Universidad Autónoma "Juan Misael Saracho"
Departamento de Investigación, Ciencia y Tecnología DICYT

2

ARTÍCULO DE INVESTIGACIÓN

ANÁLISIS DEL CUESTIONARIO MSLQ-SF MEDIANTE ALGORITMOS DE CIENCIA DE DATOS

MSLQ-SF QUESTIONNAIRE ANALYSIS BY DATA SCIENCE ALGORITHMS

Fecha de recepción: 31/12/2022 | **Fecha de aceptación:** 31/03/2023

Autor:

Echalar Flores Michael Willy[1]

[1]Ingeniero Civil, Docente Dpto. de Estructuras de la Carrera de Ingeniería Civil en la Facultad de Ciencia y Tecnología UAJMS

Correspondencia del autor: michaelechalar@gmail.com[1]

Tarija - Bolivia

RESUMEN

En el contexto educativo actual, el uso de datos provenientes de instrumentos psicométricos ha sido común para realizar correlaciones y pronósticos. Sin embargo, con la aparición de la Ciencia de Datos Educativos, se ha vuelto esencial emplear técnicas más avanzadas para el análisis de datos educativos. Este estudio se enfoca en el procesamiento del cuestionario MSLQ-SF mediante algoritmos de Ciencia de Datos utilizando Python.

El análisis se lleva a cabo en dos etapas: un Análisis Exploratorio de Datos (EDA) para comprender mejor la información disponible y la aplicación de modelos de Machine Learning para predecir si los estudiantes aprobarán o reprobarán. Se utilizan varios algoritmos, incluyendo Redes Neuronales Artificiales y Regresión Logística, y se evalúan en función de su precisión. Las Redes Neuronales Artificiales son las más precisas con un 76.32%.

El uso de Python y algoritmos de Ciencia de Datos proporciona flexibilidad y personalización en el análisis de datos educativos. Sin embargo, se señalan desafíos como la necesidad de una curva de aprendizaje para los usuarios y la falta de una ecuación matemática específica. Además, se discuten las implicaciones de la interpretación de los resultados y se sugiere que la mejora de la situación depende tanto del compromiso de los estudiantes como del apoyo institucional para abordar las causas subyacentes de los problemas identificados.

Se destaca la importancia de elegir el enfoque correcto (algorítmico o estadístico) según el contexto y subraya las ventajas y desventajas de utilizar Ciencia de Datos en el campo educativo.

ABSTRACT

In the current educational context, the use of data from psychometric instruments has been common to make correlations and forecasts. However, with the emergence of Educational Data Science, it has become essential to employ more advanced techniques for educational data analysis. This study focuses on the processing of the MSLQ-SF questionnaire through Data Science algorithms using Python.

The analysis is carried out in two stages: an Exploratory Data Analysis (EDA) to better understand the available information and the application of Machine Learning models to predict whether students will pass or fail. Various algorithms, including Artificial Neural Networks and Logistic Regression, are used and evaluated based on their accuracy. Artificial Neural Networks are the most accurate with 76.32%.

The use of Python and Data Science algorithms provides flexibility and customization in educational data analysis. However, challenges are noted such as the need for a learning curve for users and the lack of a specific mathematical equation. Furthermore, the implications of the interpretation of the results are discussed and it is suggested that improving the situation depends on both student commitment and institutional support to address the underlying causes of the identified problems.

The importance of choosing the correct approach (algorithmic or statistical) according to the context is highlighted and the advantages and disadvantages of using Data Science in the educational field are highlighted.

Palabras Clave: Ciencia de Datos Educativos, Cuestionario MSLQ-SF, Machine Learning, Python, Evaluación Educativa, Ansiedad Estudiantil.

Keywords: Educational Data Science, MSLQ-SF Questionnaire, Machine Learning, Python, Educational Assessment, Student Anxiety

1. INTRODUCCIÓN

Dentro del ambiente educativo ha sido siempre una costumbre plantear correlaciones y realizar pronósticos a partir de datos obtenidos de fuentes diversas, principalmente instrumentos psicométricos. El procesamiento de esta información se ha realizado tradicionalmente empleando paquetes de pago como Statgraphics y SPSS. En el momento actual como nunca antes, existe una gran diversidad de fuentes de información y la cantidad de datos disponibles a partir de estas es enorme, a tal punto; que como indica Romero, C., & Ventura, S. (2010) ya se emplea el término Ciencia de Datos Educativos (Educational Data Science) el cual abarca, de forma genérica, a multitud de disciplinas cuyo objetivo es aplicar las distintas técnicas de la Ciencia de Datos sobre la información generada en los sistemas educativos.

