	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022




**UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS**

# **PRY20220164 - MODELO DE ANÁLISIS PREDICTIVO PARA EL MONITOREO DE LA DESERCIÓN ESTUDIANTIL APLICANDO MACHINE LEARNING EN LA EDUCACIÓN SUPERIOR UNIVERSITARIA EN PERÚ**


## **Desarrollo del Objetivo Específico 2**

**v.0.1**

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022


### Historial de Versiones

Fecha	Versión	Autor	Descripción de los cambios
12/05/2022	0.1	Ashley Jesús Llontop Omar Jiménez Ramírez	Creación y desarrollo del documento.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

## Índice

Historial de Versiones.....	2
Índice.....	3
1    Introducción.....	4
2    Revisión de la Literatura existente .....	5
3    Recopilación de algoritmos usados .....	6
4    Benchmarking de algoritmos de predicción.....	12
5    Criterios para seleccionar variables de deserción .....	14
6    Modelo de Análisis Predictivo propuesto.....	17
7    Implementación del Modelo de Análisis Predictivo.....	18
8    Obtención de datos.....	18
9    Entrenamiento del Modelo de Análisis Predictivo.....	19
10    Desarrollo de Aplicación Web de Monitoreo de Deserión .....	19
11    Bibliografía .....	22

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022


## 1 Introducción

Nombre del proyecto	Fecha de elaboración
Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú	12/05/2022

La deserción académica en el Perú tiene diferentes impactos negativos en el sector educativo del país que van desde una disminución en la producción científica hasta la falta de profesionales que cubran puestos de trabajo requeridos. En ese sentido, el presente proyecto pretende encontrar un mecanismo que permita a las instituciones educativas conocer la situación en la que se encuentran sus estudiantes mediante un análisis predictivo que identifique oportunamente a los alumnos en riesgo de deserción.

Existen diferentes factores que explican el fenómeno de la deserción académica. Algunos autores mencionan que la salud emocional, así como los desencadenantes de estrés son algunos de los motivos por los cuales los alumnos tienden a interrumpir sus estudios. Es importante que las instituciones educativas trabajen en estrategias para la continuidad de los estudios de sus alumnos y asegurar así, su éxito profesional. (Pascoe et al., 2020).

En el presente documento, definiremos la estructuración de nuestro modelo de análisis predictivo. Es decir, la lista de variables que usaremos para la predicción y el algoritmo de Machine Learning que ejecutará el modelo. En ese sentido, procederemos a desarrollar un detallado benchmarking para identificar el mejor modelo aplicado a la solución.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

## 2 Revisión de la Literatura existente

Como parte de la investigación previa al desarrollo de la solución, el equipo de proyecto elabora un Estado del Arte. Este documento tiene por objetivo dar una visión general de la situación actual del problema planteado, así como algunos trabajos previos que aplican como técnica Machine Learning a la deserción universitaria.

El Estado de Arte del proyecto está dividida en tres categorías que corresponden a tres tipos de papers científicos:

- Problema

La inclusión de esta categoría es una respuesta a la necesidad del equipo de comprender el problema desde diferentes puntos de vista, en diferentes contextos alrededor del mundo. En ese sentido, el estado de arte incluye once (11) artículos cuyo objetivo es estudiar el problema holísticamente. En esta categoría de artículos encontramos las diferentes causas y consecuencias que circunscriben y fundamentan la existencia del problema y por qué deberíamos buscar una solución para tal situación.


- Técnica

En la categoría de técnica se incluye un total de diez (10) artículos académicos que nos ayudan a comprender la técnica que hemos elegido: machine learning. En ese sentido, los papers académicos que incluimos en esta categoría nos dicen diferentes criterios y prácticas que debemos tener en cuenta cuando trabajamos con la técnica mencionada.

Los artículos que recopilamos y analizamos tienen diferentes recomendaciones, lineamientos, buenas prácticas y otras características que podemos usar en el desarrollo de nuestra solución. Además, hemos incluido artículos que aplican machine learning como técnica, pero en otras industrias y con problemas diferentes a la deserción universitaria. De esta manera, podemos observar cómo se aborda un problema y cómo se desarrolla una solución basada en machine learning.

- Técnica Aplicada

Finalmente, en la categoría de técnica aplicada, el equipo ha incluido artículos cuyo objetivo es encontrar una solución al problema de la deserción universitaria usando machine learning. En estos artículos se han desarrollado diferentes modelos de análisis predictivo que han sido resultado de un cruce de algoritmos de machine learning y diferentes variables de deserción. En el siguiente capítulo, resumiremos los papers de esta categoría y recopilaremos cuáles fueron los algoritmos y variables utilizadas. Como punto siguiente, el equipo define qué algoritmo y variables se utilizarán en el proyecto.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

### 3 Recopilación de algoritmos usados

Artículo 01: "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques"

Este artículo pretende estudiar el fenómeno de la deserción estudiantil mediante el uso de diferentes algoritmos de machine learning y, por consiguiente, logrando diferentes resultados en precisión de predicción. El enfoque de este artículo fue estudiar el problema en un ambiente de aprendizaje virtual. Para lograr dicho objetivo, los autores tomaron una base de datos de un curso introductorio de la Universidad de Constantino el Filósofo en Nitra en la que se recopilaban los datos de más de 250 alumnos entre los años 2016 y 2020. Para cada uno de los años en el rango, se calculó el porcentaje de deserción. El desarrollo de este artículo consideró a CRISP-DM como metodología para la implementación del modelo de análisis predictivo.

