UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL SANTA FE

Trabajo Práctico Final

Autores:

- Assenza, Micaela – massenza@frsf.utn.edu.ar
- Herrmann, Cristian – cherrmann@frsf.utn.edu.ar

Profesores:

- Ale, Mariel Alejandra – male@frsf.utn.edu.ar
- Pacchiotti, Mauro José – mpacchiotti@frsf.utn.edu.ar

Fecha de entrega: 15/11/2024

Ciencia de datos – 2024

Universidad Tecnológica Nacional – FRSF

Contenido

Conjunto de datos

El conjunto de datos contiene información sobre hábitos de sueño, características físicas y condiciones de salud de individuos, incluyendo variables como género, edad, duración y calidad del sueño, niveles de actividad física y estrés, categoría de IMC, presión arterial, frecuencia cardíaca, pasos diarios y la presencia o ausencia de trastornos del sueño. El objetivo es analizar cómo estos factores se relacionan con los trastornos del sueño y predecir su presencia.

Gráficos de las variables del conjunto de datos

Elegimos algunas variables para visualizar mejor los datos y entender mejor el conjunto de datos.

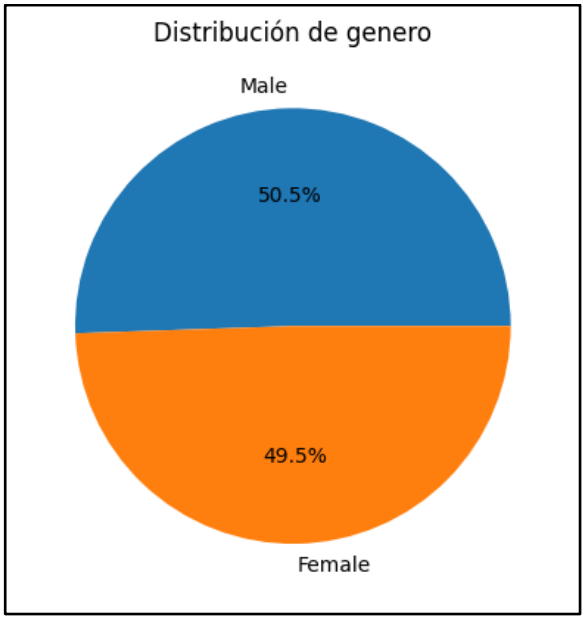


Figura 1: distribución del género en el dataset

Hay un equilibrio casi perfecto en la distribución de género, solo hay una pequeña diferencia del género masculino frente al género femenino

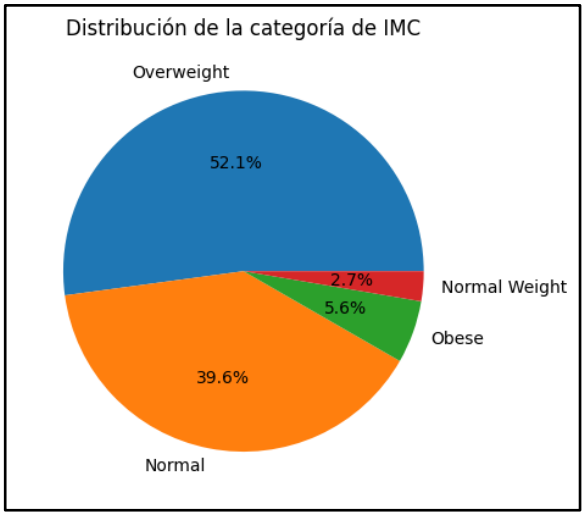


Figura 2: distribución del índice de masa corporal (IMC) del dataset

La mayoría de las personas tienen sobrepeso (52.1%). Esto significa que su peso es mayor de lo que debería ser para su altura. Un porcentaje significativo (39.6%) se encuentra en el rango de peso normal. Un grupo menor (5.6%) está considerado como obeso. Esto significa que su peso es mucho mayor de lo que debería ser para su altura. Un porcentaje muy pequeño (2.7%) está por debajo del peso normal.