Para la aplicación de estas técnicas, los paquetes antes mencionados son demasiado rígidos en su flujo de trabajo y no se pueden adecuar a la actual demanda de procesamiento especifico de gran cantidad de información, la cual exige a la comunidad educativa en general y en particular a la universitaria el empleo de herramientas con el máximo nivel posible de personalización.

La alternativa para zanjar esta dificultas es emplear software libre de código abierto, que además de ser gratuito permite el total acceso al código por parte del usuario. La opción aceptada universalmente para acometer la tarea, es el lenguaje de programación Python. A manera de introducir las técnicas de la Ciencia de Datos en nuestro medio y verificar la idoneidad del lenguaje Python, se plantea el procesamiento de información proveniente de la "Validación Preliminar del Cuestionario MSLQ-SF en Estudiantes de la Carrera de Ingeniería Civil de la Facultad de Ciencias y Tecnología de la Universidad Autónoma Juan Misael Saracho" realizada por Echalar, M. (2019).

El instrumento MSLQ-SF (Motivated Strategies for Learning Questionnaire - Short Form) la sigla en inglés de Cuestionario De Motivación y Estrategias de Aprendizaje – Forma Corta fue adaptado por Sabogal, L. F., Barraza, E., Hernández, A., & Zapata, L. (2011) a partir de la versión original desarrollada por Pintrich, P., & de Groot, E. (1990) el cual debe ser respondido empleando una escala de Likert de 7 puntos. Contiene 40 ítems que se resumen en 9 escalas agrupadas en componentes. De las escalas 2 pertenecen a la sección de Motivación y 7 a la de Estrategias de Aprendizaje. Esta estructura se muestra en la Tabla 1; se cuenta además con las notas finales de la materia en la cual fue aplicado el instrumento, conociéndose de esta forma si el alumno aprobó o reprobó la asignatura.

Tabla 1: Estructura del cuestionario MSLQ-SF

Secciones	Componentes	Escalas	Ítems
Motivación	Valor	Valoración de la tarea	20, 26, 39
	Afectivos	Test de ansiedad	3, 12, 21, 29
Estrategias de aprendizaje	Estrategias cognitivas y metacognitivas	Estrategias de elaboración	4, 5, 22, 24, 25
		Estrategias de organización	13, 14, 23, 40
		Pensamiento critico	1, 6, 15
		Autorregulación a la metacognición	16, 30, 31, 32, 34, 35,36
	Estrategias de administración de recursos	Tiempo y hábitos de estudio	2, 8, 17, 18, 33, 38
		Autorregulación del esfuerzo	7, 9, 11, 19, 27, 28
	Valor	Metas de orientación intrínseca	10, 37

Fuente: Elaboración propia

El instrumento goza de gran valides y reconocimiento a nivel mundial, el año 2022 fue objeto de adaptaciones a contextos nacionales como indica Cardeñoso Ramírez, O., Larruzea-Urkixo, N., & Bully Garay, P. (2022) y Villarreal-Fernández, J. E., & Arroyave-Giraldo, D. I. (2022). De la versión corta del instrumento (MLSQ-SF) se realizan exploraciones de su estructura interna como indica Masso Viatela, J. (2021).

El procesamiento se realiza en dos etapas, la primera consiste en un EDA (Exploratory Data Analysis) la sigla en ingles de Análisis Exploratorio de Datos que se espera sea mas flexible y permita el mejor entendimiento de la información de la cual se dispone. La segunda es el planteamiento de un modelo Machine Learning para poder predecir si los estudiantes aprobaran o reprobaran, clasificándolos en función a su motivación y estrategias de aprendizaje.

Los algoritmos de Ciencia de Datos se presentan en la Tabla 2, los mismos fueron seleccionados ya que son los que generalmente son empleados en problemas de clasificación.