Los algoritmos que se emplearon en este artículo fueron:

Tabla 1

*Algoritmos usados en el artículo (Kabathova & Drlik, 2021)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Kabathova & Drlik, 2021)	Naïve Bayes (NB)	77%
(Kabathova & Drlik, 2021)	Random Forest (RF)	93%
(Kabathova & Drlik, 2021)	Neural Network (NN)	88%
(Kabathova & Drlik, 2021)	Logistic Regression (LR)	93%
(Kabathova & Drlik, 2021)	Support Vector Machines (SVM)	92%
(Kabathova & Drlik, 2021)	Decision Tree (DT)	90%


Artículo 02: Dropout early warning systems for high school students using machine Learning

El artículo estudia los factores de deserción en la escuela secundaria aplicando el algoritmo Random Forest debido a su gran popularidad en la predicción. El objetivo del estudio es detectar tempranamente a un alumno en riesgo a desertar. La fuente de los datos fue del Sistema de Información de Educación Nacional (NEIS) en el 2014. Los autores consideraron un total de más de 165,715 datos de escolares.

Tabla 2

*Algoritmos usados en el artículo (Chung & Lee, 2019)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Chung & Lee, 2019)	Random Forest	95%

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

### Artículo 03: Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education

En el artículo el problema que se aborda es la deserción universitaria aplicando métodos y técnicas de minería de datos. El objetivo de los autores es identificar el mejor modelo mediante indicadores de rendimiento de pronóstico y prevención. Los autores contemplaron un dataset de 4365 registros de estudiantes de la Universidad Nacional de Moquegua (UNAM), Perú. Los resultados concluyen que el algoritmo RF obtuvo la mejor precisión para detectar casos de deserción universitaria en la UNAM.

Tabla 3  
*Algoritmos usados en el artículo (Flores et al., 2022)*


ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Flores et al., 2022)	Random Forest	97%
(Flores et al., 2022)	Random Tree	93%
(Flores et al., 2022)	J48	96%
(Flores et al., 2022)	REPTree	95%
(Flores et al., 2022)	JRIP	95%
(Flores et al., 2022)	OneR	79%
(Flores et al., 2022)	Bayes Net	89%
(Flores et al., 2022)	Naive Bayes	89%

### Artículo 04: Bachelor's degree student dropouts: Who tend to stay and who tend to leave?

El artículo abarca las tendencias y factores que implican la deserción académica de estudiantes de primer año. Mediante los algoritmos que se detallan, los autores buscan predecir la deserción de los universitarios. Los datos que se utilizaron fueron registros de estudiantes que se matricularon entre los años 2013 y 2014. La metodología que se empleó fue CRISP-DM.

Tabla 4  
*Algoritmos usados en el artículo (Berka & Marek, 2021)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Berka & Marek, 2021)	ZeroR	90%
(Berka & Marek, 2021)	Decision tree	94%
(Berka & Marek, 2021)	Random Forest	94%
(Berka & Marek, 2021)	Logistic regression	94%

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

#### Artículo 05: Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science

El artículo estudiar la deserción de los estudiantes pertenecientes al programa de Ciencias de la computación (CS) de los países de la Isla del Pacífico. Los autores emplearon el algoritmo RF bajo la metodología CRISP-DM y el lenguaje R. El dataset fue dividido para realizar los test, el primer modelo de 5-fold y 10-fold. Finalmente, ambas divisiones obtuvieron una precisión mayor al 80% pero los autores recomiendan el modelo de 5-folds ya que presenta un mejor rendimiento.

Tabla 5  
*Algoritmos usados en el artículo (Naseem et al., 2019)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Naseem et al., 2019)	Random Forest	80%


#### Artículo 06: A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data

El artículo abarca la deserción de estudiantes en las universidades de España para detectar de manera temprana aquellos estudiantes en riesgo de deserción. Los autores aplican los modelos presentados en cinco periodos: antes de comenzar la universidad, al finalizar el primer, segundo, tercer y cuarto semestre. Con respecto a los datos, el grupo de datos se divide en datos académicos, datos personales y datos del Sistema online de gestión de aprendizaje (LMS).

Tabla 6  
*Algoritmos usados en el artículo (Fernandez-Garcia et al., 2021)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Fernandez-Garcia et al., 2021)	Gradient Boosting	88%
(Fernandez-Garcia et al., 2021)	Random Forest	88%
(Fernandez-Garcia et al., 2021)	Support Vector Machine	90%
(Fernandez-Garcia et al., 2021)	Ensemble	89%



	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

#### Artículo 07: Forecasting Students Dropout: A UTAD University Study

El artículo se centra en la problemática de la deserción estudiantil en las universidades de Portugal. La técnica que emplean es la minería de datos para predecir la deserción universitaria. Con respecto a los datos, los autores usaron la base de datos de UTAD, cuya información involucra el periodo 2011 - 2019. Sin embargo, de acuerdo con los autores, no se priorizó datos socioeconómicos.

Tabla 7

*Algoritmos usados en el artículo (Moreira da Silva et al., 2022)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Moreira da Silva et al., 2022)	CatBoost	88%
(Moreira da Silva et al., 2022)	RF	88%
(Moreira da Silva et al., 2022)	XGBoost	90%
(Moreira da Silva et al., 2022)	Artificial Neural Networks (ANN)	84%


#### Artículo 08: Utilizing early engagement and machine learning to predict student Outcomes

El artículo tiene como objetivo de realizar una identificación temprana de estudiantes en riesgo de deserción. De acuerdo con los autores, la investigación trata de encontrar el mejor modelo de predicción sino el que detecte más temprano los casos (esto incluye a estudiantes en bajo riesgo y en riesgo de deserción). El trabajo utilizó la plataforma WEKA y los datos iniciales fue aquellos estudiantes no graduados de los años 2016-2017 de la universidad de Bangor.

