PROYECTO EDA

NURIA RUBIO FERNÁNDEZ

REPORT ANÁLISIS DE DATOS (STUDENT DEPRESSION)

INTRODUCCIÓN

Se trata de un análsis de datos sobre los elementos que afectan en la depresión y las tasas de suicidio en estudiantes de universidad desde un estudio realizado en la India, sur de Asia.

Objetivo:
Este análisis tiene como objetivo descubrir patrones e identificar posibles factores de riesgo asociados con la depresión estudiantil. Al explorar las relaciones entre los hábitos de estilo de vida, las presiones académicas y el bienestar mental, este estudio busca proporcionar información valiosa sobre los factores que afectan la salud mental de los estudiantes.
Este conjunto de datos sirve como un recurso valioso para comprender los desafíos que enfrentan los estudiantes y ayuda a diseñar intervenciones efectivas para promover el bienestar mental.

Realización:
El análisis de datos ha sido llevado a cabo utilizando cuadernillos de Jupyter Notebooks, organizados en cinco secciones diferenciadas:

1. Descarga de datos
2. Primera observación
3. Análisis exploratorio
4. Tratamiento de datos
5. Análisis descriptivo

El proceso se ha realizado en Python, utilizando diversas bibliotecas especializadas. Para el tratamiento de datos, se han empleado las librerías *NumPy* y *Pandas*, mientras que para la generación de gráficos se han utilizado *Seaborn, Matplotlib y Plotly*, entre otras. Finalmente, para los análisis estadísticos, se ha recurrido a la librería *SciPy*.

DATOS

El siguiente conjunto de datos en formato csv ha sido extraído de la página web ‘Kaggle’, este csv se compone de 27901 filas y 18 columnas (variables) en las cuales no existen jerarquías ni estructuras anidadas.
Para poder trabajar de manera más dinámica y sencilla se pasaron los datos a un Data Frame, el cual utiliza un espacio de 3.8+ MB de memoria.

Sus columnas / variables son las siguientes:

- 'id': número de clasificación único de cada estudiante
- 'Gender' : género con el que se clasifica el estudiante

- 'Age': edad exacta
- 'City': ciudad en la que está realizando sus estudios
- 'Profession': a que se dedica en ese momento cada persona (todos son estudiantes)
- 'Academic Pressure': presión académica calificado en rangos de 0 a 5, a mayor número, mayor presión
- 'Work Pressure': presión en el trabajo
- 'CGPA': promedio de calificaciones u otros puntajes académicos de cada estudiante calificado de 0 a 10
- 'Study Satisfaction': satisfacción de estudio,también calificado en rangos de 0 a 5, cuanto más alto es el número más alta es la satisfacción
- 'Job Satisfaction': satisfacción en el trabajo
- 'Sleep Duration': duración del sueño, calificado por rangos de horas
- 'Dietary Habits': dietas alimenticias, clasificadas en 3 categorías: sana (healthy), insana (unhealthy) y moderada (moderate)
- 'Degree': grado de educación y la carrera que están estudiando
- 'Have you ever had suicidal thoughts ?': columna de únicamente dos respuestas (sí o no) que contesta a la pregunta de si alguna vez han tenido pensamientos suicidas
- 'Work/Study Hours': horas empleadas al estudio o trabajos de la universidad
- 'Financial Stress': estrés financiero
- 'Family History of Mental Illness': columna de únicamente dos respuestas (sí o no) que responde a la pregunta de si tienen historia familiar de problemas de salud mental.
- 'Depression': columna de únicamente dos respuestas (sí o no) que contesta a la pregunta si sufren de depresión en ese momento.

1- PRIMERA OBSERVACIÓN

Al realizar la primera observación del data Frame se puede observar que no cuenta con datos faltantes, valores duplicados o errores evidentes, en cambio sí existen valores nulos. Son únicamente tres valores y se encuentran en la columna llamada 'Financial Stress' por lo que no supone un problema significativo para el análisis.
No obstante, en algunas variables el tipo de dato no era el adecuado. En su mayoría, el tipo de dato era 'object'. Además, en variables como la edad, se esperaba un tipo de dato numérico entero, pero se presentaban valores flotantes.

Como parte del proceso de limpieza del dataframe, se eliminaron dos columnas que se consideraron irrelevantes para el análisis y que podrían afectar negativamente los resultados.
Estas columnas, relacionadas con el puesto de trabajo, contenían principalmente valores nulos, ya que la mayoría de los estudiantes de los que se recabaron datos no se encuentran empleados. Esto hacía que dichas columnas estuvieran compuestas casi en su totalidad por ceros, lo que las hacía innecesarias para el cálculo de medias u otros análisis estadísticos.
Las columnas eliminadas fueron 'Job Satisfaction' y 'Work Pressure'.
Tras esta eliminación, el conjunto de datos fue guardado en un archivo CSV distinto al original, que será utilizado a partir de este momento.