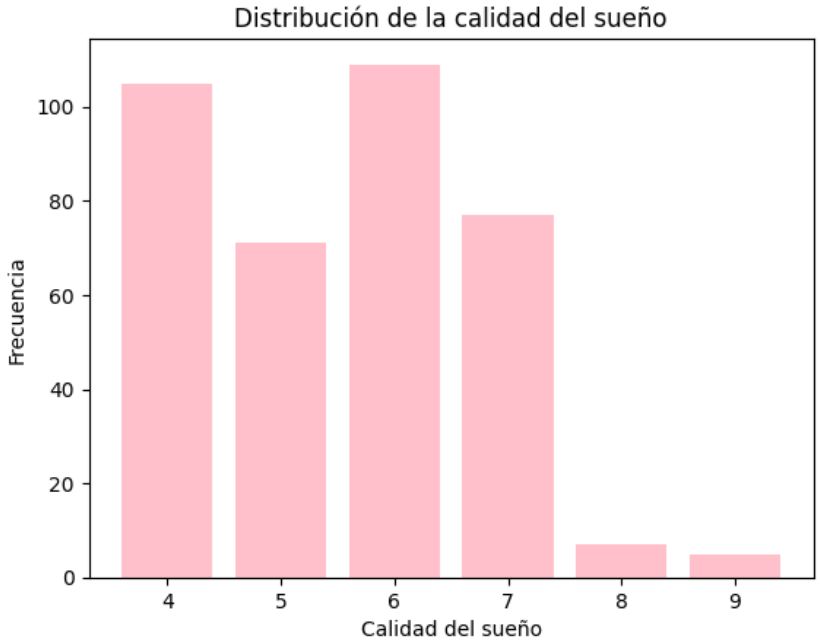


Figura 3: distribución de la calidad de sueño del dataset

El gráfico muestra que la mayoría de las personas reportan una buena calidad de sueño (entre 6 y 7 puntos en lo que suponemos es una escala del 1 al 10). Sin embargo, hay una menor cantidad de personas con sueño muy bueno o malo. La distribución general indica que, en promedio, las personas duermen bien.

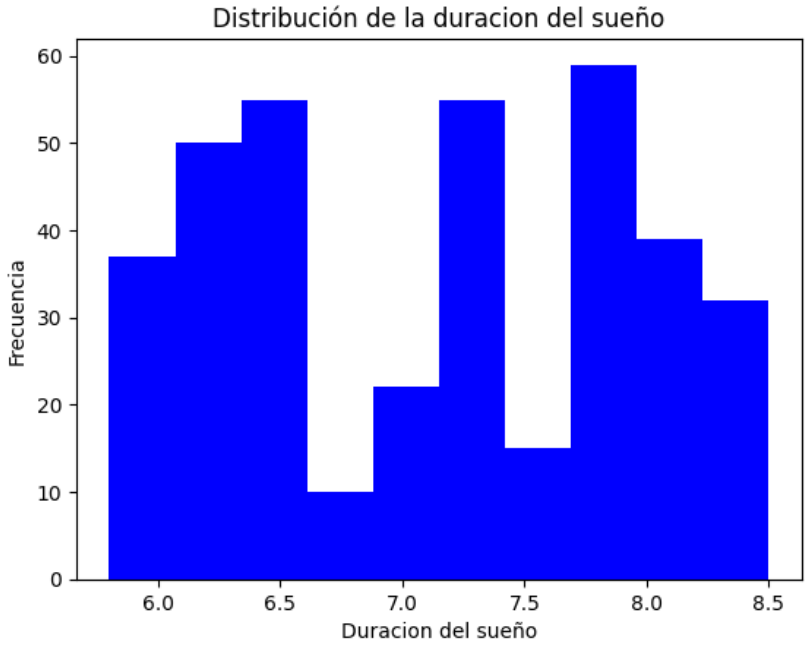


Figura 4: histograma de la duracion del sueño del dataset

La mayoría de las personas reportan que duermen entre 7 y 8 horas, ya que estas son las barras más altas del histograma. Es interesante notar que la distribución parece tener dos picos, uno alrededor de las 6 - 6.5 horas y otro alrededor de las 7.5 - 8.5 horas. Esto podría sugerir que hay dos grupos principales de personas en cuanto a sus hábitos de sueño.