Tabla 8

*Algoritmos usados en el artículo (Gray & Perkins, 2019)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Gray & Perkins, 2019)	C4.5 Trees	86%

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

Artículo 09: Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?

El objetivo del estudio es realizar un análisis de la deserción en la Universidad de Oviedo (España) mediante algoritmos de Machine Learning. Los autores obtuvieron registros de 1055 estudiante mediante una encuesta proporcionada por la universidad de los años 2010-2011. Finalmente, los autores desean que el modelo pueda ser exportado a otra universidades con similares comportamientos de deserción y, así, encontrar nueva relaciones entre las variables.

Tabla 9

*Algoritmos usados en el artículo (Rodríguez-Muñiz et al., 2019)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Rodríguez-Muñiz et al., 2019)	Classification and Regression Trees (CART)	83.50%
(Rodríguez-Muñiz et al., 2019)	C4.5	85.20%
(Rodríguez-Muñiz et al., 2019)	NB	86.20%
(Rodríguez-Muñiz et al., 2019)	RF	86.60%
(Rodríguez-Muñiz et al., 2019)	SVM	85.50%

Artículo 10: Precision education with statistical learning and deep learning: a case study in Taiwan

El artículo analiza la problemática de la deserción en las universidades de Taiwán con la finalidad de ayudar y motivar a los alumnos a seguir con sus estudios. Los autores implementan un algoritmo multicapa para detectar casos tempranos de estudiantes en riesgos. La muestra de datos fue 4748 estudiantes matriculados de los años 2012 – 2013.

Tabla 10

*Algoritmos usados en el artículo (Tsai et al., 2020)*

ARTÍCULO	ALGORITMOS USADOS	PRECISIÓN
(Tsai et al., 2020)	Multilayer perceptron	90%
(Tsai et al., 2020)	Logistic regression	88%



	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

Tabla 11

*Algoritmos usados en los artículos anteriores, tamaño de muestra y precisión lograda*

Autores	Técnica evaluada	Mejor técnica	Tamaño de dataset	Precisión
(Kabathova & Drlik, 2021)	Naïve Bayes, Random Forest, Neural Network, Logistic Regression, Support Vector Machines, Decision Tree	Random Forest	250	93%
(Chung & Lee, 2019)	Random Forest	Random Forest	165,715	95%
(Flores et al., 2022)	Random Forest, Random Tree, J48, REPTree, JRIP, OneR, Bayes Net, Naïve Bayes	Random Forest	4,365	97%
(Berka & Marek, 2021)	ZeroR, Decision tree, Random Forest, Logistic regression	Decision tree	3,339	94%
(Naseem et al., 2019)	Random Forest	Random Forest	963	80%
(Fernandez-Garcia et al., 2021)	Gradient Boosting, Random Forest, Support Vector Machine, Ensemble	Support Vector Machine	NI	90%
(Moreira da Silva et al., 2022)	CatBoost, RF, XGBoost, Artificial Neural Networks (ANN)	Random Forest XGBoost	331	88% 90%
(Gray & Perkins, 2019)	C4.5 Trees	C4.5 Trees	4970	86%
(Rodríguez-Muñiz et al., 2019)	CART, C4.5, NB, RF, SVM	C4.5	1,055	85.20%
(Tsai et al., 2020)	Multilayer perceptron, Logistic regression	Logistic regression	4,748	-


	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

## 4 Benchmarking de algoritmos de predicción

Tabla 12

*Criterios de comparación de algoritmos*

Criterio	Descripción	
Rendimiento	De acuerdo con la CMU School of Information, los algoritmos se pueden comparar según su complejidad algorítmica (BigO) o eficiencia de ejecución. <a href="https://www.cs.cmu.edu/~cburch/pgss97/slides/0715-compare.html">https://www.cs.cmu.edu/~cburch/pgss97/slides/0715-compare.html</a> Se asignará un puntaje de 5 cuando el algoritmo logra el mejor rendimiento; 1 cuando obtenga la peor eficiencia.	
Precisión	Diferentes artículos científicos han implementado modelos usando el mismo set de datos y, sin embargo, diferentes algoritmos. Se demuestra qué algoritmo presenta mejores resultados en circunstancias similares. <a href="#">(Kabathova &amp; Drlik, 2021)</a> Se asignará un puntaje de 5 cuando el algoritmo logra el porcentaje de predicciones más alto; 1 cuando obtenga la peor precisión de predicción.	
Tiempo de entrenamiento	De acuerdo con la consultora AISoma, el tiempo de entrenamiento es crucial al momento de desarrollar una solución basada en análisis predictivo. <a href="https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/">https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/</a> Se asignará un puntaje de 5 cuando el algoritmo logra el menor tiempo de entrenamiento; 1 cuando obtenga el tiempo de entrenamiento más prolongado.	
Interpretabilidad	Se refiere a la capacidad del algoritmo de ofrecer una salida (output) únicamente numérica o si puede entregar clasificaciones categóricas <a href="https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/">https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/</a> Se indicará cualitativamente (“Deficiente”, “Media” o “Buena”).	
Normalización estandarización	/	Se refiere a la capacidad del algoritmo a hacer la predicción con una estandarización de variables anterior o una conversión de variables cualitativas a numéricas <a href="https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/">https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/</a> Se indicará cualitativamente (“Sí”, “Requerida”).