2- ANÁLISIS EXPLORATORIO

Antes de empezar con este análisis, se colocó la columna ‘id’ como índice del data frame para que a la hora de graficar variables no aparezcan como datos numéricos los id ya que son irrelevantes y no tienen nada que ver con el análisis deseado.
El análisis exploratorio se divide en varias partes para que el análisis lleve más orden y claridad; esas partes son:

- Análisis univariado: se grafican las variables numéricas y después las categóricas
- Análisis bivariado: se grafican las relaciones entre variables
- Resumen detallado

2.1- Análisis Univariado

Como primera observación, el conjunto de datos se compone de un total de 18 columnas, distribuidas de la siguiente manera: 8 columnas de tipo 'object', 8 columnas de tipo 'float' y 2 columnas de tipo 'int'.
Entre las columnas de tipo 'object', dos de ellas son binarias, lo que significa que contienen solo dos posibles valores. Las columnas de tipo 'int', por su parte, están formadas por valores numéricos discretos, mientras que las columnas de tipo 'float' contienen datos numéricos continuos, con decimales.
Es importante señalar que la única columna que presenta valores faltantes es la de 'Financial Stress', la cual tiene un total de 3 valores ausentes.

A continuación, con las variables numéricas se han generado diversos gráficos con el objetivo de visualizar de manera más clara la correlación entre varias columnas. Los tipos de gráficos utilizados incluyen histogramas, diagramas de densidad, boxplots y pruebas de normalidad.
Tras realizar estas pruebas y analizar los gráficos, se observa que los datos no siguen una distribución normal, lo que lleva al rechazo de la hipótesis nula. Además, se detecta que los únicos valores atípicos significativos se encuentran en la columna 'age' (edad).

Se ha realizado un análisis similar con las variables categóricas, utilizando gráficos que representan tanto las frecuencias relativas como las absolutas, gráficos de barras, diagramas de Pareto y análisis de cardinalidad.
Las principales conclusiones obtenidas a partir de estos gráficos son las siguientes: en primer lugar, hay un mayor número de estudiantes hombres que mujeres; en segundo lugar, la mayoría de los estudiantes duerme menos de 5 horas y sigue dietas alimenticias poco saludables; finalmente, se observa que existe una proporción considerablemente mayor de estudiantes que han tenido pensamientos suicidas en comparación con aquellos que no los han experimentado, un dato que resulta alarmante y de gran relevancia para considerar en futuros análisis.

2.2 Análisis Bivariado

El análisis bivariado se ha llevado a cabo con el objetivo de examinar las relaciones entre diferentes variables. En el caso de las relaciones entre variables numéricas, se han utilizado matrices de correlación, diagramas de dispersión (scatter plots) y mapas de calor (heat maps) para visualizarlas.

Un hallazgo notable es la fuerte relación entre la depresión y la presión académica ('academic pressure'), entre la depresión y el estrés financiero ('Financial Stress'), y también entre la depresión y la edad. Estos resultados sugieren que estos factores tienen una influencia significativa en el desarrollo de la depresión.

En cuanto a las relaciones entre variables categóricas, se han utilizado tablas de contingencia, la prueba Chi-cuadrado y análisis de correspondencia para ilustrar los vínculos. A partir de estos análisis, se observa que, en general, los hombres tienden a seguir dietas menos saludables, mientras que las mujeres optan por una alimentación más equilibrada. Además, se encuentra que los estudiantes que duermen menos de 5 horas diarias y aquellos con dietas poco saludables son los que más han reportado pensamientos suicidas. Esto sugiere que tanto las horas de sueño como la alimentación tienen una influencia considerable sobre la salud mental.

Finalmente, las relaciones entre columnas numéricas y categorías se han analizado mediante diagramas de cajas (box plots) por grupo, gráficos de violín (violín plots) y el análisis de varianza (ANOVA), con el objetivo de identificar diferencias significativas entre las categorías y las variables numéricas.

2.3 Tabla resumen

Mediante el uso de una función en Python, se ha generado una tabla que presenta como índice los nombres de cada una de las variables del conjunto de datos. En las columnas de esta tabla se incluyen diversos detalles sobre cada variable, tales como el tipo de dato, el tipo de variable, la cardinalidad, la distribución, la cantidad de valores faltantes y su porcentaje, la cantidad de valores atípicos (outliers) y su porcentaje, el rango de valores, la moda, la media, la asimetría, la curtosis, así como los resultados de las pruebas de normalidad realizadas para cada variable.
Esta tabla proporciona un resumen general y visual de los gráficos y las pruebas previas realizadas en el análisis, permitiendo una comprensión más clara de las características del conjunto de datos.

3- TRATAMIENTO DE DATOS

Dado que el conjunto de datos no presenta valores faltantes, duplicados ni errores evidentes, no se ha considerado necesario eliminar, imputar ni tratar ningún dato.