Tabla 2: Algoritmos de Ciencia de Datos

Artificial Neural Network	Red Neuronal Artificial
K-Nearest Neighbors	K-Vecindades Cercanas
Ordinary Least Squares	Regresión de Mínimos Cuadrados
Logistic Regression	Regresión Logística

Fuente: elaboración Propia

La ciencia de datos ya se ha usado para la detección del desempeño de los estudiantes a partir de datos digitales generados por ellos, un caso particular bastante interesante está basado en su participación en foros de discusión como indica Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Las redes neuronales también han sido aplicadas para la detección de alumnos en riesgo y la medición de la eficiencia de centros educativos con gran éxito como indica González, D. S. (1999).

2. MATERIALES Y MÉTODOS

Esta sección detalla: recursos informáticos, la composición de la Base de Datos, los indicadores obtenidos a partir del análisis EDA, los parámetros empleados para el planteamiento de los modelos y los indicadores o métricas seleccionados para la validación de los mismos. Es interesante mencionar que la estructura básica para la realización de este trabajo se tomó de Kraus, M., Bischof, R., Kaufmann, W., & Thoma, K. (2022) los cuales realizan la aplicación de inteligencia artificial al análisis estructural.

Se realizo la investigación en un sistema de las siguientes características:

Kernel: Linux

OS: Arch

IDE: VS Code

Lenguaje: Python v.3.11.1

Librerías: numpy, pandas, matplotlib, sklearn, keras

La Base de Datos empleada por Echalar, M. (2019) consiste en 189 filas y 53 columnas, cuyos campos corresponden a: índice, sexo, colegio, 40 reactivos, 9 escalas resumen, nota final. Para este estudio se han tomado: 9 escalas resumen codificadas como [R01, R02, S01, S02, S03, S04, S05, S06, S07], nota final [Q] y se añadió el campo estado [estado] a partir de la nota final, 0 para reprobado y 1 para aprobado.

El Análisis Exploratorio de Datos se realizó todas las columnas a excepción de estado; calculando su media y desviación, generando el grafico de histograma, distribución empírica y diagrama de caja.

Para realizar el entrenamiento de los modelos se ha empleado aleatoriamente el 80% de los datos, para realizar la validación se ha empleado en resto correspondiente al 20%.

Los parámetros empleados en los modelos son los siguientes:

Red Neuronal Artificial con 9 entradas; compuesta de 3 capas escondidas de 12, 8, y 1 neuronas; 1 salida con activación sigmoide; con 150 ciclos (epochs) de entrenamiento, repetidos en varias ocasiones hasta lograr la precisión máxima.

K-Vecindades Cercanas con los hyperparametros: n_neighbors, weights y metric optimizados mediante la librería GridSearchCV que permite determinar el mejor valor a ser empleado. Regresión de Mínimos Cuadrados planteada como función lineal de los 9 predictores, el valor de regresión se lleva al entero 0 ó1 como clasificación. Regresión Logística inicializada con los valores por defecto. Dado que los modelos resuelven un problema de clasificación, para la validación de los mismos a se ha escogido como métrica de validación la función accuracy_score de la librería sklearn.metrics la cual devuelve la precisión del modelo a partir del porcentaje de acierto de los valores predichos respecto a los valores reales.

3. RESULTADOS

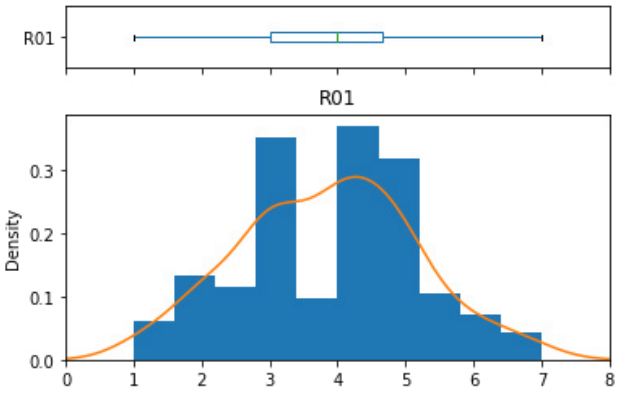
El análisis EDA se resume en la Tabla 3, se muestra la media, desviación, mínimo, cuartiles y máximo de las 9 escalas resumen y de las notas; no fue necesario identificar valores atípicos debido al empleo de la escala de Likert y no existen filas incompletas dado que las respuestas al tabularse fueron depuradas, se trata entonces de una base de datos muy limpia. A continuación de la tabla se muestran las figuras que contienen: el histograma, la distribución empírica y el diagrama de caja, de los parámetros antes mencionados; desde la Figura 1 hasta la Figura 10.