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

## Comparación de algoritmos

En este punto, comparamos los algoritmos que resultaron tener los mejores resultados según los autores de los artículos que incluimos en el estado del arte.

Tabla 13  
*Comparación de algoritmos*


Algoritmo / Criterios	Rendimiento	Precisión	Tiempo de entrenamiento	Interpretabilidad	Normalización / estandarización
Random Forest	RF se posiciona como el algoritmo más veloz 5	5	RF se posiciona como el algoritmo más veloz 5	Buena	No requiere
Decision Tree	DT es óptimo cuando se tienen muchos datos y pocas variables 3	4	DT es óptimo cuando se tienen muchos datos y pocas variables 3	Buena	No requiere
Support Vector Machine	Depende del número de vectores y el número de variables de la data 3	3.5	Depende del número de vectores y el número de variables de la data 3	Media	Requerida
XGBoost	Depende del número de de árboles y la altura de estos 3	4	Depende del número de de árboles y la altura de estos 3	Media	Requerida
Logistic regression	Mejor para aplicaciones de baja latencia 4	4	Mejor para aplicaciones de baja latencia 4	Buena	No requiere

## Fuentes de información

Para obtener la complejidad algorítmica de los algoritmos, consultamos <https://medium.com/analytics-vidhya/time-complexity-of-ml-models-4ec39fad2770>, cuyo contenido nos indica los mejores y peores algoritmos según el rendimiento.

Los puntajes en el criterio de decisión fueron obtenidos de acuerdo con lo leído en los artículos de nuestro estado del arte.

Por otro lado, para realizar la comparación de tiempos de entrenamiento, consultamos [https://marcovirgolin.github.io/extras/details\\_time\\_complexity\\_machine\\_learning\\_algorithms/](https://marcovirgolin.github.io/extras/details_time_complexity_machine_learning_algorithms/). En este blog, el autor indica cuáles son los algoritmos que cuentan con mejores tiempos de entrenamiento y podemos notar que está ligado al rendimiento de estos.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

Finalmente, para los criterios de “Interpretabilidad” y “Normalización/estandarización”, investigamos en los artículos de la consultora AISoma, una empresa alemana de se dedica a la consultoría en Machine Learning e inteligencia artificial.

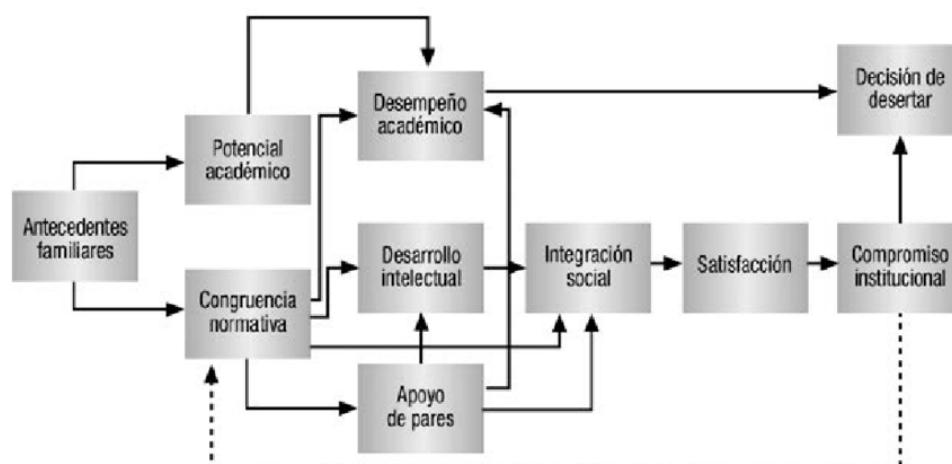
## 5 Criterios para seleccionar variables de deserción

Los modelos de análisis predictivos se componen fundamentalmente de dos elementos: el algoritmo de predicción y las variables que se usarán para determinar las salidas (outputs) del modelo. En ese sentido, es importante que el modelo haga uso de variables de deserción que realmente tengan un impacto significativo sobre el resultado final. En este capítulo, definiremos cuáles son los criterios que deben tomarse para elegir una serie de variables de predicción y confeccionaremos una lista de variables utilizables, de acuerdo con la investigación previa del proyecto y el estado del arte existente.

### Teoría de Spady (1970)


Este autor encuentra su principal argumento en la teoría del suicidio de Durkheim, que indica que una persona llega a considerar el suicidio en el momento en que pierde conexión con su entorno. De la misma manera, Spady argumenta que un alumno considera la interrupción de sus estudios cuando pierde la conexión con la institución en la que se forma. Esto deja entender que las relaciones sociales con los compañeros, así como la escasa afiliación social pueden ser críticas en el proceso de decisión de la deserción universitaria (Spady, 1970).