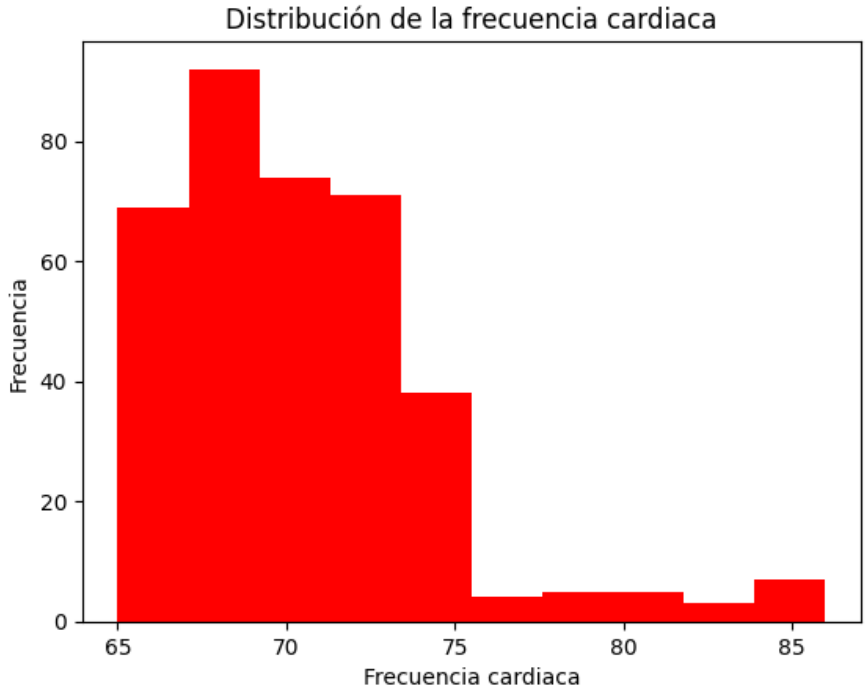


Figura 5: histograma de la frecuencia cardíaca del dataset

La mayoría de las personas tienen una frecuencia cardíaca entre 65 y 70 latidos por minuto, ya que esta es la barra más alta del histograma. Las frecuencias cardíacas registradas se encuentran principalmente entre 65 y 85 latidos por minuto.

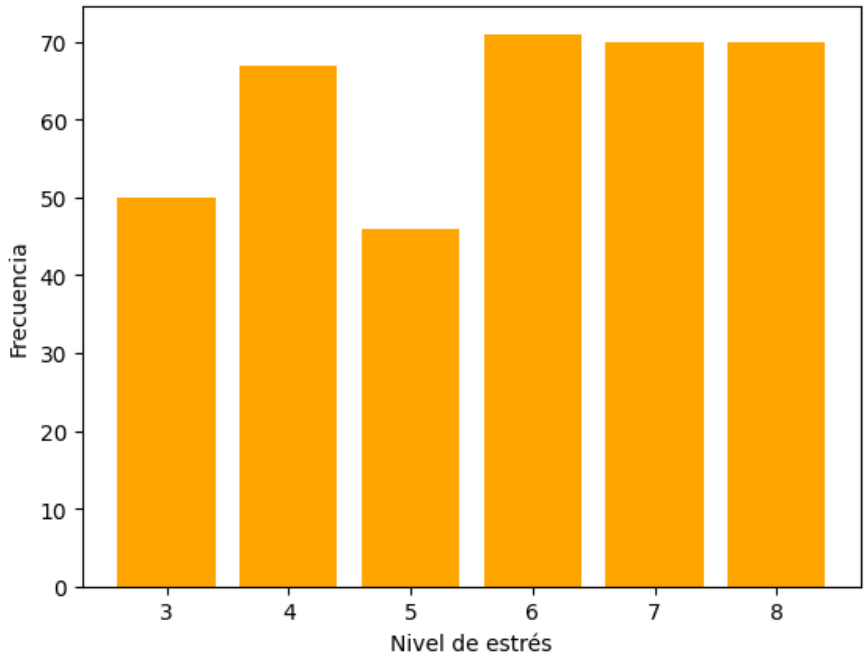


Figura 6: gráfico de barras del nivel de estrés del dataset

No hay un nivel de estrés claramente dominante. La distribución es bastante uniforme entre los niveles 6, 7 y 8, lo que indica una variabilidad considerable en los niveles de estrés entre los individuos. Lo que sí notamos es que la mayoría reporta un nivel de estrés de moderado a alto.

