53

Fecha de presentación: Septiembre, 2021
Fecha de aceptación: Noviembre, 2021
Fecha de publicación: Diciembre, 2021

ANÁLISIS

DE LAS CAUSAS DEL CONSUMO DE DROGAS MEDIANTE APRENDIZAJE AUTOMÁTICO

ANALYSIS OF THE CAUSES OF DRUG USE BY MEANS OF MACHINE LEARNING

Maikel Yelandi Leyva Vázquez[1]
E-mail: ub.c.investigacion@uniandes.edu.ec
ORCID: https://orcid.org/0000-0001-7911-5879
Remigio Edmundo Hernández Cevallos[1]
E-mail: ub.remigiohernandez@uniandes.edu.ec
ORCID: https://orcid.org/0000-0003-3782-8031
Iyo Alexis Cruz Piza[1]
E-mail: ub.iyocruz@uniandes.edu.ec
ORCID: https://orcid.org/0000-0002-9411-9672
Karina Pérez Teruel[2]
E-mail: karinaperez@uapa.edu.do
ORCID: https://orcid.org/0000-0002-1526-9913
[1] Universidad Regional Autónoma de Los Andes. Ecuador.
[2] Universidad Abierta Para Adultos. República Dominicana.

Cita sugerida (APA, séptima edición)

Leyva Vázquez, M. Y., Hernández Cevallos, R. E., Cruz Piza, I. A., & Pérez Teruel, K. (2021). Análisis de las causas del consumo de drogas mediante aprendizaje automático. *Revista Universidad y Sociedad*, 13(S3), 392-399.

RESUMEN

El consumo de drogas es un problema real, que afecta a gran parte de la sociedad adulta mundial, pero indiscutiblemente, los jóvenes son un sector muy vulnerable dentro de este escenario, situación de la cual no está exento Ecuador. Por la necesidad de estudiar esto, se estableció como objetivo fundamental de esta investigación consistió en realizar un análisis de las causas del consumo de drogas y ayudar a la prevención mediante técnicas de aprendizaje automático supervisado o no supervisado. Adicionalmente el proyecto presentó como objetivo proporcionar técnicas y herramientas sobre aprendizaje automático y su aplicación práctica en problemas relacionados. Se analizó como caso de estudio, una muestra de 3 876 jóvenes y adolescentes ecuatorianos y los resultados fueron procesados mediante el software Orange, y se obtuvo un riesgo del 85%. Las tareas definidas en la metodología CRISP DM se ejecutaron sin novedades, y se pudo indicar que los datos son suficientes para el modelado, y que las predicciones realizadas son confiables.

Palabras claves: Análisis de causas, consumo de drogas, aprendizaje automático.

ABSTRACT

Drug consumption is a real problem that affects a large part of the world's adult society, but unquestionably, young people are a very vulnerable sector within this scenario, a situation from which Ecuador is not exempt. Due to the need to study this, the main objective of this research was to analyze the causes of drug use and help prevention through supervised or unsupervised automatic learning techniques. Additionally, the project's objective was to provide techniques and tools on automatic learning and its practical application in related problems. A sample of 3,876 Ecuadorian youths and adolescents was analyzed as a case study and the results were processed using Orange software, and a risk of 85% was obtained. The tasks defined in the CRISP DM methodology were executed without novelties, and it was possible to indicate that the data are sufficient for modeling, and that the predictions made are reliable.

Keywords: Causal analysis, drug use, machine learning.

INTRODUCCIÓN

Uno de los problemas sociales y sanitarios más importantes en todo el mundo es el consumo de estupefacientes. Al menos el 5% de la población adulta del mundo ha consumido drogas al menos una vez en 2015, mientras que el daño causado está representado por 28 millones de años de vida sana perdidos como consecuencia del consumo de drogas. En Ecuador, la edad media de consumo de drogas se sitúa entre los 14 y los 15 años de edad; y que las drogas más fácilmente adquiridas son la marihuana, la heroína o H (Alarcon, 1989; Herrera, 2006; Saiz et al. 2020)

El consumo de drogas es un problema real, que afecta a la sociedad en general, pero indiscutiblemente, los jóvenes son un sector muy vulnerable dentro de este escenario (Puértolas-Gracia et al. 2021; Pera & Hernando, 2021). Por lo que resulta conveniente llevar a cabo un estudio de este fragmento de la población en el que se pueden encontrar potenciales consumidores y sobre los cuales se pueden trazar estrategias para disminuir o erradicar esta nociva conducta.