Tabla 3: Análisis de Datos Exploratorio (EDA)

Indicador	R01	R02	S01	S02	S03	S04	S05	S06	S07	N.F.
media	3.85	4.60	5.08	4.87	4.68	5.10	4.57	5.68	5.18	56.06
desv	1.29	1.41	0.97	1.15	1.03	0.80	0.97	0.80	1.13	24.97
mín	1.00	1.00	2.00	1.00	1.00	2.29	1.00	3.00	1.00	3.00
25%	3.00	3.50	4.60	4.00	4.00	4.57	4.00	5.17	4.00	51.00
50%	4.00	4.75	5.20	5.00	4.67	5.14	4.67	5.83	5.50	60.00
75%	4.67	5.50	5.80	5.50	5.33	5.71	5.17	6.17	6.00	74.00
máx	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	99.00

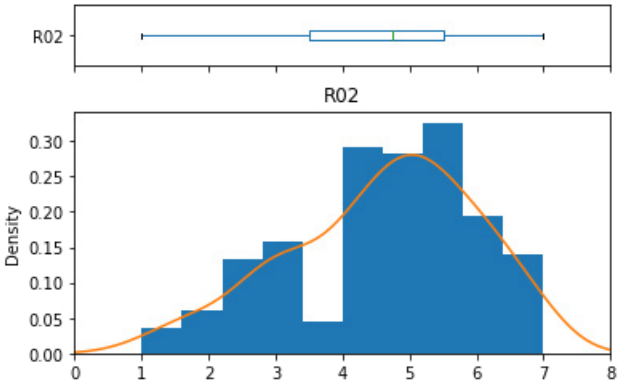
Fuente: Elaboración propia

Figura 1: Valoración de la tarea



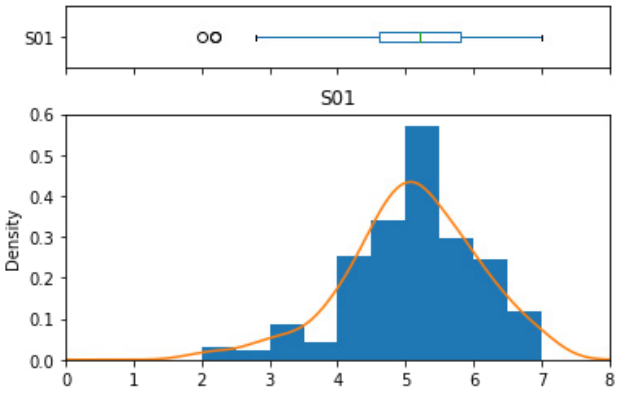
Fuente: Elaboración propia

Figura 2: Test de ansiedad



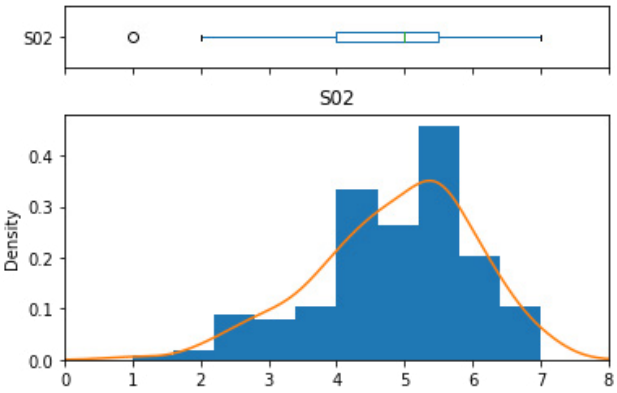
Fuente: Elaboración propia

Figura 3: Estrategias de elaboración



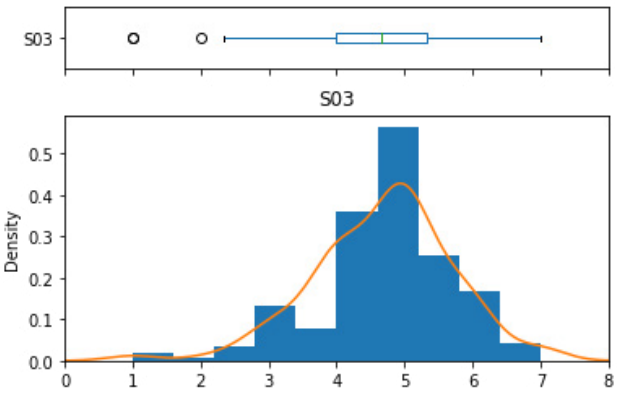
Fuente: Elaboración propia

Figura 4: Estrategias de organización



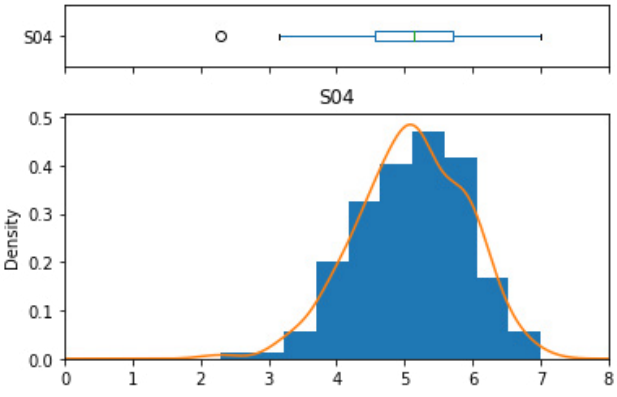
Fuente: Elaboración propia

Figura 5: Pensamiento critico



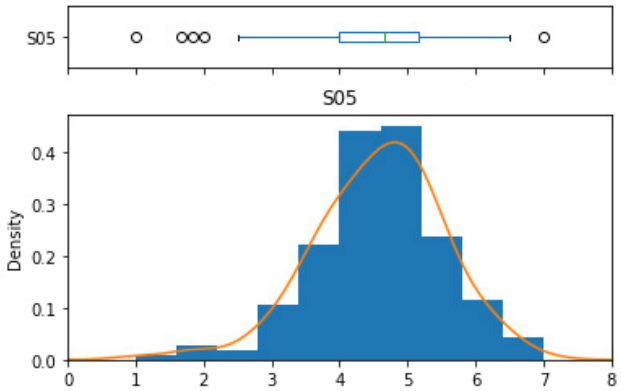
Fuente: Elaboración propia

Figura 6: Autorregulación a la metacognición



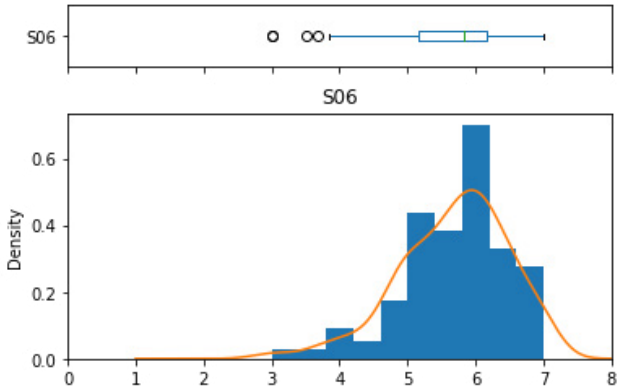
Fuente: Elaboración propia

Figura 7: Tiempo y hábitos de estudio



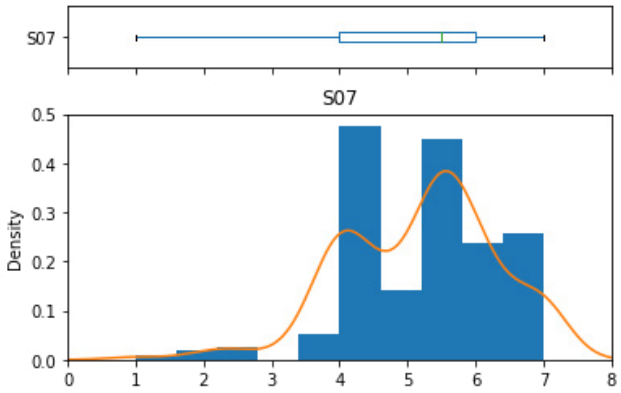
Fuente: Elaboración propia

Figura 8: Autorregulación del esfuerzo



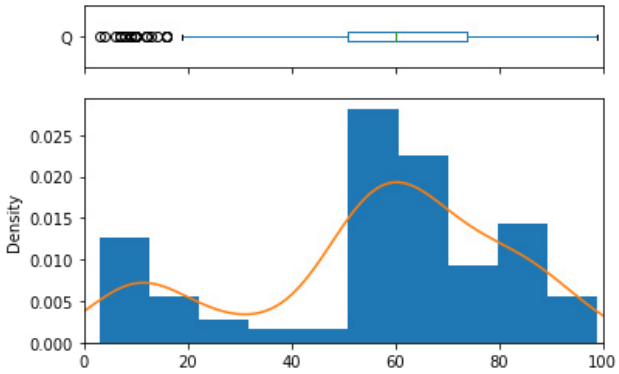
Fuente: Elaboración propia

Figura 9: Metas de orientación intrínseca



Fuente: Elaboración propia

Figura 10: Notas finales de la asignatura



Fuente: Elaboración propia

A partir de la métrica de precisión, seleccionada para la valoración de los algoritmos a continuación se muestra la Tabla 4, la cual contiene el valor obtenido por cada uno ordenados de mayor a menor.

Tabla 4: Precisión de los algoritmos

