Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Detección de casos de depresión en una población estudiantil aplicando algoritmos de aprendizaje no supervisado

Tesis presentada al

Posgrado de Ciencias de la Computación

Como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Por

Lic. Octavio Mendoza Gómez

Asesorado por

Dra. Mireya Tovar Vidal
Dr. Guillermo De Ita Luna

Puebla, Pue.
6 de Septiembre 2024

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
Maestría en Ciencias de la Computación

Detección de casos de depresión en una población estudiantil aplicando algoritmos de aprendizaje no supervisado.

Tesis que presenta el Lic. Octavio Mendoza Gómez, para obtener el grado de Maestro en Ciencias de la Computación

Autor

Lic. Octavio Mendoza Gómez

Directores de tesis

Dra. Mireya Tovar Vidal
Dr. Guillermo De Ita Luna

Puebla, Puebla
6 de Septiembre del 2024

Resumen

En la actualidad la aplicación de algoritmos de aprendizaje automático en el área médica va en aumento. Entre estas aplicaciones podemos observar la detección de tumores mediante análisis de imágenes, la observación de embarazos complicados, el seguimiento y observación de los casos de diabetes. Pero no se limitan a enfermedades fisiológicas, también se han enfocado a enfermedades mentales de tal manera que se puedan detectar en una etapa temprana y se les dé un mejor tratamiento a los pacientes que las padezcan.

En este trabajo de tesis se aplicarán algoritmos de aprendizaje no supervisado para reconocer y agrupar los posibles casos de depresión en una población estudiantil de la Facultad de ciencias de la computación de la Benemérita Universidad Autónoma de Puebla. Para realizar esta tarea se seguirá la metodología de un proyecto de ciencias de datos, donde se realizará un preprocesamiento de datos y una implementación de los algoritmos de aprendizaje no supervisado que se proponen en este trabajo.

Para evaluar la metodología y las propuestas de algoritmos se tomarán en cuenta las métricas para aprendizaje no supervisado, coeficiente de la silueta e índice de Davies-Bouldin, de los distintos modelos obtenidos de tal manera que podamos dar un veredicto del mejor modelo para detectar y categorizar los casos de depresión.

Dedicatoria

A mi padre y a mi hermana, por su amor incondicional y su apoyo constante en cada paso de mi vida. Gracias por enseñarme el valor del esfuerzo y la perseverancia.

A mi prometida, por la compañía y noches de desvelo que me acompaño. Gracias por acompañarme en esta etapa de mi vida y verme crecer.

A mis profesores y mentores, por su guía y sabiduría, que han sido fundamentales en mi formación académica y profesional.

Agradecimientos

Agradezco a mis asesores, la Dra. Mireya Tovar Vidal y el Dr. Guillermo De Ita Luna, por todas las enseñanzas que me han dado a lo largo de esta maestría así como su apoyo.

Agradezco a Ricardo por 24 años de amistad, en los cuales siempre me apoyo y tuvo confianza en mi. Agradezco a Eduardo y Víctor, amigos incondicionales que siempre han estado para mi.

Agradezco a la Vicerrectoría de Investigación y Estudios de Posgrados, por su apoyo en la realización del proyecto que se presenta en este trabajo de tesis.

Y sobre todo agradezco a la Facultad de Ciencias de la Computación por darme esta oportunidad de encontrar mi verdadera vocación y poder hacer lo que más me apasiona.

Índice general

1. Introducción **8**
1.1. Planteamiento del problema 8
1.2. Objetivos 9
1.3. Antecedentes 9
1.4. Justificación 10
1.5. Distribución de la tesis 10

2. Marco teórico **11**
2.1. Aprendizaje automático 11
2.1.1. Terminología 12
2.1.2. Entrenamiento y Prueba del modelo de aprendizaje 13
2.1.3. Enfoques de aprendizaje automático 13
2.2. Aprendizaje supervisado 14
2.2.1. Análisis de componentes principales 15
2.2.2. Redes neuronales artificiales 16
2.3. Aprendizaje no supervisado 17
2.3.1. Algoritmos de agrupamiento 18
2.3.2. Algoritmos basados en centroides 19
2.3.3. Agrupamiento jerárquico 20
2.4. Aprendizaje profundo 21
2.4.1. Redes neuronales recurrentes 23
2.4.2. Redes LSTM 24
2.4.3. Redes DKM 25
2.5. Escala CES 26
2.6. Ciencia de datos 27

3. Estado del arte **29**
3.1. Aplicación de aprendizaje supervisado y semi supervisado 30
3.2. Trabajos con aplicaciones de Aprendizaje profundo 37
3.3. Trabajos con aplicaciones de Aprendizaje no supervisado 41

4. Propuesta de solución **45**
4.1. Obtención de datos 45
4.1.1. Enfoque de la encuesta 46
4.1.2. Aplicación de las encuestas con formularios 47
4.2. Limpieza y análisis del conjunto de datos 48
4.3. Programar y comparar algoritmos de aprendizaje 48
4.3.1. Pseudocódigo de promedios K 49
4.3.2. Pseudocódigo de AGNES 50
4.3.3. Arquitectura de red neuronal DKM 50
4.4. Análisis de algoritmo para modelo de aprendizaje 52
4.5. Coeficiente de silueta 52
4.6. Índice de Davies-Bouldin 53
4.7. Optimización del modelo y Discusión de resultados 54

5. Resultados experimentales **55**
5.1. Obtención de datos 55
5.1.1. Inicio de semestre 55
5.1.2. Intermedios de semestre 56
5.1.3. Finales de semestre 56
5.2. Análisis estadístico 57
5.2.1. Síntomas a inicios de semestre 57
5.2.2. Síntomas a intermedios de semestre 58
5.2.3. Síntomas a finales de semestre 59
5.3. Resultados de la implementación de los algoritmos de aprendizaje . . 61
5.4. Evaluación de los algoritmos 64
5.5. Discusión de resultados 64

6. Conclusiones **66**

A. Encuesta con Escala CES-D **75**

Índice de tablas

2.1. Conjunto de datos 12

3.1. Trabajos relacionados con aprendizaje supervisado 34
3.2. Trabajos relacionados con aprendizaje profundo 39
3.3. Trabajos relacionados con aprendizaje no supervisado 43

4.1. Características de los datos 46

5.1. Datos de inicios de semestre. 55
5.2. Datos de intermedio de semestre. 56
5.3. Datos de finales de semestre. 56
5.4. Resultados obtenidos 61
5.5. Resultados del coeficiente de silueta 64
5.6. Resultados de índice Davies-Bouldin 64

Índice de figuras

2.1. Diagrama de tipos de aprendizaje automático 13
2.2. Relación AI, ML y DL 14
2.3. Perceptrón 17
2.4. Red Neuronal Profunda 22
2.5. Valores de activación 22
2.6. Arquitectura red DKM 26

4.1. Formulario inicio de semestre 47
4.2. Formulario intermedios de semestre 47
4.3. Formulario finales de semestre 48
4.4. Arquitectura red DKM 50

5.1. Atributos generales a principio de semestre 57
5.2. Síntomas a principio de semestre 58
5.3. Atributos generales a intermedios de semestre 59
5.4. Síntomas a intermedio de semestre 59
5.5. Atributos generales a finales de semestre 60
5.6. Síntomas a finales de semestre 60
5.7. Resultados obtenidos inicios de semestre 61
5.8. Resultados obtenidos intermedios de semestre 62
5.9. Resultados obtenidos finales de semestre 62
5.10. Comparación por algoritmo 63
5.11. Comparación por caso 63

Capítulo 1

Introducción

El problema planteado y que será observado en este trabajo, los objetivos específicos y el general, los antecedentes del trabajo, la justificación y la distribución de la tesis serán presentados en esta sección.

1.1. Planteamiento del problema

La OMS define la depresión como un trastorno mental común que se caracteriza por tristeza persistente o pérdida de interés en las actividades, acompañada de síntomas como:

- Pérdida o aumento del apetito sin intentar hacer dieta.
- Dificultad para conciliar el sueño o dormir demasiado.
- Sentirse constantemente cansado o sin energía.
- Moverse o hablar demasiado rápido o demasiado lento.
- Sentirse indeciso o tener problemas para concentrarse.
- Sentirse inútil o culpable sin motivo.

La depresión es un trastorno grave que puede afectar negativamente la forma en que una persona piensa, siente y actúa. También puede causar una serie de problemas de salud física, como problemas cardíacos, accidentes cerebrovasculares y diabetes.

Según la Organización Mundial de la Salud (OMS), la prevalencia mundial de la depresión aumentó un 25 % en el primer año de la pandemia de COVID-19. Esto significa que millones de personas más en todo el mundo están experimentando síntomas de depresión que nunca antes [1]. De entre la población mundial, alrededor del 20 % de los habitantes del mundo que asisten a consultas médicas especializadas tienen este tipo de trastorno, sin embargo, para la mayoría de ellos no hay diagnósticos o tratamientos adecuados [2].

En este trabajo de tesis nos enfocaremos en una población de adultos jóvenes y en el trastorno mental de la depresión. En investigaciones realizadas para medir la salud mental de una población universitaria se ha mostrado que los trastornos

que más se presentan son los depresivos y ansiosos. Esto debido a distintos factores presentes a lo largo de la formación profesional universitaria [3].

Reconocer los factores que llevan a un caso de depresión permitiría dar un tratamiento profesional temprano para aquellos que sufran de depresión y ansiedad. Desde el punto de vista del aprendizaje automático esta tarea se puede ver como un problema de clasificación o de agrupación, sin embargo, es necesario analizar a fondo todas las características que posee para determinar a cuál de los dos pertenece.

Se busca estimar un diagnóstico de depresión usando la escala expuesta en [4] y separarlos por casos de depresión, en caso de que exista. Al no poseer un diagnóstico profesional se busca obtener una estimación de este con los valores de la escala, esto indica que el problema es de **aprendizaje no supervisado**. La separación por casos de depresión usando aprendizaje no supervisado, se traduce a un problema de **clustering o agrupación**, entonces, categorizar casos de depresión en adultos universitarios es un problema de agrupación con aprendizaje no supervisado. La metodología que se usará para solucionar este problema será la de un proyecto de ciencia de datos y se expone en la siguiente sección.

1.2. Objetivos

Objetivo general:

- Detectar los casos de depresión en una población estudiantil aplicando algoritmos de aprendizaje automático siguiendo la metodología de un proyecto de ciencia de datos.

Objetivos específicos:

- Diseñar una encuesta apoyada en investigaciones psicológicas para medir en escalas las características que llevan a la depresión.
- Aplicar la encuesta a una población de estudiantes.
- Analizar y procesar el conjunto de datos obtenido mediante el proceso de ingeniería de datos.
- Implementar distintos algoritmos de aprendizaje automático no supervisado.
- Evaluar los modelos y sus resultados experimentales usando los valores de las métricas tales como el coeficiente de silueta y el índice de Davies-Bouldin.

1.3. Antecedentes

Desde la pandemia de Covid-19, la organización mundial de la salud ha detectado un incremento de personas que padecen de depresión, esto debido a distintos factores que se presentaron durante esta época y que siguen presentes. Esta enfermedad se

puede presentar en los distintos rangos de edad de una población de tal manera que la podemos separar en subpoblaciones.

Enfocándonos a la población de adultos jóvenes, la depresión es una enfermedad que afecta a toda población estudiantil y es complicado establecer los síntomas generales que generan depresión a los miembros de una población [5]. Esto se debe a que no todos los miembros de la población comparten los mismos síntomas que los llevan a estar deprimidos ni comparten el mismo caso de depresión.

1.4. Justificación

La investigación que se realizará en esta tesis está dirigida a estudiar la aplicación de la ciencia de datos, así como el aprendizaje automático no supervisado, para reconocer los patrones que causan depresión en estudiantes universitarios. Esto permitirá estimar su evolución en tiempo real de dichos casos.

1.5. Distribución de la tesis

La distribución de esta tesis es la siguiente:

- **Marco teórico**: es la sección donde se expone la teoría y conceptos necesarios para entender los fundamentos de este trabajo de tesis. Se presentarán los temas de aprendizaje automático, su terminología, los enfoques de aprendizaje, aprendizaje profundo, ciencia de datos, entre otros.
- **Estado del arte**: en esta sección se expondrán trabajos relacionados y que sirvieron de inspiración para la metodología de solución para este trabajo de tesis.
- **Metodología de solución**: esta sección incluirá la metodología de ciencia de datos explicando los pasos y cómo fueron llevados a cabo para cumplir los objetivos propuestos.
- **Resultados**: esta sección se hace la presentación de los resultados obtenidos de la implementación de los pasos expuestos en la metodología de solución.
- **Conclusiones**: la sección final de esta tesis presentará las conclusiones de este trabajo de tesis así como el trabajo a futuro que se le busca dar a la investigación realizada.

Capítulo 2

Marco teórico

En esta sección se explicará el aprendizaje automático, aprendizaje no supervisado, el concepto de conjunto de datos, algoritmos de agrupación, el proceso de entrenamiento, las métricas del coeficiente de silueta e índice de Davies-Bouldin, así como la definición de ciencia de datos y cómo es el proceso de aplicación que tiene.

2.1. Aprendizaje automático

Desde la creación de las primeras computadoras la programación convencional ha sido una gran herramienta para resolver ciertos problemas, sin embargo, con el avance del tiempo se han presentado problemas que necesitan apoyarse de datos históricos para solucionarse, es aquí donde se introduce el aprendizaje automático, o ML, por sus siglas en inglés.

El aprendizaje automático se puede definir como una rama de la inteligencia artificial cuya filosofía es crear algoritmos y programarlos de tal manera que una computadora simule un "aprendizaje" similar al que tenemos los humanos, para lograrlo se aplican distintas ramas de las matemáticas a conjuntos de datos históricos [6].

En el pasado, implementar un sistema inteligente consistía en algoritmos formados mayoritariamente por un conjunto de reglas y condiciones que debían ser seguidas para resolver el problema. Muchos de estos algoritmos hacían un proceso de toma de decisiones usando comandos condicionales "if" y "else", populares en todos los lenguajes de programación. Esto se traduce en tiempo de ejecución exponencial y procesamiento costoso, en su lugar se propuso el diseño, análisis e implementación de algoritmos "inteligentes" capaces de resolver problemas con mayor eficiencia [7].

La tarea principal de todos los algoritmos de aprendizaje automático es generar una función matemática relacionada con los datos de entrada que proponga una solución al problema a modelar. No todos los problemas comparten una misma manera

de solucionarse, esto lleva a separar el tipo de aprendizaje que será aplicado, este puede ser supervisado o no supervisado dependiendo de la estructura del conjunto de datos y del tipo de problemas [8].

2.1.1. Terminología

Antes de presentar los tipos de aprendizaje es necesario definir el concepto de conjunto de datos. Los datos se definen como toda observación de algún fenómeno real que puede ser representada de mediante variables numéricas o textuales, algunos ejemplos pueden ser la edad de una persona, el valor de cambio de una moneda, el id de una cuenta en una red social, etc.

Cuando se necesita manejar una gran cantidad de datos se les suele almacenar en una estructura matemática (vectores, matrices, tensores, etc) de tal manera que puedan ser manipulados de una manera más sencilla [9]. En el aprendizaje automático los datos serán almacenados en una estructura a la cual conoceremos como **conjunto de datos** y será denotado con la letra X [10]. La arquitectura del conjunto X está formada por m filas (registros) y n columnas (características o atributos).

Tabla 2.1: Conjunto de datos

$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_n$
$x_{1,1}$	$x_{1,2}$	$\dots$	$x_{1,n}$
$x_{2,1}$	$x_{2,2}$	$\dots$	$x_{2,n}$
$\vdots$	$\vdots$		$\vdots$
$x_{m,1}$	$x_{m,2}$	$\dots$	$x_{m,n}$

En el cuadro 2.1 se puede observar que el conjunto de datos esta formado por dos partes, el **vector de atributos** $\hat{\mathbf{x}}$, donde se representan todas las columnas del conjunto de datos a las cuales se les llamara atributos de ahora en adelante, y la **matriz de datos** $\mathbf{X}$, donde se almacenan datos en un formato de filas y columnas.

A pesar de tener varios atributos en un conjunto de datos, los datos almacenados en cada atributo no poseen la misma naturaleza o tipo, i.e., algunos datos pueden ser de tipo númerico, entero o decimal, de tipo texto o cadena, categóricos binarios, nominales u ordinales, etc. Dependiendo del tipo de datos que se almacene en cada atributo

Cada atributo es un vector de datos que puede ser asignado como valor de entrada para el proceso de entrenamiento y nuevamente utilizado para el proceso prueba, en la siguientes subsección serán explicados estos procesos.

2.1.2. Entrenamiento y Prueba del modelo de aprendizaje

El objetivo del aprendizaje automático es crear un modelo, o función que permita resolver problemas de mayor complejidad, para esto toma se crea una matriz mxn de datos históricos pertenecientes al conjunto X y lo relaciona con un vector objetivo perteneciente al mismo conjunto. A este proceso se le conoce como **entrenamiento de un algoritmo de aprendizaje**.

Al momento de realizar de entrenamiento se debe separar conjunto de datos en dos partes, una que será utilizada para el entrenamiento del modelo de aprendizaje, que será llamado **conjunto de entrenamiento**, y el restante será utilizado para probar el modelo de aprendizaje, este es el **conjunto de prueba**. La división depende de cuantos datos queramos usar para cada parte, suelen ser 70 % y 30 % respectivamente.

2.1.3. Enfoques de aprendizaje automático

Al aplicar el aprendizaje automático a un problema en especifico debemos conocerlo por completo, principalmente el como se usarán los datos para dar solución a dicho problema. Entre los enfoques de aprendizaje puede haber un enfoque supervisado, donde damos valores de entrada y valores de salida al algoritmo para que encuentre la relación entre ellos [8].

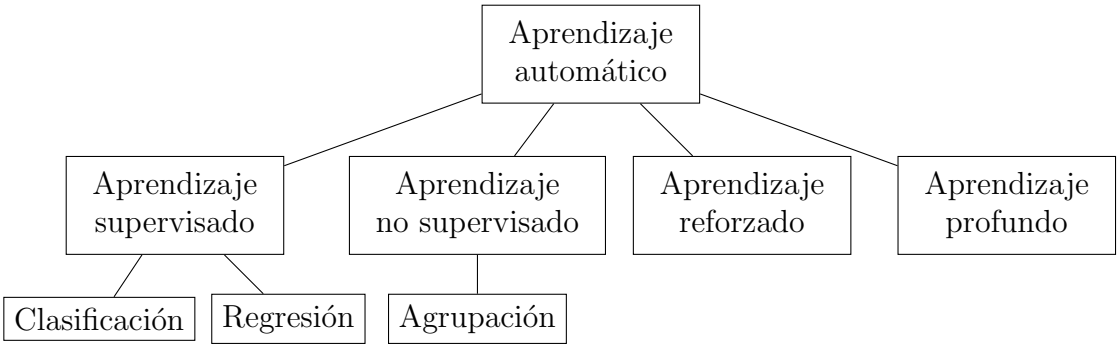

Figura 2.1: Diagrama de tipos de aprendizaje automático

Por el lado del aprendizaje no supervisado introducimos un conjunto de datos de entrada y el algoritmo de esta familia debe encontrar la relación que tienen entre si. Esto permite resolver problemas que conocemos como agrupación.

Los enfoques de aprendizaje anteriores pueden dar solución a una gran gama de problemas, sin embargo, existen problemas a los cuáles no pueden dar una solución y para lograr resolverlos se utiliza otro tipo de aprendizaje conocido como **aprendizaje profundo** (DL por sus siglas en inglés).

El aprendizaje automático y el aprendizaje profundo son dos subcampos de la inteligencia artificial (AI) que tienen mucho en común pero también algunas dife-

rencias clave. En el diagrama 2.1 podemos observar la relación entre estos dos.