En la actualidad, existen diversas técnicas, métodos y herramientas para el procesamiento de la información. Una de las más utilizadas últimamente, es el aprendizaje automático, por su gran aplicación sobre todo cuando se trata de procesar grandes volúmenes de información, así como para predecir el comportamiento de determinadas variables o patrones.

El objetivo fundamental de esta investigación consiste en realizar un análisis de las causas del consumo de drogas y ayudar a la prevención mediante técnicas de aprendizaje automático supervisado o no supervisado. Adicionalmente el proyecto presenta como objetivo proporcionar técnicas y herramientas sobre aprendizaje automático y su aplicación práctica en problemas relacionados.

Se puede enumerar, además, como objetivos específicos, el establecimiento de un caso de estudio sobre el consumo de drogas en jóvenes ecuatorianos, la confección de un diagrama de Ishikawa, la aplicación de técnicas y algoritmos de aprendizaje automático mediante el software Orange y la valoración de los resultados obtenidos.

Se considera como criterio de éxito la oportunidad de realizar predicciones sobre la probabilidad de consumo de drogas en adolescentes y jóvenes, logrando obtener resultados confiables. Adicionalmente como criterio de éxito se encuentra aumentar el conocimiento estructural de los datos disponibles mediante el aprendizaje no supervisado.

MÉTODOS

La investigación se desarrolló basada, fundamentalmente, en los siguientes métodos:

Métodos investigativos teóricos: Analítico-sintético que sirvió la elaboración del fundamento teórico. Inductivo-deductivo utilizado para inducir una respuesta particular y deducirla para un alcance general.

Métodos investigativos empíricos: Observación directa, encuestas: los datos fueron recolectados mediante encuesta a 3 876 jóvenes y adolescentes ecuatorianos con 45 variables.

Diagrama Ishikawa: Es una de las herramientas surgidas a lo largo del siglo XX en ámbitos de la industria y posteriormente en el de los servicios, para facilitar el análisis de problemas y sus soluciones en esferas como lo son; calidad de los procesos, los productos y servicios. Fue concebido por el licenciado en química japonés Dr. Kaoru Ishikawa en el año 1943 (Trujillo Bucheli, Santiago Trujillo, & Santiago Trujillo, 2020). El Diagrama de Ishikawa es también conocido con el nombre de espina de pescado (debido a su forma), o diagrama causa-efecto (CE). Constituye es una herramienta que ayuda a estructurar la información ayudando a dar claridad, mediante un esquema gráfico, de las causas que producen un problema (Trujillo Bucheli et al., 2020). Por lo que se utiliza en el presente trabajo para representar las principales causas del problema en cuestión.

Aprendizaje automático: También conocido como ***machine learning***, es un apartado de la inteligencia artificial en el cual una máquina puede analizar una gran cantidad de datos, razonar y tomar acción, pese a que la misma no fue programada para realizar dicha actividad. Su mayor particularidad es que pueden "aprender", por lo tanto sus resultados irán progresando y volviéndose más precisos con el paso del tiempo (Bustamante, 2011). De todas maneras, es importante recordar que el aprendizaje automático no es una solución para todo tipo de problemas y situaciones. Existen determinados casos en los que se pueden desarrollar soluciones sólidas sin usar técnicas de aprendizaje automático (Conrad & Schilder, 2007; Lagnado & Gerstenberg, 2017; Martínez 2017; Martínez Moncaleano & Palencia Fajardo, 2021; Muñoz & Martinez, 2020; Naoui, Lejdel, & Ayad, 2020).