Algoritmo	Precisión [%]
Red Neuronal Artificial	76.32
K-Vecindades Cercanas	75.00
Regresión de Mínimos Cuadrados	71.05
Regresión Logística	60.52

Fuente: Elaboración propia

4. DISCUSIÓN

A partir de la presentación gráfica de las escalas resumen y de acuerdo al manual de aplicación del instrumento presentado por Pintrich, P., Smith, D., Garcia, T., & McKeachie, W. (1991) el cual indica que en general si el valor está por encima de 3 se está procediendo de buena manera y se considera una buena evaluación a excepción del test de ansiedad, si se obtiene una valoración menor a 3 en más de 6 escalas se debe conversar con el docente.

En general se puede indicar que la media de todas las escalas cumple la condición anterior, sin embargo, valoración de la tarea esta solo ligeramente encima con 3.85, esta escala indica cuán importante considera el estudiante lo que esta realizando, el valor bajo indicaría que no se considera importante o de utilidad lo avanzado en la materia. El test de ansiedad presenta un valor de 4.60 que se puede considerar alto e indica que existe ansiedad y miedo en los estudiantes al momento de enfrentase a evaluaciones. Las escalas siguientes se distribuyen en valores desde 4.57 hasta 5.68 lo cual puede considerarse satisfactorio, pero no excelente; es evidente también que en varias ocasiones se forman 2 grupos de respuestas. La Figura 10 de la distribución de notas finales de la asignatura muestra los alumnos se distribuyen en 3 grupos en las materias donde se aplicó el cuestionario, la nota y denominación se muestran en la Tabla 5.

Tabla 5

Estado	Nota [%]
Destacados	90.00
Aprobación	51.00
Reprobados	10.00

Fuente: Elaboración propia