Figura 1.  
*Flujo de deserción de Spady*



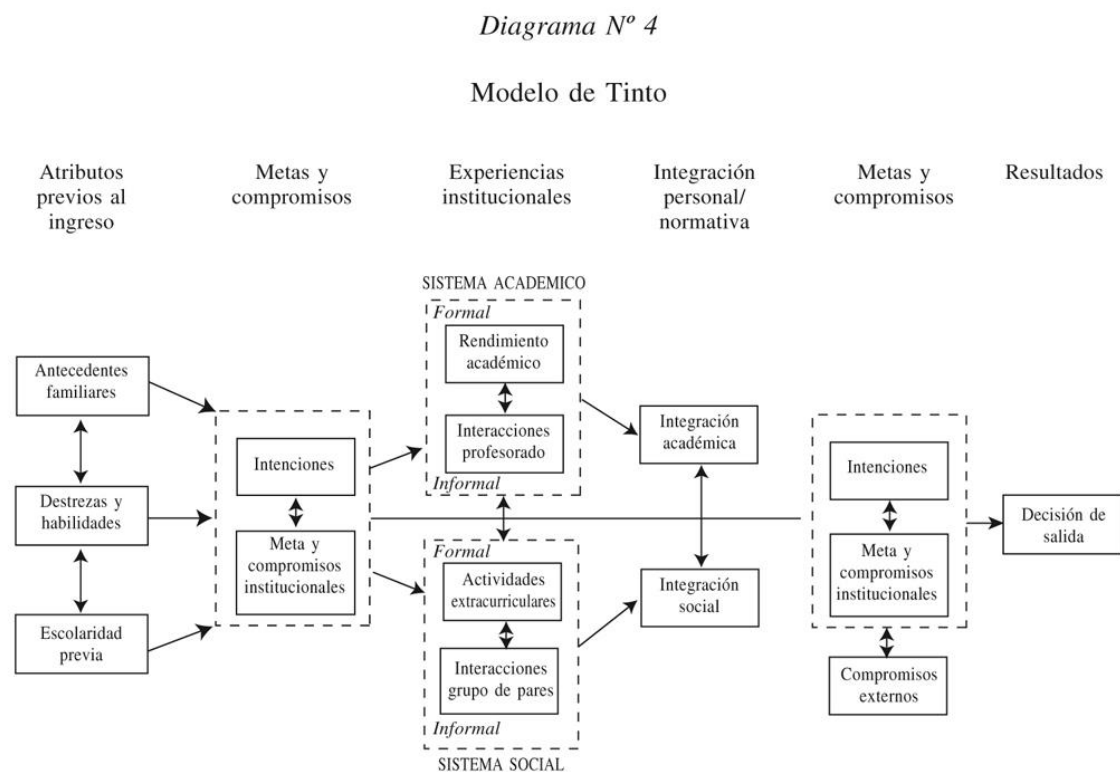
### Teoría del intercambio de Tinto (1975)

La teoría de este autor responde a la necesidad de las instituciones de la época a saber qué es lo que llevaba a los alumnos a tomar la decisión de deserción. Tinto hace una importante mención al trabajo de Spady, revisado anteriormente. Sin embargo, tiene su propio enfoque sobre las causas y motivaciones del proceso de deserción. Entonces,

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

Tinto propone un modelo longitudinal de deserción en el que se puede identificar una conspicua tendencia a variables propias del entorno del alumno que toma la decisión. El entorno familiar, la experiencia escolar, el ambiente universitario, entre otras características de otros ámbitos pueden darnos indicios de cuán susceptible es un alumno al proceso de deserción. Además, Tinto menciona una distinción importante entre retiro (*withdrawal*) y deserción (*dropout*). El primer término está más relacionado al bajo desempeño académico de los alumnos y el segundo a los factores que revisamos en las líneas anteriores.


Figura 2.  
Flujo de deserción de Tinto



Tinto, 1987, p. 114.

### Modelo conceptual de deserción en estudiantes de pregrado de Bean (1985)

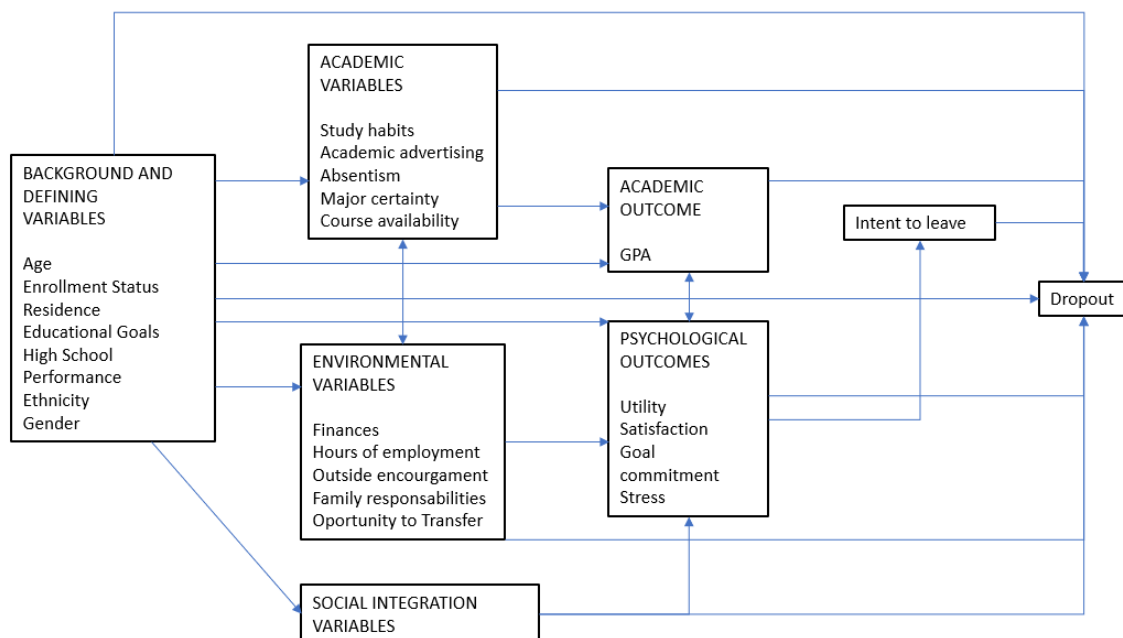
Este modelo es explícito con las categorías y las variables que se usan en él. Bean y Metzger presentan una serie de variables definidas dentro de diferentes categorías. Entre ellas se encuentran: antecedentes y variables de definición, variables académicas, variables contextuales, variables de integración social, variables psicológicas y logros académicos. Según estos autores, se puede conocer la intención de deserción si se conocen las variables pertenecientes a cada una de dichas categorías o ámbitos de la