Observando el dataset proporcionado podemos inferir que la variable de salida es el trastorno del sueño.

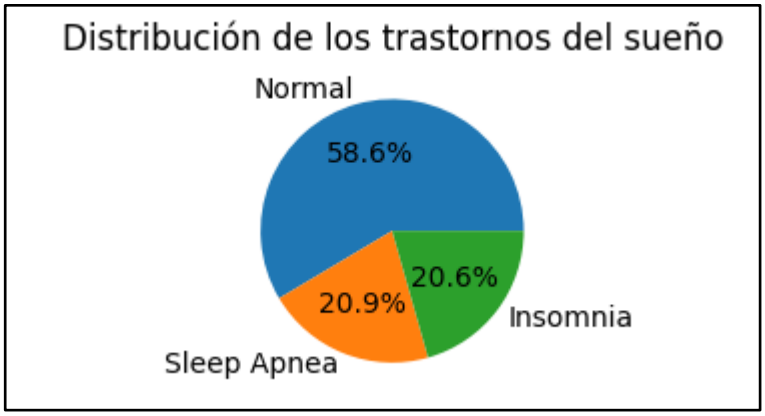


Figura 7: gráfico de torta del trastorno de sueño en el dataset

Podemos clasificarlas en 3 categorías, donde más de la mitad reporta un trastorno normal, es decir, no reporta un inconveniente con el sueño, y luego casi una quinta parte del total del conjunto experimenta insomnio y otra quinta parte padece apnea del sueño.

Procesamiento de los datos

Para poder realizar modelos de clasificación debemos hacer un tratamiento sobre el conjunto de datos.

Modificaciones al dataset

Primero, procedemos a eliminar la columna 'Person ID' ya que solo es un identificador y no proporciona información relevante.

Notamos que la columna 'Blood Pressure' mide la presión de la forma: {Sistólica}/{Diastólica}. Por lo que vemos conveniente separar estos valores en dos columnas nuevas, y eliminar la columna original para no tener redundancia en los datos.

Con respecto a los valores faltantes, notamos un faltante de datos de algunos individuos con respecto a su edad y a su ritmo cardíaco, por lo que luego de realizar estadística descriptiva hemos optado por reemplazar estos datos nulos por su media.

Con respecto a los valores atípicos, observamos lo siguiente con respecto al ritmo cardíaco.

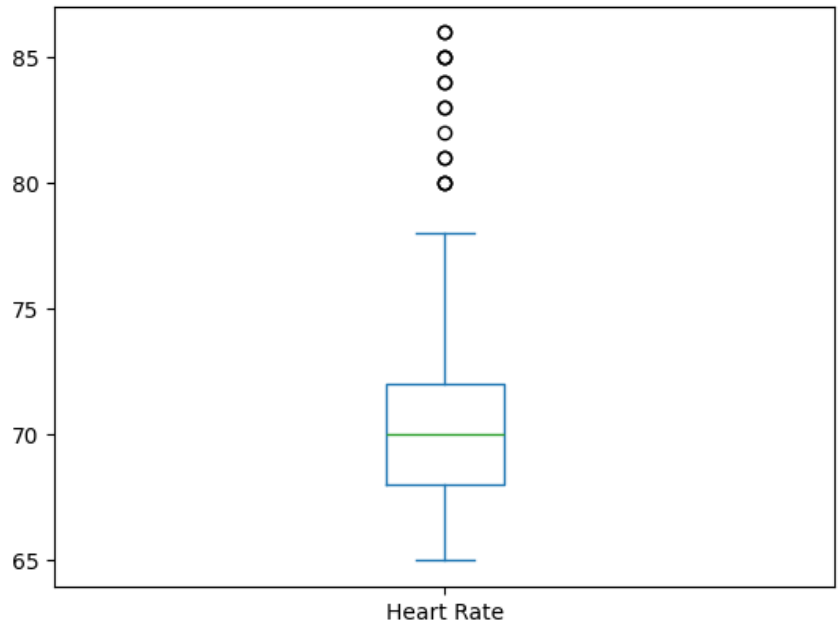


Figura 8: boxplot del ritmo cardíaco del dataset

En un principio, pensamos que los valores atípicos, es decir, aquellos que salen por fuera de los bigotes del boxplot, deberíamos reemplazarlos teniendo en cuenta alguna medida de tendencia central. Sin embargo, al no tener un adecuado conocimiento sobre los trastornos del sueño, sumado a que estos valores podrían tener una correspondencia con algunos de los posibles trastornos analizados, y que desconocemos el estado que se encontraba el individuo a la hora de informar su ritmo cardiaco, decidimos dejar estos valores tal cual, en el dataset. Es decir, no modificamos el conjunto de datos con respecto a los valores atípicos. No descartamos consultar a un especialista, pero al momento de realizar estos análisis hemos considerado que lo correcto es no modificar el conjunto.

