Modelos de Aprendizaje Automático basados en CRISP-DM para el Análisis de los niveles de Depresión en los estudiantes de la Escuela Politécnica Nacional

CRISP-DM- Based Machine Learning Models For Analyzing the Depression Level in Students of the National Polythecnic School

Jiménez, Sergio; Merino, Andrés

Sergio Jiménez
sergio.jimenez@epn.edu.ec
Escuela Poliécnica Nacional, Ecuador
Andrés Merino
aemerinot@puce.edu.ec
Pontificia Universidad Católica del Ecuador, Ecuador

Latin-American Journal of Computing
Escuela Politécnica Nacional, Ecuador
ISSN: 1390-9266
ISSN-e: 1390-9134
Periodicidad: Semestral
vol. 10, núm. 1, 2023
lajc@epn.edu.ec

Recepción: 26 Octubre 2022
Aprobación: 26 Octubre 2022

URL: http://portal.amelica.org/ameli/journal/602/6023721001/

DOI: https://doi.org/10.5281/zenodo.7503909

Resumen: El presente proyecto analiza las variables de depresión que puede tener un estudiante universitario de la Escuela Politécnica Nacional (EPN) mediante modelos de aprendizaje automático (ML). Participaron un total de 302 estudiantes de distintas carreras quienes completaron de manera voluntaria y anónima una encuesta en línea constituida por el Inventario de Depresión de Beck II (BDI-II). Las 19 preguntas de la encuesta están relacionadas al estilo de vida promedio de un estudiante de la EPN y fueron revisadas y avaladas sobre su relación con trastornos depresivos por una profesional en el campo de la psicología. Se utilizó la metodología CRISP-DM para las fases del proyecto que consistieron en el análisis de la situación actual, planteamiento de objetivos, recolección, análisis y preparación de datos, construcción de modelos de ML para predecir la severidad de depresión con base en las métricas de BDI-II y evaluación de modelos. Se obtuvo un modelo con 0.59 de exactitud y se verificó que las variables de género, edad y relaciones interpersonales son las más significativas al determinar la severidad de depresión.

Palabras clave: *Trastornos de Depresión*, *Aprendizaje Automático*, *Selección de Características*, *Ciencia de datos*, *Inventario de Depresión de Beck II*, *CRISP- DM*, *Python*.

Abstract: This project analyzes the depression rates among students from Escuela Politécnica Nacional (EPN). A total of 302 students from different EPN careers, voluntarily and anonymously completed an online survey of the Beck Depression Inventory-II (BDI-II). In addition, they were asked to answer 19 questions related to the lifestyle of an EPN student; These questions were reviewed and endorsed about their possible relationship with depressive disorders by a professional in the field of psychology. The CRISP-DM methodology was used for the project phases, which involved the analysis of the current situation, objectives setting, data collection, data preparation, and construction of ML models that allows predicting the degree of depression based on the BDI-II metrics and evaluation of the models. The model obtained has 0.59 accuracy score and shows that variables of gender, age and relationships are significant to determine severity depression.

Keywords: *Depression Disorders*, *Machine Learning*, *Feature Selection*, *Data Science*, *Beck Depression Inventory II*, *CRISP-DM*, *Python*.

I. INTRODUCCIÓN

La depresión es un trastorno psicológico, por el cual un individuo experimenta un estado de pérdida de interés prolongado en las actividades habituales. Este trastorno puede manifestarse de manera psíquica (tristeza, cansancio, falta de concentración y memoria, etc.) y física (desórdenes alimenticios, trastornos de sueño, fatiga corporal, dolores de cabeza, etc.) [1]. De acuerdo con la OMS, el trastorno de depresión se estima que afecta al 5% de los adultos, con mayor prevalencia en mujeres y personas jóvenes a nivel global [2]; índice que iría en incremento debido al impacto del COVID-19 en la salud mental [2] [3]. Este trastorno psicológico es uno de los más frecuentes motivos de consulta en centros de bienestar estudiantil en universidades [4]. Los factores que influyen en la depresión de estudiantes universitarios están relacionados con los ámbitos académicos, sociales, familiares, económicos, culturales y personales [4]. Estos factores llegan a afectar a los estudiantes, sumiéndolos en un estado prolongado de desánimo y tristeza; haciendo que su rendimiento académico, social y personal vayan disminuyendo. En casos graves de depresión, puede provocar la deserción de asignaturas académicas, rupturas sentimentales, pérdida de trabajo y, en casos extremos, llevar a la autolesión y el suicidio [4] [2].

Con base en esta situación, se plantea la necesidad de realizar un análisis de los posibles factores que pueden afectar psicológicamente a un estudiante promedio de la Escuela Politécnica Nacional (EPN), centrándose en aquellos aspectos que puedan provocar un estado de depresión prolongado de mínimo dos semanas. La recolección de datos se la realizó durante las 2 últimas semanas del semestre 2022A, en donde, se sabe que los casos de estrés, ansiedad y depresión pueden llegar a manifestarse con mayor frecuencia.

Específicamente, se plantea la elaboración de un análisis estadístico descriptivo y la implementación de modelos de aprendizaje automático para determinar cuáles son las variables que inciden en el nivel de depresión de los estudiantes.

Los objetivos de principales del presente trabajo consisten en el análisis, desarrollo y evaluación de modelos de aprendizaje automático, capaces de predecir el nivel de depresión que puede padecer un estudiante de la EPN. Los resultados de la evaluación del rendimiento de estos modelos permitirán el análisis estadístico de las variables que pueden incidir más en la detonación de trastornos depresivos, además de desarrollar un modelo de aprendizaje automático (ML por sus siglas en inglés) que sea capaz de predecir un posible cuadro de depresión en base a las variables analizadas.

El análisis estadístico de los datos mostrará las proporciones de estudiantes hombres, mujeres y de otros géneros, de diferentes rangos de edad, que presenten distintos niveles de depresión. Por otro lado, se presentará un contraste de los resultados de los niveles de depresión y las variables relacionadas con aspectos académicos, sociales, personales y hábitos de consumo.

El modelo de aprendizaje automático no busca reemplazar el diagnóstico profesional de parte de un experto en la salud mental, sino ser una herramienta que proporcione información útil para la toma de decisiones en tratamiento y/o campañas para disminuir o prevenir los casos de depresión es estudiantes de la EPN.

II. TRABAJOS RELACIONADOS

De acuerdo con estudios realizados en universidades ecuatorianas, en el año 2015 por la universidad de Cuenca y en 2018 por la Pontificia Universidad Católica del Ecuador, existe una prevalencia de casos de depresión en estudiantes mujeres y jóvenes de entre 19 a 24 años [5] [6]. Sin embargo, otro estudio desarrollado en 2018 muestra ausencia de diferencias en cuanto al género respecto a trastornos de depresión, ansiedad o estrés [7]. Por otro lado, un estudio desarrollado en 2020, y aplicado en la Pontificia Universidad Católica del Ecuador, mostró que los hombres tenían más alteraciones en la salud mental que las mujeres [8].

Un evento que influye en este análisis es la pandemia provocada por el COVID-19 donde, de acuerdo con los resultados de una evaluación de estrés, ansiedad y depresión en Ecuador en el año 2021, la población ecuatoriana sufre de estrés en un 41%, depresión en 39%, y ansiedad en 46% [9].

El enfoque de utilizar herramientas de Aprendizaje Automático para detectar trastornos de depresión se ha llevado a cabo a través de distintas técnicas que van desde el procesamiento del lenguaje natural a través de textos en redes sociales o procesamiento de metadatos de publicaciones [10,11], utilizar distintas técnicas de Aprendizaje Automático y Aprendizaje profundo con base en métricas de uso de dispositivos móviles [12,13] o creando distintos modelos de acuerdo con los resultados de cuestionarios de evaluación de depresión estandarizados, además de cuestionarios con preguntas relacionadas al medio en que los participantes se encuentran involucrados [14].

III. MARCO TEÓRICO

Con el objetivo de recolectar datos, se realizaron encuestas a voluntarios cuyas respuestas mostrarán una estimación de si padecen un estado de depresión y puntuarán aproximadamente en que escala se encuentra. Para esto, se analizaron las pruebas estandarizadas para medir el nivel de depresión y ansiedad; de entre ellas, se escogió el Inventario de depresión de Beck (BDI-II por sus siglas en inglés).

1. Inventario de depresión de Beck II:

Desarrollado por Aaron T. Beck y siendo una versión mejorada del BDI (1961) y BDI-AI (1978). El BDI-II es un inventario para evaluar la severidad de depresión desarrollado con base en los criterios del American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition (DSM-IV). Este inventario está compuesto de 21 elementos de evaluación, cada uno con 4 puntos de escala de 0-3. El puntaje que se puede obtener puede ser de entre 0 a 63 puntos [15] [16].

Las escalas de severidad del estado de depresión que puede padecer una persona son:

- Depresión mínima, 0-13
- Depresión leve, 14-19
- Depresión moderada, 20-28
- Depresión severa, 29-63 [17]

El inventario es una herramienta que ayuda al diagnóstico de un trastorno de depresión, el cual se lo deberá realizar de manera objetiva por un profesional de la salud mental [16]. Asimismo, cabe señalar el formulario no se encuentra orientado a descartar un caso de depresión, por lo tanto, queda a consideración del profesional clasificar si una persona que se encuentra en el grupo de depresión mínima es en realidad un caso de un paciente sin depresión.

Para el proceso de análisis de la situación, análisis de datos, procesamiento de estos, construcción de modelos de ML y despliegue de resultados se tomó con base a la metodología CRISP-DM.