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

vida personal y académica del estudiante. Las categorías de variables que considera Bean son las siguientes:

- Entorno y variables de definición
- Variables académicas
- Variables de contexto
- Logro académico
- Variables de integración social
- Resultados psicológicos

Figura 3  
*Flujo de deserción de Bean y Metzger*




Nota: adecuado del gráfico original por motivos de legibilidad

## Resumen

Estos trabajos previos proponen una serie de variables que, aún hoy en día, son aplicables para definir la tendencia de deserción de un alumno de formación profesional. Hemos decidido definir algunas categorías de variables, las cuales son:

- Demografía del alumno
- Formación preuniversitaria y admisión del alumno
- Responsabilidades del alumno externas al estudio
- Integración social y desempeño académico del alumno
- Variables cognitivas y emocionales del alumno



	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022


#### Definición de variables

Categoría	Variables
Demografía del alumno	<ul style="list-style-type: none"> <li>- Residencia</li> <li>- Nivel socioeconómico</li> <li>- Edad</li> <li>- Localidad</li> </ul>
Formación preuniversitaria y admisión del alumno	<ul style="list-style-type: none"> <li>- Carrera elegida</li> <li>- Modalidad de admisión</li> </ul>
Entorno familiar	<ul style="list-style-type: none"> <li>- Estado civil</li> <li>- Ingreso familiar</li> <li>- Instrucción del padre</li> <li>- Instrucción de la madre</li> </ul>
Integración social y desempeño académico del alumno	<ul style="list-style-type: none"> <li>- Grupo de vulnerabilidad</li> <li>- Calificación en cursos de carrera</li> <li>- Calificación en cursos generales</li> </ul>
Variables cognitivas y emocionales del alumno	<ul style="list-style-type: none"> <li>- Grado de satisfacción del alumno con la institución educativa</li> </ul>

## 6 Modelo de Análisis Predictivo propuesto

En los dos puntos anteriores se analizó a detalle los diferentes algoritmos de predicción y las variables propuestas para el modelo de análisis predictivo.

Por un lado, el algoritmo seleccionado para implementar el modelo será Random Forest por las bondades que ofrece y que explicamos en puntos anteriores. Por otro lado, se ha fundamentado la elección de ciertas variables que son capaces de describir integralmente el contexto y el desempeño de cada alumno. Ese aspecto de la información a utilizar nos da la seguridad de que el algoritmo sea capaz de identificar patrones de comportamiento que, posteriormente, se convertirá en conocimiento para el modelo de análisis predictivo.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

## 7 Implementación del Modelo de Análisis Predictivo

De acuerdo con lo definido en el Objetivo Específico 1 del proyecto, el desarrollo del modelo de análisis predictivo utiliza 'python' como principal lenguaje de programación, debido a su escalabilidad, fácil implementación y aprendizaje ágil. La librería que estamos empleando para implementar el modelo es 'sci kit learn', una librería de Python que permite importar algoritmos de clasificación ya codificados.

En primer lugar, luego de configurar nuestro entorno de desarrollo, procedimos a importar las librerías y objetos correspondientes:

```
from sklearn.model_selection import train_test_split as tts
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

Luego, declaramos el algoritmo y lo almacenamos en una variable:

```
#Declarar el algoritmo como variable
clf = RandomForestClassifier()
```

Como siguiente paso, debemos entrenar el modelo mediante una de las funciones propias el objeto: 'fit' o 'ajuste'. Esta función toma como parámetro un set de datos y los resultados reales de estos registros. Mediante este aprendizaje supervisado, el algoritmo creará conocimiento y será capaz de predecir.

```
clf.fit(train_feats, train_labels)
```

Finalmente, para realizar predicciones, es necesario llamar a la función 'predict' que toma como parámetro una serie de valores (variables) para obtener los resultados.