Software Orange: Desarrollado por el Laboratorio de Bioinformática de la Facultad de Informática y Ciencias de la Información de la Universidad de Liubliana, Eslovenia. Es un software libre de aprendizaje automático y ***data-mining***. Sus características principales residen en sus funcionalidades como la programación ***visual front-end*** para

explorar datos y la visualización de resultados. Aunque también puede usarse como una biblioteca Python (Cáliz, 2018). En este caso se utiliza para la visualización de los datos.

A continuación, se describen las acciones de mitigación sobre los posibles escenarios de riesgo.

Tabla 1. Contingencia del Proyecto.

Riesgos identificados	Contingencia
Incumplimiento de tiempo	Elaborar un cronograma y realizar un seguimiento de las actividades.
Alcance del Proyecto mal delimitado	Reuniones permanentes con el tutor un adecuado dimensionamiento del proyecto.
Desconocimiento de herramientas de ML	Reuniones con expertos en técnicas de Machine Learning.
Falta de recursos informáticos	Utilizar recursos informáticos en la nube.
No contar con datos	Realizar copias de seguridad de la base de datos en la nube.
Modelo predictivo no confiable	Auto Machine Learning ayudará a generar modelos confiables.
Mala Calidad de los datos	Realizar un análisis previo de los datos y tomar las medidas adecuadas como normalización e imputación de los datos.

Fuente: Datos de la investigación

Elaborar el Plan de Proyecto

A continuación, se detallan las fases a realizar en el proyecto, relacionándolas con sus respectivas tareas, tiempo de duración, entradas y salidas obtenidas.

Tabla 2. Plan del Proyecto

Fases	Tareas	Recursos	Tiempo	Entradas	Salidas
Comprensión del negocio	Objetivos del Negocio. Criterios de éxito del negocio. Valoración de la situación actual. Inventario de recursos. Riesgos y Contingencias. Elaboración del plan de proyecto.	Investigador	3 semanas	Recopilar información	Analizar y los datos
Comprensión de los datos	Recopilación de los datos. Descripción de los datos Exploración de los datos Verificación de la calidad de los datos.	Investigador	1 semana	Reunir la información necesaria relacionada con el consumo de droga	Obtener la base de datos.
Preparación de los datos	Seleccionar los datos a analizar Limpieza de datos	Investigador	1 semana	Procesamiento de los datos obtenidos mediante encuesta	Datos elegidos y listos para aplicar técnicas de aprendizaje automático
Modelado	Elegir las técnicas de modelado más acertadas Desarrollar un modelo de comprobación. Construcción del modelo	Investigador	4 semanas	Técnicas de Machine Learning tanto supervisadas como no supervisadas	Técnica que aplicar y generación de modelo óptimo.
Evaluación	Evaluación de resultados	Investigador	2 semanas	Modelo de Machine learning	Evaluar los resultados obtenidos aplicando ML

RESULTADOS Y DISCUSIÓN

Se procede a aplicar los métodos descritos para estudiar la problemática planteada: causas del consumo de drogas y cómo ayudar a su prevención.

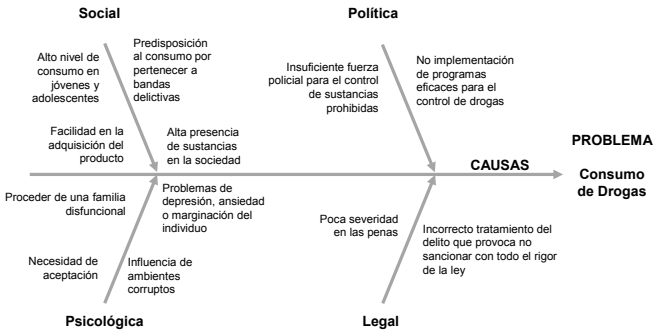

Figura 1. Diagrama de Ishikawa

Comprensión de los datos: Esta fase permite tener un primer acercamiento con los datos proporcionados, entenderlos en su totalidad, conocer la calidad que tienen, con el propósito de garantizar que sean datos relevantes para la investigación que se desarrolla.