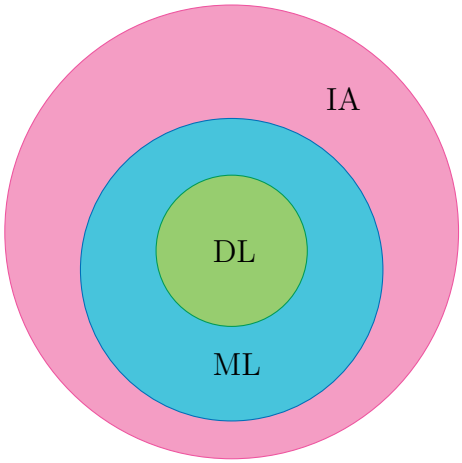

Figura 2.2: Relación AI, ML y DL

El aprendizaje profundo es una subcategoría del aprendizaje automático. Utiliza **redes neuronales artificiales** (NN por sus siglas en inglés) con muchas capas (de ahí el término "profundo") para modelar y entender estructuras complejas en los datos. A este tipo de redes neuronales las llamaremos **redes neuronales profundas** (DNN por sus siglas en inglés) pueden aprender características de nivel superior a partir de los datos de entrada, lo que las hace particularmente útiles para tareas como el reconocimiento de imágenes y el procesamiento del lenguaje natural.

Para resumir, el aprendizaje profundo es aprendizaje automático, pero no todo el aprendizaje automático es aprendizaje profundo. El aprendizaje profundo es simplemente una forma de implementar el aprendizaje automático que resulta ser muy efectiva para ciertos tipos de problemas.

En las siguientes secciones explicaremos más a fondo estos tipos de aprendizaje, los problemas a los que se pueden aplicar y algunos algoritmos de cada tipo.

2.2. Aprendizaje supervisado

Ya definidos los conceptos de conjunto de datos y columna objetivo podemos explicar el tipo de aprendizaje llamado supervisado y las tareas a las cuales es aplicado.

De aquí podemos definir los dos principales problemas a los que se aplica el aprendizaje supervisado: la clasificación y la regresión [11]. Para ambas tareas se debe declarar una columna objetivo y el modelo de aprendizaje relacionará todas las demás columnas, o solo unas seleccionadas, con esta mediante una función matemática que será generada en el proceso de entrenamiento del algoritmo de aprendizaje automático.

Los problemas de **clasificación** tienen el objetivo de asignar etiquetas a cada uno de los registros dentro del conjunto de prueba. Esto se hace con el modelo de aprendizaje obtenido del conjunto de entrenamiento que generan una ecuación del tipo exponencial para categorizar los datos en dos o más categorías, dependiendo del problema [12].

Los principales problemas a los que se dedica el aprendizaje supervisado son: clasificación y regresión. Entre los algoritmos que se pueden utilizar para crear modelos de aprendizaje supervisado que den soluciones a este tipo de problemas encontramos los siguientes:

- Regresión logística
- Máquina de vector de soporte
- Árboles de decisión
- K vecinos cercanos
- Redes neuronales artificiales: Perceptrón para clasificación binaria

Por el lado de la **regresión**, los algoritmos dedicados a este problema generan un modelo capaz de predecir valores futuros, por ejemplo, valores de ventas o valores de inversión [13].

Los problemas de regresión usan ecuaciones lineales que se apoyan de constantes obtenidas, por columna, gracias a ecuaciones estadísticas tales como la varianza y covarianza [14]. Algunos algoritmos que se utilizan para este tipo de problemas se presentan a continuación:

- Regresión lineal
- Regresión de Ridge
- Regresión polinomial

Estos problemas no son ejemplos de todos los que se puede solucionar con la aplicación de algoritmos de aprendizaje automático. En esta sección presentaremos algunos algoritmos de aprendizaje supervisado que serán utilizados en este trabajo de tesis.

2.2.1. Análisis de componentes principales

El análisis de componentes principales PCA es una técnica estadística que se utiliza para describir un conjunto de datos en términos de nuevas variables no correlacionadas, llamadas componentes principales. Estas variables capturan la mayor parte de la variabilidad original de los datos, y pueden ayudar a reducir la dimensionalidad y eliminar la redundancia de los datos.

PCA es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Supongamos que tenemos un conjunto de datos X de dimensión $n \times m$ y que-

remos construir a partir de él otro conjunto de datos X' de menor dimensión $n \times l$ con la menor pérdida de información útil posible utilizando para ello la matriz de covarianza.

Los datos para el análisis tienen que estar centrados a media 0 (restándoles la media de cada columna) y/o autoescalados (centrados a media 0 y dividiendo cada columna por su desviación estándar).

El PCA se basa en la descomposición en vectores propios de la matriz de covarianza, la cual se calcula con la ecuación 2.1:

$$A = cov(X) = \frac{X^T X}{(n-1)} \tag{2.1}$$

Una vez obtenida está matriz de covarianza, se crea la relación

$$Ap_a = \lambda_a p_a \tag{2.2}$$

Donde λ_a es el valor propio asociado al vector propio p_a. Por último, definimos el vector de vectores propios con la ecuación 2.3.

$$t_a = X_{p_a} = (\hat{x_1}, \hat{x_2}, \cdots, \hat{x_z}) \tag{2.3}$$

Cada una de estas z nuevas variables recibe el nombre de componente principal. En el método PCA, cada una de las componentes se corresponde con un vector propio, y el orden de componente se establece por orden decreciente de valores propios. Así pues, la primera componente es el vector propio con el valor propio asociado más alto. Hasta ahora se ha observado el poder que tiene el aprendizaje supervisado para resolver cierto tipo de problemas, más no puede resolver todo tipo de problemas, en la siguiente sección presentaremos el aprendizaje no supervisado y el tipo de problemas que pueden ser solucionados con ese aprendizaje.

2.2.2. Redes neuronales artificiales

Una red neuronal artificial (NN) es un nombre elegante para referirse a una función matemática. La inspiración de este algoritmo de aprendizaje son las neuronas biológicas en el cerebro mamífero y su estructura se debe separar por capas [15]:

- Capa de entrada o axón de entrada, por donde los datos accederán a la neurona.
- Sinapsis, encargadas de cargar los pesos.
- Dendritas, la conexión entre los valores de la sinapsis a la neurona.
- Cuerpo o neurona, donde recibe la suma ponderada y una función de activación.
- Cuello de axón, encargada de llevar el resultado final a una salida.

- Axón de salida, es la capa de salida que nos dará el resultado final.

La arquitectura más básica del aprendizaje profundo es un algoritmo de aprendizaje supervisado que se conoce cómo perceptron. En la figura 2.3 podemos observar esta arquitectura.

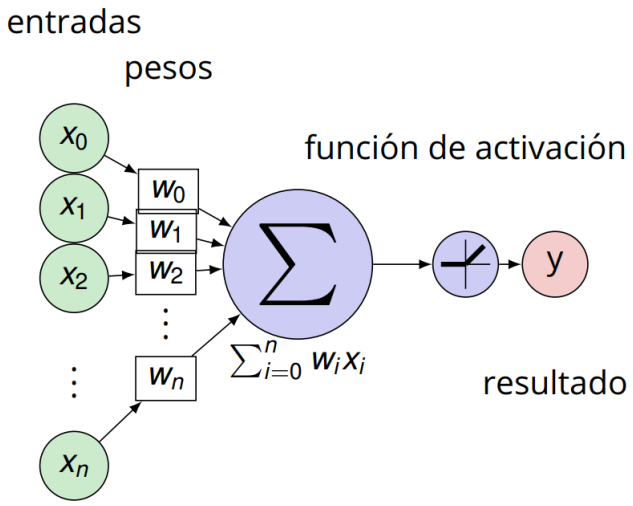

Figura 2.3: Perceptrón

Ya mostrada la arquitectura del perceptron podemos definir los conceptos de pesos, neurona y función de activación.

Los pesos son valores numéricos que empiezan siendo aleatorios y existe uno por cada valor de entrada, es decir, si tenemos n valores de entrada y k pesos. Al pasar por la sinapsis se hace el producto de los valores de entrada con los pesos, al pasar a la neurona se aplica una función de activación σ a la suma ponderada de estos productos, esto se muestra en la ecuación 2.4.

$$y = \sigma\left(\sum_{i=1}^{n} w_i x\right) \tag{2.4}$$

El valor obtenido de está función σ será el resultado, también conocido como valor de activación y. Esta arquitectura anteriormente presentada es la más básica de las redes neuronales artificiales y se le conoce como perceptron que es considerado como un algoritmo de aprendizaje automático [16].

2.3. Aprendizaje no supervisado

En los tipos de problema anteriores observamos que era necesario tener ejemplo con una clase dada, pero en el aprendizaje no supervisado se puede extraer información de ejemplos que no posean una clase definida, es decir, no es necesario tener un valor objetivo al cual se relacionarán las demás características, ahora se relacionarán una con otra para generar grupos [17].

Los grupos se forman a partir de datos que comparten características similares a los centroides, los cuales representan el centro de cada grupo y, por lo general, se encuentran separados entre sí. Esto permite definir el problema de agrupamiento, también conocido como clustering [18].

Existen distintos tipos de agrupamiento, a continuación, presentamos los que se usarán en este trabajo de tesis:

- Agrupamiento basado en centroides.
- Agrupamiento jerárquico.

Cada uno tiene su funcionamiento propio que lo diferencia y caracteriza, en este trabajo de tesis se aplicarán los siguientes agrupamiento.

- agrupamiento basado en centroides (promedios K)
- Agrupamiento jerárquico (AGNES)
- Agrupamiento con aprendizaje profundo (KNN)

A continuación, explicaremos cada uno de ellos.

2.3.1. Algoritmos de agrupamiento

Los algoritmos de agrupamiento son una técnica común en el análisis de datos estadísticos que se utiliza a menudo en diversos campos. Su utilidad es, principalmente, categorizar datos que no hayan sido previamente etiquetados.

El funcionamiento de los algoritmos de agrupación se basa en la búsqueda de grupos dentro de los datos, con un número de grupos que se representa a través de la letra K. Tras esto, procede a asignar cada punto de datos a uno de los K grupos, dependiendo de las características que se le hayan proporcionado.

Existen diferentes tipos de algoritmos de agrupamiento que manejan todo tipo de datos únicos:

1. **Basados en la densidad**: en él, los datos se agrupan según las áreas de altas concentraciones de datos que se rodean de áreas bajas de concentración de puntos de datos. A estos lugares los llama grupos.
2. **Algoritmos de distribución**: se basa en la distribución y funciona detectando un punto central de datos que, a medida que aumenta la distancia a este desde el centro, la posibilidad de que se haga parte del grupo, decrece.
3. **Algoritmos basados en centroides**: es el más común, rápido y eficiente. Su función es separar puntos de datos, asignando cada uno de ellos a la distancia al cuadrado del centroide.
4. **Algoritmos de agrupamiento jerárquico**: es un método de análisis de datos que busca construir una jerarquía de grupos. Se pueden utilizar dos enfoques: aglomerativo (ascendente) y divisivo (descendente).

En términos matemáticos, los algoritmos de agrupamiento se basan en la mini-

mización de una función objetivo, por ejemplo, la suma de las distancias cuadradas dentro del mismo clúster (para el caso de los algoritmos basados en centroides).

En el caso de los agrupamientos por densidad y distribución tenemos un funcionamiento similar, estos estudian la densidad de la dispersión en el espacio de los datos puntuales y cómo es posible agruparlos dependiendo de la densidad, por eso se les conoce como métodos basados en mallas. Su aplicación posee una gran ventaja sobre cualquier otro método, pues pueden construir grupos con formas no regulares, alguno de estos métodos son el DBSCAN y Sting [19].

2.3.2. Algoritmos basados en centroides

En el agrupamiento basado en centroides se generan grupos conformados por distintos datos que compartan alguna similitud, después se asigna un dato puntual de manera aleatoria como el centroide de su respectivo grupo, se comparan las distancias de todos los datos con cada uno de los centroides y cuando cumplan una condición establecida se dice que pertenecen a un grupo en concreto [20]. Uno de los más populares, y que se utilizará en este trabajo de tesis, es el algoritmo de agrupamiento por promedios K.

Como se mencionó anteriormente, los algoritmos de agrupamiento se aplican en conjuntos de datos no etiquetados, el método conocido por promedios K es capaz de separar los datos en K grupos de manera eficiente y rápida, sin necesidad de sobreponer grupos [7].

El procedimiento del agrupamiento por promedios K se puede expresar matemáticamente, sean $C_1, \cdots, C_k$ conjuntos cuyos elementos son todos los datos de contenidos en un grupo, se deben satisfacer dos propiedades.

1. $C_1 \cup C_2 \cup \ldots \cup C_k = 1, \ldots,$ n, es decir, cada dato pertenece a uno de los K grupos
2. $C_k \cap C_{k'} = \emptyset \quad \forall\, k \neq k'$. Esto significa que los grupos no deben sobreponerse entre sí y que una observación pertenece únicamente a un grupo.

Ya definidas las propiedades de los grupos, podemos definir la distancia interna de un grupo W y la distancia externa de los datos, que pueden ser definidas dependiendo del objetivo, en el caso de promedios K se utilizan distancias euclídeas o de Manhattan. El objetivo es encontrar los centroides μ_i que minimizan la función objetivo J [21]. Para este trabajo se utilizará el algoritmo de agrupamiento por centroides de promedios K cuya función objetivo se describe en la ecuación 2.5:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2.5}$$

donde:

- J es la función objetivo (también conocida como inercia).
- k es el número de clústeres.
- C_i es el conjunto de puntos de datos en el clúster i.
- x es un punto de datos.
- μ_i es el centroide del clúster i.
- $||x - \mu_i||^2$ es la distancia cuadrada entre el punto de datos x y el centroide μ_i.

2.3.3. Agrupamiento jerárquico

En el agrupamiento jerárquico, se construye un árbol de agrupación (dendrograma) para representar datos, donde cada grupo (o nodo) se vincula a dos o más grupos sucesores.

Los grupos están anidados y organizados como un árbol, lo que idealmente termina como un esquema de clasificación significativo. Existen dos tipos principales de agrupamiento jerárquico:

- Aglomerativo: Este es un enfoque de abajo hacía arriba, cada dato comienza en su propio grupo y los pares de grupos se fusionan a medida que se sube en la jerarquía.
- Divisivo: Este es un enfoque de arriba hacía abajo, al principio todos los datos pertenecen en un grupo único y se realizan divisiones a medida que se baja en la jerarquía.

El agrupamiento jerárquico se caracteriza por el uso de medidas de similitud $S(x,y)$ y de diferencia $D(x,y)$ entre dos objetos (datos) x y y, las cuales se generalizan para ser obtenidas entre dos grupos G y G', es decir, obtenemos las similitudes y diferencias $S(G,G')$ y $D(G,G')$ entre grupos para así diferenciar los datos que forman parte de los mismos [19]. Para esto se utiliza el siguiente proceso iterativo:

1. Se genera un grupo por cada dato puntual, si hay N y K grupos entonces $K=N$.
2. Se buscan los grupos pares G y G' que maximicen la similitud.
3. Se unen dichos grupos que forman el par anteriormente mencionado para crear un nuevo grupo.
4. Mantener un recuento de las uniones en un dendograma para comprobar la cantidad final de grupos por iteración.
5. Definir nuevos valores de similitud para los nuevos grupos creados.
6. Repetir el proceso hasta llegar a un valor umbral o un límite de grupos.

El agrupamiento jerárquico se utiliza en una variedad de campos, como la biología para mostrar la agrupación entre genes o muestras, pero puede representar cualquier tipo de datos agrupados. En este trabajo de tesis se implementará el algoritmo AGNES.

El algoritmo AGNES (AGglomerative NESting) es un método de agrupamiento jerárquico aglomerativo. Comienza tratando cada objeto de datos como un grupo individual. Luego, calcula la distancia entre todos los pares de grupos y fusiona los dos grupos más cercanos. Este proceso se repite hasta que todos los objetos de datos están en un solo grupo o hasta que se cumple una condición de terminación. La *distancia* entre los grupos puede definirse de varias maneras, como la distancia mínima, máxima, media entre los puntos de los grupos, o la distancia entre los centroides de los grupos. El resultado del algoritmo AGNES se puede visualizar utilizando un dendrograma, que es un árbol que muestra cómo los grupos se fusionan en cada paso.

2.4. Aprendizaje profundo

El aprendizaje profundo (DL) es una rama de la inteligencia artificial que se centra en el entrenamiento de algoritmos para que puedan aprender a realizar tareas más complejas. Su objetivo es el estudio y construcción de sistemas de cómputo capaces de *aprender* a partir de la experiencia. Estos sistemas deben ser entrenados a partir de ejemplos conocidos.

Se inspira ligeramente en algunos principios del funcionamiento del cerebro animal. Utiliza redes neuronales artificiales que imitan la manera en que el cerebro humano procesa la información para obtener conocimiento y crear predicciones.

A este tipo de aprendizaje automático se le llama profundo porque presenta una estructura jerárquica que extrae diferentes niveles de detalle de los datos en cuestión.

La arquitectura de redes neuronales artificiales donde se realiza un aprendizaje *profundo* es conocido como **red neuronal profunda** o de conexión completa (DNN), la cual posee un funcionamiento similar al perceptron, pero en este caso se utilizan capas formadas por una cantidad m_i de neuronas por capa oculta p_j y puede haber cuantas capas se desee, esto hace que este tipo de aprendizaje sea profundo.

En la figura 2.4 se presenta la arquitectura de una DNN así como las capas p_j, que son un conjunto de neuronas las cuales se relacionan una con otras mediante el par ordenado (p_{j-1}, p_j) de esta manera los pesos ahora deben tener una forma $w_{i,j}$ donde i es el índice que representa el número de la neurona sobre la capa y j hace referencia a la capa donde se encuentra el peso [16].

De la misma manera que el perceptrón los pesos suelen iniciar con un valor aleatorio, pero ahora se cuenta con un método de actualización y un valor de error que están relacionados entre sí, dependiendo de dicha similitud se afectaran los valores de activación que se generan entre las capas ocultas.

En la figura 2.5 se observa el comportamiento de los valores de activación y en la ecuación 2.5 se muestra su notación matemática.

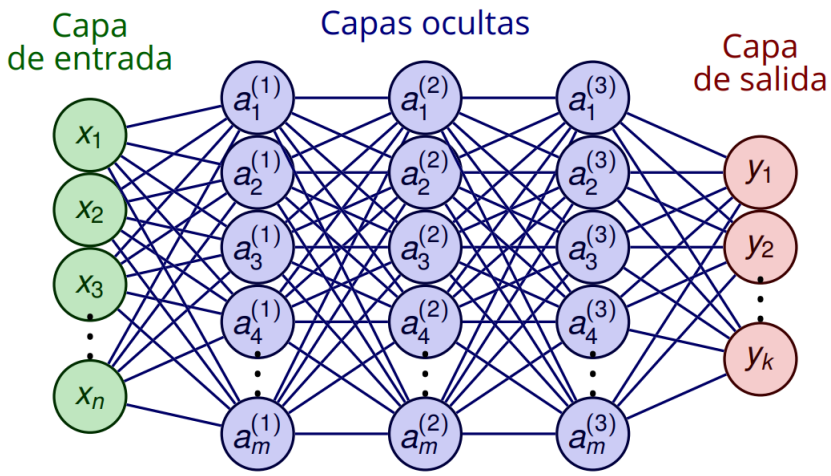

Figura 2.4: Red Neuronal Profunda

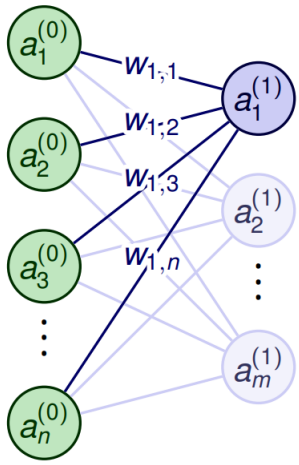

Figura 2.5: Valores de activación

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma \left[\begin{pmatrix} w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ w_{2,0} & w_{2,1} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \dots & w_{m,n} \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix} \right] \tag{2.6}$$

La ecuación 2.6 se puede expresar en su notación matricial debemos definir a la matriz de pesos $\mathbf{W^{(p)}}$, el vector de valores de activación $\hat{\mathbf{a}}^{(p)}$ y el vector de valores de sesgo $\hat{\mathbf{b}}^{(p)}$. Esto se observa en la ecuación2.7

$$\hat{\mathbf{a}}^{(1)} = \sigma \left(\mathbf{W}^{(0)} \hat{\mathbf{a}}^{(0)} + \hat{\mathbf{b}}^{(0)} \right) \tag{2.7}$$

El valor de error se mide de acuerdo un porcentaje de similitud entre la salida y

con un valor objetivo t, se calcula usando alguna ecuación tales como error medio cuadrático, raíz cuadrática media, suma de errores cuadrados, etc. Si el error supera un umbral entonces se activa el método de actualización.