2. Metodología CRISP-DM

Esta metodología es usada en proyectos de ciencia de datos que cuenta con seis fases de desarrollo e iterativas como se ve en la Fig. 1. Estas seis fases son [18]:

- Entendimiento del negocio
- Entendimiento de datos
- Preparación de datos
- Construcción del modelo
- Evaluación del modelo
- Despliegue del modelo

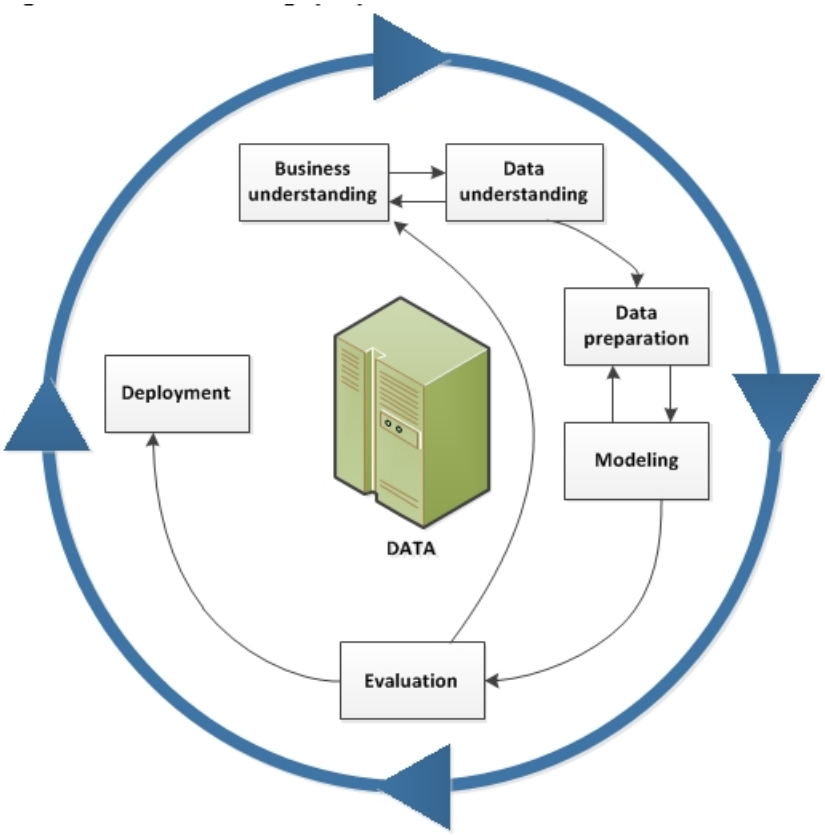


FIG. 1.
Ciclo de vida de CRISP-DM de la minería de datos [18]

Para el desarrollo del presente proyecto se utilizaron herramientas de programación y análisis de datos utilizadas en proyectos de ciencia de datos.

3. Herramientas

a. Python 3: Lenguaje de programación de alto nivel, utilizado en el desarrollo de multiples proyectos de Inteligencias Artificial, Aprendizaje Automático y Ciencia de Datos, el cual cuenta con multiples librerías

para el procesamiento de datos, visualización de datos, construcción y evaluación de modelos [19]. Las librerías de este lenguaje utilizadas para este proyecto fueron: Numpy, Pandas, Matplotlib, Seaborn, Sklearn y LazyPredict

b. Google Colab: Es un entorno de desarrollo integrado (IDE) de jupyter notebook que se encuentra en la nube y no requiere configuración. Es capaz de integrarse al servicio de Google Drive para el manejo de archivos o fuentes de datos (.csv, xlsx, .text, .sql, entre otros) [20].

c. Microsoft Forms: Herramienta de Microsoft para el desarrollo de formularios simples, personalizables a nivel visual y de contenido que tienen la capacidad de ser compartido por medio digitales [21].

d. Tableau: Programa de visualización de datos con multiples herramientas tanto de manipulación, vizualización y presentación de datos.

e. Modelos de Machine Learning

i. DecisionTree Classifier: Modelo de predicción de aprendizaje deductivo mediante construcciones lógicas como los sistemas basados en reglas donde a través de nodos se forma la estructura de decisiones [22].

ii. BernoulliNB: Este modelo se basa en el teorema de bayes el cual toma valores binarios 0 o 1 con una suposión "ingenua" de la independencia condicional entre dos elementos [23].

iii. LGBM Classifier: Este algoritmo permite el uso de variables categóricas y utiliza una técnica de gradiente de muestreo unilateral donde las instancias de los datos con mayores gradientes en cada iteración [24].

iv. Bagging Classifier: Modelo basado en un algoritmo conjunto que combina predicciones de distintos clasificadores donde se ajustan a diferentes modelos cada uno con diversos datos de entrenamiento [25].

v. Nearest Centroid: es un modelo basado en distancias, donde se calcula el centroide de cada clase como el promedio de todas sus muestras en el conjunto de entrenamiento para predecir el centroide más cercano [26].

f. Métricas de rendimiento de modelos de clasificación.

i. Confusion Matrix (Matriz de confusión): utilizada para evaluar la precisión de un modelo de clasificación, valorando entre verdaderos positivos, falsos negativos, falsos positivos y verdaderos negativos [25].

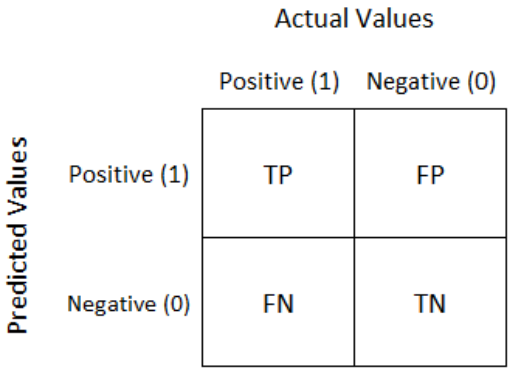


FIG. 2.
Matriz de confusión [27]

ii. Accuracy (Exactitud): Muestra el número de elementos clasificados correctamente en comparación al número total de los elementos clasificados como se ve en (1) [25]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

iii. Precision (Precisión): Es el número de verdaderos positivos sobre el número de valores predichos como verdaderos [25]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

iv. Recall (Exahustividad): Es la cantidad de verdaderos positivos con base en los valores totales de los valores positivos [25]:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

v. F1-Score (Puntuación F1): es la media armónica de las métricas Recall y Precision [25]:

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

IV. METODOLOGÍA

A continuación, se detallan las fases del proyecto de acuerdo con la metodología CRISP-DM mencionada anteriormente.

1. Entendimiento del negocio

Esta fase se enfocó en el entendimiento de objetivos y requerimientos del proyecto [28]. La cual estuvo compuesta por:

- Investigación: Análisis de los fundamentos del trastorno de depresión en población estudiantil universitaria investigando en las fuentes de proyectos, libros, informes y publicaciones relacionadas
- Evaluación de situación actual: Determinación de la disponibilidad de recursos, análisis de los factores implicados, determinación de la población dentro de la EPN a la cual se aplicó el estudio.
- Definición de objetivos del proyecto: definición de los objetivos de éxito del proyecto que encaminarán las etapas subsiguientes:
    - Un análisis estadístico de los datos recolectados desde una perspectiva de género, rango de edad y nivel de depresión
    - Desarrollo de un modelo de ML para determinar el nivel de depresión con base en datos académicos, sociales, económicos, familiares, culturales, personales y de consumo de un estudiante promedio de la EPN.
- Plan de proyecto: Definición de estrategias de recolección y tratamiento de datos, tecnologías, herramientas y pasos a realizar en cada fase del proyecto siguiendo la metodología.

2. Análisis de los datos

Esta fase se centró en la comprensión de datos, identificación, extracción y recopilación de estos [28]. En esta fase se desarrolló:

- Formulación de datos a recopilar: con la investigación realizada en la fase anterior, se estableció un total de 19 preguntas relacionadas a ámbitos académicos, culturales, sociales, económicos, personales

y de consumo. Estas preguntas fueron evaluadas y aprobadas por una profesional en el campo de la psicología

- Recopilación de datos: Se creó un cuestionario en línea usando la herramienta de Microsoft Forms el cual estuvo compuesto por dos partes: la primera consiste en los 21 elementos de BDI-II y la segunda basada en las 19 preguntas relacionadas con el estilo de vida promedio de un estudiante de la EPN. Estas preguntas estuvieron relacionadas con (entre paréntesis se señala el código asignado a cada pregunta):

    1. El género que se identificaban (q22)
    2. La edad que tienen (q23)
    3. Con quienes viven (q24)
    4. Si eran responsables del cuidado de alguna persona (q25)

    En cuanto al ámbito académico las preguntas fueron:

    5. Que tan de acuerdo está con que le guste su carrera universitaria (q26)
    6. Si toma alguna asignatura básica de cálculo y si representa la mayoría de las que asignaturas que está tomando (q27). Las asignaturas de formación básica son comunes para todos los estudiantes de la universidad e históricamente, suelen ser consideradas las más demandantes.
    7. Cuántas asignaturas considera exigentes (q28)
    8. Si está repitiendo alguna materia y el nivel que le preocupa este hecho (q29)

    La carga horaria de trabajo sea académico o laboral, de ser el caso, también fue un factor a considerar. Las preguntas fueron:

    9. Cuántas horas de clases dedicas a la semana (q30)
    10. Cuántas horas de estudio dedicas fuera de las horas de clase a la semana (q31)
    11. Cuántas horas dedicas a actividades de ocio, entretenimiento o relajación (q32)
    12. Si trabaja, cuántas horas dedicaba a esta actividad semanalmente (q33)

    Respecto al ámbito financiero solo se consideró únicamente:

    13. Cómo financia mayoritariamente sus estudios (q34)

    Otros factores fueron el consumo de alimentos altos en grasas y azúcares, alcohol y sustancias psicoactivas:

    14. Cuántas veces consume alimentos altos en grasas y azúcares (q35)
    15. Con qué frecuencia consume alcohol (q36)
    16. Si consume alguna sustancia psicoactiva y la frecuencia con que lo hace (q37)

    Por último, se establecieron preguntas respecto a la relación con las personas que convive en distintos ámbitos:

    17. Cómo es su relación con sus compañeros de universidad (q38)
    18. Cómo es su relación sentimental en caso de tener una (q39)
    19. Cómo es su relación con las personas que vive (q40)
- Descripción de datos: Examinación inicial del formato, tipo de valores, distribución y volumen de datos. En esta fase, se exploró la cantidad de los datos, así como la calidad de estos. Se recolectaron un total 302 respuestas de los participantes de los cuales fueron 202 hombres, 97 mujeres y 3 personas que se identifican con otro género.
- Exploración de datos: Profundización de datos en ambas secciones del cuestionario a través de tablas, visualización de datos y correlación
- Evaluación de calidad de los datos: Identificación de datos relevantes, nulos, repetidos, de tipo incorrectos y valores atípicos

3. Preparación de datos

En esta fase, se recopilaron los datos para el análisis y construcción del modelo. Esta fase puede llevar entre un 50% a un 80% del proyecto y se interrelaciona con la fase de modelamiento en el sentido de establecer el modelo con el mejor resultado [18]. Las fases que lo conformaron fueron:

- Selección de datos: construcción del conjunto de datos de entrenamiento del modelo donde las primeras 21 columnas de fueron seleccionadas para el cálculo del nivel de depresión según las métricas de BDI-II, el resto de las 19 columnas representaron las variables para el análisis de índice de depresión.
- Limpieza de datos: aplicando de técnicas de tratamiento de datos, se eliminaron datos erróneos y atípicos tanto en el conjunto de datos de las 19 variables como el de BDI-II
- Transformación y formateo de datos: Se transformaron los datos categóricos de tipo texto a tipo entero en una escala positiva desde el cero.
- Construcción de datos: para determinar las puntuaciones de resultados de BDI-II, se sumó los puntos de cada elemento para después agruparlo por las escalas respectivas de severidad de depresión.
- Integración datos: Creación de nuevos conjuntos de distintas fuentes de datos ya estandarizados, transformados y limpios en un solo conjunto de datos
- Análisis Exploratorio y Estadístico de datos: Esta fase comprendió el ver las relaciones entre las variables que podrían incidir en la depresión, tales como:
    - Severidad de depresión por edad y género.
    - Severidad de depresión por personas con quienes viven, parejas sentimentales y compañeros de universidad.
    - Severidad de depresión por nivel de agrado de la carrera universitaria, cantidad de materias de formación básicas (las cuales tienen altos índices de reprobación en la EPN), número de materias exigentes y si se encuentran repitiendo alguna materia.
    - Severidad de depresión por carga horaria de clases, estudio, descanso y trabajo si es que tuviese uno.
    - Severidad de depresión por consumo de comida chatarra, alcohol y sustancias psicoactivas.
    - Severidad de depresión por financiamiento de estudios
    - Severidad de depresión por personas quienes son responsables de una persona

4. Modelamiento

En esta fase, se determinó cuáles serían los posibles modelos que arrojen los mejores resultados, se iteró con el conjunto de datos de entrenamiento hasta determinar el que obtiene las mejores métricas. Para conseguirlo, fue necesario ajustar los parámetros de cada modelo y se tuvo que restructurar y formatear los datos de entrenamiento para que se adapten mejor a los modelos. Cuenta con las siguientes etapas:

- Técnicas de selección de modelos: Con base en los datos disponibles, la cantidad y la calidad de estos, se optó por analizar el resultado de los mejores algoritmos devueltos por la librería LazyPredict de Python con el objetivo de escoger y desarrollar el mejor modelo ajustando sus parámetros a los requisitos.
- Generación del diseño de prueba: Los modelos fueron evaluados en función de las métricas de algoritmos de aprendizaje supervisado, las cuales son: matriz de confusión (Confusion matrix), Exactitud (Accurancy), Precisión (Precision) y Puntuación F1 (F1-Score).

- Construcción de modelos, evaluación y aceptación: Los datos fueron separados en datos de entrenamiento y pruebas; posteriormente, fueron cargados en la función LazyClassifier de la librería LazyPredict para entrenamiento de los modelos. Adicionalmente, se realizaron diferentes técnicas de procesamiento de datos, como normalización y estandarización; de manera iterativa. Se utilizaron los datos con los diferentes procesamientos para entrenar a los algoritmos que destacará LazyClassifier. Finalmente, los 3 mejores puntuados fueron escogidos para su posterior configuración de parámetros con el fin de obtener los mejores modelos, con base a las métricas antes descritas.
- Se desarrolló una permutación de las variables de entre 2 a 15 para entrenar los modelos de LazyClassifier para de ese modo comprobar las variables tuviesen un mejor resultado en entrenar modelos, verificando así el análisis estadístico anteriormente desarrollado de las variables.

5. Evaluación del modelo

En esta fase del proyecto, se evaluó el desempeño de los modelos construidos mediante matrices de confusión. También, se analizó si los modelos y el análisis estadístico fue satisfactorio respecto a los objetivos planteados.

- Evaluación de resultados: Los modelos fueron analizados en cuanto a la exactitud que reflejaban en la matriz de confusión en donde se observó el contraste de los valores predichos y los valores reales.

6. Despligue del modelo

En este proyecto, se contempló el desplegar los mejores modelos si las métricas de rendimiento eran superiores o iguales a las puntuaciones esperadas, además de señalar las variables más significativas de los modelos seleccionados. En cuanto al análisis estadístico se desarrollaron los gráficos respectivos y representativos, como se estableció en el Análisis Exploratorio y Estadístico de datos de la fase de preparación de datos.

V. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

En el análisis exploratorio y estadístico de los datos, se realizó la partición de las preguntas de la evaluación por los 21 elementos de BDI-II y las 19 variables que podrían incidir en la depresión.

De los 302 resultados de BDI-II se calculó el puntaje de cada participante y se definió su severidad de depresión siguiendo los rangos establecidos, como se muestra en la Tabla I y la Fig. 3:

TABLA I.
Severidad de depresión por número de participantes

Severidad de depresión	Rango de puntuación	No. de participantes	%
Depresión severa	[29-63]	100	33.11%
Depresión moderada	[20-28]	81	26.82%
Depresión ligera	[14-19]	53	17.55%
Depresión mínima	[0-13]	68	22.52%

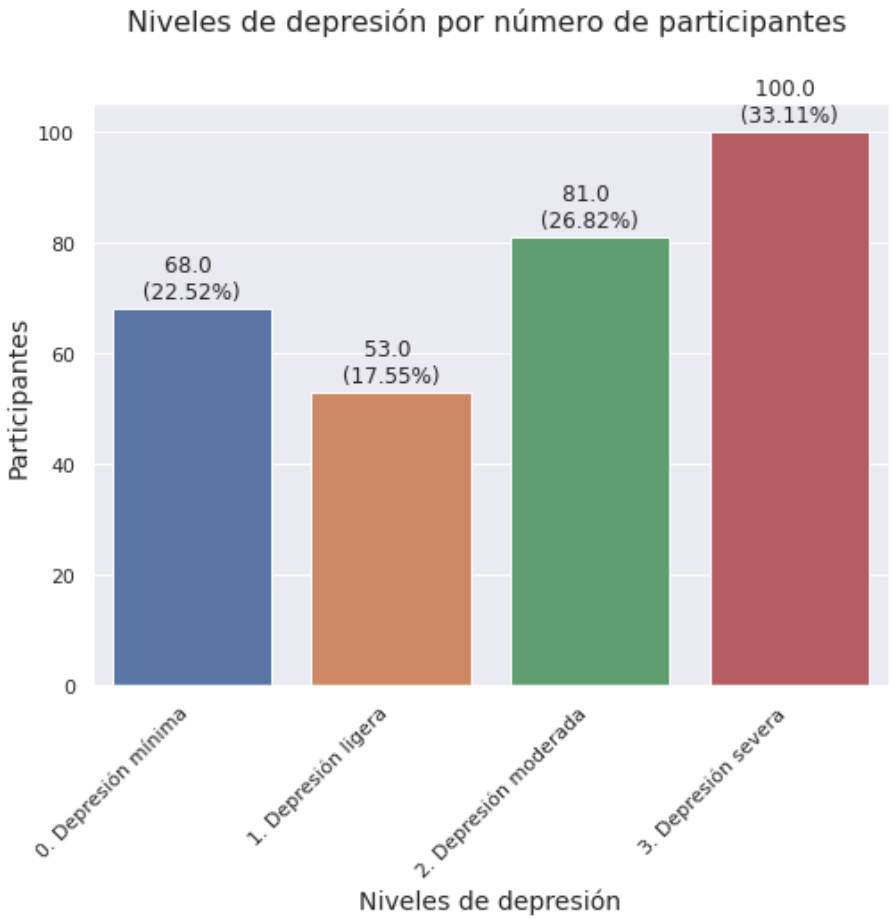


FIG. 3.
Severidad de depresión por número de participantes

En cuanto al análisis del número de personas que son responsables de algún tercero fueron 58 (19.2 %) de los estudiantes. Los detalles del número de los estudiantes se muestras en la Tabla II

TABLA II.
Severidad de depresión por cuidado de algún tercero

Severidad de Depresión	Cuidado	No. Estudiantes
0. Depresión mínima	No	61
	Si	7
1. Depresión ligera	No	41
	Si	12
2. Depresión moderada	No	66
	Si	15
3. Depresión severa	No	77
	Si	23

El análisis del financiamiento de los estudios muestra que 263 (87.08%) de los estudiantes financian su estudio a través de sus padres, tutores u otros familiares. El número de estudiantes por severidad de depresión y modo de financiamiento se muestra en la Tabla III.