Descripción de los datos: La base contiene los siguientes datos y el código utilizado.

Tabla 3. Descripción de la Base de Datos

Código	
1.1	Nombre Encuestado:
	Riesgos
2.1	¿Dónde ha recibido información sobre drogas?
2.2	¿Sabía que el consumo de drogas afecta su salud?
2.3	¿Ha consumido algún tipo de drogas?
2.4	¿Razones para Consumir?
	Información del Hogar
3.1	Número de personas en el hogar
3.2	¿Su padre vive en el hogar?
3.3	¿Su madre vive en el hogar?
3.4	Otros Familiares
	Relación Familiar
4.1	¿Existe agresión física en su hogar?
4.2	¿Sufre maltratos físicos por parte de su familia?
4.3	¿Cuándo tiene algún problema, lo cuenta a su familia?
4.4	¿Se siente escuchado por parte de su familia?
	Entorno Social
5.1	¿Tiene amigos cercanos que consumen drogas?
5.2	¿Pertenece a algún grupo juvenil?
5.3	¿En su barrio existen zonas de recreación?
	Entorno Educativo
6.1	Tipo de Escuela
6.2	¿Mantiene comunicación con sus docentes?
6.3	¿Su unidad educativa tiene DECE?
6.4	¿Tiene confianza al DECE de su Centro Educativo?
6.5	¿Confiaría sus secretos a algún docente?
6.6	¿Se siente respaldado por su centro educativo?
6.7	¿Ha sido víctima de maltrato escolar?
	Detalles de vivienda
7.1	Tipo de vivienda
7.2	El material predominante de las paredes exteriores de la vivienda es de: _____
7.3	El agua que recibe la vivienda es
7.4	El servicio de energía eléctrica de la vivienda proviene de
7.5	Dispone este hogar de servicio de teléfono convencional
7.6	Algún miembro de este hogar dispone de servicio de teléfono celular
7.7	La vivienda que ocupa este hogar es
7.8	TV
7.9	Cocina a gas con horno
7.10	Microondas
7.11	Refrigeradora
7.12	Lavadora
7.13	Ducha eléctrica
7.14	Ventilador
7.15	Aire acondicionado
7.16	Computadora en tu Hogar
7.17	Consolas de Video Juegos
7.18	Vehículo en el hogar

Exploración de los datos: En esta etapa se realiza la exploración de los datos de la base con el objetivo de conocer el tipo de dato de cada variable, aplicar estadística descriptiva, y determinar que si los datos se encuentran balanceado o no es especial la variable predictora. Se puede apreciar en los resultados del análisis que no encuentran datos faltantes, pero si hay datos que no están balanceados.

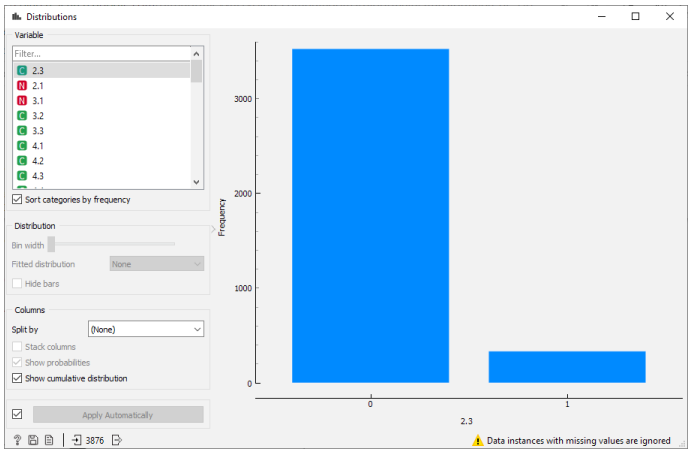

Figura 2. Histograma de la variable.

En este caso se encuentran 3 876 jóvenes y adolescentes que afirman no haber consumido drogas y 335 que sí.

