UNIVERSIDAD TECNOLÓGICA DE CAMPECHE

INGENIERÍA EN DESARROLLO Y GESTIÓN DE SOFTWARE

ACTIVIDAD:

PRÁCTICA

ASIGNATURA:

EXTRACION DE CONOCIMIENTO EN BASE DE DATOS

PRESENTADO POR:

Magaly Monceraht Graniel Peralta

Damaris del Mar Ochoa Damian

Erick Valenzuela Lagunas

Fausto Diaz Vazquez

NOMBRE DEL PROFESOR:

MTRA. MARIA DE LOURDES CÁRDENAS MALDONADO

GRADO Y GRUPO

9 “B”

PERIODO

MAYO – AGOSTO 2023

JUSTIFICACIÓN DEL ALGORITMO ULTILIZADO

Este caso de estudio busca diseñar un algoritmo que pueda recabar y caracterizar la salud mental de los estudiantes de universidad de las áreas de sistemas e informática después de la experiencia de aislamiento enforzada por el COVID-19.

Se usa el algorimo K-Means finalmente debido a su estatus como un algoritmo de aprendizaje no supervisado que existe dentro de los modelos de agrupamiento. La razón tras esto debería resultar evidente, pero es porque se busca encasillar a las diferentes entradas de información en grupos para su manejo como segmentos del espectro de la salud mental.

El objetivo de este estudio se centra en la utilización de técnicas de minería de datos y la aplicación de un algoritmo no supervisado K-Means para la agrupación y caracterización de los niveles de depresión, ansiedad, y estrés a través de la Escala DASS-21, en estudiantes universitarios del área de sistemas e informática de las universidades de Abancay, Apurímac-Perú.

MODELO UTILIZADO

- **Algoritmo K-Means (Modelo de Agrupación):**

Se aplicó la metodología CRISP-DM, Cross Industry Standard Process for Data Mining, que proporciona una descripción general del ciclo de vida de un proyecto de minería de datos. Contiene las siguientes fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

Comprensión del negocio:

El objetivo de negocio de este estudio es caracterizar las condiciones de salud mental de los estudiantes universitarios que podrían verse afectadas por el confinamiento obligatorio por el COVID-19, de las Escuelas Profesionales de Sistemas e Informática de la Universidad Tecnológica de los Andes y la Universidad Nacional Micaela Bastidas de Apurímac. El objetivo de la minería de datos es utilizar el algoritmo K-Means, para describir los

niveles de trastornos psicológicos como la Depresión, Ansiedad y Estrés, utilizando la Escala DAAS-21, la misma que se muestra en la Tabla 1.

DASS-21 La escala de calificación es la siguiente: **0: No me ha ocurrido; 1: Me ha ocurrido un poco, o durante parte del tiempo; 2: Me ha ocurrido bastante, o durante una buena parte del tiempo; 3: Me ha ocurrido mucho, o la mayor parte del tiempo.**					
1	Me ha costado mucho descargar la tensión	0	1	2	3
2	Me di cuenta que tenía la boca seca	0	1	2	3
3	No podía sentir ningún sentimiento positivo	0	1	2	3
4	Se me hizo difícil respirar	0	1	2	3
5	Se me hizo difícil tomar la iniciativa para hacer cosas	0	1	2	3
6	Reaccioné exageradamente en ciertas situaciones	0	1	2	3
7	Sentí que mis manos temblaban	0	1	2	3
8	He sentido que estaba gastando una gran cantidad de energía	0	1	2	3
9	Estaba preocupado por situaciones en las cuales podía tener pánico o en las que podría hacer el ridículo.	0	1	2	3
10	He sentido que no había nada que me ilusionara	0	1	2	3
11	Me he sentido inquieto	0	1	2	3
12	Se me hizo difícil relajarme	0	1	2	3
13	Me sentí triste y deprimido	0	1	2	3
14	No toleré nada que no me permitiera continuar con lo que estaba haciendo	0	1	2	3
15	Sentí que estaba al punto de pánico	0	1	2	3
16	No me pude entusiasmar por nada	0	1	2	3
17	Sentí que valía muy poco como persona	0	1	2	3
18	He tendido a sentirme enfadado con facilidad	0	1	2	3
19	Sentí los latidos de mi corazón a pesar de no haber hecho ningún esfuerzo físico	0	1	2	3
20	Tuve miedo sin razón	0	1	2	3
21	Sentí que la vida no tenía ningún sentido	0	1	2	3