En cuanto a los valores atípicos (outliers), estos solo se encuentran en dos columnas: 'age' (edad) y 'CGPA' (promedio de calificaciones u otros puntajes académicos). Dado que dichos valores atípicos son relevantes para el análisis y se consideran importantes a la hora de realizar mediciones, se ha optado por dejarlos sin modificar ni eliminar.

En relación con la columna 'age', se observa que la mayor concentración de edades de los estudiantes se encuentra entre los 21 y los 30 años, siendo 25 años la mediana. Sin embargo, también existen estudiantes mayores, con edades que oscilan entre los 45 y los 60 años, lo que resulta notable y constituye un dato relevante.

Por otro lado, en la columna 'CGPA', el promedio de calificaciones de los estudiantes varía entre 6 y 8,5, aunque se ha identificado un valor atípico significativo con un puntaje de 0, lo cual también resulta destacable.

4- ANÁLISIS DESCRIPTIVO

La última sección de este análisis se compone de estadísticas descriptivas, que incluyen medidas de tendencia central, dispersión, forma, así como el análisis de percentiles, y visualizaciones, como series temporales, gráficos de composición, diagramas matriciales y mapas de calor.
En primer lugar, se han calculado las medidas de tendencia central para cada una de las columnas del conjunto de datos, específicamente la media, la mediana y la moda. Una observación destacada es que la columna que contabiliza los casos de depresión presenta un valor de 1 tanto como moda como mediana, lo que indica que existe un mayor número de estudiantes que padecen depresión en comparación con aquellos que no la sufren.
A continuación, se han analizado las medidas de dispersión.
Se observa que, en términos de varianza y desviación estándar, la variable con mayor dispersión de valores es la edad, seguida por las horas de estudio. Dado que se ha decidido conservar los valores atípicos (outliers), se puede apreciar esta dispersión en el rango intercuartílico, especialmente en la columna de edades.

En el análisis de la forma, se han calculado la asimetría y la curtosis de las variables. Se observa que existen diferentes tipos de asimetría según los valores de las variables. Por ejemplo, la variable "age" (edad), "work pressure" (presión laboral) y "job satisfaction" (satisfacción laboral) presentan una asimetría positiva, mientras que la variable "CGPA" (promedio de calificaciones) muestra una asimetría negativa. Las demás variables presentan distribuciones simétricas. Además, también se han calculado los percentiles correspondientes.
Este análisis detallado de las estadísticas descriptivas permite una comprensión más profunda de la distribución y características de los datos en el conjunto de estudio.

En cuanto a las visualizaciones, se ha graficado la correlación entre diversas columnas con el fin de facilitar un análisis más comprensible mediante el uso de soportes visuales. Para ello, se han generado varios gráficos de composición segmentados por sexo, depresión y pensamientos suicidas. Asimismo, se han elaborado diagramas de dispersión matriciales y mapas de calor interactivos, que han permitido observar, de manera destacada, la clara correlación entre la presión académica, la edad y el estrés financiero con la depresión.
Estos resultados proporcionan una base sólida para extraer conclusiones generales sobre los factores que más influyen en el desarrollo de la depresión, según los datos de este estudio.

CONCLUSIONES GENERALES

El análisis de los factores que afectan la depresión y las tasas de suicidio en estudiantes universitarios revela hallazgos clave sobre los riesgos asociados con la salud mental en este grupo. Se observa una fuerte correlación entre la depresión y variables como la presión académica, el estrés financiero y la edad. En particular, los estudiantes que enfrentan alta presión académica y dificultades económicas son más propensos a experimentar depresión, mientras que la edad también juega un papel significativo en el desarrollo de trastornos emocionales.
Los hábitos de vida, como la calidad del sueño y la dieta, también son factores cruciales. Se encontró que los estudiantes que duermen menos de cinco horas al día y siguen dietas poco saludables tienen una mayor prevalencia de pensamientos suicidas. Además, la historia familiar de enfermedades mentales se asocia con un mayor riesgo de depresión en los estudiantes, sugiriendo un componente hereditario.

En cuanto al tratamiento de los datos, se identificaron valores atípicos en las columnas de edad y promedio de calificaciones, los cuales fueron conservados por su relevancia en el análisis. Las visualizaciones, como los gráficos de composición y mapas de calor, confirmaron las relaciones entre estas variables y proporcionaron una comprensión más clara de los factores que influyen en la salud mental de los estudiantes.

En resumen y dado que la depresión afecta a los pensamientos y actos suicidas este estudio subraya la importancia de abordar tanto las presiones académicas como los factores socioeconómicos y los hábitos de vida para promover el bienestar mental de los estudiantes universitarios. Estos hallazgos pueden ser fundamentales para el diseño de intervenciones efectivas en el ámbito académico.