**Informe**

***Origen datos:***

Los datos se han extraído de una web de bases de datos regularizadas por lo que los datos son fiables.

https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset?resource=download

***Problema:***

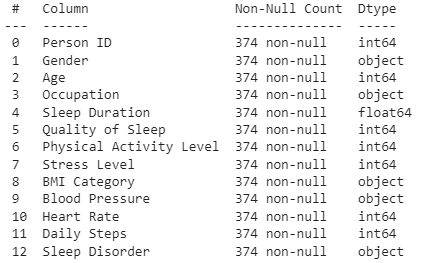
Hoy en día entre el estilo de vida acelerado de la sociedad actual con vidas ocupadas y estresantes, el uso de dispositivos electrónicos antes de dormir que emiten luz azul y que impide la producción de la hormona del sueño (melatonina), los trabajos nocturnos con horarios irregulares y la ingesta de cafeína cada vez más normalizada en nuestra sociedad, creemos que un modelo que prediga si puedes tener alguna enfermedad del sueño podría ser muy útil.

***Aplicación:***

Medicina y salud: Un modelo que prediga la probabilidad de tener un trastorno del sueño puede ayudar a los médicos a evaluar y diagnosticar diferentes hábitos para mejorar esto, también puede ser utilizado en estudios de investigación del sueño.

Dispositivos y aplicaciones para el sueño: Muchos dispositivos y aplicaciones inteligentes de seguimiento del sueño registran datos sobre los patrones de sueño de los usuarios, como la duración del sueño, la frecuencia cardíaca y los movimientos. Un modelo que prediga la probabilidad de tener un trastorno del sueño puede utilizar estos datos para proporcionar retroalimentación y recomendaciones personalizadas para mejorar la calidad del sueño.

***Descripción variables y feature engineering:***



Como vemos en la foto tenemos 5 columnas object 1 float y las demás int y sobre la que queremos predecir es la float entonces veremos los datos y haremos algunas transformaciones:

* La columna ‘BMI Category’ tiene un valor duplicado que es Normal y Normal Weight por lo que los unimos en una sola variable ‘Normal weight’

['Overweight' 'Normal' 'Obese' 'Normal Weight']

['Overweight' 'Normal Weight' 'Obese']

* La columna ‘Blood Pressure’ tiene 2 valores separados por ‘ / ’ por lo que dividiremos esta columna en 2 columnas una con presión sistólica o máxima y otra con presión diastólica o mínima

['126/83' '125/80' '140/90' '120/80' '132/87' '130/86' '117/76' '118/76' '128/85' '131/86' '128/84' '115/75' '135/88' '129/84' '130/85' '115/78' '119/77' '121/79' '125/82' '135/90' '122/80' '142/92' '140/95' '139/91' '118/75']



* Para el entrenamiento pasaremos las columnas: "Sleep Disorder", "BMI Category", "Occupation", "Gender" por un encoder en este caso pd.get\_dummies para convertirlas en diferentes columnas con valores numéricos
* Pondremos de índice la columna ‘Person ID’ para no tener un índice duplicado y que el índice sea el ID de cada persona.

***Modelos usados tiempo de entrenamiento comparar métricas y***

***elección modelo:***

Hemos probado 4 modelos de clasificación y como tenemos pocos datos el tiempo de entrenamiento es de segundos y sus métricas de entrenamiento por cada clase son estás:

SVC

Precisión: [0.88 0.89 0.91] 0.8920977195909812

Recall: [0.78 0.97 0.8 ] 0.8514736405147364

SGDClassifier

Precisión: [0.21 0.56 0. ] 0.25711569132621764

Recall: [0.2 0.8 0. ] 0.33378995433789954

RandomForest

Precisión: [0.79 0.93 0.89] 0.870485778203899

Recall: [0.8 0.95 0.83] 0.8617960426179604

XGBClassifier

Precisión: [0.81 0.95 0.86] 0.8762492228687287

Recall: [0.78 0.97 0.85] 0.868140307181403

ANN

Accuracy on train set: 90.54%

Con estas métricas de entrenamiento nos quedaremos con el RandomForest, con el XGBClassifier, con el SVC y la red neuronal para probarlos con el test porque dan muy bien los 3, ahora los probaremos con el test:

SVC

Precisión: [0.9 0.91 0.78] 0.8614814814814814

Recall: [0.82 0.93 0.78] 0.8424888150915547

F1: [0.86 0.92 0.78] 0.8512798512798513

RandomForest

Precisión: [0.91 0.93 0.78] 0.872791845394585

Recall: [0.91 0.93 0.78] 0.872791845394585

F1: [0.91 0.93 0.78] 0.872791845394585

XGBClassifier

Precisión: [0.87 0.93 0.76] 0.8552593163531047

Recall: [0.91 0.93 0.72] 0.8542733268760666

F1: [0.89 0.93 0.74] 0.854417627020366

ANN

Accuracy on test set: 90.27%

De los modelos probados elegiremos la red neuronal porque tiene mejor precisión.