Tabla 1

Comprensión de los datos:

La encuesta aplicada consta de 25 ítems, los 5 primeros corresponde al nombre de la Universidad y datos personales de los estudiantes como código, correo institucional, fecha de nacimiento y sexo. Los otros 21 ítems corresponden a la presencia e intensidad de cada trastorno psicológico que evalúa la Escala DAAS-21, estos ítems se dividen en 3 grupos de 7 ítems: Depresión (ítems: 3, 5, 10, 13, 16, 17 y 21), Ansiedad (ítems: 2, 4, 7, 9, 15, 19 y 20) y Estrés (ítems: 1, 6, 8, 11, 12, 14 y 18).

La escala de respuesta es del tipo Likert de 0 a 3 puntos, el puntaje total se calcula con la suma de los ítems pertenecientes a cada escala y varía entre 0 y 21 puntos. La interpretación de cada sintomatología se realizará de acuerdo a la siguiente puntuación que se muestra en la Tabla 2.

Sub Escala DASS21	Puntuación	Escala
Depresión	5-6	Depresión leve
	7-10	Depresión moderada
	11-13	Depresión severa
	14 a más	Depresión extremadamente severa
Ansiedad	4	Ansiedad leve
	5-7	Ansiedad moderada
	8-9	Ansiedad severa
	10 a más	Ansiedad extremadamente severa
Estrès	8-9	Estrès leve
	10-12	Estrès moderada
	13-16	Estrès severa
	17 a más	Estrès extremadamente severa

Tabla 2

Preparación de los datos:

El objetivo de esta fase es adaptar los datos para el modelado de minería, según la técnica a desarrollar. Se seleccionaron los atributos para el modelo descriptivo. Los datos de la encuesta se obtuvieron en un archivo Excel y luego se guardó en un archivo con la extensión CSV. Se estandarizó los valores de cada ítem en formato de tipo numérico, tomando la siguiente escala que se muestra en la Tabla 3.

Campo de origen	Campo de flujo
No me ha ocurrido	0
Me ha ocurrido un poco, o durante parte del tiempo.	1
Me ha ocurrido bastante, o durante una buena parte del tiempo.	2
Me ha ocurrido mucho, o la mayor parte del tiempo.	3

Tabla 3

Se elaboraron 9 vistas minables en formato csv. tres vistas que representan el total del conjunto de datos constituido por 189 instancias y 7 atributos, para evaluar los trastornos psicológicos de Depresión, Ansiedad y Estrés. Se dividió el conjunto de datos por sexo y se obtuvieron otras 6 vistas minables que corresponden a cada escala con 7 atributos y 155 instancias que corresponden al género masculino y otras de 34 instancias que corresponde al género femenino, esto con la finalidad de establecer comparaciones.

Modelado:

En esta fase se utilizó la técnica descriptiva de agrupamiento de datos (clustering) con el algoritmo KMeans, donde se optó por la creación cuatro

clusters, esto por la clasificación de la encuesta en sus cuatro niveles de sintomatología: leve, moderada, severa y extremadamente severa. Se utilizó la herramienta WEKA para el modelado de minería de datos

Evaluación:

-Depresión:

Para la elaboración del modelo de depresión con el algoritmo K-Means, con 189 instancias que representa el total de estudiantes encuestados, 7 atributos que corresponden a los 7 ítems, y con 10 iteraciones, se puede obtener los siguientes resultados en cuatro clusters, los mimos que se muestran en la Tabla 4.

- Cluster 0, sus características representan un nivel de depresión extremadamente severa, con 19 puntos y corresponde al 12% de estudiantes.
- Cluster 1, los datos corresponden a características de un nivel de depresión extremadamente severa, con 21 puntos y corresponde al 5% de estudiantes.
- Cluster 2, los niveles de depresión están por debajo de los puntos de corte, por lo tanto, no sufren de depresión y es equivalente al 40% de estudiantes.
- Cluster 3, sus características representan un nivel de depresión moderada, con 8 puntos y corresponde al 43% de estudiantes.