TABLA III.
Severidad de depresión por financiamiento de estudios

Severidad de Depresión	Financiamiento de estudios	No. Estudiantes
0. Depresión mínima	A través de becas y/o ayudas económicas	2
	A través de mi trabajo	6
	A través de mis padres u otros familiares	60
1. Depresión ligera	A través de becas y/o ayudas económicas	4
	A través de mi trabajo	5
	A través de mis padres u otros familiares	44
2. Depresión moderada	A través de becas y/o ayudas económicas	3
	A través de mi trabajo	8
	A través de mis padres u otros familiares	70
3. Depresión severa	A través de becas y/o ayudas económicas	7
	A través de mi trabajo	8
	A través de mis padres u otros familiares	84
	A través de un préstamo estudiantil	1

El análisis de las personas con quien viven los estudiantes mostró que 257 (85.1%) viven con sus padres. Su desglose por la severidad de trastorno depresivo se ve en la Fig. 4.

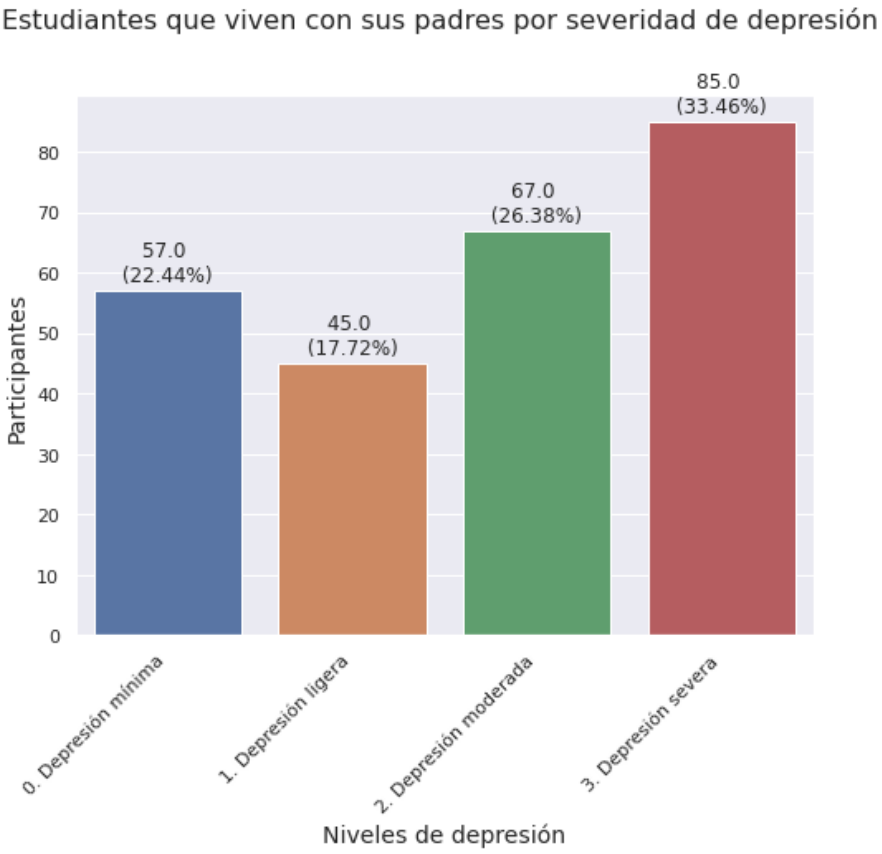


FIG. 4.
Estudiantes que viven con sus padres por severidad de depresión

Por otro lado, la distribución de las personas que no viven con sus padres, por su severidad de depresión se observa en la Tabla IV, donde no se muestra una diferencia considerable de la severidad de la depresión.

TABLA IV.
Estudiantes que no viven con sus padres por severidad de depresión

Severidad de Depresión	Con quien vives	No. Estudiantes
0. Depresión mínima	Vivo con amigos/compañeros/personas que no son mis familiares	1
	Vivo con mi pareja	3
	Vivo con otros familiares que no son mis padres	5
	Vivo solo o sola	2
1. Depresión ligera	Vivo con mi pareja	1
	Vivo con otros familiares que no son mis padres	4
	Vivo solo o sola	3
2. Depresión moderada	Vivo con amigos/compañeros/personas que no son mis familiares	4
	Vivo con mi pareja	1
	Vivo con otros familiares que no son mis padres	7
	Vivo solo o sola	2
3. Depresión severa	Vivo con amigos/compañeros/personas que no son mis familiares	2
	Vivo con mi pareja	1
	Vivo con otros familiares que no son mis padres	6
	Vivo solo o sola	6

En cuanto a los datos académicos, se tuvo la siguiente relación entre severidad de depresión con las variables de: agrado por la carrera, materias de formación básicas, materias exigentes y materias repetidas como se muestra en las tablas: Tabla V, Tabla VI, Tabla VII y Tabla VIII. Estos datos no cuentan con una diferencia significativa en cuanto a la severidad de depresión.

TABLA V.
Agrado por la carrera por severidad de depresión

Severidad de Depresión	Gusto por la carrera	No. Estudiantes
0. Depresión mínima	De acuerdo	23
	Ni de acuerdo ni en desacuerdo	10
	Totalmente de acuerdo	33
	Totalmente en desacuerdo	2
1. Depresión ligera	De acuerdo	17
	En desacuerdo	2
	Ni de acuerdo ni en desacuerdo	13
	Totalmente de acuerdo	19
	Totalmente en desacuerdo	2
2. Depresión moderada	De acuerdo	30
	En desacuerdo	1
	Ni de acuerdo ni en desacuerdo	26
	Totalmente de acuerdo	20
	Totalmente en desacuerdo	4
3. Depresión severa	De acuerdo	33
	En desacuerdo	7
	Ni de acuerdo ni en desacuerdo	43
	Totalmente de acuerdo	12

TABLA VI.
Materias básicas por severidad de depresión.

Severidad de Depresión	Materias básicas	No. Estudiantes
0. Depresión mínima	No	31
	Sí pero no son la mayoría de las materias que tomo	16
	Sí y son la mayoría de las materias que tomo	21
1. Depresión ligera	No	28
	Sí pero no son la mayoría de las materias que tomo	13
	Sí y son la mayoría de las materias que tomo	12
2. Depresión moderada	No	47
	Sí pero no son la mayoría de las materias que tomo	10
	Sí y son la mayoría de las materias que tomo	24
3. Depresión severa	No	47
	Sí pero no son la mayoría de las materias que tomo	20
	Sí y son la mayoría de las materias que tomo	33

TABLA VII.
Materias exigentes por severidad de depresión

Severidad de Depresión	Materias exigentes	No. Estudiantes
0. Depresión mínima	Dos	25
	Más de tres	4
	Ninguna	7
	Tres	10
	Una	22
1. Depresión ligera	Dos	13
	Más de tres	7
	Ninguna	4
	Tres	13
	Una	16
2. Depresión moderada	Dos	35
	Más de tres	6
	Ninguna	6
	Tres	18
	Una	16
3. Depresión severa	Dos	36
	Más de tres	12
	Ninguna	2
	Tres	33
	Una	17

TABLA VIII.
Materias repetidas por severidad de depresión

Severidad de Depresión	Repetir materia	No. Estudiantes
0. Depresión mínima	No, ninguna	49
	Si y me preocupa mucho	8
	Si y me preocupa un poco	4
	Si, pero no me preocupa	7
1. Depresión ligera	No, ninguna	38
	Si y me preocupa mucho	8
	Si y me preocupa un poco	3
	Si, pero no me preocupa	4
2. Depresión moderada	No, ninguna	58
	Si y me preocupa mucho	16
	Si y me preocupa un poco	6
	Si, pero no me preocupa	1
3. Depresión severa	No, ninguna	56
	Si y me preocupa mucho	29
	Si y me preocupa un poco	11
	Si, pero no me preocupa	4

Por el lado del análisis de la carga horaria, se separó en grupos: aquellos estudiantes que trabajan con un total de 88 estudiantes y los que no trabajan con un total de 214 estudiantes.

En el caso de las horas de clases semanales, se observa que existen pocas variaciones respecto a la distribución de porcentajes de severidad de depresión tanto de estudiantes que trabajan como aquellos que no trabajan como se observa en la Fig. 5.

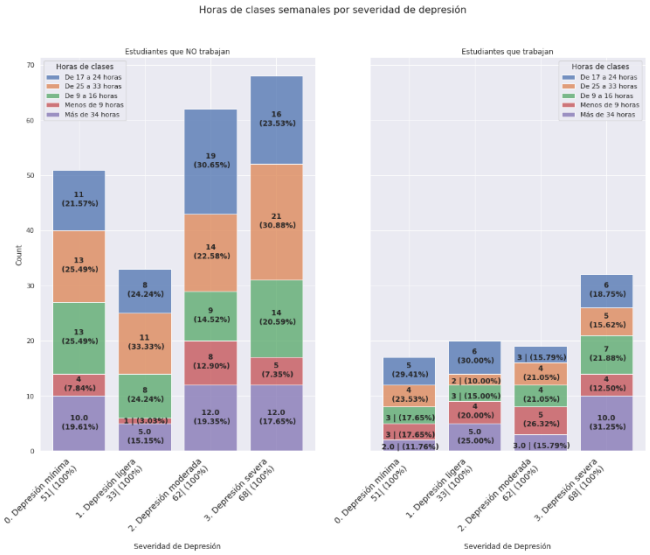


FIG. 5.
Horas de clases semanales por severidad de depresión

En cuanto a las horas de estudio, se puede notar que la mayoría de los estudiantes tiende a estudiar entre 11 a 20 horas semanales independiente de su severidad de depresión o si se encuentran trabajando, ver Fig. 6.