```
clf.predict(feats.values)
```

## 8 Obtención de datos

Para la obtención de datos, el equipo investigó en diferentes repositorios de datasets especializados para este tipo de proyecto. La consigna era encontrar un set de datos que incluyera variables para cada una de las categorías de variables definidas en el presente documento. En ese sentido, a continuación, presentamos una lista de repositorios que visitamos.

Kaggle

Carnegie Mellon University Libraries

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

ZBW Journal Data Archive

Google Data Search

En este último repositorio, se encontró un dataset que contiene un total de 4424 registros que representan datos de alumnos. Este es el set de datos que se utilizó para entrenar el modelo

## 9 Entrenamiento del Modelo de Análisis Predictivo

Estrategia de entrenamiento

El set de datos que el equipo usó para el desarrollo del proyecto cuenta con un total de X registros. La estrategia de entrenamiento del modelo consistió en fragmentar los datos en dos grupos. El primer grupo, con una representación del 75% del total de registros del dataset, se usó para el entrenamiento del modelo de análisis predictivo. Esta etapa de la implementación del modelo es de suma importancia, pues consiste en la identificación de patrones de comportamiento implícitos en las variables seleccionadas. Luego de esta primera etapa de entrenamiento, el modelo es capaz de clasificar cualitativamente a los registros del set de datos, según los valores de sus variables de predicción.

El segundo grupo, con una representación del 25% del total de registros del dataset, se usó para la validación del modelo predictivo. El modelo se ejecutó con este último grupo para validar si el modelo ha aprendido del grupo de entrenamiento y es capaz de identificar casos de alumnos en riesgo de deserción. En este punto, tenemos dos resultados: el resultado conocido incluido en el set de datos y el resultado de la ejecución del modelo. La validación será exitosa cuando ambas salidas sean iguales.

## 10 Desarrollo de Aplicación Web de Monitoreo de Deserción

Para mostrar el funcionamiento del modelo de análisis predictivo que desarrollamos, incluimos en el alcance del proyecto el desarrollo de una aplicación web que permita visualizar los resultados de este modelo.


De acuerdo con los mostrado en la Capa de Aplicación de nuestra Arquitectura propuesta, la solución tendría cuatro módulos principales. Sin embargo, decidimos incluir uno dedicado exclusivamente al entrenamiento del modelo. De esta manera, el usuario de la aplicación será capaz de cargar datos históricos al modelo y lograr una mayor precisión en su predicción.

Entonces, los módulos de la aplicación son:

- Módulo de entrenamiento
- Módulo de predicción
- Módulo de reportería
- Módulo de Seguimiento

Adicionalmente, se tienen características no funcionales como:

- Módulo de seguridad (login)
- Módulo de documentación

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

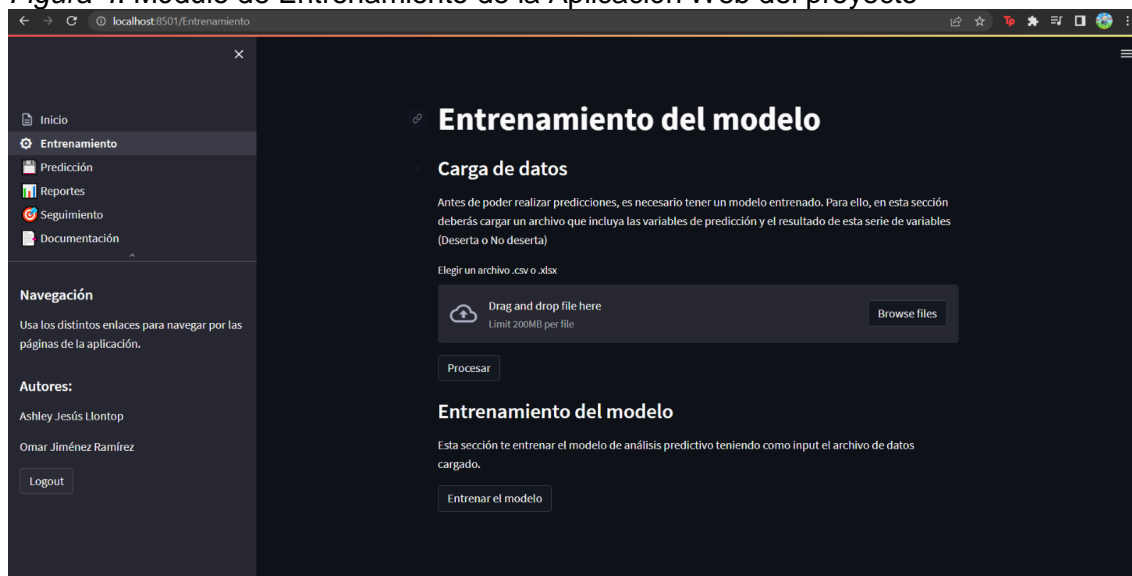
La aplicación web fue desarrollada mediante 'streamlit', un framework de desarrollo de aplicaciones de Python que está diseñado para proyectos de ciencias de datos y aprendizaje automático.

#### Módulo de entrenamiento

Este primer módulo permite al usuario de la aplicación cargar un archivo separado por comas (.csv) en el cual se encuentren registros históricos de alumnos que hayan estudiado en cierta institución educativa y el resultado que tuvo respecto a la interrupción o continuidad de sus estudios.