Verificar la calidad de los datos: Una vez realizada la exploración de los datos, se verifica que la base no presenta datos incompletos o faltantes, sin embargo, se han evidenciado que la variable que se desea predecir no está balanceada.

Seleccionar los datos: A continuación, se muestran los datos seleccionados tomando como target la variable 2.3 e ignorando la variable 2.4

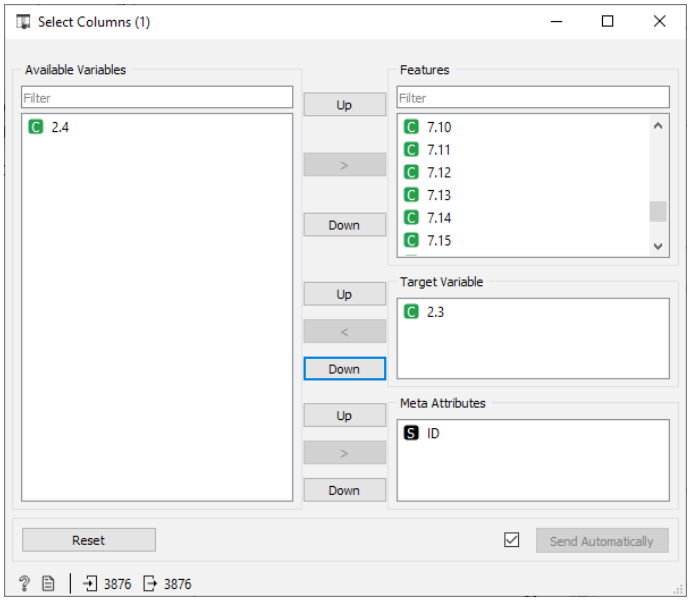

Figura 3. Histograma de la variable.

Limpieza de datos: La fase de limpieza de datos implica corregir datos faltantes, datos atípicos que se registren fuera de rango de las variables elegidas para realizar la predicción. Se utilizó como estrategia para trabajar con conjuntos de datos desbalanceados el demuestre. Dentro de esta estrategia se empleó el sobremuestreo que consiste en aumentar la presencia de la clase minoritaria.

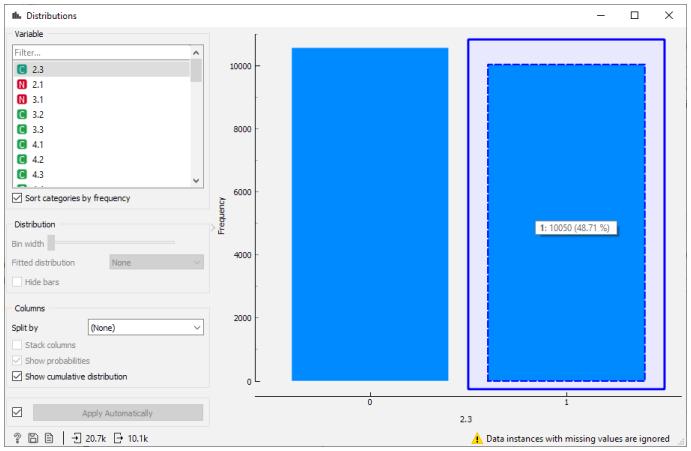

Figura 4. Histograma de la variable.

Posteriormente, se transformaron las variables numéricas a tipo categórico (2.1, 7.1, 7.2, 7.3, 7.4, 7.7)

Formateo de los datos: Durante esta fase se realiza la preparación de los datos, para evitar que los Algoritmos de Machine Learning presenten inconvenientes al encontrarse con datos fuera de rango o variables de tipo carácter. Para ello se normalizaron los datos en el intervalo [0,1] (Guzmán, Sánchez, & Uprimny, 2010).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Modelado: La fase de modelado consiste en elegir e implementar los algoritmos que tributen de forma efectiva al cumplimiento de objetivos que apunta la presente investigación. Se compararon distintos modelos de acuerdo al a la precisión (De Vergara et al. 2006):