Procesamiento de variables categóricas

Con respecto a las variables ‘Gender’ ‘Occupation’, ‘BMI Category’ y ‘Sleep Disorder’, las hemos codificado de forma que cada valor único este representado con una categoría numérica. De esta forma podemos representarlas de forma matemática.

Correlación entre variables

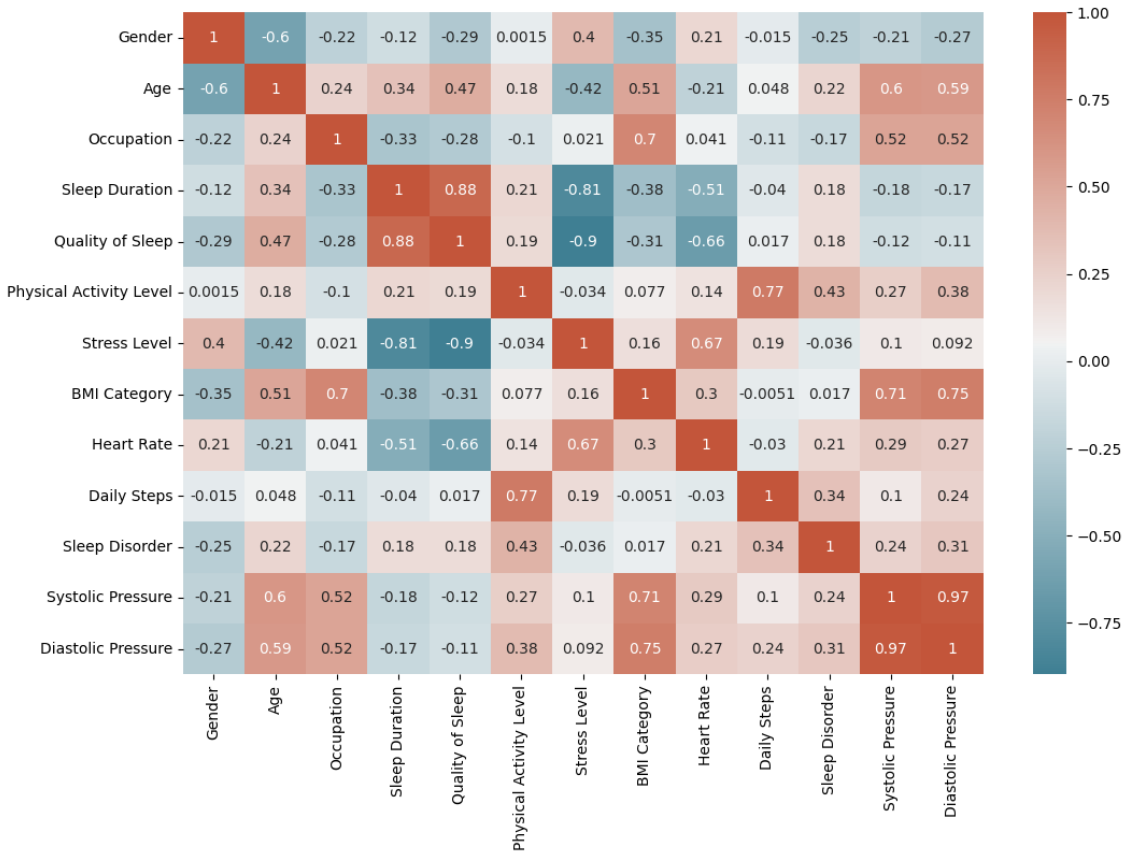


Figura 9: matriz de correlación de las variables

En la figura 9 vemos la matriz de correlaciones de las variables. Notamos variables con alta correlación. Decidimos establecer un umbral de 0.85 para la eliminación de variables correlacionadas, y de cada par de variables, hemos elegido la que tiene mayor promedio de correlación.