Pero ¿cómo sabe un método de actualización que pesos debe modificar? Para esto se usa un algoritmo de reconocimiento conocido como retro propagación, presentado en la ecuación 2.8, cuyo objetivo es señalar los pesos que afectan los valores de activación que requieren ser actualizados de tal manera que den un mejor resultado y, sin afectar otros pesos.

$$\frac{\partial E}{\partial w_{j,i}^{(p-1)}} = \frac{\partial E}{\partial a_{j,i}^{(p)}} \frac{\partial a_{j,i}^{(p)}}{\partial z_{j,i}^{(p)}} \frac{\partial z_{j,i}^{(p)}}{\partial w_{j,i}^{(p-1)}} = \delta_i^{(p)} \hat{a}_i^{(p-1)} \tag{2.8}$$

Una vez encontrados los mejores pesos se aplica un optimizador de tal manera que los resultados obtenidos por las funciones de activación siempre sean los mejores posibles [22].

Antes mencionamos la flexibilidad que tienen las NN para resolver cualquier problema, esto se hace al agregar capas que tienen distintas funciones y características, algunas pueden ser: capas convolucionales, capas de memoria a corto-largo plazo, capas de promedio K, etc. Dependiendo del problema se genera una arquitectura distinta de una NN [23]. Un tipo de redes neuronales artificiales que se usarán en este trabajo serán las redes neuronales recurrentes y presentaremos su funcionamiento en la siguientes sección.

2.4.1. Redes neuronales recurrentes

Las redes neuronales recurrentes (RNN) son un tipo de red neuronal que se caracteriza por manejar datos cuyo formato es secuencial, es decir, tienen un orden único e inmutable. Los datos pueden ser de tipo numérico así como de texto. Se aplican para resolver tareas que requieren una inteligencia enfocada a la lingüística, es decir, la traducción de idiomas, el reconocimiento de voz y subtítulos de imágenes. Estos son problemas en los que se necesita puedan reconocer texto y entender su significado sobre una o varias oraciones [23].

Las RNN se distinguen por su "memoria", ya que obtienen información de entradas anteriores para influir en la entrada y salida actuales. Mientras que las redes neuronales profundas tradicionales asumen que las entradas y salida son independientes entre sí, la salida de las redes neuronales recurrentes depende de los elementos anteriores dentro de la secuencia.

Utilizan el algoritmo de retropropagación a través del tiempo (BPTT) para determinar los gradientes, que es ligeramente diferente de la retropropagación tradicional, ya que es específico para secuenciar los datos. Los principios del BPTT son

los mismos que los de la retropropagación tradicional, donde el modelo se entrena a sí mismo calculando errores desde la capa de salida hasta la capa de entrada [24].

Otra característica distintiva de las RNN es que comparten parámetros en cada capa de la red. Mientras que las redes de propagación hacia delante tienen diferentes pesos en cada nodo, las redes neuronales recurrentes comparten el mismo parámetro de peso en cada capa de la red.

2.4.2. Redes LSTM

Las redes LSTM (*Long-Short Term Memory Network*) son un tipo de redes neuronales recurrentes que pueden procesar datos secuenciales, es decir, datos donde el orden cronológico es importante. Las redes LSTM se diferencian de las redes RNN tradicionales en que tienen celdas de memoria que pueden almacenar y recuperar información a largo plazo. Esto les permite aprender de secuencias más largas y complejas, y evitar el problema del desvanecimiento o la explosión del gradiente que afecta a las redes RNN. Las redes LSTM se utilizan para tareas como el reconocimiento de voz, la traducción automática, el análisis de sentimientos, la generación de texto, etc [24].

Para entender mejor cómo funcionan las redes LSTM, se procederemos a explicar algunos conceptos clave:

- **Bloque funcional:** Es una unidad básica que realiza una operación sobre sus entradas y produce una salida. En las redes LSTM, los bloques funcionales son los **bloques de memoria**, que contienen una o más celdas de memoria. Cada bloque de memoria tiene un vector **memoria a corto plazo** (STM) y un vector **memoria a largo plazo** (LSTM). El STM almacena temporalmente la información relevante para el bloque, mientras que el LSTM almacena permanentemente la información importante para toda la secuencia.
- **Celda:** Es una unidad individual dentro de un bloque de memoria. Cada celda tiene un peso asociado al vector STM y al vector LSTM. El peso determina cómo se combina la información del STM con la del LSTM para generar la salida del bloque.
- **Capa oculta:** Es una capa intermedia entre la capa externa y la capa interna de una red neuronal. En las redes LSTM, la capa oculta contiene varios bloques funcionales conectados entre sí por pesos sinuales. Los pesos sinuales permiten el paso bidireccional entre los bloques funcionales, lo que facilita el aprendizaje paralelo y distribuido.
- **Capa interna:** Es una capa final donde se produce la predicción o el salto final. En las redes LSTM, la capa interna contiene uno o más bloques funcionales con pesos sinuales conectados a los bloques funcionales correspondientes en la

capa oculta[25].

El proceso general para entrenar una red LSTM es el siguiente:

- Se toma un conjunto de datos dividido en varias secuencias o títulos.
- Se extrae cada secuencia como entrada y se pasa por todos los bloques funcionales hasta llegar a la capa interna.
- Se calculan las salidas parciales para cada secuencia usando los pesos sinuales entre los bloques funcionales.
- Se calculan las pérdidas usando alguna función objetivo como el error cuadrático medio (MSE) o el error absoluto medio (MAE).
- Se actualizan los pesos usando algún método como el descenso del gradiente estocástico (SGD) o el descenso del gradiente conjugado (CG).
- Se repite este proceso hasta alcanzar un criterio de parada como un número máximo de iteraciones o un valor mínimo de pérdida[26].

2.4.3. Redes DKM

Las redes DKM (Differential K-means) son un tipo de redes neuronales artificiales que se utilizan para el aprendizaje automático y la inteligencia artificial. Las redes DKM se basan en el concepto de **redes de conocimiento**, que son redes que almacenan y procesan información en forma de conocimiento, es decir, en términos de hechos, reglas, principios y relaciones. Las redes DKM se diferencian de las redes neuronales convencionales en que tienen una estructura jerárquica y organizada, donde cada nodo representa un nivel de abstracción o generalización del conocimiento. Los nodos se conectan entre sí mediante enlaces que indican la dependencia o la relación entre ellos. Los nodos pueden contener tanto datos como reglas o hechos sobre los datos. Las redes DKM se utilizan para tareas como el reconocimiento de patrones, la clasificación, la inferencia, la generación, etc[27].

En la figura 2.6 podemos observar la arquitectura de una RNN con una capa DKM. Para entender mejor cómo funcionan las redes DKM, te voy a explicar algunos conceptos clave:

- **Nodo:** Es una unidad básica que almacena o procesa información. Cada nodo tiene un valor asociado al dato o al conocimiento que contiene. Los nodos pueden ser de diferentes tipos según su función o su nivel de abstracción. Por ejemplo, hay nodos simples que solo contienen datos numéricos o categóricos, nodos compuestos que contienen datos y reglas o hechos sobre ellos, nodos especiales que contienen reglas o hechos generales sobre los datos, etc.
- **Enlace:** Es una conexión entre dos nodos que indica la dependencia o la relación entre ellos. Los enlaces pueden ser de diferentes tipos según su naturaleza o su función. Por ejemplo, hay enlaces simples que solo indican una asociación

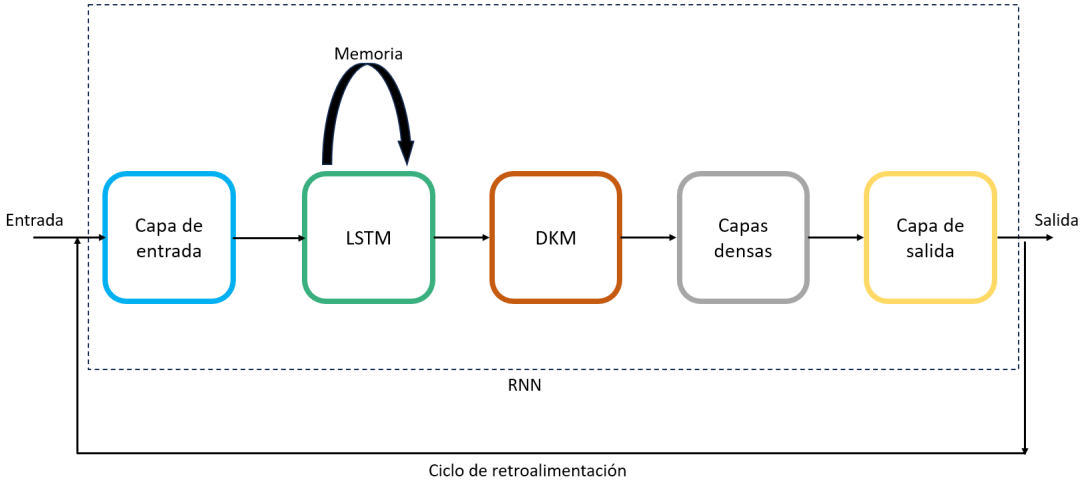

Figura 2.6: Arquitectura red DKM

entre dos valores numéricos o categóricos, enlaces compuestos que indican una asociación entre un valor numérico y una regla o hecho sobre él, enlaces especiales que indican una asociación entre una regla o hecho general y un valor numérico específico, etc.

- **Arquitectura:** Es la estructura global de la red DKM. La arquitectura define cómo están organizados los nodos y los enlaces dentro de la red. La arquitectura determina el nivel de abstracción y generalización del conocimiento almacenado y procesado por la red.
- **Aprendizaje:** Es el proceso por el cual la red DKM adquiere nuevos conocimientos a partir de los datos y las reglas existentes. El aprendizaje implica modificar los valores asociados a los nodos y a los enlaces mediante algún método como el descenso del gradiente estocástico (SGD) o el descenso del gradiente conjugado (CG). El aprendizaje permite mejorar el rendimiento y la precisión de la red.

2.5. Escala CES

La escala CES es una herramienta que se utiliza para evaluar los síntomas de depresión en la población general. Fue desarrollada por Laurie Radloff en 1977 y revisada por William Eaton y otros en 2004. La escala CES consta de 21 ítems que se responden con una escala de 0 a 3, donde 0 significa nunca y 3 significa todo el tiempo o la mayoría del tiempo. La suma de los puntos obtenidos indica el nivel de depresión, siendo menor a 5 puntos normal, igual o mayor a 5 puntos síntomas depresivos significativos, y mayor a 21 puntos depresión severa. La escala CES es una herramienta útil y económica para el tamizaje de la depresión, ya que no requiere personal especializado ni tiempo prolongado para su aplicación. Además,

es autoaplicable y fácil de entender. La escala CES se ha utilizado en diversos países y contextos, incluyendo México, donde ha demostrado tener una buena validez y fiabilidad.

La aplicación de la escala CES se vio modificada en México en el trabajo [4], donde agregaron más preguntas a la escala y evaluaron con una escala del 1 al 5 donde se mide presencia de sensaciones o situaciones en un periodo máximo de dos semanas. Para este trabajo de tesis se aplicará la escala expuesta en dicho trabajo con 15 preguntas expuestas en dicho trabajo.

2.6. Ciencia de datos

La ciencia de datos es el estudio de los datos con el fin de extraer información valiosa para las empresas, la sociedad o la ciencia. La ciencia de datos combina diversas técnicas, desde el procesamiento y la limpieza de los datos hasta la construcción de modelos predictivos, utilizando herramientas como la estadística, la programación, la inteligencia artificial y el aprendizaje automático. La ciencia de datos permite responder a preguntas como qué pasó, por qué pasó, qué pasará y qué se puede hacer con los resultados [28].

La ciencia de datos es importante porque ayuda a generar conocimiento a partir de los datos, que son cada vez más abundantes y complejos. Los datos provienen de diversas fuentes, como dispositivos, sistemas, redes sociales, imágenes, audio, vídeo, etc. Estos datos pueden contener información relevante para mejorar los procesos, los productos, los servicios, las decisiones o las políticas de las organizaciones. La ciencia de datos también puede contribuir al avance de la investigación científica, la innovación tecnológica, la educación, la salud, el medio ambiente y otros ámbitos de la sociedad [29].

La ciencia de datos es un campo multidisciplinario que requiere de diferentes habilidades y competencias. Algunas de las principales son [30]:

- Conocimiento del dominio: se refiere al entendimiento del contexto, los objetivos y los problemas que se quieren resolver con los datos. Es necesario tener una visión global y crítica de la realidad y de las necesidades de los usuarios o clientes.
- Conocimiento técnico: se refiere al manejo de las herramientas, los métodos y las tecnologías que se utilizan para trabajar con los datos. Incluye aspectos como la recopilación, el almacenamiento, el procesamiento, el análisis, la visualización y la comunicación de los datos. También implica el uso de lenguajes de programación, bases de datos, sistemas operativos, plataformas en la nube, etc.
- Conocimiento analítico: se refiere a la capacidad de aplicar técnicas estadísti-

cas, matemáticas, de inteligencia artificial y de aprendizaje automático para extraer información de los datos. Implica el diseño, la construcción, la evaluación y la optimización de modelos predictivos, descriptivos, prescriptivos o explicativos que permitan responder a las preguntas planteadas.

- Conocimiento comunicativo: se refiere a la habilidad de transmitir los resultados y las conclusiones de la ciencia de datos de forma clara, precisa y persuasiva. Implica el uso de lenguajes, formatos y medios adecuados para cada audiencia y situación. También implica el respeto de los principios éticos, legales y sociales relacionados con el uso de los datos.

Al realizar un proyecto de ciencia de datos se debe seguir ciertos pasos que no suelen formar parte de una metodología, pero son parte de todo proyecto que conlleve la aplicación de ciencia de datos. Los pasos son los siguientes:

1. El primer paso es definir claramente el problema que se está tratando de resolver. Esto implica comprender los objetivos del proyecto, las preguntas que se están tratando de responder y las métricas que se utilizarán para medir el éxito.
2. Una vez que se ha definido el problema, el siguiente paso es recopilar los datos necesarios para resolverlo. Esto puede implicar recopilar datos de una variedad de fuentes, como bases de datos, encuestas y experimentos.
3. Los datos recopilados a menudo deben limpiarse y prepararse antes de poder usarse para análisis. Esto puede implicar abordar valores faltantes, eliminar valores atípicos y transformar variables.
4. Una vez que los datos estén limpios y preparados, el siguiente paso es explorarlos para comprender sus características y patrones. Esto puede implicar crear visualizaciones de datos, calcular estadísticas descriptivas y realizar pruebas de hipótesis.
5. El siguiente paso es desarrollar un modelo que pueda usarse para resolver el problema definido en el paso 1. Esto puede implicar una variedad de técnicas de modelado, como regresión, clasificación y aprendizaje automático.
6. Una vez que se ha desarrollado un modelo, debe evaluarse para ver qué tan bien funciona. Esto implica usar el modelo para hacer predicciones en un conjunto de datos de prueba y comparar esas predicciones con los valores reales.
7. Si el modelo se evalúa con éxito, se puede implementar en un entorno de producción. Esto implica implementar el modelo para que pueda usarse para hacer predicciones en nuevos datos.

Capítulo 3

Estado del arte

En los últimos años se han realizado investigaciones enfocadas en la depresión presente en adultos jóvenes y adolescentes, este es un tema de investigación al cuál se le ha dado mucha atención y puede variar según la población de estudio.

En 2022 se realizó la investigación [31] donde se estudia el problema de la depresión en adolescentes. En este trabajo se muestra el bajo nivel de reconocimiento y apertura ante esta enfermedad, así como la relación que tiene con el alto nivel de depresión, esto debido a la estigmatización que tiene en la población. Los investigadores llevaron a cabo una recaudación de datos para realizar medidas estadísticas sobre los mismos donde querían destacar relaciones entre los valores del género, el promedio, año de estudio con un diagnóstico profesional de cada uno de los miembros de la población de estudio.

Al siguiente año, 2023, el trabajo [32] realiza una investigación sobre la depresión ahora enfocada a la población de profesores de universidad durante la tercera ola de Covid-19. De la misma manera que en el trabajo [31] se realizó una recaudación de datos mediante cuestionario a los profesores y de los datos obtenidos destacan las relaciones entre depresión con la carrera a la que imparten clase, cantidad de alumnos, género y estado familiar, donde destacaban más los miembros de la población con familia.

Los trabajos anteriores muestran un enfoque directo a la enfermedad de la depresión, pero también se pueden estudiar otras enfermedades. En el trabajo [33] realizado en 2023 se exploró el dominio de trastornos mentales tales cómo depresión, ansiedad y burnout, que es un tipo de agotamiento mental, en una población de estudiantes de medicina en Sudáfrica.

En el mismo año, en [34] se miden los niveles de depresión en poblaciones universitarias donde se observa la diferencia entre los niveles de depresión de los estudiantes universitarios que pertenecen a las carreras de ciencias naturales y música de tal manera que destaquen las diferencias en el comportamiento de sus respectivos miembros estudiantiles.

Estos trabajos son investigaciones psicológicas, pero es posible aplicar aprendizaje automático a trabajos para reconocer y medir depresión. En esta sección se presentarán investigaciones enfocadas a la aplicación del aprendizaje automático a la salud mental y se separan en aprendizaje supervisado en el estudio de la salud mental, clasificación de trastornos mentales con algoritmos de aprendizaje profundo y finalmente aprendizaje no supervisado para agrupamiento de casos de salud mental. Las investigaciones se presentarán en orden cronológico y se destacará la similitud que tengan con este trabajo de tesis.

3.1. Aplicación de aprendizaje supervisado y semi supervisado

El aprendizaje supervisado aplicado para la observación de salud mental en adultos jóvenes y adolescentes es un estudio que se encuentra en crecimiento, esto debido a los distintos resultados que han otorgado distintas investigaciones.

En trabajos realizados en 2020 se pueden observar estudios de casos de salud mental relacionados con la pandemia de COVID-19. En [35] notaron un aumento en problemas de salud mental tales como la depresión, el estrés y la ansiedad, y para predecir el crecimiento de estas aplicaron algoritmos de aprendizaje automático de clasificación, como árboles de decisión, bosques aleatorios, máquina de vectores de soporte, vecinos cercanos y estadística de Naïve Bayes para crear modelos de aprendizaje que permitan una predicción de dichas enfermedades y se les pueda dar acompañamiento adecuado a los pacientes que las sufran.

En el año 2021 encontramos una investigación enfocada a la depresión post parto en mujeres [36][31] donde se hace un estudio socio demográfico acompañado de aprendizaje estadístico para medir los niveles de depresión en mujeres después del parto. Utilizaron cuestionarios psicométricos que evaluaban distintas características de las pacientes con respecto a su salud mental y aplicaron algoritmos de aprendizaje tales como regresión de Ridge y Lasso, árboles aleatorios extremos y distribuidos, Naïve Bayes y modelos de ensamblado en pila. Dependiendo de la naturaleza de los datos, distintos algoritmos otorgaban los mejores resultados para una predicción, entre los más destacados están la regresión de Ridge y los algoritmos de ensamblado en pila.

En el año 2022 se realizó la investigación expuesta en [37] que muestra el impacto de las horas extra de trabajo y un ambiente de trabajo en la salud mental de los trabajadores. Muestra un análisis exploratorio de datos donde se relacionan las características generales de los trabajadores (genero, edad, estado civil, años de estudio, salud mental y salud física), con los valores de trabajo a sobre tiempo y el

ambiente laboral en el que se encuentran. Los modelos de aprendizaje los realizaron sobre las medidas estadísticas promedio, error estándar y tamaño de muestra usando ecuaciones especificas al trabajo de tal manera que las relaciones se mostrarán por cada una de las características del trabajador que hayan sido capturadas.