Cluster 0	2,3,3,3,3,2,3
Cluster 1	3,3,3,3,3,3,3
Cluster 2	0,0,0,0,0,0,0
Cluster 3	1,1,1,1,1,1,2

Tabla 4

-Ansiedad:

Se evaluó la ansiedad con 189 instancias, 7 atributos que corresponden a los 7 ítems, y 5 iteraciones, obteniendo los siguientes resultados en cuatro clusters, los mimos que se muestran en la Tabla 5.

- Cluster 0, los datos corresponden a características de un nivel de ansiedad extremadamente severa, con 12 puntos y corresponde al 10% de estudiantes.
- Cluster 1, los datos corresponden a características de un nivel de ansiedad extremadamente severa, con 21 puntos, correspondiente al 3% de los estudiantes.
- Cluster 2, los niveles de ansiedad están por debajo de los puntos de corte, por lo tanto, no sufren de depresión y es equivalente al 54% de estudiantes.
- Cluster 3, sus características representan un nivel de ansiedad moderada, con 6 puntos, correspondiente al 33% de los estudiantes.

Cluster 0	1,1,0,3,2,3,2
Cluster 1	3,3,3,3,3,3,3
Cluster 2	1,0,0,0,0,0,0
Cluster 3	1,1,1,0,1,1,1

Tabla 5

-Estrés:

Se evaluó el estrés con 189 instancias, 7 atributos que corresponden a los 7 ítems, y 5 iteraciones, obteniendo los siguientes resultados en cuatro clusters.

- Cluster 0, sus características representan un nivel de estrés moderado, con 11 puntos y corresponde al 17% de estudiantes.
- Cluster 1, sus características representan un nivel de estrés extremadamente severo, con 21 puntos, correspondiente al 3% de los estudiantes.
- Cluster 2, los estudiantes no sufren de estrés, debido a que los niveles se encuentran debajo de los puntos de corte y corresponde al 40% de estudiantes.
- Cluster 3, los niveles de estrés están por debajo de los puntos de corte, por lo tanto, no sufren de estrés y corresponde al 40% de estudiantes.

EVALUACIÓN Y OPTIMIZACIÓN DEL MODELO

Como resultado de la aplicación de la encuesta DASS-21 en estudiantes universitarios, el algoritmo K-Means fue ejecutado para el agrupamiento de los datos, con cuatro clústers por los cuatro niveles de sintomatología: leve, moderada, severa y extremadamente severa, que se puede dar según la puntuación obtenida en la evaluación.

-Depresión:

De acuerdo a los resultados obtenidos de los clusters, se puede describir que en la población de estudiantes universitarios el 60% presenta niveles de depresión, teniendo mayor prevalencia en los niveles de depresión moderada, mientras el 40% restante no sufre de depresión. la misma que se muestra en la Figura 1.

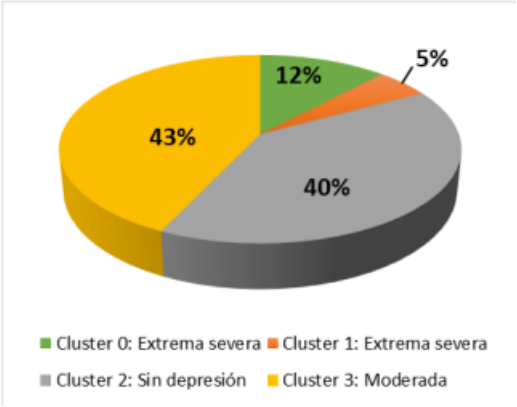

Figura 1

-Ansiedad:

De acuerdo a los resultados obtenidos, se puede describir que en la población de estudiantes el 54% no sufren de ansiedad, mientras el 46% restante presenta niveles de ansiedad, con mayor prevalencia en niveles de ansiedad moderada, tal como se muestra en la Figura 2.

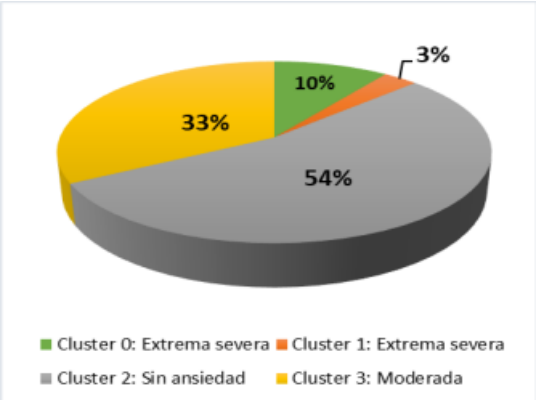

Figura 2

-Estrés:

De acuerdo a los resultados obtenidos, se puede describir el 80% de estudiantes no sufren de estrés, mientras el 20% restante presenta niveles de estrés, con mayor incidencia en niveles de estrés moderado, tal como se muestra en la Figura 3.