El resultado de este análisis determinó que se eliminan las variables ‘Quality of Sleep‘ y 'Diastolic Pressure' (variable que resultó de separar en dos ‘Blood Pressure’).

Balance del conjunto de datos

Si vemos el gráfico de torta de la variable de salida (figura 7) podemos observar que hay un notorio desbalance en los datos, que se aprecia mejor en la figura 10

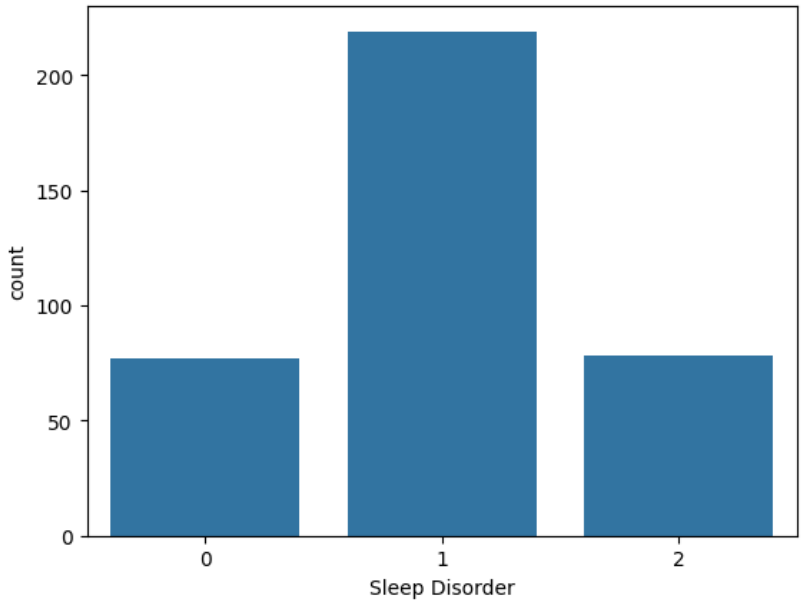


Figura 10: balance del dataset previo a un tratamiento

Optamos por utilizar OverSample para aumentar la cantidad de muestras de las clases minoritarias para alcanzar a la más abundante. El resultado se observa en la figura 11

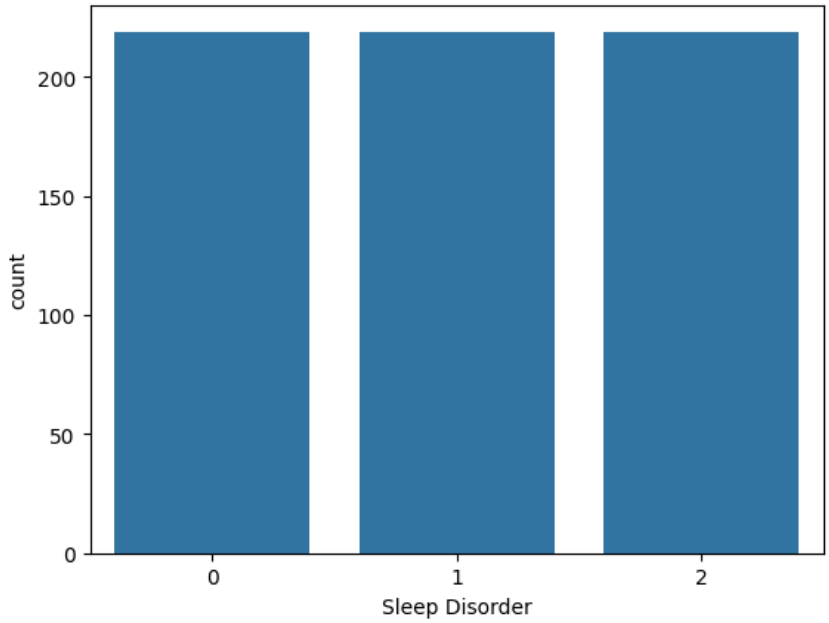


Figura 11: balance del dataset luego del oversampling

Escalado de características

En este paso decidimos realizar un escalado min-max para que todas las columnas estén normalizadas en el rango de 0 a 1.