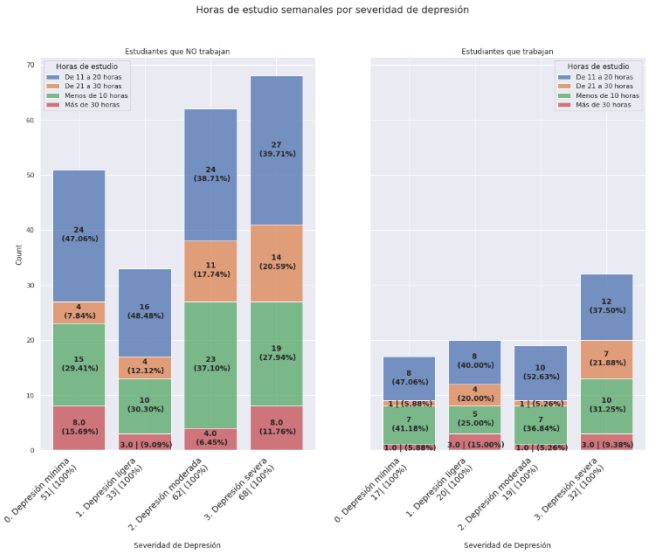


FIG. 6.
Horas de estudio semanales por severidad de depresión

En cuanto a las horas de descanso semanal, se puede notar que más del 50% de los estudiantes afirman tener menos de 4 horas para actividades de ocio en la semana. Como se muestra en la Fig. 7.

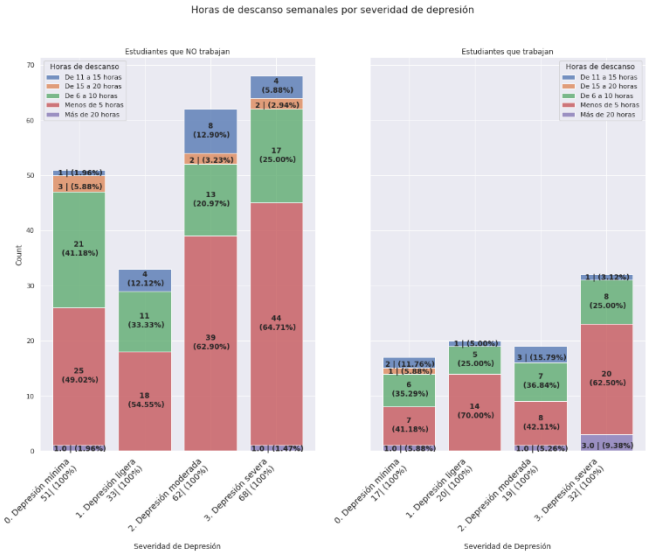


FIG. 7.
Horas de descanso semanales por severidad de depresión

En el caso de la dedicación a las horas de trabajo, se contó únicamente a aquellos estudiantes que trabajan, donde se ve que la mayoría tiene un trabajo que ocupa de 1 a 10 horas de su tiempo a la semana, ver Fig. 8.

El consumo de alcohol habitual y muy frecuente tiene el mayor porcentaje en el nivel de depresión severa, es más, esta categoría es la que mayor proporción de consumo frecuente tiene. Por otro lado, el nunca consumir, en conjunto con el consumir poco, abarca en gran medida al grupo de depresión mínima cómo se observa en la Fig. 10.

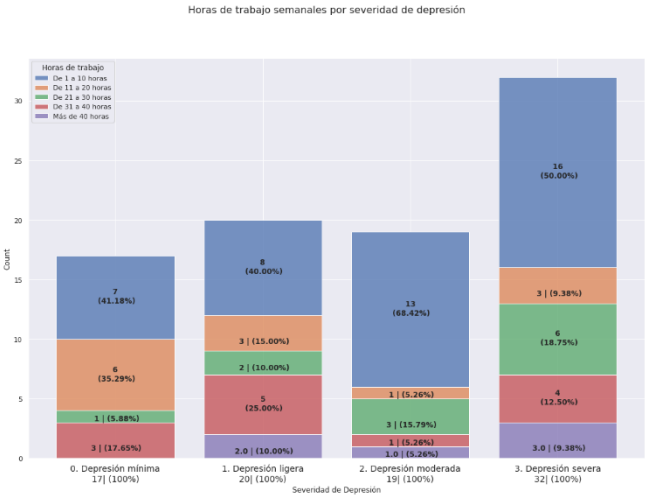


FIG. 8.
Horas de trabajo semanales por severidad de depresión

Por último, el consumo de sustancias psicoactivas no parece existir un porcentaje relevante entre las frecuencias de consumo y la severidad de depresión salvo el hecho que existe un 3% del grupo de estudiantes con depresión severa quienes consumirían sustancias psicoactivas de manera habitual, ver Fig. 11.

En cuanto a las variables de relación interpersonal, tienen una mayor diferencia en cuanto a los niveles de severidad de depresión las relaciones con compañeros de universidad, se evidencia que el porcentaje de relaciones indiferentes u hostiles al no llevarse bien con los compañeros es superior en los niveles de depresión moderada y severa, como se observa en la Fig. 12.

Lo mismo ocurre en caso de tener una relación sentimental donde el porcentaje de no sentirse bien es superior con un 6.17% respecto al número estudiantes con depresión moderada y un 9% respecto a estudiantes con depresión severa, ver Fig. 13.

Por último, el porcentaje "no sentirse o llevarse bien con las personas quien se convive" tiene un marcado porcentaje de 21% respecto al número de estudiantes con depresión severa y 16.05% en estudiantes con depresión moderada. También la respuesta "Me es indiferente" tiene un 32% respecto al número de estudiantes con depresión severa. Estos porcentajes son superiores en comparación con los niveles de depresión mínima y ligera como se ve en la Fig. 14. En este punto, se podría realizar una desagregación según las personas con las que vive, sin embargo, dado que la mayoría de la población (85.1%) vive con sus padres, esto no aportaría al análisis.