Otro trabajo presentado en 2022 [38] investiga la detección temprana de la depresión mediante un análisis de sentimientos usando algoritmos de aprendizaje automático. En esta investigación se usaron datos obtenidos de publicaciones de Twitter y se implementaron cuatro algoritmos de clasificación binaria que son el potenciado del gradiente extremo, bosque aleatorio, regresión logística y la máquina de vector de soporte. Para medir el mejor modelo de aprendizaje se utilizó la métrica de exactitud (accuracy) de todos los modelos que se obtiene usando los valores de los verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN), y usando la ecuación de exactitud. El mejor modelo de aprendizaje es aquel obtenido del entrenamiento del algoritmo de regresión logística con una exactitud del 96.3 %, mientras que los otros algoritmos tienen una exactitud cercana siendo el mejor siguiente el algoritmo de máquina de vectores de soporte.

Otra enfermedad que se ha estudiado es la ansiedad, en el trabajo [39] realizado en 2022. El objetivo de este trabajo era predecir ataques de ansiedad, para tal tarea usaron el conjunto de datos IMAGEN que está conformado por neuroimágenes de una población de adolescentes de 14 años y también algunos cuestionarios que evalúan la salud mental, las emociones y el consumo de sustancias nocivas. Al ser una tarea de categorización optaron por algoritmos de aprendizaje supervisado los cuales son regresión logística, máquina de vector de soporte y bosque aleatorio. La exactitud de estos modelos dependía de la cantidad de población con la cual se había entrenado el algoritmo de aprendizaje, la exactitud promedio se encuentra en el rango de 60 % a 90 %.

A finales de 2022 y principios de 2023 se realizó el trabajo [40] donde investigan los factores de mayor riesgo que detonan una depresión en adolescentes y proponen dos métodos; uno puramente teórico y la aplicación de técnicas de aprendizaje de máquinas. Este trabajo explora características únicas a la cultura de cada adolescente, su manera de ser criados, situación familiar, situación económica, ingresos familiares, estado de vivienda y seguridad. El algoritmo que implementaron en este trabajo fueron la máquina de vector de soporte, aunque también consideraron un bosque aleatorio, de esto obtuvieron una exactitud entre el 60 % a 80 % de predicción correcta.

En 2023 se realizaron investigaciones relacionadas con otras enfermedades y trastornos mentales. En [41] se realizó una investigación del espectro autista para la predicción de los posibles espectros que pueda tener un paciente con autismo. Para esta tarea utilizaron los algoritmos de máquina de vectores de soporte, bosque aleatorio,

Naïve Bayes y un perceptrón simple para clasificación binaria. Entre los modelos de aprendizaje obtenidos, el que poseía las mejores métricas de exactitud, precisión y puntaje F fue el obtenido del entrenamiento del algoritmo de bosques aleatorios. Otro estudio del autismo se observa en [42] donde usan algoritmos de aprendizaje supervisado para observar posibles espectros de autismo en bebes de 3 a 6 meses, donde obtuvieron métricas de Precisión: 70 % y Puntuación F1: 69 %.

El rango de edad no es importante en la depresión, en 2005-2018 se hizo una recaudación de datos para examinar la salud y nutrición en estados unidos. En esta misma se capturaron 2546 registros de veteranos de guerra. El trabajo [43] propone la aplicación de algoritmos de aprendizaje, automático y profundo para compararlos y observar las características que pueden ser factores de riesgo para la depresión en veteranos. Los algoritmos propuestos por parte del aprendizaje profundo fueron redes neuronales recurrentes y de aprendizaje automático fueron algoritmos de clasificación tales como potenciado del gradiente extremo, árboles de decisión, máquina de vector de soporte, K vecinos cercanos y bosque aleatorio. Los modelos obtenidos por los modelos de aprendizaje profundo obtuvieron una exactitud del 83 % siendo mayor a la obtenida por cualquier algoritmo de aprendizaje automático.

En el mismo año 2023 se publicaron trabajos donde se aplicaron algoritmos de aprendizaje automático encargados a medir la depresión. El trabajo [44] estudia cómo se relaciona la depresión con los hábitos alimenticios y la nutrición mediante algoritmos de aprendizaje automático. Propone la aplicación de los algoritmos de clasificación binaria y los seleccionados fueron regresión lineal, máquina de vectores de soporte, bosques aleatorios, arboles de decisión y un potenciado de gradiente extremo. Para el entrenamiento se realizó una validación cruzada con los datos que se separaron para esta etapa para eliminar un posible sobreajuste en el modelo. Para evaluar el desempeño de los modelos se utilizaron las métricas de exactitud, precisión, sensibilidad y puntaje F1, donde los mejores algoritmos fueron el potenciado de gradiente extremo y el bosque aleatorio, para ambos la exactitud tuvo un valor de 86.18 % y 84.76 % respectivamente.

Anteriormente se expuso un trabajo realizado en 2020 que está enfocado a la observación activa de la salud mental durante la cuarentena de la pandemia de Covid-19 [33]..

En el trabajo [45] se utilizan sensores portátiles y aprendizaje automático para detectar trastornos mentales en niños. Se analizaron datos de 84 participantes durante tareas de inducción del estado de ánimo. Se descubrieron agrupaciones latentes en los datos, que estaban más relacionadas con el género que con la edad. Los casos de estudio indicaron que el alto deterioro y los subtipos diagnósticos podrían explicar a los niños más distintivos conductualmente. Se necesitan más investigaciones para mejorar las características y los enfoques de modelado.

El trabajo expuesto en [46] propone un método para visualizar la estructura laminar de la corteza humana usando MRI y aprendizaje automático. Este método, que combina varias técnicas, permite explorar las variaciones en las capas corticales. Los resultados muestran que las capas corticales producen firmas distintas en la MRI y que el método puede distinguirlas automáticamente. Se observó una buena concordancia con la segmentación histológica y se destacó la importancia de T2 en la diferenciación cortical. El estudio sugiere que el método podría usarse en estudios in vivo.

La investigación [47] presenta un método para clasificar el estrés mental en tres categorías usando características neurofisiológicas y aprendizaje no supervisado. Se reclutaron cuatro participantes para un experimento, y se aplicó el algoritmo K-means a los datos recogidos para crear tres clusters representando niveles de estrés. Los resultados mostraron una buena consistencia de los clusters y permitieron identificar qué cluster corresponde a qué nivel de estrés.

Los autores del trabajo [48] proponen aplicar el análisis de agrupamiento difuso a la salud mental de los estudiantes universitarios. Presenta un método mejorado basado en el algoritmo de luciérnaga y realiza experimentos comparativos, demostrando que el algoritmo propuesto tiene un mejor rendimiento. El estudio también analiza los factores que afectan la salud mental de los estudiantes universitarios.

En el trabajo [49] los investigadores utilizan varios algoritmos de aprendizaje automático para identificar el estado de salud mental de un individuo. Se diseñó un cuestionario y se aplicaron técnicas de aprendizaje no supervisado para extraer etiquetas de grupos. Los autores sugieren direcciones para trabajos futuros.

El objetivo del estudio realizado en [50] es comparar métodos de aprendizaje supervisado y semi-supervisado para monitorear la salud mental, específicamente en el trastorno bipolar. Se mayor enfoque es en uso de datos acústicos de un paciente con trastorno bipolar recolectados a través de una aplicación móvil. Usan modelos clasificadores semi-supervisados como la propagación de etiquetas, la difusión de etiquetas, la máquina de soporte vectorial semi-supervisada y el clasificador de autoentrenamiento. Los algoritmos de aprendizaje semi-supervisado observados en el trabajo superaron a los supervisados en la predicción de episodios de trastorno bipolar.

La investigación realizada en [51] Se utilizó la metodología ágil SCRUM y herramientas tecnológicas como Python, SQL Server, Android Studio y Marvel para el diseño de prototipos. La aplicación utiliza machine learning para analizar datos de redes sociales y enviar notificaciones de alerta a personas de confianza si se detecta un comportamiento de riesgo. Este trabajo tiene como objetivo mejorar la eficiencia de los planes de salud mental en Perú, ofreciendo un servicio accesible y preventivo para la población.

En la tabla 3.1 se presentan los trabajos observados en esta sección, así como el año en el que se publicaron, su objetivo, los algoritmos de aprendizaje utilizados y la mejor métrica obtenida.

Tabla 3.1: Trabajos relacionados con aprendizaje supervisado

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[35]	2020	Observación de la salud mental y depresión durante la pandemia Covid-19	▪ Árboles de decisión ▪ Bosques aleatorios ▪ SVM ▪ Naïve Bayes	No especificado
[36]	2021	Predicción de la depresión postparto	▪ Regresión de Ridge ▪ Regresión de Lasso ▪ Árboles aleatorios ▪ Naïve Bayes ▪ Ensamblado en pila	Exactitud: 99.66 %
[38]	2022	Algoritmos de aprendizaje automático (ML) para detectar síntomas de depresión en datos textuales de redes sociales.	▪ Árboles de Decisión ▪ Máquina de Vectores de Soporte ▪ Redes Neuronales simples ▪ LDA ▪ CNN	▪ Precisión: 88 % ▪ Puntuación F: 61 %
[37]	2023	Relación de trabajo con horas extra y depresión	Análisis mediante robustez y bootstrapping	No aplica
[39]	2023	Clasificar depresión relacionada con los hábitos alimenticios y nutrición personal	▪ Regresión logística ▪ SVM ▪ Bosques aleatorios ▪ SVM	▪ Exactitud: 66.7 % ▪ Precisión: 100 % ▪ Sensibilidad: 75 % ▪ Puntuación F1: 89.36 %
[40]	2023	Predecir y clasificar depresión mediante el análisis de datos	▪ SVM ▪ Bosque aleatorio	No especificado

Continua en la siguiente página

Tabla 3.1 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[41]	2023	Investigación del espectro autista para la predicción de los posibles espectros que pueda tener un paciente con autismo	▪ Naïve Bayes ▪ K vecinos cercanos ▪ Árboles de decisión	No especificado
[43]	2023	Detección de depresión en veteranos de Estados Unidos	▪ Potenciado del gradiente extremo ▪ Árboles de decisión ▪ SVM ▪ K vecinos cercanos ▪ Bosque aleatorio	No especificado
[44]	2023	Nutrición y salud relacionada con depresión	▪ Gradiente potenciado ▪ Bosque aleatorio	No especificado
[33]	2023	Determinar la prevalencia y los factores asociados con la depresión, ansiedad y agotamiento entre los estudiantes de medicina de la Universidad de Namibia	Es utilizaron instrumentos de captura de datos especializados para la depresión y otras enfermedades mentales.	No aplica.
[52]	2023	Observación de la salud mental durante la pandemia de Covid-19	▪ K vecinos cercanos ▪ Árboles de decisión ▪ Naïve Bayes	No especificado
[45]	2023	Observación temprana de la salud mental de infantes con aprendizaje automático	▪ Naïve Bayes ▪ K vecinos cercanos ▪ Árboles de decisión	Exactitud: 92.32 %

Continua en la siguiente página

Tabla 3.1 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[46]	2023	Desarrollar una técnica no invasiva para mapear la estructura laminar de la corteza humana utilizando MRI multidimensional y aprendizaje automático no supervisado.	Agrupamiento por Promedios K	Índice de Rand: 0.98, 0.82, 0.91
[47]	2022	Analizar la salud mental de los estudiantes universitarios utilizando un algoritmo de agrupamiento difuso mejorado basado en el algoritmo de luciérnagas.	Agrupamiento por Promedios K	Coeficiente de Silueta: 76 %
[48]	2021	Estudio de salud mental de estudiantes universitarios con aprendizaje difuso	Método de Agrupamiento Difuso por Promedios C	Exactitud en el rango: 45 % a 93 %
[49]	2018	Identificar el estado de salud mental de individuos utilizando algoritmos de aprendizaje automático para predecir enfermedades mentales en diferentes grupos de población	▪ Máquina de Vectores de Soporte ▪ Naïve Bayes ▪ Regresión Logística ▪ K vecinos cercanos ▪ Árboles de decisión	Exactitud: 90 %

Continua en la siguiente página

Tabla 3.1 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[50]	2023	Comparar métodos de aprendizaje supervisado y semi-supervisado para monitorear la salud mental, específicamente en pacientes con trastorno bipolar, utilizando características acústicas del habla.	▪ Árboles de decisión ▪ Naïve Bayes ▪ K vecinos cercanos ▪ Máquinas de Vectores de Soporte ▪ Propagación de etiquetas ▪ S3VM ▪ Difusión de etiquetas (LS) ▪ Clasificador de Autoentretenimiento (STC)	▪ Exactitud: 96 % ▪ Precisión: 100 % ▪ Sensibilidad: 96 % ▪ Puntuación F1: 98 %
[51]	2022	Desarrollar una aplicación móvil para la detección temprana de problemas de salud mental en Perú, utilizando inteligencia artificial y análisis de sentimientos.	Máquina de Vectores de Soporte	▪ Precisión: 95 % ▪ Sensibilidad: 92 % ▪ Puntuación F1: 92 %

3.2. Trabajos con aplicaciones de Aprendizaje profundo

La aplicación de algoritmos aprendizaje automático ha sido explorada en los trabajos presentados en la sección anterior, a continuación, se presentarán trabajos donde se aplican algoritmos de aprendizaje profundo al reconocimiento y clasificación de trastornos mentales.

En 2021 fue publicado el trabajo [53] donde usan un conjunto de datos etiquetado que contiene datos de texto provenientes de publicaciones en la red social Twitter, la fecha de publicación y el nivel de depresión que se diagnosticó a cada publicación. Los modelos de aprendizaje profundo fueron obtenidos de entrenar redes cuyas arquitecturas están formadas por una red de memoria a largo-corto plazo (LSTM) y una red de unidad recurrente cerrada (GRU), ambas pueden ser uni o bidirecciona-

les. De entre los modelos el mejor fue obtenido por la red neuronal de arquitectura LSTM bidireccional con una exactitud de 81.32 %. Este porcentaje es destacable debido al tipo de datos que se manejan y el tipo de problema que pertenece al procesamiento de lenguaje natural.

Otro trabajo que usa datos de tipo texto es [54] e investiga la implementación de arquitecturas con redes de memoria a largo-corto plazo (LSTM) y redes neuronales recurrentes (RNN) para detectar la depresión en fragmentos de texto obtenidos de publicaciones de la red social Twitter disponibles en la página Kaggle. El algoritmo propuesto en este trabajo que obtuvo la mejor exactitud y que pertenece al aprendizaje profundo fue la red LTSM con 99.66 % en la décima época.

Un trabajo publicado en 2022 [55] busca detectar la depresión en adultos mayores mediante el uso de datos almacenados en audio. Este modelo usa biomarcadores para reconocer los niveles de depresión mediante las frecuencias que emiten las cuerdas vocales, que son transformadas en datos numéricos y sirven de entrada para la red neuronal que está formada por una capa convolucional, una capa de agrupación máxima seguida de una red neuronal recurrente (RNN) y para finalizar una red de conexión completa. Esta red recibe el nombre DepAudioNet. Este modelo obtuvo una exactitud máxima de 66.07 % con una sensibilidad del 75 % y una especificidad de 89.36 %.

La aplicación de redes neuronales cuya arquitectura está formada por redes de memoria a largo-corto plazo (LTSM), redes neuronales convolucionales (CNN) y redes de unidad recurrente cerrada (GRU) han ido en aumento, a dicha arquitectura se le conoce como modelo SSCL. En el trabajo [56] proponen agregar una red neuronal recurrente a la arquitectura de un modelo SSCL para procesar datos en formato de cadena, mismos que provienen de publicaciones realizadas en la red social Twitter. El objetivo de aplicar este modelo es detectar la ansiedad de manera similar que el modelo MDHAN, propuesto en [51], detecta la depresión. Para el modelo propuesto en este trabajo la arquitectura que obtuvo la mejor exactitud fue 96.7 % y está formada por redes CNN+LSTM+GRU.

En el trabajo [52], publicado en 2023, realizan un modelo de aprendizaje profundo para detectar depresión en la pandemia de covid-19 mediante el entrenamiento de distintas redes neuronales cuyas arquitecturas están formadas por la combinación de una red neuronal convolucional (CNN) una red de memoria a largo-corto plazo (LSTM) uni o bidireccional o también una red de unidad recurrente cerrada (GRU). El modelo con la mayor exactitud, de valor 97.4 %, fue obtenido por el entrenamiento de una red con arquitectura combinada de una CNN, una LSTM y una GRU, y también fue apoyada por atención humana para guardar los mejores pesos.

En 2023 se realizó el trabajo [57] donde se obtuvo un modelo de aprendizaje profundo para apoyar al diagnóstico de depresión en distintos pacientes. Para esto

usaron el conjunto de datos DAIC-WOZ que consiste en entrevistas clínicas con un total de 189 registros, el objetivo es implementar una red de memoria de largo-corto plazo bidireccional para crear un modelo de aprendizaje profundo capaz de reconocer casos de depresión mediante datos de tipo texto.

La investigación [58] realizada en 2023 expone la comparación del rendimiento que tienen los algoritmos de aprendizaje profundo contra los algoritmos de aprendizaje automáticos dedicados a la tarea de detectar enfermedades y trastornos mentales, casos de depresión y ansiedad, la viabilidad de dar diagnóstico y el apoyo necesario mediante un futuro tratamiento. Este trabajo muestra las desventajas que posee el aprendizaje automático para el reconocimiento de enfermedades tales como Alzheimer, pero con una fuerte ventaja en cuanto a la detección de los casos de distintos trastornos.

Un trabajo que propone el diagnostico psicológico apoyado por el aprendizaje automático fue realizado en 2023 [59], presenta la implementación de los algoritmos de aprendizaje automático y aprendizaje profundo para generar modelos basados en distintos enfoques, tales como tratamientos psicológicos y tratamientos farmacológicos. Esto les permitió reconocer nuevos factores que pueden llevar a un caso de depresión, así como también los pasos que llevan a un mejor tratamiento, sea cual sea de los anteriormente mencionados.

En el trabajo [60] Se exploran técnicas de aprendizaje profundo, especialmente redes neuronales convolucionales (CNN), para el análisis de emociones y detección de salud mental en estudiantes. Los modelos implementados pueden procesar entradas multimodales (texto, audio, visual) para interpretar mejor los estados emocionales y el bienestar mental de los estudiantes. Los hallazgos obtenidos contribuyen al desarrollo de sistemas inteligentes que proporcionan apoyo personalizado y oportuno a los estudiantes, mejorando su bienestar mental y éxito académico.

En la tabla 3.2 presentaremos los trabajos expuestos en esta sección, su año de publicación, el objetivo que tenían, los algoritmos de aprendizaje profundo utilizados y la mejor métrica de exactitud que obtuvieron.

Tabla 3.2: Trabajos relacionados con aprendizaje profundo

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[53]	2021	Detectar depresión usando datos de redes sociales	▪ LSTM ▪ GRU	Exactitud: 91.32 %

Continua en la siguiente página

Tabla 3.2 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[54]	2022	Detección de depresión mediante análisis de texto	▪ CNN ▪ LSTM ▪ GRU	Exactitud: 99.66 %
[55]	2022	Depresión en adultos mayores mediante el uso de datos almacenados en audio	▪ RNN ▪ DAN	▪ Exactitud: 66.7 % ▪ Sensibilidad: 75 % ▪ Esp. 89 %
[56]	2023	Detectar depresión mediante variables en formato cadena	▪ CNN ▪ LSTM ▪ GRU	▪ ROC: 86 % ▪ Sensibilidad: 75 % ▪ Esp. 89 %
[42]	2023	Investigar el potencial de las grabaciones de ECG como biomarcadores para predecir la probabilidad de autismo en bebés de 3 a 6 meses utilizando algoritmos de aprendizaje automático	▪ *K* Vecinos Cercanos ▪ Bosque Aleatorio ▪ Ada Boost ▪ Árboles de Decisión ▪ Gradiente Potenciado ▪ Perceptrón Multícapa	▪ Precisión: 70 % ▪ Puntuación F1: 69 %
[58]	2023	Apoyo para el diagnóstico de depresión con modelos de aprendizaje profundo	No especificado	No especificado

Continua en la siguiente página

Tabla 3.2 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[57]	2023	Comparar modelos de aprendizaje automático con aprendizaje profundo para detectar enfermedades mentales	No especificado	No especificado
[59]	2023	Implementación de los algoritmos de aprendizaje automático y aprendizaje profundo para generar modelos enfocados en tratamientos	No especificado	No especificado
[60]	2023	El estudio explora el uso de análisis de emociones y detección de salud mental en estudiantes mediante técnicas de aprendizaje profundo, específicamente redes neuronales convolucionales (CNN)	Redes Neuronales Convolucionales	Puntaje F: 83 %

3.3. Trabajos con aplicaciones de Aprendizaje no supervisado

En todos los trabajos mencionados anteriormente se ha presentado la capacidad que poseen los algoritmos de aprendizaje supervisado y profundo para separar un conjunto de datos usando las características como principal componente, pero también se pueden aplicar algoritmos de aprendizaje no supervisado para la agrupación de datos similar a una clasificación binaria, es decir, podemos resolver un problema de aprendizaje supervisado, sin la necesidad de implementar algoritmos pertenecientes a este tipo de aprendizaje.