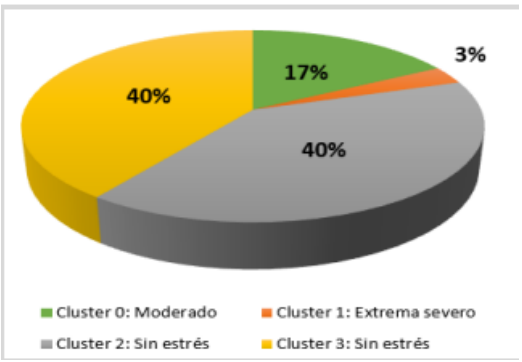

Figura 3

La Figura 4, muestra resultados más específicos de los trastornos psicológicos que presentan las mujeres y hombres; donde se puede determinar que las mujeres tienen un mayor porcentaje de depresión, ansiedad y estrés en comparación de los hombres.

Los resultados de la depresión reflejan una prevalencia de 79%, donde el nivel de depresión leve fue la más frecuente (41%) en el género femenino. En cuanto a la ansiedad se presenta niveles de ansiedad moderada en ambos géneros, con una mayor prevalencia en el género femenino (64%).

En cuanto al estrés se observa que el género masculino Memorias de la Vigésima Primera Conferencia Iberoamericana en Sistemas, Cibernética e

Informática (CISCI 2022) presenta niveles de estrés severos (19%) en comparación al nivel de estrés leve que presenta el género femenino con una mayor prevalencia (33%).

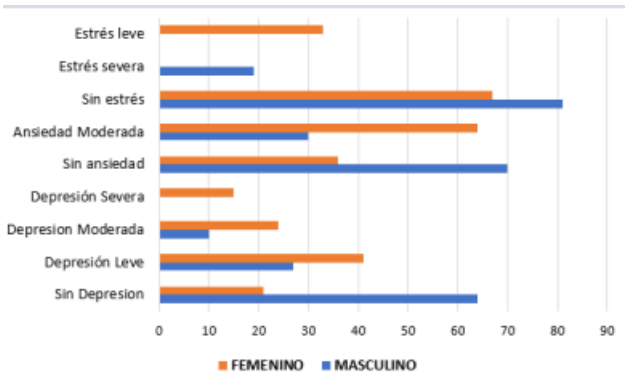

Figura 4

Para cada una de la pruebas se ejecutaron, los algoritmos K-means, K-means++ y RENTOL, haciendo variaciones del valor de K, desde $K = 2,\ldots, 9$. Se utilizó la distancia Euclidiana, además se midió la calidad de agrupamiento utilizando índices de validación internos: PS, S_Dbw, CS.

CONCLUSIÓN

En conclusión podemos decir que los algoritmos de Aprendizaje no Supervisado manejan datos sin entrenamiento previo,en otras palabras es una función que hace su trabajo con los datos a su disposición,es decir,se deja a su suerte para que resuelva las cosas a su antojo.

Los algoritmos no supervisados funcionan con datos no etiquetados,así como su propósito es la exploración, explorar la estructura de la información y detectar patrones distintos, así como extraer ideas valiosas y aplicarla en su funcionamiento con el fin de aumentar la eficacia del proceso de toma de decisiones.

El algoritmo que se utilizó para esta práctica fue de agrupación (algoritmo k-means) el cual nos dice que es uno de los algoritmos de agrupación más utilizados, este se encarga de dividir los datos en k grupos o clusters, donde k es el número de clusters dado. El algoritmo comienza con k centros seleccionados al azar y asigna el centro más cercano a cada punto de datos.

Este algoritmo se aplicó en un grupo de estudiantes universitarios para evaluar su salud mental, es decir,se tiene como objetivo que este estudio ayude a utilizar las técnicas de minería de datos y la aplicación de un algoritmo no supervisado K-Means para la agrupación y de esta manera poder saber los niveles de depresión, ansiedad, y estrés a través de la Escala DASS-21, en estudiantes universitarios del área de sistemas e informática de las universidades de Abancay, Apurímac-Perú