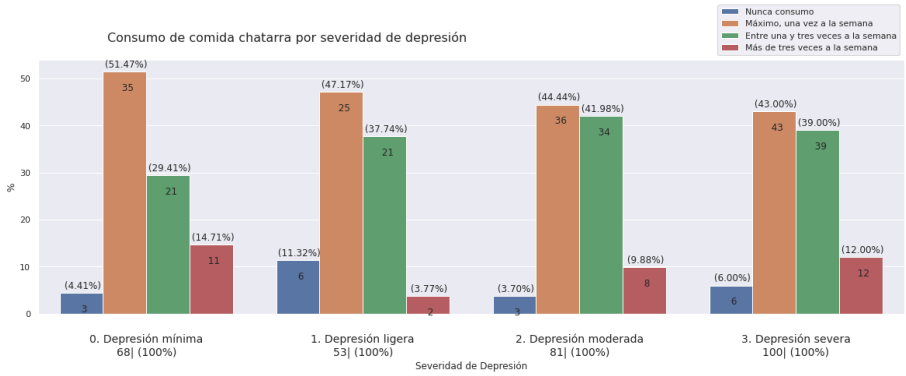


FIG. 9.
Consumo de comida chatarra por severidad de depresión

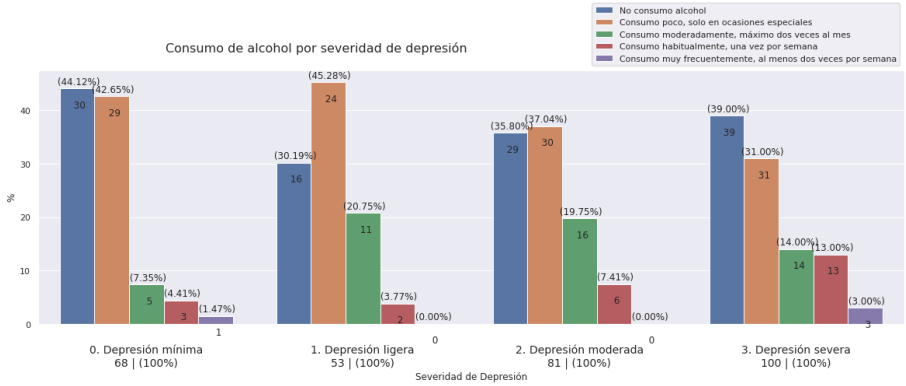


FIG. 10.
Consumo de alcohol por severidad de depresión

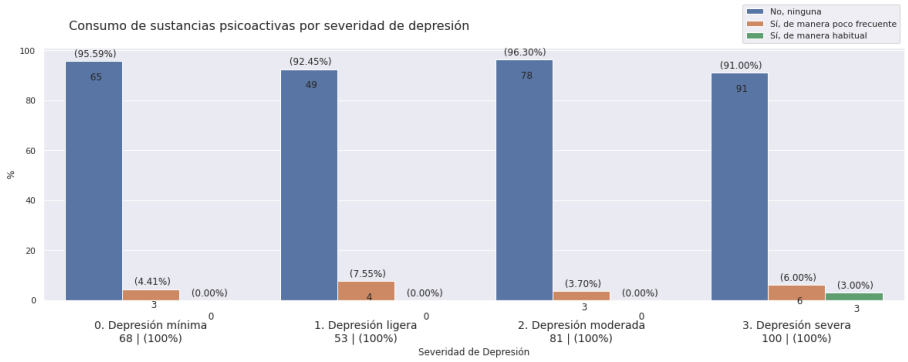


FIG. 11.
Consumo de sustancias psicoactivas por severidad de depresión

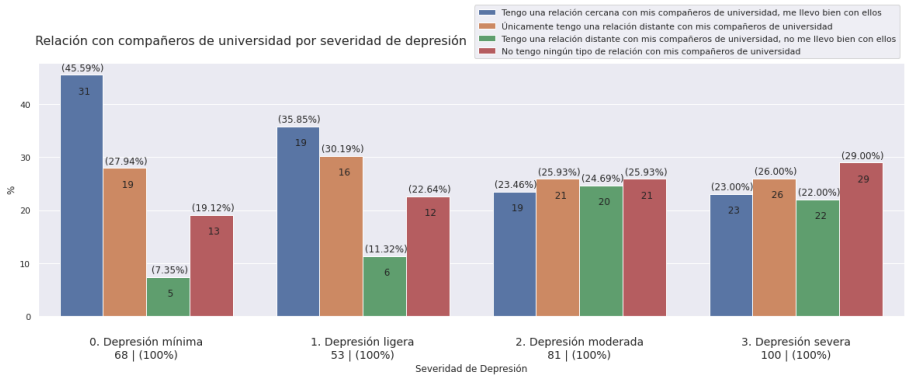


FIG. 12.
Relación con compañeros de universidad por severidad de depresión

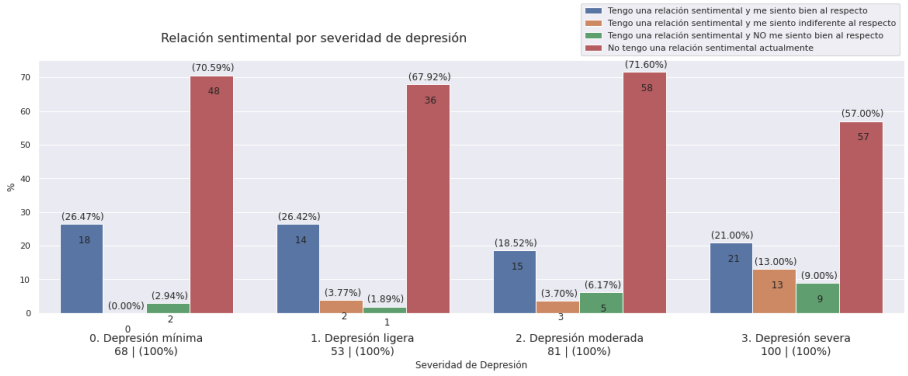


FIG. 13.
Relación sentimental por severidad de depresión

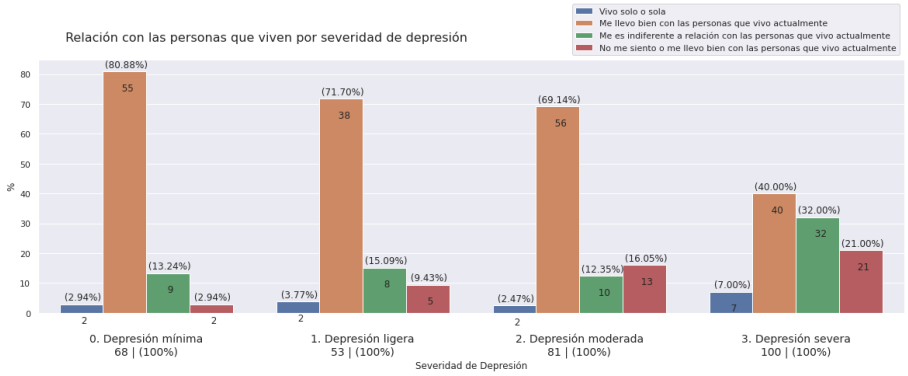


FIG. 14.
Relación con las personas que viven por severidad de depresión

El resumen de la distribución de los datos se muestra la Fig. 15, en donde se muestra el número de estudiantes por la puntuación obtenida en el BDI-II agrupados por los niveles de severidad de depresión. En esta figura, se destaca el número de personas con depresión severa como la más numerosa.

Por último, en la Fig. 16 se muestra como están distribuidos por edad y género el número de estudiantes por severidad de depresión en donde se muestra que existe un mayor número de Hombres que Mujeres en la muestra, además de que el número de Otros tiene solo 3 personas. En cuanto a la edad se ve que la muestra cuenta mayormente con estudiantes de entre 20 a 23 seguido por los estudiantes de entre 16 a 19 años. Estos dos grupos muestran un mayor número de estudiantes con depresión moderada y severa tanto en hombres, mujeres y otros.

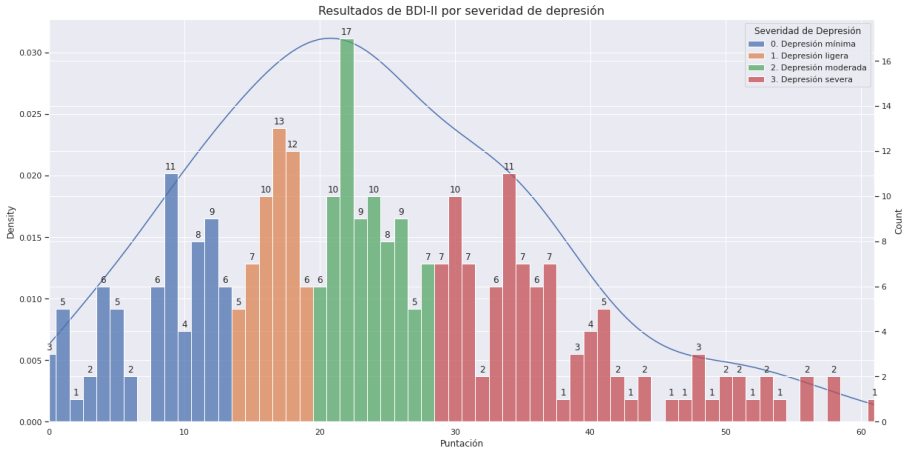


FIG. 15.
Resultados de BDI-II por severidad de depresión

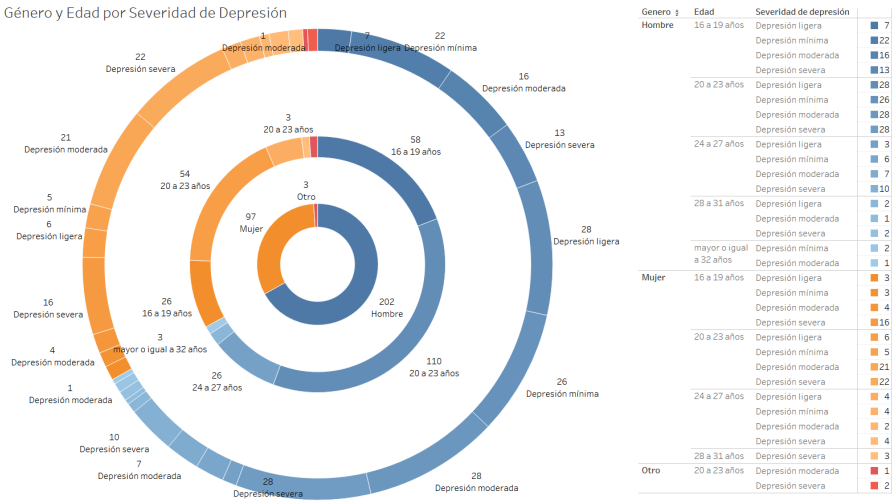


FIG. 16.
Análisis de segregación por género, edad y severidad de depresión

Luego de este análisis de variables, se corroboró estas hipótesis y planeamientos con la construcción, entrenamiento y evaluación de los modelos de ML.

En cuanto a los resultados de rendimiento y evaluación, la función LazyClassifier de LazyPredict mostró los 3 mejores algoritmos puntuados para cada tipo de dato preprocesado que se realizó, donde se destacó los datos normalizados L1 con las métricas de rendimiento respectivas.

Los resultados se obtuvieron de haber entrenado a LazyClassifier con las 19 variables del conjunto de datos normalizados L1, destacando el modelo de DecisionTreeClassifier con un Balanced Accuracy de 0.59 y un F1-Score de 0.58. Seguido por el modelo de BernoulliNB con 0.42 de Balanced Accuracy y 0.41 de F1-Score. Finalmente, el modelo BagginClassifer con 0.39 de Balanced Accurancy y 0.39 de F1-Score como se observa en la Fig. 17.

DATA_NORMALIZED_L1

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
DecisionTreeClassifier	0.57	0.59	None	0.58
BernoulliNB	0.43	0.42	None	0.41
BaggingClassifier	0.41	0.39	None	0.39

FIG. 17.
Modelos seleccionados de ML por LazyPredict para datos normalizados L1

Los resultados de la matriz de confusión para estos modelos se muestran en las Fig. 18, 19, 20, donde se observa la eficacia de los clasificadores en función del Accuracy puntuado.