Un trabajo realizado en 2019 [61] expone cómo se realiza un agrupamiento mediante el aprendizaje profundo aplicando análisis de sentimientos. Como se menciona en la sección 2.7.5 este tipo de aprendizaje usa un método por capas, esto significa que el procesamiento de la información se hace por partes y, hasta terminar el recorrido sobre la red neuronal, se pone en conjunto esta información. Para el entrenamiento del modelo de aprendizaje profundo se busca extraer características de los datos tales como el estado civil, problemas de enfermedad, educación, situación familiar, entre otras. Algunos de los textos estaban escritos en distintos lenguajes, por lo que fue necesario un proceso de traducción en la limpieza y análisis de datos. El modelo entrenado obtuvo una exactitud del 98 % medida mediante una matriz de confusión.

En 2021 se realizó un trabajo [62] donde se presenta la aplicación de una red neuronal de memoria a corto-largo plazo (LSTM) basada en una red neuronal recurrente (RNN) aplicada al reconocimiento de depresión en conjuntos de datos grandes formados por datos de tipo texto. El conjunto de datos fue obtenido mediante la recaudación de publicaciones de la red social Facebook y se usa para entrenar y probar la red neuronal propuesta, el objetivo de entrenarla es que fuera capaz de separar datos en dos diferentes grupos: no deprimidos y deprimidos. La exactitud final después de un entrenamiento de cien épocas fue de 84.20 %.

En 2023 el trabajo [63] se realiza una investigación que propone reconocer los factores que llevan a un desorden de depresión mayor (MDD) aplicando algoritmos de aprendizaje automático y aprendizaje profundo. El conjunto de datos está formado por registros capturados de la actividad en cada hemisferio del cerebro, de tal manera que midieron distintas características de un cerebro deprimido. En este trabajo midieron la exactitud mediante una matriz de confusión y para el modelo de aprendizaje profundo propuesto obtuvo una precisión de 97.66 % y 99.13 %.

El trabajo [64] realizado en 2023 presenta una investigación donde se aplican algoritmos de aprendizaje para encontrar la relación entre la depresión, el consumo de tabaco y la obesidad con enfermedades del corazón en jóvenes adultos.

Para el entrenamiento del modelo de aprendizaje utilizaron un conjunto de datos sobre la nutrición y salud a nivel nacional de Korea (KNHANES) y realizaron análisis de una variable y multivariable con redes neuronales donde encontraron que el mayor riesgo de presentar enfermedades que resulten en casos de mortalidad en los niveles altos de consumo de tabaco, aunque no haya registro de obesidad mórbida en algunos miembros de esa población. La conclusión de este trabajo fue que en niveles medios y altos de consumo de tabaco hay mayor presencia de adultos jóvenes.

En [65] se identifican tres subgrupos de estudiantes universitarios que usan una aplicación de salud mental durante la pandemia de COVID-19. Los hallazgos sugieren que estos subgrupos son únicos y tienen diferentes objetivos de atención de

salud mental. El estudio enfatiza la necesidad de personalizar las aplicaciones de salud mental para mejorar el compromiso del usuario.

Los autores del trabajo [66] revisa los patrones de agrupamiento de la dieta, actividad física y comportamiento sedentario en jóvenes, y su impacto en la salud. Se identificaron 172 clusters, clasificados como saludables, no saludables y mixtos. Los clusters no saludables, asociados con peores resultados de salud, fueron prevalentes en familias de bajo nivel socioeconómico.

El trabajo [67] propone un algoritmo mejorado de K-means para analizar la educación de la salud mental de los estudiantes universitarios. El algoritmo mejora la selección de los centroides iniciales y determina el número óptimo de grupos. Se aplica a varios conjuntos de datos para probar su efectividad y precisión. El trabajo también explora la teoría de la autodeterminación y el diseño de la intervención.

La investigación realizada en el trabajo [68] presenta un método para identificar subtipos clínicos de la enfermedad de Alzheimer usando registros electrónicos de salud. El método descubre cinco clusters con diferentes perfiles clínicos y demuestra que el aprendizaje no supervisado puede ser utilizado para identificar subtipos de condiciones heterogéneas.

A continuación, se mostrará la tabla 3.3 con los trabajos expuestos en esta sección, año de publicación, el objetivo, los algoritmos de aprendizaje profundo utilizados y métrica de exactitud que obtuvieron.

Tabla 3.3: Trabajos relacionados con aprendizaje no supervisado

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[61]	2019	Agrupamiento mediante aprendizaje profundo con analisis a datos de sentimientos	▪ LSTM ▪ RNN ▪ DKM	No especificado
[62]	2021	Reconocimiento de depresión en grandes conjuntos de datos	▪ LSTM ▪ RNN	Exactitud: 99.66 %
[63]	2023	Detección de MDD	▪ Potenciado del gradiente extremo ▪ Bosque aleatorio ▪ Regresión logística ▪ SVM	No especificado

Continua en la siguiente página

Tabla 3.3 – *Continua de la página anterior*

	Año	**Objetivo**	**Algoritmos**	**Puntuación**
[64]	2023	Depresión relacionada con nutrición y salud en Korea	Análisis de una variable y multivarialbe con redes neuronales	No especificado
[65]	2023	Estudio de la depresión durante pandemia de COVID-19 mediante una aplicación y algoritmos de aprendizaje no supervisado.	▪ Aprupamiento por promedios K ▪ SDA ▪ CART	Exactitud: 88 %
[66]	2023	Patrones de salud mental relacionados con actividad física y sedentarismo	Agrupamiento por promedios K	No especificado
[67]	2022	Algoritmo de agrupamiento por promedios K mejorado para analizar la educación en salud mental de los estudiantes universitarios	Agrupamiento por promedios K	Comparación de grupos y tiempo de entrenamiento mejorados en algoritmo de promedios K mejorado
[68]	2020	Identificar y caracterizar subtipos clínicos de la enfermedad de Alzheimer utilizando registros electrónicos de salud (EHR) del Reino Unido mediante enfoques de aprendizaje no supervisado.	▪ Aprupamiento por promedios K ▪ MCA	Se mide la progresión en cada paciente

Capítulo 4

Propuesta de solución

Este documento seguirá los pasos de una **metodología de proyecto de ciencia de datos** para resolver el problema planteado, dichos pasos son los siguientes[6]:

1. Obtención de datos.
2. Limpieza y análisis del conjunto de datos.
3. Programar y comparar algoritmos de aprendizaje.
4. Análisis de algoritmo para modelo de aprendizaje.
5. Optimización del modelo de aprendizaje seleccionado.
6. Discusión de resultados y conclusión del modelo.

En esta sección explicaremos los pasos y qué acciones se llevarán a cabo en cada uno.

4.1. Obtención de datos

El entrenamiento del modelo se realiza al introducir un conjunto de datos a los algoritmos de aprendizaje y para obtenerlos se pueden usar distintos métodos tales como minería de datos, captura de imágenes, aplicación de encuestas, etc.

Para la obtención de datos necesarios para este trabajo de tesis se realizaron encuestas mediante formularios de *Google Forms*, los cuales incluyen las preguntas expuestas en el apéndice A.

Dichos formularios se aplicaron en tres etapas del semestre de otoño 2023 en la Facultad de Ciencia de la Computación de la Benemérita Universidad Autónoma de Puebla. Las etapas de aplicación de las encuestas fueron las siguientes:

- Inicio de semestre
- Intermedios de semestre
- Finales de semestre

La población mínima requerida para una correcta representación del fenómeno de estudio propuesto en este trabajo es de **210** miembros de la población. Sin embargo,

en el caso de obtener mayor cantidad de respuestas en alguna de las aplicaciones de la encuesta, se conservarán los datos.

4.1.1. Enfoque de la encuesta

La encuesta A se diseñó para capturar datos de distintas características de los miembros de la población, las cuales separaremos en dos, **Información general** y **Síntomas de depresión**. En la tabla 4.1 se presentan las características del conjuntos de datos y su división.

Tabla 4.1: Características de los datos

Tipo de dato	Definición	Atributos
Información general	Son todos los datos que describen la información personal del miembro de la población.	▪ Edad ▪ Genero ▪ Promedio ▪ Año de ingreso ▪ Localidad ▪ Estado civil ▪ Materias atrasadas ▪ Horas de estudio
Síntomas de depresión	Son los síntomas medidos mediante la escala CES, estos se expresan en una escala de días: ▪ Escasamente (0 a 1 días) ▪ Algo (1 a 2 días) ▪ Ocasionalmente (3 a 4 días) ▪ Mayormente (5 a 7 días) ▪ Diariamente (10 a 14 días)	▪ Entumecimiento ▪ Tristeza ▪ Insomnio ▪ Mala persona ▪ Perdida de interés ▪ Horas de sueño ▪ Movimientos lentos ▪ Disgusto personal ▪ Perdida de peso ▪ Temeroso ▪ Fracaso ▪ Poco amigable ▪ Menos habla ▪ Poco placer ▪ Fácil estrés ▪ Concentración ▪ Soledad ▪ Molestia repentina

4.1.2. Aplicación de las encuestas con formularios

En la figura 4.1 se puede observar el formulario que se realizó para la obtención de datos a inicios de semestre.

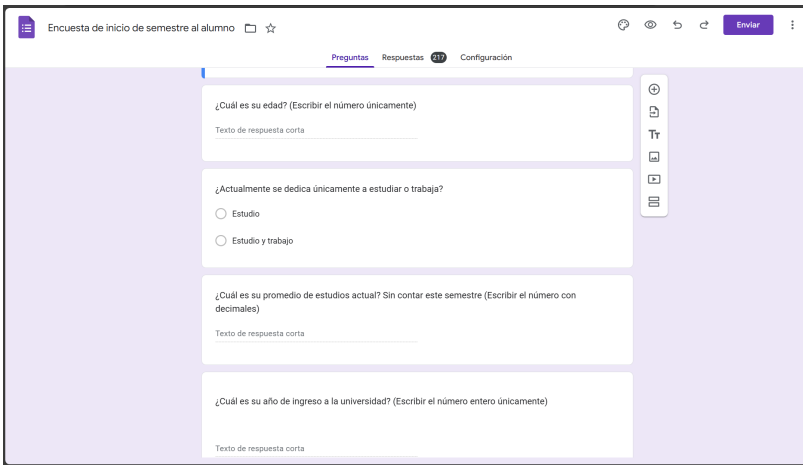

Figura 4.1: Formulario inicio de semestre

De esta encuesta se obtuvo un total de 217 registros, de los cuales uno no estaba completo, por lo mismo se le ignorará. En esta encuesta se permitió a los estudiantes a contestar con mayor libertad a las preguntas, la mayoría eran de respuesta abierta.

La razón principal para admitir este tipo de respuestas fue para generar rangos de los posibles valores en los que puedan estar los datos, principalmente los datos generales.

En la figura 4.2 se presenta el formulario que se aplico para la encuesta de intermedios de semestre.

Figura 4.2: Formulario intermedios de semestre

En esta encuesta obtuvimos un total de 228, más que en el caso anterior y se

sigue respetando la población mínima. Para la encuesta aplicada a intermedios de semestre se cambio el estilo de varias respuestas, ahora se implementó un estilo desplegable de respuesta donde se incluían varias opciones para contestar.

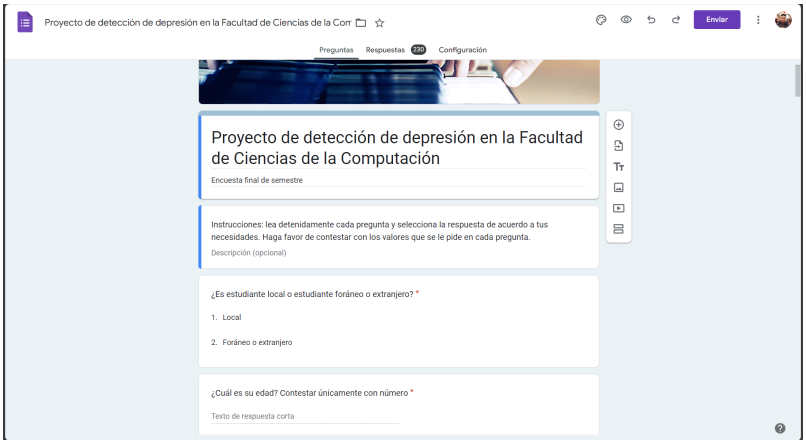

Figura 4.3: Formulario finales de semestre

4.2. Limpieza y análisis del conjunto de datos

Siempre que se tiene un conjunto de datos, provenga de dominio público o se hayan realizado una minería de datos, se debe realizar un proceso donde se eliminen datos atípicos o datos de valores nulos, así como también reconocer el tipo de datos que predomina en un conjunto de datos y encontrar relaciones que puedan mejorar las métricas del modelo de aprendizaje. Esta parte de la metodología es la primera que se hace sobre los datos, pues es necesario tener una mejor calidad de datos para tener un buen modelo de aprendizaje.

4.3. Programar y comparar algoritmos de aprendizaje

El siguiente paso después de analizar los datos es la implementación de los algoritmos de aprendizaje automático propuestos usando el lenguaje de programación Python. Se hará uso de las librerías scikit learn, numpy y matplotlib pues son las herramientas más comúnes para realizar esta tarea. En el listado 4.1 se presentarán los modulos de python usados.

```
# Tratamiento de datos
import numpy as np
import pandas as pd

#==============================================================
```

```
6 # Graficas
7 import matplotlib.pyplot as plt
8 from scipy import stats as sts
9  # animar graficas 3D
10 from matplotlib import animation
11 # graficas 3D
12 from mpl_toolkits.mplot3d import Axes3D
13 import ptitprince as pt
14 import seaborn as sns
15 sns.set(style="darkgrid")
16
17 #==============================================================
18 # Aprendizaje automatico
19 from sklearn.decomposition import PCA
20 from sklearn.pipeline import make_pipeline
21 from sklearn.preprocessing import StandardScaler
22 from sklearn.preprocessing import scale
23 from sklearn.model_selection import train_test_split
24 from sklearn.cluster import KMeans
25 from sklearn import manifold
26
27 #==============================================================
28 # Aprendizaje profundos
29 import tensorflow as tf
30 from tensorflow.keras.layers.experimental import preprocessing
31 from tensorflow import keras
```

Listado 4.1: Librerías de python

A continuación presentaremos los pseudocodigos y su implementación para los algoritmos de agrupamiento por promedios K y AGNES, así como la aquitectura de la red neuronal DKM que fue diseñada.

4.3.1. Pseudocódigo de promedios K

El algoritmo de agrupamiento por promedios K se presentará a continuación:

```
datos_entrada = [x1,x2,x3,...,xn];
inicilizar_promedios_k = [x1,x2,...,xk];
for all(n-k) samples:
    buscar_distancia_minima;
    for all k selected promedios:
        calcular distancias de las muestra desde todos los
        promedios K seleccionados;
    asignar muestra al grupo en la cual la distancia sea minima;
for all k promedios:
```

```
        calcular los valores promedios actualizados
    Repetir ciclos hasta que no sea posible actualizar el valor
    de los centroides;
    Salida = K grupos.
```

Su implementación en python se realizará con la librería de Scikit learn que ya incluye un método para implementar este algoritmo.

4.3.2. Pseudocódigo de AGNES

Para implementar el algoritmo de agrupamiento jerárquico AGNES primero se presentará el pseudocódigo de este algoritmo.

```
datos_entrada = [x1,x2,x2,...,x2]
k_grupos = "valor deseado de grupos"
for all k = len(datos_entrada):
    calcular distancias entre grupos D(s,r);
    for all D(s,r):
        Encontrar grupos con min(D(s,r));
            for all grupos cercanos B_i:
                Unir grupos B_i;
    for all grupos B_i:
        calcular distancia actualizada;
Repetir proceso hasta obtener grupos definidos;
Salida = k grupos.
```

La implementación en python se realizará mediante la librería de Scikit learn el cuál ya incluye una implementación del algoritmo de agrupamiento por AGNES.

4.3.3. Arquitectura de red neuronal DKM

Para un mejor entendimiento de la arquitectura de la red neuronal DKM diseñada para este trabajo de tesis se presentará el diagrama de la red neuronal y también se explicará cada uno de los bloques en el diagrama presentado en la figura 4.4.

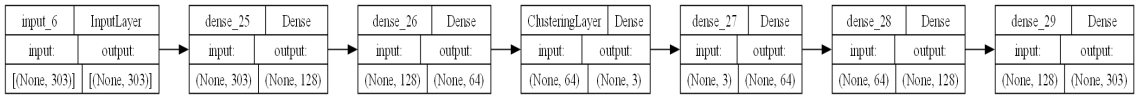

Figura 4.4: Arquitectura red DKM

Las partes más importantes en este diagrama son:

- Input_layer6: representación de la capa de entrada de la red neuronal. Tendrá 303 neuronas de entrada.

- Dense_25: Capa oculta compuesta por 128 neuronas, tendrá una salida de 128 valores de activación.
- Dense_26: Capa oculta compuesta por 64 neuronas, tendrá una salida de 64 valores de activación.
- ClusteringLayer: Capa de agrupación DKM, aplica el proceso de segregación y agrupamiento de los valores de activación. Tendrá como salida 3 valores.
- dense_27: Capa oculta que recibe 3 valores de activación y regresa 64 valores de activación como salida. Aquí asigna a cada dato puntual a un grupo.
- dense_28: Capa oculta que recibe 64 valores de activación. Actualiza los valores que se encuentran en los grupos. Da 128 valores de activación como salida.
- Dense_29: Capa oculta que recibe 128 valores de activación para actualizar los valores de los grupos. Su salida son 303 valores de activación pero separados por los 3 grupos obtenidos por la capa de agrupamiento.

El pseudocódigo en python que implementa el diagrama mostrado en la figura 4.4 se mostrará en el listado 4.2:

```
1 # Muestra de entrada
2 full_dim = two_sample_fitted.shape.as_list()[1]
3
4 # Cantidad de nueronas por capa y cantidad de grupos
5 encoding_dim1 = 128
6 encoding_dim2 = 64
7 encoding_dim3 = 3
8
9 # Codificacion de la entrada
10 encoder_input_data = keras.Input(shape=(full_dim,))
11
12 # Representacion de las capas ocultas
13 encoded_layer1 = keras.layers.Dense(encoding_dim1, activation='relu
    ')(encoder_input_data)
14 encoded_layer2 = keras.layers.Dense(encoding_dim2, activation='relu
    ')(encoded_layer1)
15 # Capa de agrupamiento
16 encoded_layer3 = keras.layers.Dense(encoding_dim3, activation='relu
    ', name="ClusteringLayer")(encoded_layer2)
17
18 encoder_model = keras.Model(encoder_input_data, encoded_layer3)
19
20 # Valores de activacion de cada capa
21 decoded_layer3 = keras.layers.Dense(encoding_dim2, activation='relu
    ')(encoded_layer3)
22 decoded_layer2 = keras.layers.Dense(encoding_dim1, activation='relu
    ')(decoded_layer3)
23 decoded_layer1 = keras.layers.Dense(full_dim, activation='sigmoid')
    (decoded_layer2)
```

```
24
25 # Codificacion de los valores de activacion
26 autoencoder_model = keras.Model(encoder_input_data, outputs=
     decoded_layer1, name="Encoder")
27
28 # Compilacipn del modelo
29 autoencoder_model.compile(optimizer="RMSprop", loss=tf.keras.losses
     .mean_squared_error,metrics=['accuracy'])
```

Listado 4.2: Arquitectura de red neuronal DKM en Python

4.4. Análisis de algoritmo para modelo de aprendizaje

La evaluación del rendimiento de los algoritmos de agrupamiento es un desafío debido a la naturaleza no supervisada del problema. A diferencia del aprendizaje supervisado, donde se dispone de etiquetas de verdad básica para comparar las predicciones del modelo, en el aprendizaje no supervisado, la evaluación del rendimiento debe basarse en medidas intrínsecas de la calidad del agrupamiento.