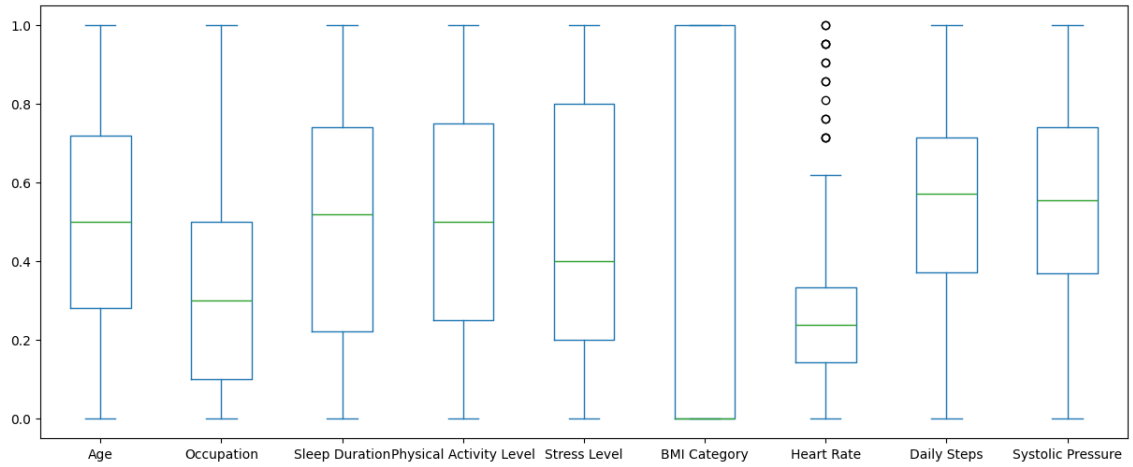


Figura 12: botplox de las variables

Implementación y evaluación de los modelos

Luego del procesamiento de datos implementaremos dos modelos de clasificación. Teniendo en cuenta que nuestro conjunto de datos no es grande creemos que las mejores opciones son:

- K-Nearest Neighbors (KNN): Adecuado para pocos datos, fácil de implementar y entender. Para entrenar el modelo definimos una grilla donde establecemos los vecinos.
- Bosques Aleatorios (Random Forest): Al tener pocos datos, el entrenamiento no requiere mucho tiempo por lo que podremos definir una grilla donde modificaremos hiperparametros como la profundidad de los árboles, la cantidad de aprendices débiles, etc.

KNN

Hiperparámetros: { 'n_neighbors': [3, 5, 7, 9] }

Bosques Aleatorios

hiperparametros: {

'n_estimators' : [100, 200, 300],

'max_features': [5,6,7,8],

'max_depth' : [2,4,5,6],

'random_state' : [18]

}

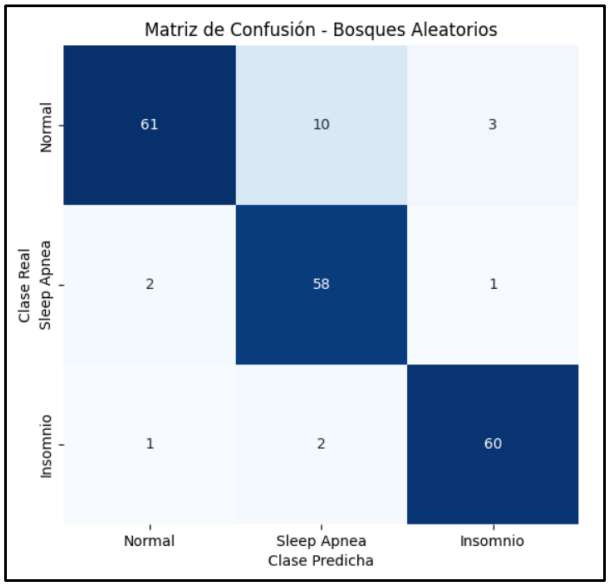
Métricas

Utilizamos f1-Score y accuracy para establecer las métricas. Además para los bosques aleatorios se anexa un apartado de código para graficar el árbol que se genera.

Resultados Obtenidos

A continuación detallamos los resultados que obtuvimos al hacer pruebas en los modelos.