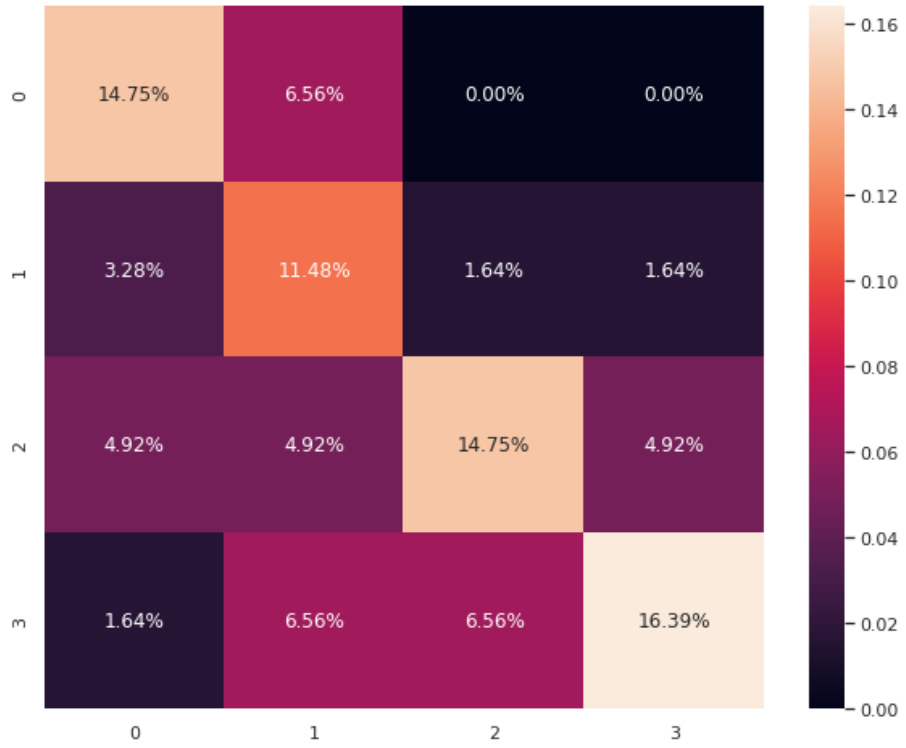


FIG. 18.
Matriz de confusión de DecisionTreeClassifier

Por la parte del análisis de las variables para determinar su relevancia, se desarrolló una permutación de entre 2 a 15 variables que podrían incidir en la depresión destacando que las preguntas q23 y q24, relacionados con el género y edad respectivamente, están presentes como factores a incidir en estudiantes universitarios, como se observa en la Fig. 21

Por otro lado, las preguntas q38, q39, q40, asociadas con relaciones interpersonales, las cuales fueron analizadas como se vio en las Fig. 12, 13 y 14, tienen una mayor incidencia en determinar la severidad de depresión de un estudiante y se llegó a una exactitud balanceada de hasta 0.53 usando un modelo BaggingClassifier.

Mientras que las variables de q30, q31, q32 y q33, relacionadas con la carga horaria, cuya exactitud balanceada muestra valores entre 0.31 y 0.32 con distintos modelos de aprendizaje automático. Esto sustenta el análisis planteado en las Fig. 5, 6, 7 y 8. El detalle de los modelos, el Balanced Accuracy y las variables se muestran en la Fig. 24

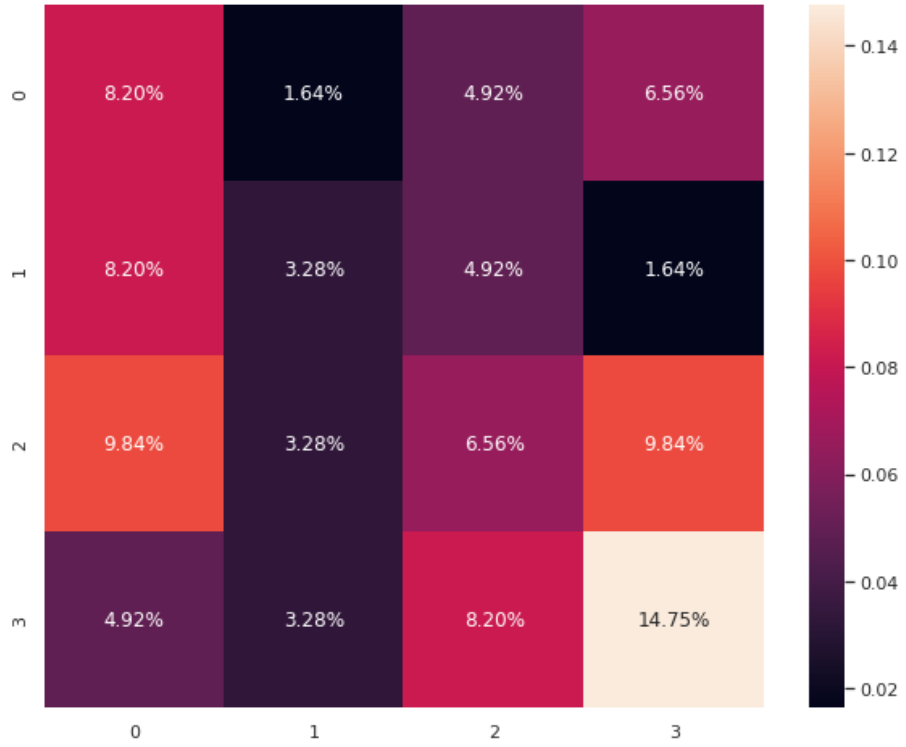


FIG. 19.
Matriz de confusión de BernoulliNB

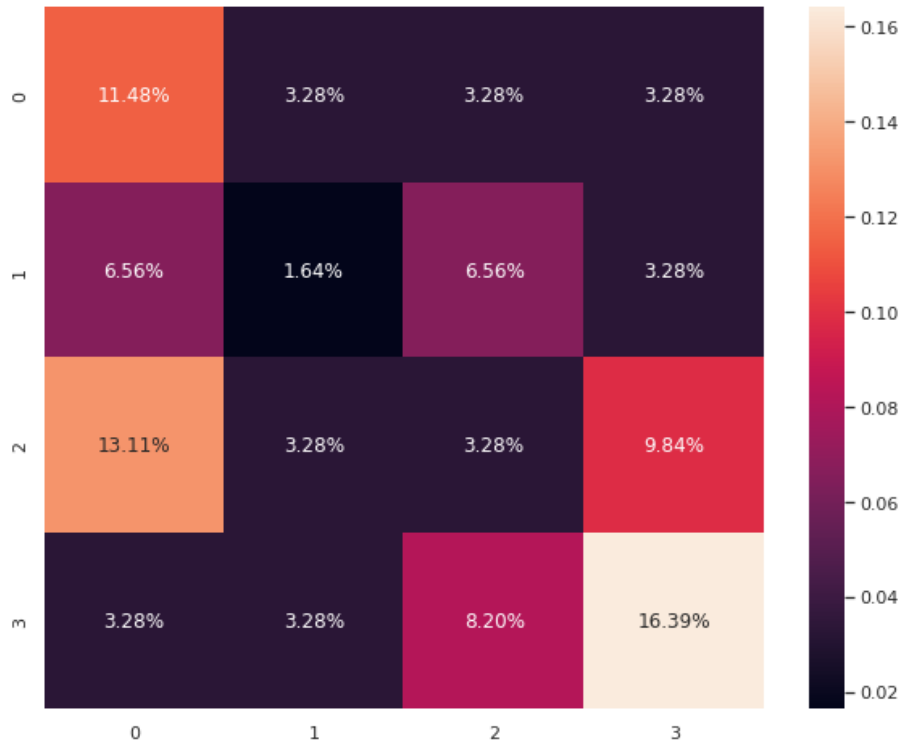
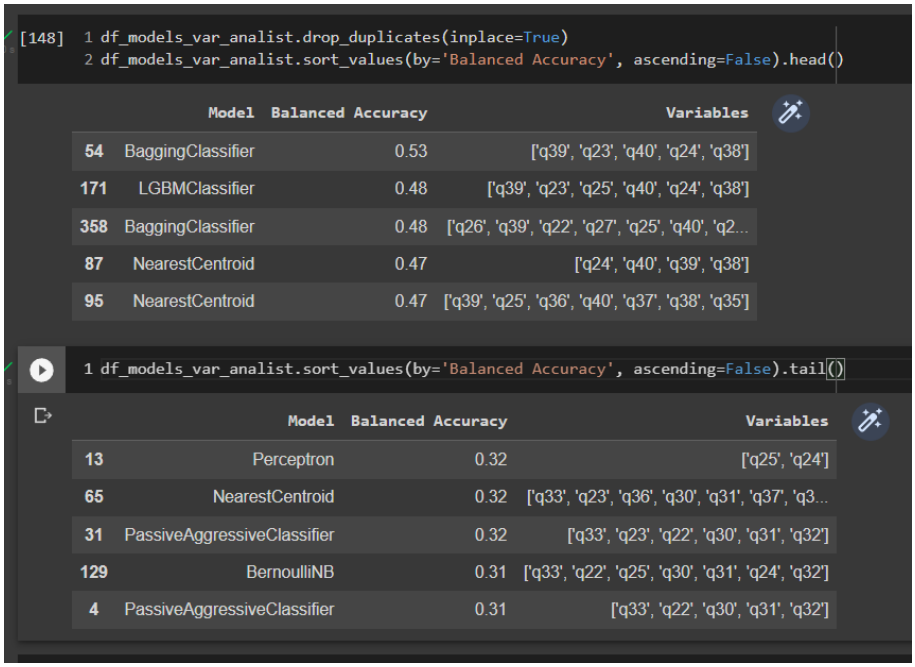


FIG. 20.
Matriz de confusión de BaggingClassifier



```
[148] 1 df_models_var_analist.drop_duplicates(inplace=True)
      2 df_models_var_analist.sort_values(by='Balanced Accuracy', ascending=False).head()
```

	Model	Balanced Accuracy	Variables
54	BaggingClassifier	0.53	['q39', 'q23', 'q40', 'q24', 'q38']
171	LGBMClassifier	0.48	['q39', 'q23', 'q25', 'q40', 'q24', 'q38']
358	BaggingClassifier	0.48	['q26', 'q39', 'q22', 'q27', 'q25', 'q40', 'q2...
87	NearestCentroid	0.47	['q24', 'q40', 'q39', 'q38']
95	NearestCentroid	0.47	['q39', 'q25', 'q36', 'q40', 'q37', 'q38', 'q35']

```
1 df_models_var_analist.sort_values(by='Balanced Accuracy', ascending=False).tail()
```

	Model	Balanced Accuracy	Variables
13	Perceptron	0.32	['q25', 'q24']
65	NearestCentroid	0.32	['q33', 'q23', 'q36', 'q30', 'q31', 'q37', 'q3...
31	PassiveAggressiveClassifier	0.32	['q33', 'q23', 'q22', 'q30', 'q31', 'q32']
129	BernoulliNB	0.31	['q33', 'q22', 'q25', 'q30', 'q31', 'q24', 'q32']
4	PassiveAggressiveClassifier	0.31	['q33', 'q22', 'q30', 'q31', 'q32']