Una de las métricas más comunes utilizadas para evaluar la calidad de un agrupamiento es el Coeficiente de Silueta. Esta métrica calcula la cohesión y la separación de los clusters, proporcionando una medida de cuán bien se agrupan las muestras dentro de su propio cluster y cuán bien se separan de las muestras en otros clusters.

Otra métrica comúnmente utilizada es el Índice de Davies-Bouldin (DBI). El DBI compara la distancia media dentro de los clusters y entre los clusters para proporcionar una medida de la similitud entre los clusters. Un valor bajo del DBI indica que los clusters están bien separados y son compactos, lo que sugiere un buen rendimiento del algoritmo de agrupamiento.

Es importante destacar que ninguna de estas métricas es definitiva y cada una tiene sus limitaciones. Por lo tanto, la evaluación del rendimiento de los algoritmos de agrupamiento debe basarse en una combinación de diferentes métricas y técnicas de evaluación, teniendo en cuenta el contexto y los objetivos específicos del análisis.

4.5. Coeficiente de silueta

El coeficiente de silueta es una métrica esencial en el análisis de agrupamiento que proporciona una evaluación cuantitativa de la cohesión y separación de los clusters. Para cada muestra en el conjunto de datos, se calcula la distancia media intra-cluster (a) y la distancia media al cluster más cercano (b). La distancia intra-cluster es la distancia promedio entre una muestra y todas las demás muestras en el mismo

cluster. Por otro lado, la distancia al cluster más cercano es la distancia promedio entre una muestra y todas las muestras en el cluster más cercano. El coeficiente de silueta para una muestra específica se calcula entonces mediante la ecuación 4.1.

$$s = \frac{b - a}{\text{máx}(a, b)} \tag{4.1}$$

Este coeficiente oscila entre -1 y +1. Un valor cercano a +1 indica que la muestra está bien agrupada y lejos de los clusters vecinos. Un valor cercano a 0 sugiere que la muestra está cerca de un límite de decisión entre dos clusters. Un valor cercano a -1 implica que la muestra está incorrectamente agrupada.

El coeficiente de silueta puede ser utilizado para justificar la elección de un algoritmo de agrupamiento sobre otro. Además, puede ser útil para determinar el número óptimo de clusters en un conjunto de datos, ya que un número de clusters que maximiza el coeficiente de silueta promedio puede ser considerado como una buena elección. Sin embargo, es importante recordar que el coeficiente de silueta es sólo una de las muchas métricas que se pueden utilizar para evaluar la calidad de un agrupamiento, y su interpretación debe ser complementada con otras técnicas de evaluación y validación.

4.6. Índice de Davies-Bouldin

El Índice de Davies-Bouldin (DBI) es una métrica que se utiliza para evaluar la calidad de un agrupamiento. Se calcula utilizando las siguientes ecuaciones:

Para cada cluster, se calcula la distancia media intra-cluster (S_i), que es la distancia promedio entre cada punto en el cluster y el centroide del cluster, esto se calcula con la ecuacón 4.2

$$S_i = \frac{1}{n_i} \sum_{x \in C_i} d(x, c_i) \tag{4.2}$$

donde n_i es el número de puntos en el cluster C_i, x es un punto en el cluster C_i, c_i es el centroide del cluster C_i, y $d(x, c_i)$ es la distancia entre el punto x y el centroide c_i.

Luego, para cada par de clusters, se calcula una medida de 'similitud' (R_{ij}) (ecuación 4.3) que es la suma de las distancias medias intra-cluster de cada cluster dividida por la distancia entre los centroides de los dos clusters.

$$R_{ij} = \frac{S_i + S_j}{d(c_i, c_j)} \tag{4.3}$$

donde S_i y S_j son las distancias medias intra-cluster de los clusters C_i y C_j,

respectivamente, y $d(c_i, c_j)$ es la distancia entre los centroides c_i y c_j.

Finalmente, para cada cluster, se calcula la máxima 'similitud' de ese cluster a todos los demás clusters (D_i) usando la ecuación 4.4a. El DBI (ecuación 4.4b) es el promedio de estos valores máximos para todos los clusters donde N es el número total de clusters.

$$D_i = \max_{j \neq i} R_{ij} \tag{4.4a}$$

$$DBI = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{4.4b}$$

El DBI varía de 0 a infinito, donde valores más bajos indican una mejor calidad de agrupamiento. Un valor bajo del DBI indica que los clusters están bien separados (es decir, la distancia entre los clusters es grande) y son compactos (es decir, la distancia entre los puntos dentro de un cluster es pequeña). Puede ser una herramienta útil para comparar la calidad de los agrupamientos generados por diferentes algoritmos o configuraciones de algoritmos. Sin embargo, al igual que con cualquier métrica, es importante recordar que el DBI tiene sus limitaciones y debe utilizarse junto con otras técnicas de evaluación y validación.

4.7. Optimización del modelo y Discusión de resultados

Aquel algoritmo que tenga mejores valores de exactitud, precisión, sensibilidad y tiempo de ejecución será el de mejor modelo, pero se necesita realizar una optimización de su funcionamiento. Esto se puede hacer aplicando el paso de limpieza y análisis del conjunto de datos, así como también aplicar una función de optimización en el proceso de entrenamiento.

Para finalizar, se realiza una discusión de las diferencias notables de cada algoritmo propuesto, las razones por las que son distintas las métricas, se realizan las conclusiones y se discute el trabajo a futuro que se puede realizar con las bases de este trabajo de tesis.

Capítulo 5

Resultados experimentales

En esta sección se presentarán los resultados obtenidos de la aplicación de la metodología de ciencia de datos. Entre estos datos encontramos: los conjuntos de datos obtenidos por la aplicación de la encuesta diseñada para este trabajo de tesis en tres etapas del semestre, los conjuntos de datos después de haber realizado un proceso de limpieza y transformación de distintas variables para su manejo posterior, el resultado obtenido del análisis de los datos, las gráficas obtenidas por la implementación de los algoritmos de aprendizaje no supervisado presentados en la sección 3.

5.1. Obtención de datos

A continuación presentaremos los formularios realizados junto con una tabla que resume las características de los conjuntos de datos obtenidos.

5.1.1. Inicio de semestre

Tabla 5.1: Datos de inicios de semestre.

	Edad	Promedio	Año de ingreso	$\cdots$	Dificultad de concentración	Soledad	Molestía repentina
0	25	7.9	2016	$\cdots$	1	0	2
1	18	9.4	2022	$\cdots$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
215	19	8.7	2022	$\cdots$	1	3	2
216	24	9.0	2019	$\cdots$	0	0	1

Del formulario presentado en la figura 4.1 se obtuvo el conjunto de datos presentado en la tabla 5.1.Este conjunto de datos pasó por un proceso de limpieza de datos para ser utilizado para entrenar los algoritmos de aprendizaje.

5.1.2. Intermedios de semestre

Gracias al formulario mostrado en la figura 4.2 se obtuvo el conjunto de datos 5.2, este conjunto de datos estuvo bajo un procedimiento de limpieza de datos para ser utilizado para el entrenamiento del algoritmo de aprendizaje.

Tabla 5.2: Datos de intermedio de semestre.

	Edad	Promedio	Año de ingreso	$\cdots$	Dificultad de concentración	Soledad	Molestía repentina
0	18	8.0	2023	$\cdots$	4	1	3
1	25	7.8	2016	$\cdots$	5	2	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
225	23	9.7	2019	$\cdots$	3	3	0
226	19	7.8	202	$\cdots$	2	4	0

5.1.3. Finales de semestre

Los datos obtenidos de la aplicación del formulario en la figura 4.3 serán presentados en la tabla 5.3.

Tabla 5.3: Datos de finales de semestre.

	Edad	Promedio	Año de ingreso	$\cdots$	Dificultad de concentración	Soledad	Molestía repentina
0	23	9.5	2021	$\cdots$	4	3	3
1	22	7	2018	$\cdots$	4	4	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
229	21	8	2019	$\cdots$	3	5	3
230	22	8.8	202	$\cdots$	4	4	2

Ya obtenidos estos datos se da por concluida la etapa de recaudación de datos y se realizará la última parte de la limpieza y análisis de datos, se procederá a la implementación de los algoritmos de aprendizaje que será expuesta en la sección 5.3.

5.2. Análisis estadístico

En esta sección de los resultados analizaremos el comportamiento de los atributos del conjunto de datos, dichos atributos se mostrarán en una gráfica de nube de lluvia la cual es similar a una gráfica de caja con bigotes pero con más información de la distribución de los datos.

5.2.1. Síntomas a inicios de semestre

En la figura 5.1 observamos la gráfica de nube de lluvia de los atributos generales (edad, promedio, materias atrasadas, horas de estudio y año de ingreso) del conjunto de datos obtenido por una encuesta aplicada a principios de semestre.

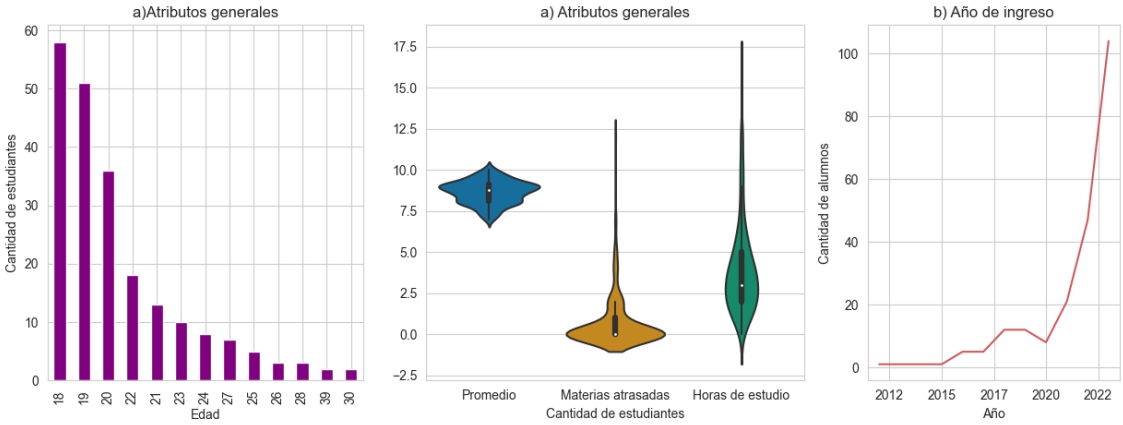

Figura 5.1: Atributos generales a principio de semestre

En esta gráfica se puede observar lo siguiente:

- La mayoría de los estudiantes tienen una edad entre el rango de 18 a 23 años y tiene un promedio entre 8 a 10.
- La mayoría de los estudiantes que respondieron esta encuesta son de nuevo ingreso o tiene poco tiempo en la universidad por lo que hay pocas materias atrasadas.

En la gráfica 5.2 se muestra el comportamiento de los síntomas que se capturan en la encuesta mediante las preguntas provenientes de la escala CES.
La gráfica 5.2 nos permite observar lo siguiente:

- La mayoría de los síntomas se encuentra en la escala más baja al principio de semestre.
- La mayoría de los síntomas tienen un comportamiento similar al principio de semestre.
- Existe una gran cantidad de alumnos que han perdido el interés en sus actividades cotidianas.

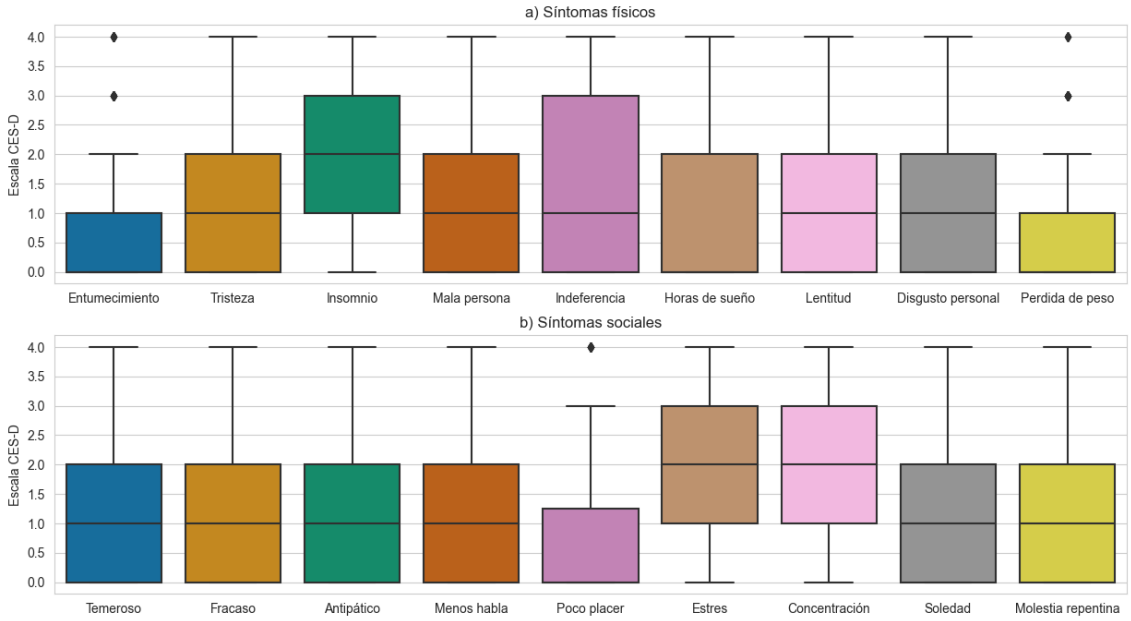

Figura 5.2: Síntomas a principio de semestre

- Puede existir una relación entre la facilidad para estresarse y la dificultad de concentración.

Los puntos anteriores son observaciones que se hacen sobre los datos obtenidos en la captura de datos de principios de semestre, por lo que es importante observar su evolución en la etapa de intermedios de semestre.

5.2.2. Síntomas a intermedios de semestre

La siguiente etapa es intermedios de semestre, donde observaremos la evolución de los casos de depresión, desde los atributos generales hasta los síntomas más destacables.

En la figura 5.3 se puede observar que hay un comportamiento distinto en la población de estudiantes universitarios.

- Ahora el rango de edad tiene un limte de 25 años.
- Hay mayor cantidad de población perteneciente a generaciones de mayor antigüedad.
- Aumentó el número de materias atrasadas.

Por la parte de los síntomas, en la figura 5.4 se observa cambio en todos los síntomas, principalmente en aquellos presentes en la segunda fila. En este figura observamos:

- Un aumento en síntomas que reflejan el distanciamiento social, el estrés y la forma de conllevar el día a día.
- No hay un aumento significativo en los síntomas que refieren a los hábitos de sueño y sensaciones corporales.

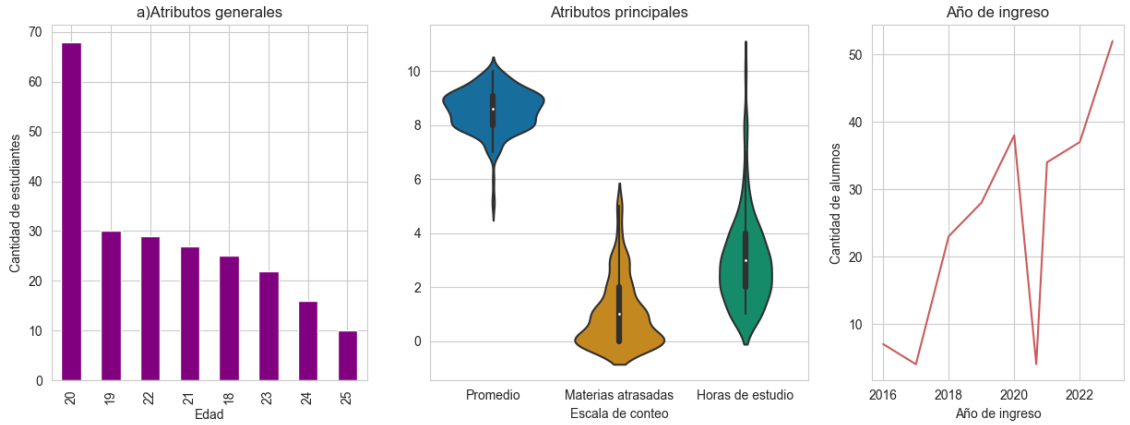

Figura 5.3: Atributos generales a intermedios de semestre

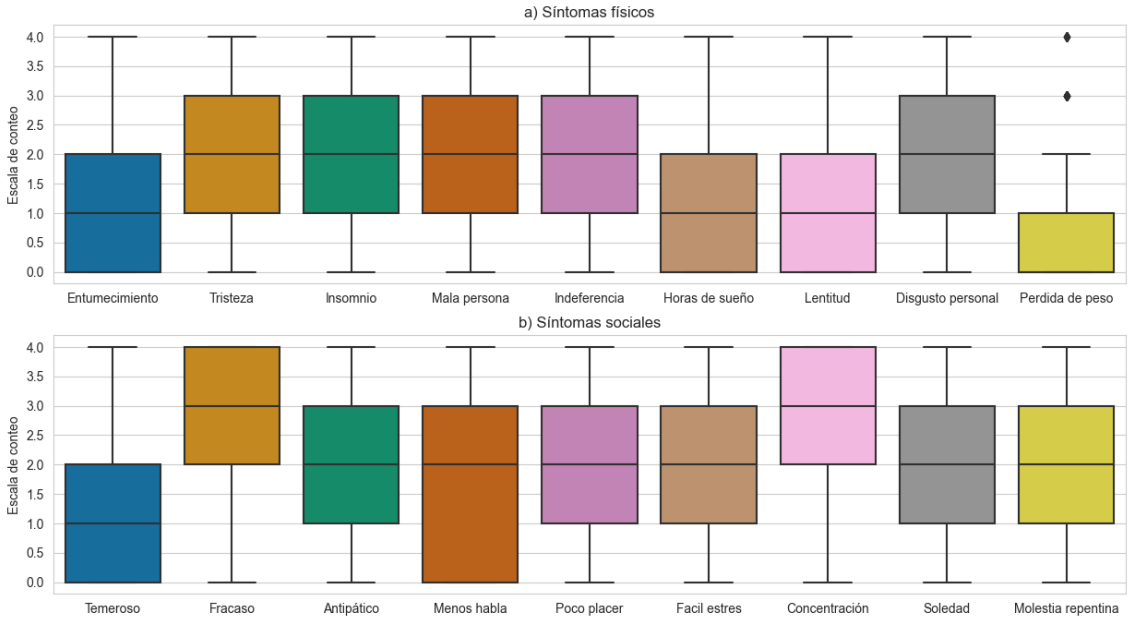

Figura 5.4: Síntomas a intermedio de semestre

5.2.3. Síntomas a finales de semestre

Hasta ahora se han presentado los análisis realizados sobre los conjuntos de datos de inicio de semestre e intermedios de semestre.

En la figura 5.5 se observa que en la población a estudiar existe un registro de edad que ronda los 40 años, igual que en la etapa de inicios de semestre se ignorará este registro para el proceso de entrenamiento. Otras observaciones en esta etapa son:

- El promedio de las materias atrasadas ronda el rango de 2 a 3 materias por registro.
- Las edades de la población tienen un promedio de 23 años.
- La muestra de esta etapa tienen una mayor pertenencia a los últimos años de inscripción, siendo el más reciente el año 2023.