Para Bosques Aleatorios



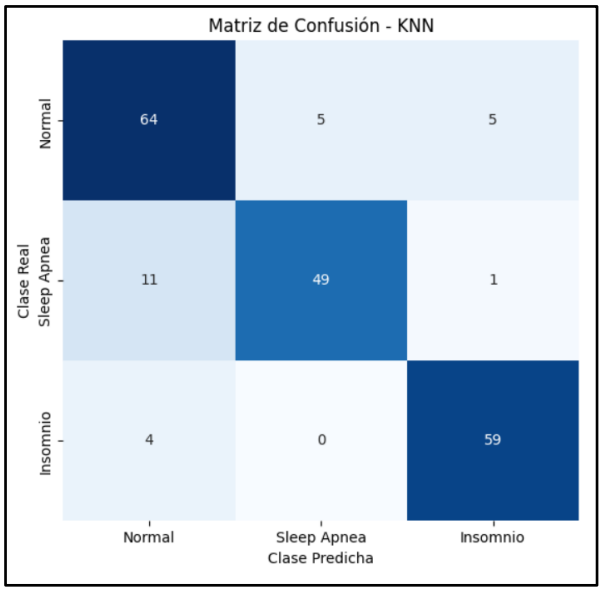
	precision	recall	f1-score	support
0	0.95	0.82	0.88	74
1	0.83	0.95	0.89	61
2	0.94	0.95	0.94	63
accuracy			0.90	198
macro avg	0.91	0.91	0.90	198
weighted avg	0.91	0.90	0.90	198

El modelo muestra un rendimiento satisfactorio en la clasificación de las tres categorías: Normal, Sleep Apnea e Insomnio.

- Clase Normal: El modelo identificó correctamente 61 de 74 casos, lo que representa una tasa de acierto del 82.4%. Sin embargo, hubo cierta confusión con la clase Sleep Apnea, ya que 10 casos fueron clasificados erróneamente en esta categoría.
- Clase Sleep Apnea: Esta clase fue correctamente clasificada en 58 de 61 ocasiones, alcanzando una precisión del 95.1%. Solo 3 casos fueron mal clasificados, con dos identificados incorrectamente como Normal y uno como Insomnio.
- Clase Insomnio: Se observó un alto desempeño en la clasificación de Insomnio, con 60 de 63 casos correctamente identificados (una precisión del 95.2%). Solo 3 instancias fueron clasificadas incorrectamente, con un leve sesgo hacia la clase Sleep Apnea y Normal.

El modelo demuestra un rendimiento sólido en general, con una precisión particularmente alta para las clases Sleep Apnea e Insomnio, ambas superiores al 95%. No obstante, el modelo presenta un mayor grado de confusión al diferenciar entre las clases Normal y Sleep Apnea, donde se observó un número considerable de falsos positivos y negativos

Para KNN



	precision	recall	f1-score	support
0	0.81	0.86	0.84	74
1	0.91	0.80	0.85	61
2	0.91	0.94	0.92	63
accuracy			0.87	198
macro avg	0.88	0.87	0.87	198
weighted avg	0.87	0.87	0.87	198

- Clase Normal: El modelo identificó correctamente 64 de 74 casos, lo que equivale a una precisión del 86.5%. Sin embargo, hay cierta confusión con las clases Sleep Apnea e Insomnio, ya que hubo 5 casos clasificados erróneamente en cada una de estas categorías.
- Clase Sleep Apnea: La clase Sleep Apnea fue correctamente clasificada en 49 de 61 instancias, lo que representa una precisión del 80.3%. Sin embargo, se observó una cantidad considerable de errores al clasificar como Normal, con 11 casos confundidos.
- Clase Insomnio: El modelo muestra un excelente rendimiento en la clasificación de la clase Insomnio, con 59 de 63 casos correctamente identificados, lo que equivale a una precisión del 93.7%. Hubo solo 4 casos clasificados erróneamente como Normal, sin confusión con Sleep Apnea.

En general, el modelo tiene un buen desempeño, particularmente en la identificación de la clase Insomnio, con una alta precisión. Sin embargo, presenta desafíos al diferenciar entre Normal y Sleep Apnea, como se evidencia en el mayor número de falsos positivos y negativos entre estas dos categorías.

Al analizar las Matrices de Confusión y comparar a su vez los valores dados de f1 y accuracy notamos un mejor rendimiento en el Bosque Aleatorio contra el KNN con mejores porcentajes de aciertos.