$$P = \frac{Tp}{Tp + Fp} \quad (2)$$

Tp⇒ Verdaderos positivos
Fp⇒ Falsos positivos

En este caso la mejor precisión la alcanza el método de Naive Bayes

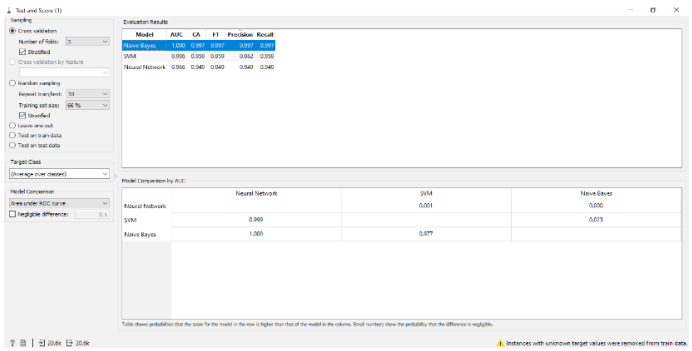

Figura 5. Evaluación de los modelos

Se puede apreciar la matriz de confusión obtenida

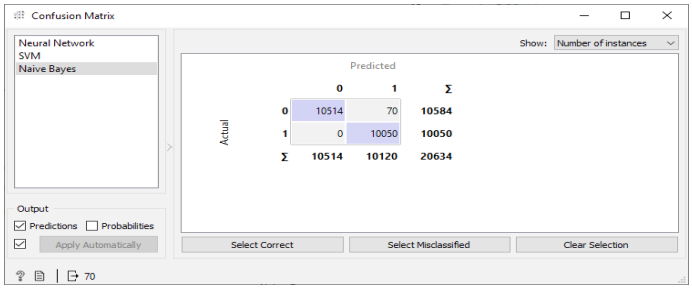

Figura 6. Matriz de Confusión.

Selección Técnica del Modelado: A partir de estos resultados se empleó el algoritmo Naive Bayes o Clasificador bayesiano ingenuo.

Explicación del modelo: Para comprender el modelo se empleó el índice de Shap (Man & Chan, 2021). A continuación, se muestran las 5 variables de mayor impacto en la predicción

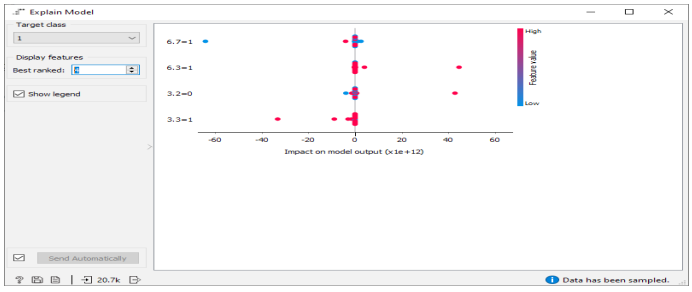

Figura 7. Explicación del modelo.

En este caso 6.1 (Tipo de escuela), 6.3 (Si la unidad educativa tiene DECE), 3.2 (si el padre vive en el hogar), 3.3 (si la madre vive en el hogar). Finalmente, para calcular la probabilidad de consumir drogas se emplea los métodos de explicación de la predicción.

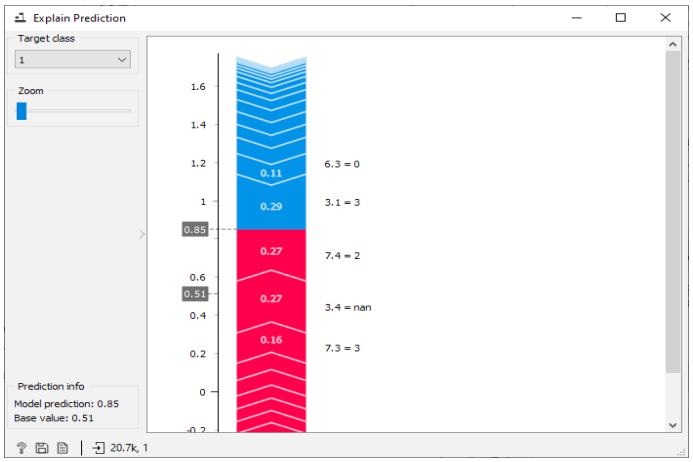

Figura 8. Explicación de la predicción.