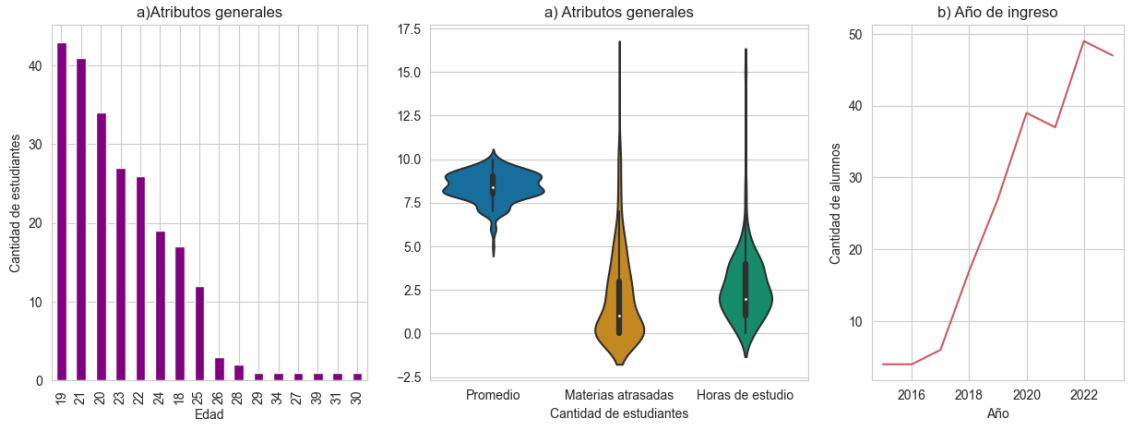

Figura 5.5: Atributos generales a finales de semestre

Por parte del análisis de los síntomas, cuyo análisis se presenta en la figura 5.6, se puede observar lo siguiente:

- La mayoría de los síntomas tienen una mayor presencial temporal.
- La sensación de fracaso, poca concentración y la sensación de soledad son síntomas que se presentan por más tiempo durante el final de semestre.

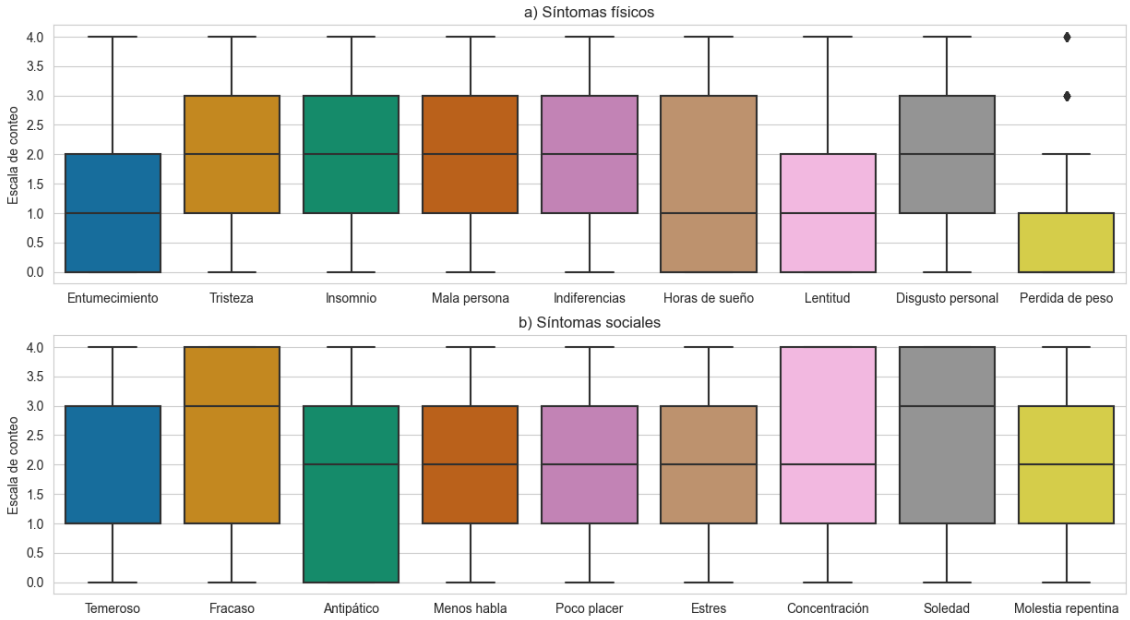

Figura 5.6: Síntomas a finales de semestre

En la sección 5.3 se presentarán los resultados obtenidos de la implementación de los algoritmos de agrupamiento propuestos.

5.3. Resultados de la implementación de los algoritmos de aprendizaje

Después de haber realizado los pasos de obtención, limpieza y análisis de datos se procedió a la implementación de los algoritmos propuestos en las secciones 4.3.1, 4.3.2 y 4.3.3 en el lenguaje de programación Python. Los resultados obtenidos fueron tres grupos por cada periodo del semestre, estos serán mostrados en un histograma con las escalas representantes de cada grupo de acuerdo a la tabla 5.4

Tabla 5.4: Resultados obtenidos

Rango	Caso de depresión
0 a 1	Depresión leve
1.25 a 1.5	Depresión moderada
1.75 a 2	Depresión severa

De acuerdo a la escala de la tabla 5.4 las columnas que representan a los grupos serán repartidas en un rango de 0 a 2 y por cada algoritmo de agrupamiento se tendrán tres columnas que representan las etapas del semestre.

Las columnas generadas por el algoritmo AGNES son aquellas de color azul, las columnas generadas por promedios K son de color naranja y las columnas generadas por la red neuronal con capa DKM son de color verde.

La figura 5.7 muestra nueve columnas, que pertenecen a los distintos casos de depresión y que fueron generadas por los algoritmos de agrupamiento.

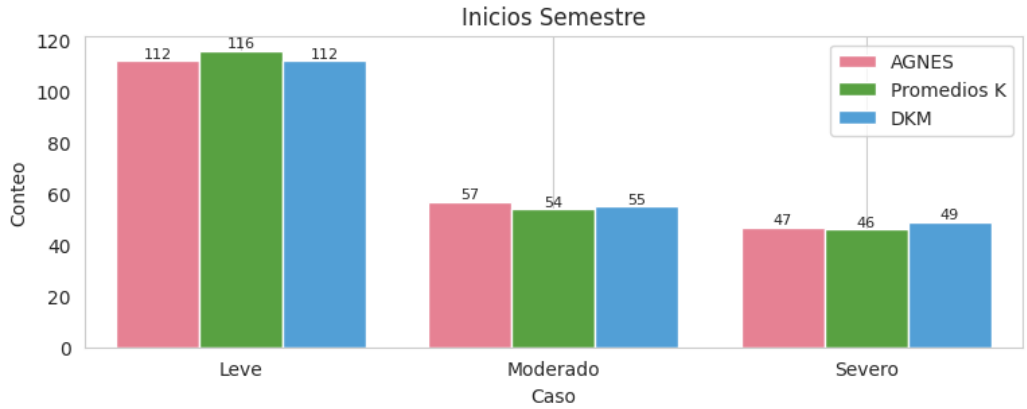

Figura 5.7: Resultados obtenidos inicios de semestre

De la figura 5.7 podemos notar que los tres algoritmos crearon un grupo de casos de depresión leve con mayor población que los otros dos casos, esto puede deberse al hecho de que los datos fueron capturados a inicios de semestre. Por otro lado el comportamiento de los casos de depresión moderada y severa tiene un comportamiento similar en esta etapa.

En la figura 5.8 se presenta la evolución de los casos en la etapa intermedia del semestre.

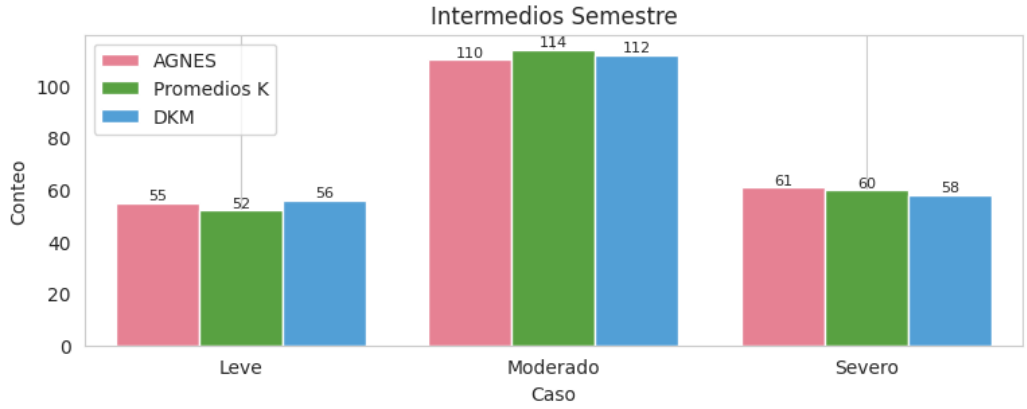

Figura 5.8: Resultados obtenidos intermedios de semestre

Ahora el grupo predominante es aquel que perteneces al caso de la depresión intermedia y se nota una gran disminución en la población del caso de depresión leve mientras que la depresión severa empieza a tener un aumento de su población.

La siguiente es la observación de la evolución de los casos de depresión a finales de semestre universitario, los grupos presentados se presentarán los resultados de la figura 5.9.

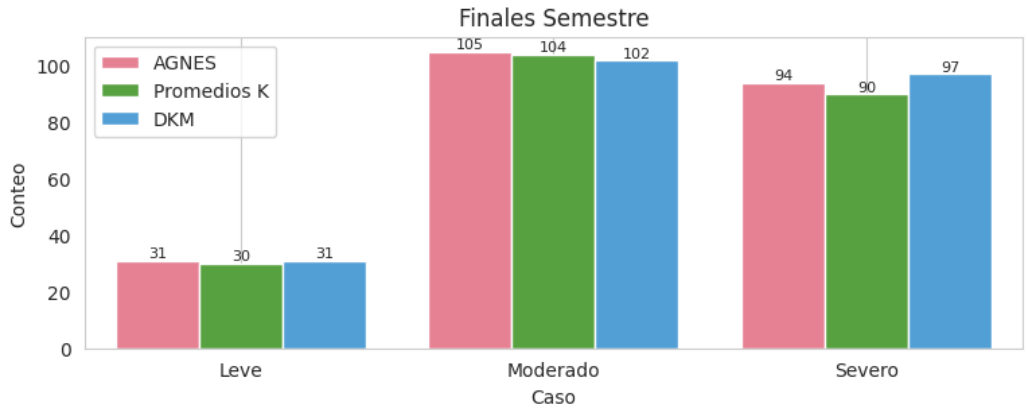

Figura 5.9: Resultados obtenidos finales de semestre

Las poblaciones que poseen una gran cantidad de miembros son de casos de depresión moderada y severa, a pesar de que en la etapa pasada se lograba notar una gran distinción en esta etapa esa separación se ve acortada y se observa un gran decremento en la población de casos de depresión leve, que para los tres algoritmos tiene un mismo comportamiento.

Otra manera de observar los grupos realizados por lo tres algoritmos es mediante su evolución por etapa de semestre, es decir, cómo va cambiando cada una de las

poblaciones respectiva a un caso de depresión. En la figura 5.10 se presenta la evolución de cada una de las poblaciones con respeto al avance del semestre, así como los distintos grupos por algoritmo.

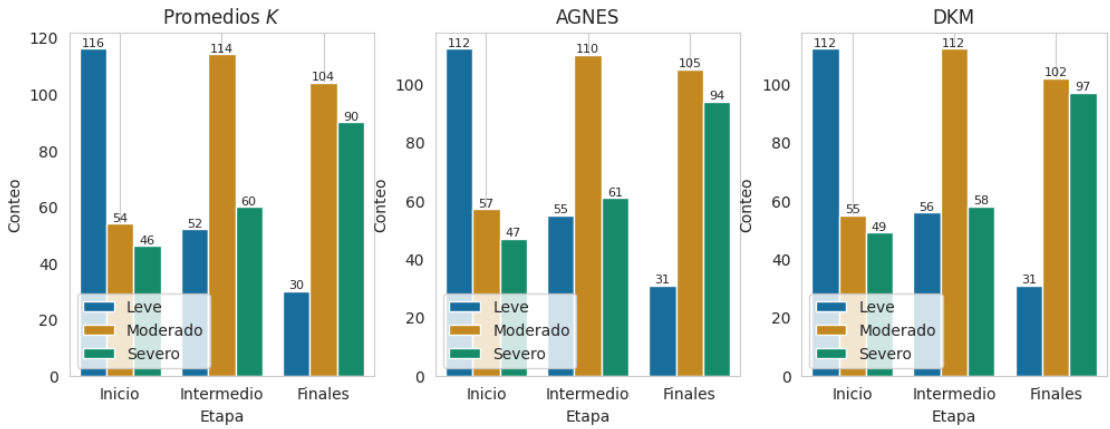

Figura 5.10: Comparación por algoritmo

Se puede observar que el comportamiento de los grupos es similar para todos los algoritmos, siendo los grupos más grandes los que fueron generados por el algoritmo de agrupamiento por Promedios K, mientras que el algoritmo AGNES generó grupos más pequeños.

Otro punto desde el cuál se puede observar el comportamiento de los grupos es por tipo de caso. En la figura 5.11 se puede observar que el caso leve tiene una tendencia hacia abajo, los casos leves tienen tendencia a una formación de campana y los casos severos tienen una tendencia hacía arriba.

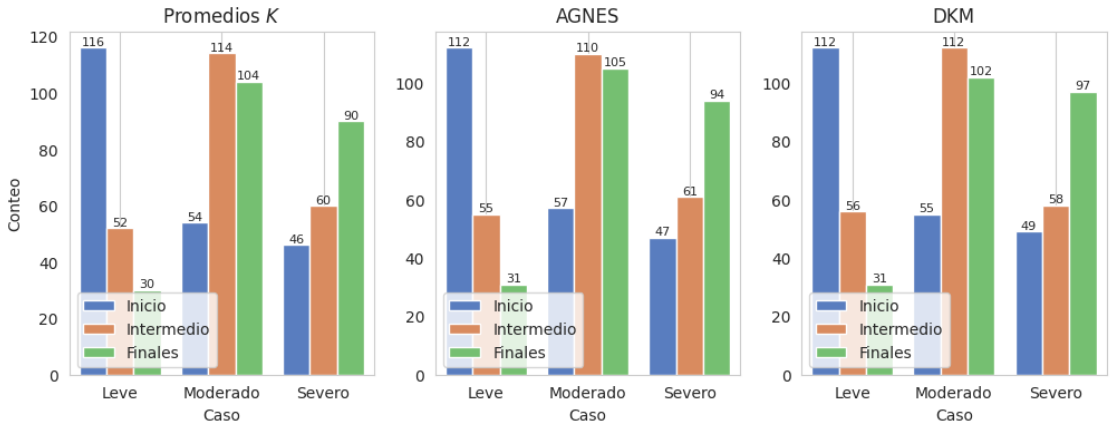

Figura 5.11: Comparación por caso

Cada uno de los algoritmos generó grupos diferentes, sin embargo es importante la evaluación de los distintos algoritmos, para esto se usarán las métricas del coeficiente de silueta y el índice de Davies-Bouldin. En la siguiente sección se presentarán los resultados obtenidos de las evaluaciones a cada algoritmo.

5.4. Evaluación de los algoritmos

A pesar de que se ha cumplidó el objetivo de la implementación de los algoritmos, haber generado tres grupos por cada etapa del semestre, es necesario evaluar la precisión de cómo fueron generados estos grupos, para eso se usarán las métricas expuestas en las secciones 4.5 y 4.6.

La evaluación realizada por cada métrica será presentada en las tablas 5.5 y 5.6. La evaluación de los tres algoritmos, uno por columna donde las etapas de aplicación de la encuesta serán las filas de las tablas.

Tabla 5.5: Resultados del coeficiente de silueta

Etapa	**Agnes**	**Promedios *K***	**DKM**
Inicio de semestre	0.62054	0.59128	0.60413
Intermedios de semestre	0.67210	0.63990	0.67901
Finales de semestre	0.67009	0.65406	0.67099

Por parte del coeficiente de silueta (5.5) los valores se encuentran en un rango de [-1, 1], mientras más cerca estén al 1 la evaluación el algoritmo de agrupamiento será mejor.

Los resultados de este método de evaluación muestran al algoritmo de agrupamiento jerárquico AGNES pues en las tres etapas obtuvo el mejor valor de la métrica.

En el caso de los resultados del índice de Davies-Bouldin (DBI), el rango del mejor valor se encuentra en [0,1] mientras más pequeño sea el valor será mejor.

Tabla 5.6: Resultados de índice Davies-Bouldin

Etapa	**Agnes**	**Promedios *K***	**DKM**
Inicio de semestre	0.2530	0.2902	0.2561
Intermedios de semestre	0.2706	0.3110	0.2681
Finales de semestre	0.2517	0.2761	0.2421

Por parte del DBI los algoritmos que mejor resultados obtuvo fue el algoritmo de agrupación por redes neuronales DKM, seguido por el algoritmo de agrupamiento jerárquico AGNES. Esto porque los valores de esta métrica aplicada a este algoritmo están más cercanas al cero.

5.5. Discusión de resultados

Los algoritmos propuestos en este trabajo de tesis obtuvieron buenos resultados gracias a que los datos recaudados no tenían una variación o distancia tan grande, esto les permitió distinguir tres grupos sin mayor problema. Es destacable que los

algoritmos AGNES y DKM son los mejores evaluados en ambas métricas, a pesar de tener un funcionamiento diferente entre si, mientras que el algoritmo de promedios K es el de menor evaluación entre los tres. Esto puede indicar un mejor funcionamiento en algoritmos de agrupamiento que funcionan mediante procesos iterativos.

Capítulo 6

Conclusiones

En este trabajo de tesis, se estudió la depresión de una población estudiantil universitaria aplicando algoritmos de aprendizaje no supervisado siguiendo una metodología de ciencia de datos. Los resultados obtenidos hasta la fecha han mostrado que existe una correlación entre los distintos casos de depresión que pueden tener miembros de un alumnado de una facultad universitaria con la época del semestre en la que se encuentren.

Durante la investigación realizada hasta la fecha se ha observado que los casos de depresión en los estudiantes pueden evolucionar de manera drástica en cuestión de meses. También se observaron las distintas características generales más comunes que conllevan a un caso de depresión moderada o severa, siendo una de las principales los promedios académicos altos y ciertos síntomas emocionales relacionados al ámbito social de los estudiantes.

El proceso de investigación realizado también mostró el gran desempeño que tienen los algoritmos de aprendizaje no supervisado enfocados al problema de agrupamiento para reconocer los patrones (síntomas y características) que llevan a los estudiantes a pertenecer a cierto caso de depresión.

Esto nos permite ver a la ciencia de datos como una herramienta más para resolver problemas que abarcan temas delicados o complicados de estudiar, pero deben tener el apoyo de profesionales de la disciplina a la cuál se enfoque el trabajo de ciencia de datos.

La investigación realizada en este trabajo de tesis puede llevarse a una aplicación en tiempo real como trabajo a futuro. En la misma se puede generar un programa o aplicación web donde se implementen los modelos de aprendizaje no supervisado realizados en este trabajo de tesis a una interfaz gráfica, donde se aplique la encuesta realizada para la obtención de datos usados en este trabajo de tesis.

Bibliografía

[1] Katrina Lloyd, Dirk Schubotz, Rosellen Roche, Joel Manzi y Martina McKnight. «A Mental Health Pandemic? Assessing the Impact of COVID-19 on Young People's Mental Health». En: *International Journal of Environmental Research and Public Health* 20 (ago. de 2023), pág. 6550. DOI: 10.3390/ijerph20166550.

[2] Ricardo Sánchez, Heidy Cáceres y Dora Gomez. «Ideación suicida en adolescentes universitarios: prevalencia y factores asociados». En: *Biomédica* 22 (dic. de 2002), págs. 407-416. DOI: 10.7705/biomedica.v22iSupp2.1189.