Vemos que el grupo más numeroso es el segundo, estando los dos restantes con un número similar de personas. Esta distribución de los estudiantes es muy lejana a la ideal en la cual el 90% de los alumnos aprueban el 90% de las materias con el 90% de la nota. La interpretación por defecto es que la mayoría de los alumnos solo cursan la materia por la aprobación, la segunda opción es que el sistema se evaluación no es adecuado, existiendo la posibilidad de que el momento para la realización de la prueba no sea el adecuado o incluso la prueba misma y el sistema de ponderación no sean los adecuados. La tercera interpretación es que la metodología del proceso enseñanza aprendizaje planteado por el docente es deficiente.

5. CONCLUSIÓN

Se puede concluir que el empleo del lenguaje de programación Python y el uso de algoritmos de Ciencia de Datos es en general mucho mas flexible y personalizable al momento de realizar la tarea de análisis e interpretación; producto de esto ha surgido la posibilidad de plantear más de una opción al momento de interpretar las notas finales de los estudiantes, quedando por determinar cual seria la interpretación mas verídica. Para cualquiera de las opciones no se están realizando esfuerzos para determinar la causa correcta y de la misma forma tampoco se están tomando acciones para encontrar el remedio. La mejora de esta situación le corresponde en gran medida a los estudiantes, pero requiere que exista un compromiso institucional de la universidad para monitorear estas situaciones, tratar de identificar las causas y palear las que le sean atribuibles.

Respecto a los modelos conseguidos, estos pueden ser empleados durante la etapa de la evaluación inicial para determinar qué tan grande es el grupo de estudiantes con problemas, para poder realizar la programación de cómo se llevará a cabo su nivelación y ayudar a la planeación del desarrollo de la materia a lo largo del semestre.