Es este caso el riesgo del 85%

Agrupamiento: Se emplearon los métodos de agrupamiento para comprender y segmentar los adolescentes. Utilizando el coeficiente de silueta se calcula utilizando la distancia media entre Clúster (a) y la distancia media entre Clúster (b) para cada muestra (Benavides Sarango, 2019).

$$CS = (b - a) / max(a,b) \quad (3)$$

El coeficiente de silueta de una muestra, en este caso el valor máximo se alcanza en el Clúster 7.

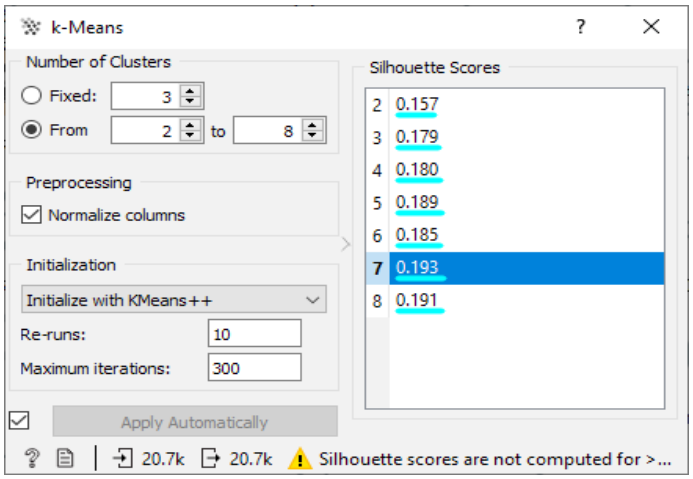

Figura 9. Coeficiente de silueta.

A partir de estos datos se analiza la distribución de las variables en los distintos segmentos en que se han dividido los datos.

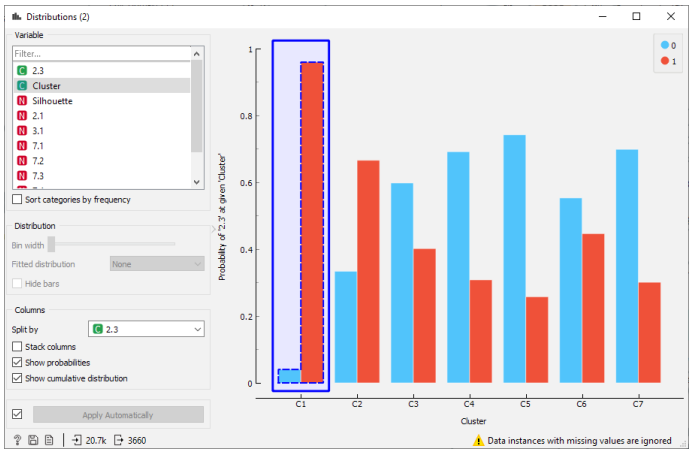

Figura 10. Análisis por clúster.

Se puede apreciar que en el Clúster 1 existe mayor probabilidad de encontrar individuos que consumen drogas.

Las tareas definidas en la metodología CRISP DM se ejecutaron sin novedades, podemos indicar que los datos son suficientes para el modelado, y que las predicciones son confiables.

CONCLUSIONES

La aplicación de algoritmos de Aprendizaje Automático, como el clasificador Naive Bayes a una muestra de datos, constituye una gran herramienta para predecir la clase de la instancia de prueba con la mayor probabilidad posterior.

Las causas del consumo de drogas tienen orígenes muy diversos. Estas pueden ser de índole social, político, psicológico o legal.

Después de procesada la muestra de 3 876 jóvenes ecuatorianos, se puede decir que las cinco variables de mayor impacto son: 6.1 (Tipo de escuela), 6.3 (Si la unidad educativa tiene DECE), 3.2 (si el padre vive en el hogar), y 3.3 (si la madre vive en el hogar).