[3] Katherine Vergara, Shyrley Díaz-Cárdenas y Farith Gonzalez. «Síntomas de depresión y ansiedad en jóvenes universitarios: prevalencia y factores relacionados». En: *Revista Clínica de Medicina de Familia* 7 (feb. de 2014), págs. 14-22. DOI: 10.4321/S1699-695X2014000100003.

[4] Catalina González-Forteza, José Alberto Jiménez-Tapia, Luciana Ramos Lira y Fernando A. Wagner. En: *Salud Pública de México* 50.4 (jul. de 2008), págs. 292-299. URL: https://saludpublica.mx/index.php/spm/article/view/6831.

[5] Marianne Kastrup y Armando Ramos. «Global Mental Health». En: *Danish medical bulletin* 54 (mar. de 2007), págs. 42-3.

[6] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019. ISBN: 9781492032618. URL: https://books.google.com.mx/books?id=HHetDwAAQBAJ.

[7] Andreas C. Müller y Sarah Guido. *Introduction to Machine Learning with Python*. Sebasstopol, Rusia: O´Reilly, 2017.

[8] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor. *An Introduction to Statistical Learning: with Applications in Python*. Ene. de 2023, págs. 23-27. ISBN: 978-3-031-38746-3.

[9] Peter Harrington. *Machine Learning in Action*. USA: Manning Publications Co., 2012. ISBN: 1617290181.

[10] A. Zheng y A. Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2018. ISBN: 9781491953198. URL: https://books.google.com.mx/books?id=sthSDwAAQBAJ.

[11] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor. *An Introduction to Statistical Learning: with Applications in Python*. Ene. de 2023, págs. 520-525. ISBN: 978-3-031-38746-3.

[12] K. Kolodiazhnyi. *Hands-On Machine Learning with C++: Build, Train, and Deploy End-To-end Machine Learning and Deep Learning Pipelines*. Packt Publishing, Limited, 2020. ISBN: 9781789955330. URL: https://books.google.com.mx/books?id=0CeHzQEACAAJ.

[13] Peter Harrington. *Machine Learning in Action*. USA: Manning Publications Co., 2012, págs. 18-20. ISBN: 1617290181.

[14] M. Kubat. *An Introduction to Machine Learning*. Springer International Publishing, 2021. ISBN: 9783030819354. URL: https://books.google.com.mx/books?id=cshEEAAAQBAJ.

[15] G. Ciaburro y B. Venkateswaran. *Neural Networks with R: Smart Models Using CNN, RNN, Deep Learning, and Artificial Intelligence Principles*. Packt Publishing, 2017. ISBN: 9781788397872. URL: https://books.google.com.mx/books?id=6YIbtAEACAAJ.

[16] J. Patterson y A. Gibson. *Deep Learning: A Practitioner's Approach*. O'Reilly Media, 2017. ISBN: 9781491914212. URL: https://books.google.com.mx/books?id=rLcuDwAAQBAJ.

[17] C.C. Aggarwal y C.K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2018. ISBN: 9781315360416. URL: https://books.google.com.mx/books?id=cH5ODwAAQBAJ.

[18] A.A. Patel. *Hands-on Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*. O'Reilly Media, 2019. ISBN: 9781492035640. URL: https://books.google.com.mx/books?id=8dMpvgEACAAJ.

[19] S. Miyamoto. *Theory of Agglomerative Hierarchical Clustering*. Behaviormetrics: quantitative approaches to human behavior. Springer, 2022. URL: https://books.google.com.mx/books?id=Ch4izwEACAAJ.

[20] M. Kubat. *An Introduction to Machine Learning*. Springer International Publishing, 2021, págs. 273-278. ISBN: 9783030819354. URL: https://books.google.com.mx/books?id=cshEEAAAQBAJ.

[21] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor. *An Introduction to Statistical Learning: with Applications in Python*. Ene. de 2023, págs. 39-42. ISBN: 978-3-031-38746-3.

[22] F. Chollet. *Deep Learning with Python*. Manning, 2017. ISBN: 9781638352044. URL: https://books.google.com.mx/books?id=wzozEAAAQBAJ.

[23] C.C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, 2023. ISBN: 9783031296420. URL: https://books.google.com.mx/books?id=0-rIEAAAQBAJ.

[24] F.M. Salem. *Recurrent Neural Networks: From Simple to Gated Architectures*. Springer International Publishing, 2022. ISBN: 9783030899295. URL: https://books.google.com.mx/books?id=bJpXEAAAQBAJ.

[25] A. Glassner. *Deep Learning: A Visual Approach*. No Starch Press, 2021. ISBN: 9781718500730.

[26] I. Drori. *The Science of Deep Learning*. Cambridge University Press, 2022. ISBN: 9781108890441.

[27] Minsik Cho, Keivan Alizadeh-Vahid, Saurabh Adya y Mohammad Rastegari. «DKM: Differentiable K-Means Clustering Layer for Neural Network Compression». En: *CoRR* abs/2108.12659 (2021). arXiv: 2108.12659. URL: https://arxiv.org/abs/2108.12659.

[28] K.S. Nandi Dr. Rupam Dr. Gypsy. *Data Science Fundamentals and Practical Approaches*. BPB Publications, 2020. ISBN: 9789389845679.

[29] J.M.O. Candel. *Big data, machine learning y data science en python*. RA-MA S.A. Editorial y Publicaciones, 2022. ISBN: 9788419444592.

[30] U. Qamar y M.S. Raza. *Data Science Concepts and Techniques with Applications*. Springer Nature Singapore, 2020. ISBN: 9789811561337.

[31] Yasmin Al-Shannaq, Sajeda Darwish, Anas Mohammad y Diana Jaradat. «Depression and Depression Literacy among Adolescent School Students». En: 2 (mar. de 2023), págs. 55-68. DOI: 10.14525/JJNR.v2i1.08.

[32] Hamid Saeed, Amna Qureshi, Muhammad Rasool, Muhammad Islam, Furqan Hashmi, Amna Saeed, Rimsha Asad, Arfa Arshad y Azba Qureshi. «Determinants of anxiety and depression among university teachers during third wave of COVID-19». En: *BMC Psychiatry* 23 (abr. de 2023). DOI: 10.1186/s12888-023-04733-9.

[33] Nelao Mhata, Vuyokazi Ntlantsana, Andrew Tomita, Kissah Mwambene y Shamima Saloojee. «Prevalence of depression, anxiety and burnout in medical students at the University of Namibia». En: *South African Journal of Psychiatry* 29 (mayo de 2023). DOI: 10.4102/sajpsychiatry.v29i0.2044.

[34] Michaela Korte, Deniz Cerci, Roman Wehry, R. Timmers y Victoria Williamson. «The same but different. Multidimensional assessment of depression in students of natural science and music». En: *Health Psychology Research* 11 (mayo de 2023). DOI: 10.52965/001c.74879.

[35] Anu Priya, Shruti Garg y Neha Tigga. «Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms». En: *Procedia Computer Science* 167 (ene. de 2020), págs. 1258-1267. DOI: 10.1016/j.procs.2020.03.442.

[36] Sam Andersson, Deepti Bathula, Stavros Iliadis, Martin Walter y Alkistis Skalkidou. «Predicting women with depressive symptoms postpartum with machine learning methods». En: *Scientific Reports* 11 (abr. de 2021), pág. 7877. DOI: 10.1038/s41598-021-86368-y.

[37] Wei Wanqing y LinYu Li. «The Impact of Artificial Intelligence on the Mental Health of Manufacturing Workers: The Mediating Role of Overtime Work and the Work Environment». En: *Frontiers in Public Health* 10 (abr. de 2022), pág. 862407. DOI: 10.3389/fpubh.2022.862407.

[38] Danxia Liu, Farooq Ahmed, Muhammad Shahid, Jing Guo y Xing Lin Feng. «Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review». En: *JMIR Mental Health* (mar. de 2022). DOI: 10.2196/27244.

[39] Alice Chavanne, Marie Martinot, Jani Penttilä, Yvonne Grimmer, Patricia Conrod, Argyris Stringaris, Betteke van Noort, Corinna Isensee, Andreas Becker, Tobias Banaschewski, Arun Bokde, Sylvane Desrivières, Herta Flor, Antoine Grigis, Hugh Garavan, Penny Gowland, Andreas Heinz, Frauke Nees y Hélène Vulser. «Anxiety onset in adolescents: a machine-learning prediction». En: *Molecular Psychiatry* 28 (dic. de 2022). DOI: 10.1038/s41380-022-01840-z.

[40] William Rothenberg, Andrea Bizzego, Gianluca Esposito, Jennifer Lansford, Suha Al-Hassan, Dario Bacchini, Marc Bornstein, Lei Chang, Kirby Deater-Deckard, Laura Giunta, Kenneth Dodge, Sevtap Gurdal, Qin Liu, Qian Long, Paul Oburu, Concetta Pastorelli, Ann Skinner, Emma Sorbring, Sombat Tapanya y Liane Alampay. «Predicting Adolescent Mental Health Outcomes Across Cultures: A Machine Learning Approach». En: *Journal of Youth and*

Adolescence 52 (abr. de 2023), págs. 1-25. DOI: 10.1007/s10964-023-01767-w.

[41] M. Qureshi, Muhammad Qureshi, Junaid Asghar, Fatima Alam y Ayman Aljarbouh. «Prediction and Analysis of Autism Spectrum Disorder Using Machine Learning Techniques». En: *Journal of Healthcare Engineering* 2023 (jul. de 2023), págs. 1-10. DOI: 10.1155/2023/4853800.

[42] Deepa Tilwani, Jessica Bradshaw, Amit Sheth y Christian O'Reilly. «ECG Recordings as Predictors of Very Early Autism Likelihood: A Machine Learning Approach». En: *Bioengineering* 10 (jul. de 2023), pág. 827. DOI: 10.3390/bioengineering10070827.

[43] Zihan Qu, Yashan Wang, Dingjie Guo, Guangliang He, Chuanying Sui, Yuqing Duan, Xin Zhang, Linwei Lan, Hengyu Meng, Yajing Wang y Xin Liu. «Identifying depression in the United States veterans using deep learning algorithms, NHANES 2005–2018». En: *BMC Psychiatry* 23 (ago. de 2023). DOI: 10.1186/s12888-023-05109-9.

[44] Payam Hosseinzadeh Kasani, Jung Lee, Chihyun Park, Cheol-Heui Yun, Jae Won Jang y Sang-Ah Lee. «Evaluation of nutritional status and clinical depression classification using an explainable machine learning method». En: *Frontiers in Nutrition* (mayo de 2023). DOI: 10.3389/fnut.2023.1165854.

[45] Bryn Loftness, Donna Rizzo, Julia Halvorson-Phelan, Aisling O'Leary, Shania Prytherch, Carter Bradshaw, Anna-Jane Brown, Nicholas Cheney, Ellen Mcginnis y Ryan McGinnis. «Toward Digital Phenotypes of Early Childhood Mental Health via Unsupervised and Supervised Machine Learning». En: (feb. de 2023). DOI: 10.1101/2023.02.24.23286417.

[46] Shinjini Kundu, Stephanie Barsoum, Jeanelle Ariza, Amber Nolan, Caitlin Latimer, Christopher Keene, Peter Basser y Dan Benjamini. «Mapping the individual human cortex using multidimensional MRI and unsupervised learning». En: *Brain Communications* 5 (nov. de 2023), fcad258. DOI: 10.1093/braincomms/fcad258.

[47] Moumita Bhowmik, Naim Al Bhuyain, Md. Rokonuzzaman Reza, Nafiz Imtiaz Khan y Muhammad Nazrul Islam. «Neurophysiological Feature Based Stress Classification Using Unsupervised Machine Learning Technique». En: oct. de 2022, págs. 603-614. ISBN: 978-981-19-2444-6. DOI: 10.1007/978-981-19-2445-3_42.

[48] Qinghua Tang, Yixuan Zhao, Yujia Wei y Lu Jiang. «Research on the Mental Health of College Students Based on Fuzzy Clustering Algorithm». En: *Security and Communication Networks* 2021 (sep. de 2021), págs. 1-8. DOI: 10.1155/2021/3960559.

[49] M. Srividya, Mohanavalli Subramaniam y Bhalaji Natarajan. «Behavioral Modeling for Mental Health using Machine Learning Algorithms». En: *Journal of Medical Systems* 42 (abr. de 2018), pág. 88. DOI: 10.1007/s10916-018-0934-5.

[50] Gabriella Casalino, Giovanna Castellano, Olgierd Hryniewicz, Daniel Leite, Karol Opara, Weronika Radziszewska y Katarzyna Kaczmarek-Majer. «Semi-supervised vs. supervised learning for mental health monitoring: A case study on bipolar disorder». En: *International Journal of Applied Mathematics and Computer Science* 33 (sep. de 2023). DOI: 10.34768/amcs-2023-0030.

[51] Edwin Ponce, Melissa Cruz y Laberiano Andrade-Arenas. «Machine Learning Applied to Prevention and Mental Health Care in Peru». En: *International Journal of Advanced Computer Science and Applications* 13 (ene. de 2022). DOI: 10.14569/IJACSA.2022.0130196.

[52] Sofia Arora, Arun Malik, Dr. Mohammad Shabaz y Evans Asenso. «Machine Learning based Model for Detecting Depression During Covid-19 Crisis». En: *Scientific African* 20 (mayo de 2023), e01716. DOI: 10.1016/j.sciaf.2023.e01716.

[53] Manar Elshazly, Mohamed Haggag y Soha Mohamed. «A Depression Detection Model using Deep Learning and Textual Entailment». En: *International Journal of Computer Science and Information Security,* 19 (dic. de 2021). DOI: 10.5281/zenodo.5852684.

[54] Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya y Mueen Uddin. «Deep Learning for Depression Detection from Textual Data». En: *Electronics* (feb. de 2022). DOI: 10.3390/electronics11050676.

[55] Yunhan Lin, Biman Liyanage, Yutao Sun, Tianlan Lu, Zhengwen Zhu, Yundan Liao, Qiushi Wang, Chuan Shi y Weihua Yue. «A deep learning-based model for detecting depression in senior population». En: *Frontiers in Psychiatry* 13 (nov. de 2022). DOI: 10.3389/fpsyt.2022.1016676.

[56] Aleena Nadeem, Muhammad Naveed, Muhammad Islam, Hammad Afzal, Tanveer Ahmad y Ki-Il Kim. «Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to

Social Networking Data». En: *Sensors* 22 (dic. de 2022), pág. 9775. DOI: 10.3390/s22249775.

[57] Clinton Lau, Xiaodan Zhu y Wai-Yip Chan. «Automatic depression severity assessment with deep learning using parameter-efficient tuning». En: *Frontiers in Psychiatry* 14 (jun. de 2023). DOI: 10.3389/fpsyt.2023.1160291.

[58] Narpinder Singh, Mostafa Fouda, Luca Saba y Jasjit Suri. «Attention-Enabled Ensemble Deep Learning Models and Their Validation for Depression Detection: A Domain Adoption Paradigm». En: *Diagnostics* 13 (jun. de 2023), pág. 2092.

[59] Matthew Squires, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou, U Rajendra Acharya y Yuefeng Li. «Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment». En: *Brain Informatics* 10 (abr. de 2023). DOI: 10.1186/s40708-023-00188-6.

[60] Jagneet Kaur y Raino Bhatia. «Utilizing Emotion Analysis for Suicide Prediction and Mental Health Detection in Students with Deep Learning». En: 12 (mayo de 2024), págs. 729-738.

[61] Doaa Mohey El-Din, Mohamed Hamed y Nour Eldeen. «SentiNeural: A Depression Clustering Technique for Egyptian Women Sentiments». En: *International Journal of Advanced Computer Science and Applications* 10 (ene. de 2019). DOI: 10.14569/IJACSA.2019.0100572.

[62] Md. Zia Uddin, Kim Dysthe, Asbjørn Følstad y Petter Brandtzaeg. «Deep learning for prediction of depressive symptoms in a large textual dataset». En: *Neural Computing and Applications* 34 (ene. de 2022), págs. 1-24. DOI: 10.1007/s00521-021-06426-4.

[63] Shreeya Garg, Urvashi Shukla y Linga Reddy Cenkeramaddi. «Detection of Depression Using Weighted Spectral Graph Clustering With EEG Biomarkers». En: *IEEE Access* 11 (ene. de 2023), págs. 57880-57894. DOI: 10.1109/ACCESS.2023.3281453.

[64] Choon-Young Kim, Cheolmin Lee, Seungwoo Lee, Jung Yoo, Heesun Lee, Hyo Park, Kyungdo Han y Su-Yeon Choi. «The Association of Smoking Status and Clustering of Obesity and Depression on the Risk of Early-Onset Cardiovascular Disease in Young Adults: A Nationwide Cohort Study». En: *Korean circulation journal* 53 (nov. de 2022). DOI: 10.4070/kcj.2022.0179.

[65] Artur Shvetcov, Alexis Whitton, Suranga Kasturi, Wu-Yi Zheng, Joanne Beames, Omar Ibrahim, Jin Han, Leonard Hoon, Kon Mouzakis, Sunil Gupta, Svetha Venkatesh, Helen Christensen y Jill Newby. «Machine learning identifies a COVID-19-specific phenotype in university students using a mental health app». En: *Internet Interventions* 34 (sep. de 2023), pág. 100666. DOI: 10.1016/j.invent.2023.100666.

[66] Noura Alosaimi, Lauren Sherar, Paula Griffiths y Natalie Pearson. «Clustering of diet, physical activity and sedentary behaviour and related physical and mental health outcomes: a systematic review». En: *BMC Public Health* 23 (ago. de 2023). DOI: 10.1186/s12889-023-16372-6.

[67] Jing Lei. «An Analytical Model of College Students' Mental Health Education Based on the Clustering Algorithm». En: *Mathematical Problems in Engineering* 2022 (sep. de 2022), págs. 1-11. DOI: 10.1155/2022/1880214.

[68] Nonie Alexander, Daniel Alexander, Frederik Barkhof y Spiros Denaxas. «Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records». En: *Studies in health technology and informatics* 270 (jun. de 2020), págs. 499-503. DOI: 10.3233/SHTI200210.

Apéndice A

Encuesta con Escala CES-D

La encuesta que se diseño para ser implementada en los formularios de Google expuestos en la sección 4.1.2 posee distintas preguntas, en este apéndice se presentarán las preguntas que contenían los formularios y apoyaron a la medición de los casos de depresión.

Las preguntas y escalas pueden ser observadas en el trabajo [2]. **Encuesta de datos del alumno**

1. ¿Es estudiante local o foráneo?
2. ¿Cuál es su edad? (Escribir el número únicamente)
3. ¿Actualmente se dedica únicamente a estudiar o trabaja?
4. ¿Cuál es su promedio de estudios actual? Sin importar este semestre (Escribir el número con decimales)
5. ¿Cuál es su año de ingreso? (Escribir el número entero únicamente)
6. ¿Cuenta con materias atrasadas a tu semestre? En caso de que sí indicar cuántas (Escribir el número únicamente)
7. ¿Cuál es su estado civil?
8. ¿Cuál es su género?
9. ¿Tiene dependientes económicos?
10. En promedio ¿Cuántas horas dedica diariamente al estudio de sus materias? (Escribir el número únicamente)

Encuesta propuesta inspirada en CES-D-R

La escala CES-D-R se enfoca en medir situaciones en un periodo de tiempo, es decir, cuánto tiempo una persona ha sentido o ha estado en alguna de las siguientes situaciones:

- Tengo sensaciones de entumecimiento o temblor en mi cuerpo
- Me siento triste
- Duermo pero no descansaba
- Me siento una mala persona
- Pierdo interés en mis actividades

- Duermo más de lo habitual
- Sentía mis movimientos lentos
- Me sentía disgustado conmigo mismo
- Perdía peso sin intentarlo
- Me siento temeroso
- Pienso que soy un fracaso en mi vida
- He sido poco amigable
- Hablo menos de lo usual
- Nada me divierte o me da placer
- Me estreso con facilidad
- Me siento solo
- Tengo dificultad para concentrarme
- Me molesto de cosas que no me molestaban