Respecto al lenguaje Python y los algoritmos empleados se pueden indicar las siguientes ventajas:

- Se personalizan al problema y los datos
- La flexibilidad es superior a todos los programas de pago
- El algoritmo y los datos son fácilmente puestos disponibles al público mediante GitHub, para el caso de este trabajo se pueden encontrar en el siguiente repositorio: https://github.com/michaelechalar/mslq-sf.

Se deben también enumerar las desventajas encontradas al realizar la investigación:

- Se debe transitar una curva de aprendizaje.
- No se obtiene una ecuación o expresión matemática.
- Cada interacción de las redes neuronales genera un modelo diferente.
- La distribución de los componentes requiere la disposición expresa del autor a publicarlos.

De lo enumerado anteriormente, al comparar la ciencia de datos con la estadística surgen dos posiciones definidas por Boulesteix, A.-L., & Schmid, M. (2014), la primera seria algorítmica y predictiva y la segunda estocástica y explicativa.

Como mencionan los autores la primera considera los datos generados por un mecanismo desconocido que únicamente se puede aproximar mediante algoritmos, la segunda considera los datos generados por un modelo estocástico el cual puede llegar a aproximarse de manera matemática. Dando esto la opción de usar ciencia de datos en casos en los que la predicción sea mas necesaria que la comprensión y el empleo de la estadística en caso contrario.

A esto se debe añadir que, si bien se ha mencionado la facilidad de publicación en internet de los datos y modelos, para que otros investigadores puedan hacer uso de ellos; deben primero obtener una copia de estos y disponer del programa adecuado para abrirlos. Estas condiciones dependen de la voluntad del autor para liberar los objetos digitales y manejarlos en una plataforma accesible. Dada la naturaleza aleatoria de la selección de los datos de entrenamiento, incluso contando con los insumos mencionados, es posible no lograr exactamente el mismo modelo logrado por el autor.

6. BIBLIOGRAFÍA

- Boulesteix, A.-L., & Schmid, M. (2014). Machine learning versus statistical modeling. Biometrical Journal, 56(4), 588–593.
- Cardeñoso Ramírez, O., Larruzea-Urkixo, N., & Bully Garay, P. (2022). Adaptación al contexto universitario español y propiedades psicométricas del MSLQ: Contribución a la medida y análisis de las diferencias de género del aprendizaje autorregulado. Anales de Psicologia, 38(2).
- Echalar, M. (2019). Validación preliminar del cuestionario MSLQ-SF en estudiantes de la carrera de Ingeniería Civil de la Facultad de Ciencias y Tecnología de la Universidad Autónoma Juan Misael Saracho. SEC Ciencia, 2(3).
- González, D. S. (1999). Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales (Documentos de Trabajo de la Facultad de Ciencias Económicas y Empresariales). Universidad Complutense de Madrid, Facultad de Ciencias Económicas y Empresariales.
- Kraus, M., Bischof, R., Kaufmann, W., & Thoma, K. (2022). Artificial Intelligence - Finite Element Method - Hybrids for efficient nonlinear analysis of concrete structures. In Acta Polytechnica CTU Proceedings (Vol. 36). https://doi.org/10.14311/APP.2022.36.0099
- Masso Viatela, J. (2021). Cuestionario de motivación y estrategias de aprendizaje forma corta–MSLQ SF en estudiantes universitarios: análisis de la estructura interna. Los Libertadores Fundacion Univerisitaria.

- Pintrich, P., & de Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82(1), 33–40.
- Pintrich, P., Smith, D., Garcia, T., & McKeachie, W. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). The University of Michigan.
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618.
- Sabogal, L. F., Barraza, E., Hernández, A., & Zapata, L. (2011). Validación del cuestionario de motivación y estrategias de aprendizaje forma corta MSLQ-SF, en estudiantes universitarios de una Institución Pública-Santa Marta. Psicogente, 14(25), 36–50.
- Villarreal-Fernández, J. E., & Arroyave-Giraldo, D. I. (2022). Adaptación y validez de la escala de motivación del Motivated Scale Learning Questionnaire (MSLQ) en universitarios colombianos. Electronic Journal of Research in Education Psychology, 20(56), 119–150.