FIG. 21.
Resultado de variables que inciden en la depresión

VI. CONCLUSIONES

- La recolección de datos se la realizó durante las últimas 2 semanas del semestre académico 2022A de la EPN, esto pudo repercutir en el resultado de los estados de estrés, ansiedad y depresión de muchos de los voluntarios, de los cuales, aproximadamente un tercio mostró depresión severa
- Los datos género y edad representaron variables significadas en determinar la severidad de depresión de acuerdo con el análisis estadístico y los resultados del Accuracy de los modelos de ML que tenían estas variables.
- A parte del género y la edad, las variables que inciden en mayor medida fueron aquellas de relaciones interpersonales (q38, q39 y q40), seguido por las variables de consumo (q35, q36 y q37).
- Las variables poco significativas resultaron ser las variables académicas (q26, q27, q28 y q29) y especialmente las variables de carga horaria (q30, q31, q32 y q33) ya que no representaron una diferencia sustancial en los porcentajes de severidad de depresión por número de estudiantes. Lo cual se pudo constatar al aplicar el entrenamiento de los modelos de ML con el conjunto de las variables permutadas, donde muestra que aquellos modelos con Balanced Accuracy más bajo fueron entrenados con estas variables.
- El mejor modelo fue el de Decisión Tree Classifier con un Accuracy de 0.59, sin embargo, para considerar desplegar un modelo óptimo de ML se debería tener un Accuracy más cercano a 1.0.

VII. RECOMENDACIONES

- En caso de realizar un análisis a una población universitaria, relacionado con situaciones de estrés, se debe ejecutar la recolección de información de los participantes en periodos del semestre en los cuales no se tenga carga excesiva de trabajo en el ámbito académico, como lo son la cercanía a exámenes parciales o finales ya que pueden alterar la precepción del estudiante sobre su situación psicológica.

- Se debe seleccionar las variables más significativas en cuanto a la construcción de cualquier modelo que se busque desarrollar. Esta selección a su vez dependerá de la cantidad de datos que se haya recolectado y servirá para entrenar el modelo, obteniendo mejores resultados en sus métricas de rendimiento.
- Se debe determinar las variables a evaluar, para posteriormente realizar el análisis exploratorio y estadístico de los datos, en función de los objetivos planteados en la fase de análisis de la situación o entendimiento del negocio, para evitar iteraciones innecesarias en el ciclo de vida del proyecto.
- Se deberá considerar el recopilar una cantidad de registros (filas) significativa para el entrenamiento de los modelos de ML. Aproximadamente, 100 veces el número de variables (columnas) del conjunto de datos. Por ejemplo, el conjunto de datos de este proyecto contó con 19 variables y la cantidad de 1900 registros sería la óptima.
- Considerar la evaluación conjunta de otros cuestionarios estandarizados para la estimación de severidad de trastorno de depresión.

Agradecimientos

Agradecemos a la Escuela Politécnica Nacional, la Facultad de Ingeniería en Sistemas, su subdecana Monserrate Intriago y al PhD. Denys Flores por su ayuda al distribuir el Cuestionario de variables de índice de depresión, vital para la recolección de datos y sus posteriores fases. Al equipo de Ciencia de Datos del Club de Software de la Facultad de Ingeniería en Sistemas de la EPN con quienes realizamos las fases de análisis de la situación y entendimiento del negocio. A la Asociación de Estudiantes de Ingeniería en Sistemas (AEIS) y la Federación de Estudiantes de la Politécnica Nacional (FEPON) en compartir el cuestionario. Y por último, a todos los estudiantes quienes participaron, de manera anónima, respondiendo el cuestionario.

REFERENCIAS

P. Retamal, Depresión - Guías para el paciente y la familia, Santiago de Chile: Editorial Universitaria, 1999.

World Health Organization (WHO), «COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide,» World Health Organization (WHO), 02 03 2022. [En línea]. Available: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide. [Último acceso: 19 09 2022].

World Health Organization (WHO), «Depression,» World Health Organization (WHO), 2022. [En línea]. Available: https://www.who.int/health-topics/depression#tab=tab_1. [Último acceso: 19 09 2022].

D. Agudelo, C. Claudia y S. Diana, «CARACTERÍSTICAS DE ANSIEDAD Y DEPRESIÓN EN ESTUDIANTES UNIVERSITARIOS,» International Journal of Psychological Research, vol. 1, nº 1, pp. 34-39, 2008.

A. M. Juan, B. Nora, C. A. Paola y M. R. Fray, Prevalencia de Depresión y Factores Asociados en Estudiantes Universitarios de la Ciudad de Cuenca-Ecuador, Cuenca: Universidad de la Ciudad de Cuenca-Ecuador, 2015.

L. Alomoto y G. Cañarejo, ASOCIACIÓN ENTRE EL APOYO SOCIAL Y SÍNTOMAS DE ANSIEDAD Y DEPRESIÓN EN ESTUDIANTES UNIVERSITARIOS DE PRIMER NIVEL DE LA PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR, SEDES QUITO, IBARRA, SANTO DOMINGO Y PORTOVIEJO DURANTE EL AÑO 2018., Quito: PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR, 2018

I. Gaibor y R. Moreta, «Optimismo disposicional, ansiedad, depresión y estrés en una muestra del Ecuador. Análisis inter-género y de predicción,» Actualidades en Psicología, vol. 34, nº 129, pp. 17-31, 2018.

R. Moreta, J. Zambrano, H. Sánchez y S. Naranjo, «Salud mental en universitarios del Ecuador: síntomas relevantes, diferencias por género y prevalencia de casos,» Pensamiento Psicológico, vol. 19, nº 1, pp. 1-26, 2021.

S. d. R. Puchaicela, J. Lozam, I. Fiallo, A. Benítez y A. Amaya, «Evaluación de estrés, ansiedad y depresión en Ecuador durante la pandemia de COVID-19,» La Ciencia al Servicio de la Salud y la Nutrición, vol. 13, nº 1, pp. 13-25, 2022.

R. Chiong, G. SatiaBudhi, S. Dhakal y F. Chiong, «A textual-based featuring approach for depression detection using machine learning classifiers and social media texts,» Computers in Biology and Medicine, vol. 135, 2021.

L. Danxia, F. Xing Lin, A. Farooq, S. Muhammad y G. Jing, «Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review,» JMIR Ment Health, vol. 9, nº 3, 2022.

R. Razavi, A. Gharipour y M. Gharipour, «Depression screening using mobile phone usage metadata: a machine learning approach,» Journal of the American Medical Informatics Association, vol. 24, nº 4, pp. 522-530, 2020.

X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankoff y A. K. Dey, «Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students,» Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, nº 3, pp. 1-33, 2019.

M. Zhao y Z. Feng, «Machine Learning Methods to Evaluate the Depression Status of Chinese Recruits: A Diagnostic Study,» Neuropsychiatric Disease and Treatment, vol. 16, p. 2743 – 2752, 2020.

A. Beck, R. Steer y G. Brown, «BECK DEPRESSION INVENTORY-SECOND EDITION,» National Child, 2022. [En línea]. Available: https://www.nctsn.org/measures/beck-depression-inventory-second-edition. [Último acceso: 31 08 2022].

G. A. Mehmet Taha, «Beck Depression Inventory-II: A Study for Meta Analytical Reliability Generalization,» Pegem Journal of Education and Instruction, vol. 11, nº 3, pp. 88-101, 2021.

K. L. Smarr y A. Keefer, «Measures of Depression and Depressive Symptoms,» Arthritis Care & Research, vol. 63, nº 11, pp. 454-466, 2011.

IBM, «CRISP-DM Help Overview,» IBM, 17 08 2021. [En línea]. Available: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview. [Último acceso: 19 09 2022].

S. Raschka, J. Patterson y C. Nolet, «Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence,» information, vol. 11, nº 4, pp. 1-44, 2020.

Google Colab, «Te damos la bienvenida a Colab,» Google Colab, 2022. [En línea]. Available: https://colab.research.google.com/#scrollTo=5fCEDCU_qrC0. [Último acceso: 19 09 2022].

Microsoft 365, «Microsoft Forms,» Microsoft 365, 2022. [En línea]. Available: https://www.microsoft.com/en-us/microsoft-365/online-surveys-polls-quizzes. [Último acceso: 19 09 2022].

R. Barrientos, N. Ramírez, H. Acosta, I. Suárez, M. Trejo, P. León y S. Blázquez, «Árboles de decisión como herramienta en el diagnóstico,» Revista Médica de la Universidad Veracruzana, vol. 9, nº 2, pp. 19-24, 2009.

H. Coronado, A. Han y L. García, Detección Automática de Sitios Web Fraudulentos, Madrid: Universidad Complutense de Madrid, 2020.

J. Domínguez, Inteligencia artificial para la detección de fraude en transacciones realizadas con tarjetas de crédito, Sevillar: Universidad de Sevilla, 2021.

C. Chamat, Modelo Predictivo de Deserción Estudiantil de Educación Preescolar, Básica y Media en el Municipio de Medellín, Medellín: Universidad de Antioquia, 2021.

C. Sánchez, Selection Heuristics on Semantic Genetic Programming for Classification Problems, Aguascalientes: INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN, 2020

F. Izco, «Base de datos corporativa de personas,» bookdown.org, 27 11 2018. [En línea]. Available: https://bookdown.org/f_izco/BDC-POC/metricas.html. [Último acceso: 20 09 2022].

N. HOTZ, «What is CRISP DM?,» Data Science Process Alliance, 08 08 2022. [En línea]. Available: https://www.datascience-pm.com/crisp-dm-2/. [Último acceso: 31 08 2022].