Luego de la carga de datos, el sistema es capaz de interpretar el archivo y reconoce las variables de predicción. Cada registro del archivo sirve para entrenar al modelo y, cuando se ha creado el conocimiento, este se guarda en un archivo binario.

*Figura 4. Módulo de Entrenamiento de la Aplicación Web del proyecto*

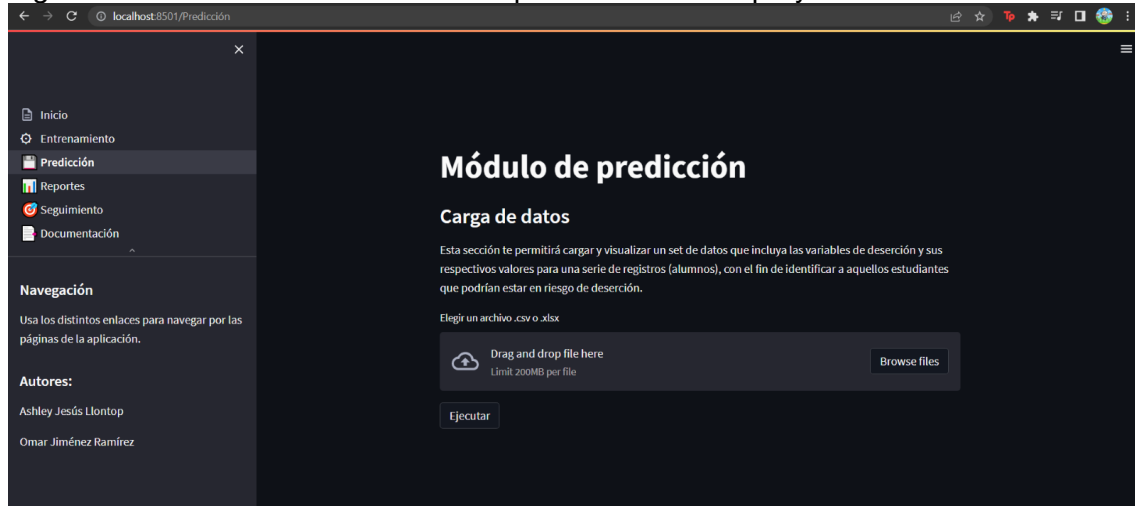


#### Módulo de predicción

Este módulo permite al usuario cargar un archivo de datos que incluya los datos de un alumno y su desempeño en cierto periodo académico. La aplicación recupera el archivo binario creado en el módulo anterior y es capaz de realizar una predicción para identificar alumnos en riesgo de deserción.

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

*Figura 5. Módulo de Predicción de la Aplicación Web del proyecto*



### Módulo de reportería

Esta sección permite al usuario visualizar los resultados de los entrenamientos y ejecuciones del modelo. Se encuentran métricas de desempeño como el número de entrenamientos, el porcentaje de precisión logrado y otros.

*Figura 6. Módulo de Reportería de la Aplicación Web del proyecto*



### Módulo de seguimiento

En este módulo, el usuario de la aplicación consulta por las predicciones realizadas para cierto alumno mediante una consulta usando el código del alumno.


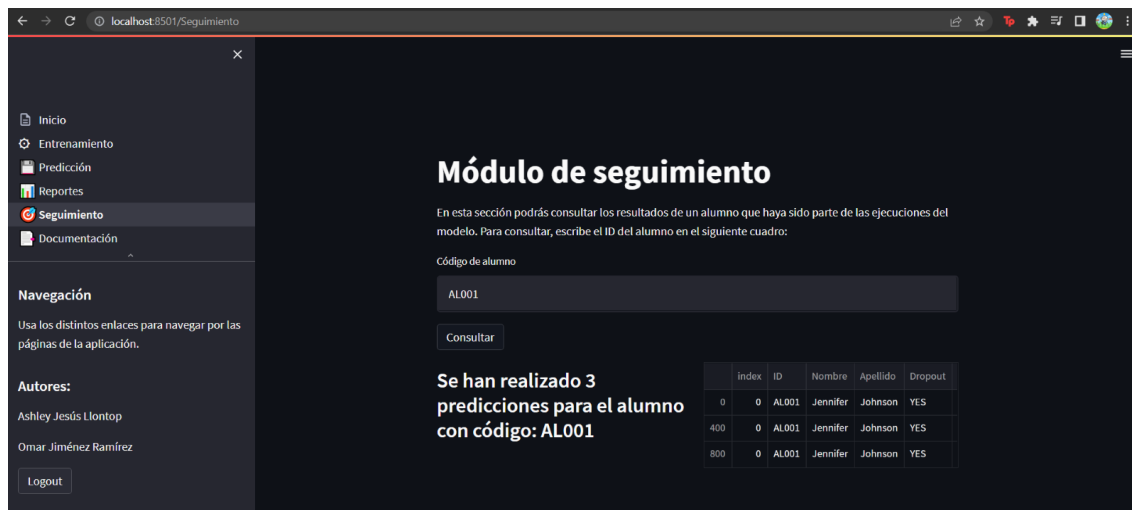

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

Figura 7. Módulo de Seguimiento de la Aplicación Web del proyecto



## 11 Bibliografía

- Berka, P., & Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave? *Studies in Educational Evaluation*, 70(January). <https://doi.org/10.1016/j.stueduc.2021.100999>
- Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346–353. <https://doi.org/10.1016/j.childyouth.2018.11.030>
- Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. *IEEE Access*, 9, 133076–133090. <https://doi.org/10.1109/ACCESS.2021.3115851>
- Flores, V., Heras, S., & Julian, V. (2022). Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. *Electronics (Switzerland)*, 11(3). <https://doi.org/10.3390/electronics11030457>
- Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers and Education*, 131(January), 22–32. <https://doi.org/10.1016/j.compedu.2018.12.006>
- Kabathova, J., & Drlík, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences*, 11(7), 3130. <https://doi.org/10.3390/app11073130>
- Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting Students Dropout: A UTAD University Study. *Future Internet*, 14(3), 1–14. <https://doi.org/10.3390/fi14030076>
- Naseem, M., Chaudhary, K., Sharma, B., & Lal, A. G. (2019). Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science. *2019 IEEE Asia-*

	<b>Proyecto:</b>	PRY20220164 - Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando Machine Learning en la educación superior universitaria en Perú		
	<b>Documento:</b>	Análisis de objetivo específico 2	<b>Autor:</b>	Ashley Jesús Llontop Omar Jiménez Ramírez
	<b>Versión:</b>	0.1	<b>F. Creación:</b>	12/05/2022

- Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–8.  
<https://doi.org/10.1109/CSDE48274.2019.9162389>
- Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth*, 25(1), 104–112. <https://doi.org/10.1080/02673843.2019.1596823>
- Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M., & Díaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *PLoS ONE*, 14(6), 1–20. <https://doi.org/10.1371/journal.pone.0218796>
- Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1), 64–85. <https://doi.org/10.1007/BF02214313>
- Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education*, 17(1). <https://doi.org/10.1186/s41239-020-00186-2>
- <https://www.cs.cmu.edu/~cburch/pgss97/slides/0715-compare.html>