De manera general, se puede afirmar que los jóvenes ecuatorianos están siendo afectados por el consumo de drogas y de acuerdo con los datos analizados en la encuesta mediante el software Orange, el clúster o grupo que es más propenso al consumo de estas sustancias es el número 1, con un riesgo del 85%.

REFERENCIAS BIBLIOGRÁFICAS

Alarcón, J. (1989). Daños corporales: Concepto y bases determinantes para la fijación del "quantum" indemnizatorio. ***Revista de Derecho de la Circulación***, *5*, 231; 1-23.

Benavides Sarango, M. S. (2019). Importancia de la educación formal en la prevención de la delincuencia juvenil. Universidad Regional Autónoma de Los Andes.

Bustamante, L. T. (2011). El femicidio Género, Diversidad Violencia Intrafamiliar. Jurídica del Ecuador.

Cáliz, D. H. (2018). Teoría y Práctica del Femicidio en Ecuador. Editorial Ecuador.

Conrad, J. G., & Schilder, F. (2007). Opinion mining in legal blogs. In ***Proceedings of the 11th international conference on Artificial intelligence and law*** (pp. 231-236).

De Vergara, L. C., & Santiago, A. M. (2006). Análisis del proceso de toma de decisiones en las grandes empresas de Barranquilla utilizando el análisis por conglomerados. Pensamiento & Gestión, (20), 55-109.

Guzmán, D., Sánchez, N., & Uprimny, R. (2010). Las Victimas y la Justicia Transicional. Fundación para el Debido Proceso Legal.

Herrera, A. S. (2006). Criterios para determinar el indemnizatorio en el daño moral un estudio de la jurisprudencia española. Revista Chilena de Derecho Privado, (7), 51-87.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. ***Oxford handbook of causal reasoning***, 565-602.

Man, X., & Chan, E. P. (2020). The best way to select features. Comparing MDA, LIME and SHAP.

Man, X., & Chan, E. P. (2021). The Best Way to Select Features? Comparing MDA, LIME, and SHAP. ***The Journal of Financial Data Science***, *3*(1), 127-139.

Martínez Moncaleano, C. J., & Palencia Fajardo, O. (2021). Modelo de minería de datos para el análisis de la productividad y crecimiento personal en las mujeres emprendedoras: el caso de la Asociación las Rosas. Suma de Negocios, 12(26), 23-30.

Martínez, J. J. (2017). Minería de opiniones mediante análisis de sentimientos y extracción de conceptos en Twitter. Universidad Complutense de Madrid-España.

Muñoz, S. X. S., & Martinez, M. A. Q. (2020). A Framework for Selecting Machine Learning Models Using TOPSIS. In ***Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2020 Virtual Conferences on Software and Systems Engineering, and Artificial Intelligence and Social Computing,*** 1213, 119). Springer Nature.

Naoui, M. A., Lejdel, B., & Ayad, M. (2020). Using K-means algorithm for regression curve in big data system for business environment. ***Revista Cubana de Ciencias Informáticas***, ***14***(2), 34-48.

Pera, C. P., & Hernando, A. G. (2021). Intoxicación aguda por drogas de abuso. Manejo en atención primaria. FMC-Formación Médica Continuada en Atención Primaria, 28(2), 94-100.

Puértolas-Gracia, B., Juárez, O., Ariza, C., & Villalbí, J. R. (2021). La prevención universal del consumo de drogas en el entorno escolar: el valor de la monitorización continua. Gaceta Sanitaria, (2032), 1-3

Saiz, M. S., Chacón, R. F., Abejar, M. G., Parra, M. S., ValentÍn, M. D., & Yubero, S. (2020). Perfil de consumo de drogas en adolescentes. Factores protectores. Medicina de Familia. SEMERGEN, 46(1), 33-40.

Trujillo Bucheli, B., Santiago Trujillo, A., & Santiago Trujillo, P. F. (2020). Adolescente Infractor. Derecho Ecuador.