Tecnológico de Monterrey

# Análisis y Reporte sobre el desempeño del modelo

Maria Fernanda Argueta Wolke A00830194

Inteligencia artificial avanzada para la ciencia de datos I

(Gpo 102)

# Tabla de contenidos

# 1. Introducción

En este portafolio se encuentra el proceso que se llevó a cabo para la optimización de un modelo de regresión logística. Con el fin de realizar predicciones sobre trastornos del sueño utilizando el conjunto de datos Sleep Health. El objetivo principal es mejorar el rendimiento del modelo utilizando diferentes técnicas y procesos de limpieza de datos, ajuste de estos y refinamiento del modelo propuesto.

### Descripción del Conjunto de Datos

Se estará trabajando con el conjunto de datos *Sleep Health, extraído de la plataforma Kaggle. El cual* contiene información sobre varios factores relacionados con la salud del sueño de los individuos. Las características incluyen variables como la duración del sueño, calidad del sueño, nivel de actividad física, nivel de estrés, categoría de IMC (Índice de Masa Corporal), presión arterial, frecuencia cardíaca y otros indicadores de salud. La variable objetivo es la presencia o ausencia de un trastorno del sueño, la cual se transformó en una variable binaria para facilitar la modelización.

### Objetivo

El propósito de este proyecto es identificar y aplicar los cambios necesarios al modelo de regresión logística para maximizar su precisión y capacidad de generalización. Para lograr esto, se llevaron a cabo varias etapas clave como la limpieza de datos, ingeniería de variables, creación y refinamiento del modelo. Se compararon los resultados del modelo optimizado con los de un modelo base para determinar si los ajustes realizados contribuyeron a una mejora significativa en el rendimiento, implementando técnicas de regularización y ajustes alternativos para evitar el sobreajuste y asegurar un modelo robusto.

# 2. Solución

### Versión Anterior

En el portafolio anterior se trabajó con este mismo set de datos y se realizaron predicciones utilizando regresión logística. Sin embargo, en esta primera versión se utilizó de forma diferente los datos para realizar predicciones. Se procuró predecir la calidad del sueño en base a la duración del sueño y nivel de estrés. Sin embargo, la variable de calidad de sueño era un valor en escala de 1 a 10. En el que 10 define a la mejor calidad de sueño posible.

Se hicieron los ajustes necesarios para poder utilizar la calidad del sueño como una variable binaria. Sin embargo, se observó que el modelo tenía un nivel de accuracy muy elevado lo cual se le atribuyó a la manipulación de la variable ¨calidad de sueño¨. Por otro lado, se tomaron en cuenta muy pocas variables para crear el modelo de regresión logística, dichas variables fueron escogidas sin mucho fundamento y por conveniencia para no tener que hacer muchos cambios en la base de datos.

### Cambios principales Nueva Versión

En base a los resultados obtenidos, y los nuevos conocimientos adquiridos en cuanto a ciencia de datos y Machine Learning, se decidió hacer ciertos cambios para esta nueva versión del modelo. Primero que nada se estará realizando predicciones en base a la variable de Sleep Disorder. Se observó que esta variable si bien indica el tipo de enfermedad de sueño que posee cada persona, también indica si tiene o no una enfermedad. Por lo que esta variable tiene mucho potencial para ser utilizada para determinar si un paciente tiene o no enfermedad del sueño.

Por otro lado, se hizo un análisis del impacto de cada una de las variables en el resultado de la variable que indica si tiene o no una enfermedad del sueño. Dicho análisis se realizó utilizando la herramienta de Decision Tree, la cual se detalla más adelante en el documento.

Finalmente, se utilizaron métricas de Precisión global, precisión, sensibilidad, puntuación F1, soporte. Las cuales fueron analizadas a profundidad para evaluar el desempeño del modelo. Por otra parte, se realizaron ajustes al modelo y se compararon con dichas métricas para evaluar si estas eran factibles o no. Otro punto muy importante es que se estuvo trabajando con una base de datos dividida en 3, entrenamiento, testeo y validación.

## 2. