))
```

7. Calculating and displaying the confusion matrix and the ROC curve.

```
A = rf_bch(d,n1,n2,'pass')
print("Arquitectura: %dx%d, Accuracy: %0.4f" % (n1, n2, A ))
```

After running the algorithm with the data we can see the level of accuracy by random forest architecture. By selecting the best fitting architecture: n_estimator= 40, max_depth = 20, Accuracy: 0.8031 we can see the following confusion matrix.

Table 3: *Confusion Matrix - Random Forest*

		Prediction	
		0	1
Actual Value	0	19	31
	1	6	71

The model has been successful in 71 cases of passing the subject and 19 cases of failing the subject. However, 6 cases have been identified in which the model incorrectly classified approvals as non-approvals (false positives) and 31 cases in which the model incorrectly classified non-approvals as approvals (false negatives). In addition, an accuracy (Accuracy) of the model has been calculated: 0.71 and the ROC curve can be observed.

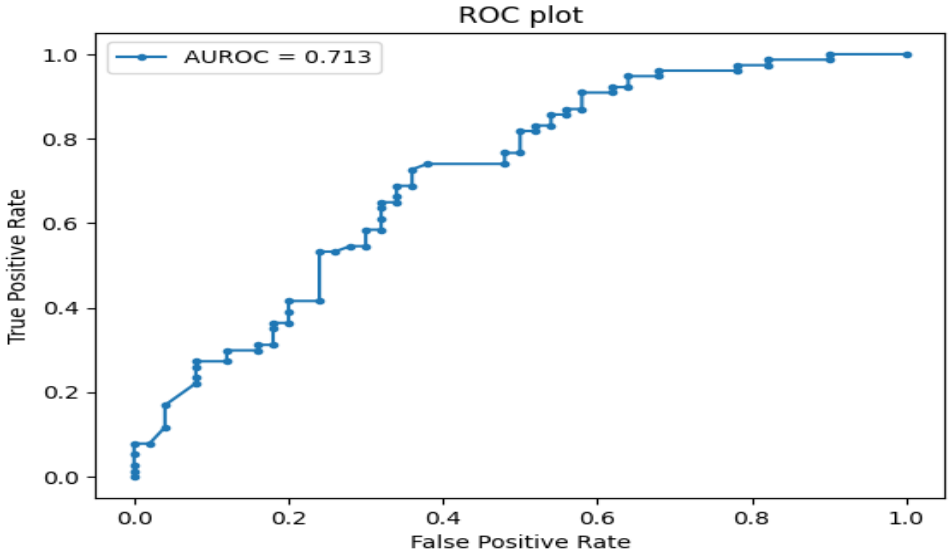

Fig 5: ROC curve - Random Forest

5. Discussion

In order to compare the 2 machine learning algorithms used, we will use the most usual performance measures in supervised learning, which are the Confusion Matrix, the accuracy and the ROC (Receiver Operating Characteristic) curve.

Algorithm	Accuracy	AUROC
Decision tree	**0.68**	**0.69**
Random forest	**0.77**	**0.71**

It can be observed that using the Random Forests algorithm with hyperparameters of 40 trees and a maximum depth of 20 levels, a higher accuracy of 0.77 is achieved. In addition, the value of the AUROC indicator in the ROC curve of the Random Forests algorithm is greater than the threshold of 0.7, reaching an acceptable estimate compared to the Decision Tree algorithm, whose value is 0.69.

Maching Learning applications for the professional schools of the Faculty of Industrial Process Engineering of the National University of Juliaca (UNAJ, 2021).

Textile and Apparel Engineering

- Fabric quality analysis
- Textile product demand forecasting
- Production optimization
- Consumption pattern analysis
- Quality control in textile processes

Industrial engineering

- Optimization of production processes
- Supply chain planning and management
- Quality control
- Predictive maintenance
- Production data analysis

Food Industry Engineering

- Quality control
- Optimization of production processes
- Production data analysis
- Automatic food labeling
- New product development

Bibliographic References

[1] Abu Bakar, N., Yusop, H., Ali, N. M., Fauziah, N., & Bakar, A. (2023). Determinants of Students' Academic Performance in Higher Learning Institutions in Malaysia. *International Journal of Academic Research in Business and Social Sciences*, *13*(2), 1496–1508. https://doi.org/10.6007/IJARBSS/v13-i2/16250

[2] Adext. (2017). *¿Qué es machine learning?* Adext Blog. https://blog.adext.com/machine-learning-guia-completa/

[3] Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review. *Applied Computational Intelligence and Soft Computing*, *2022*. https://doi.org/10.1155/2022/8924028

[4] Ayesha, S., Mustafa, T., Sattar, A. R., & Khan, M. I. (2010). Data Mining Model for Higher Education System. *European Journal of Scientific Research*, *43*, 24–32.

[5] Breiman, L. (2004). *Consistency for a simple model of random forests*. https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf

[6] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

[7] Garbanzo Vargas, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*, 4363. https://doi.org/10.31206/rmdo072018

[8] García-Tinisaray, D., Mejias, J. L. P., & Pichardo, J. M. M. (2018). Learning Analytics as an analysis factor of university academic performance. *CEUR Workshop Proceedings*, *2231*.

[9] Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22–32. https://doi.org/10.1016/J.COMPEDU.2018.12.006

[10] Hasan, R., Palaniappan, S., Mahmood, S., Sarker, K. U., & Abbas, A. (2020). Modelling and predicting student's academic performance using classification data mining techniques. *International Journal of*

Business Information Systems, *34*(3), 403–422. https://doi.org/10.1504/IJBIS.2020.108649

[11] IBM. (2020). *¿Qué es un árbol de decisión?* https://www.ibm.com/es-es/topics/decision-trees

[12] IBM. (2022). *¿Qué es Machine Learning?* https://www.ibm.com/mx-es/analytics/machine-learning

[13] IBM. (2023). *¿Qué es el random forest?* https://www.ibm.com/es-es/topics/random-forest?mhsrc=ibmsearch_a&mhq=bosque aleatorio

[14] Kumar Yadav, S., & Saurabh, P. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification Saurabh Pal. *World of Computer Science and Information Technology Journal (WCSIT)*, *2*(2), 51–56.

[15] Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant Applications of Machine Learning for COVID-19 Pandemic. *Journal of Industrial Integration and Management*, *05*(04), 453–479. https://doi.org/10.1142/S2424862220500268

[16] Lonn, S., Aguilar, S. J., & Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Computers in Human Behavior*, *47*, 90–97. https://doi.org/10.1016/j.chb.2014.07.013

[17] Pacheco, S. (2017). *Camino hacia la sociedad de la inteligencia artificial*. La Vanguardia. https://www.lavanguardia.com/economia/20170119/413499547095/sociedad-inteligencia-artificial-the-valley-digital.html

[18] Page, M. A., Monreal, M. J. B., Sopeña, J. A. C., Victoria, J. C., Cubias, M. J. E., López, C. G., Soto, J. L. G., Bueno, C. G., Suárez, S. C. J., Pérez, B. M., Romero, L. M.-J., CebaUos, A. L. M., Ruiz, A. S., & Marco, C. T. (1990). *Hacia un Modelo Causal del Rendimiento Académico. Madrid. Editorial* (Issue January). https://www.researchgate.net/publication/39127951

[19] Rodríguez Ayán, M. N., & Ruiz Díaz, M. Á. (2011). Indicadores de rendimiento de estudiantes universitarios versus créditos acumulados. *Revista de Educación, ISSN 0034-8082, Nº 355, 2011, Págs. 467-492*, *355*, 467–492. https://dialnet.unirioja.es/servlet/articulo?codigo=3639395&info=resumen&idioma=ENG

[20] Rojas Torres, L. (2013). Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de colegio como variables control. *Actualidades Investigativas En Educación*, *13*(1), 1–24. https://doi.org/10.15517/aie.v13i1.8562

[21] Salah Hashim, A., Akeel Awadh, W., & Khalaf Hamoud, A. (n.d.). *Student Performance Prediction Model based on Supervised Machine Learning Algorithms*. https://doi.org/10.1088/1757-899X/928/3/032019

[22] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/J.PROCS.2015.12.157

[23] Touron, J. (1985). La predicción del rendimiento académico: procedimientos, resultados e implicaciones. *Revista Española de Pedagogía.*, 1–23. https://dadun.unav.edu/handle/10171/18774

[24] UNAJ. (2021). *Facultad de Ingeniería de Procesos Industriales*. https://unaj.edu.pe/facultad-procesos-industriales

[25] Vidal García, J. (1999). *Plan nacional de evaluación de la calidad de las universidades. Indicadores en la universidad: información y decisiones - Publicaciones - Ministerio de Educación y Formación Profesional*. https://sede.educacion.gob.es/publiventa/detalle.action?cod=7873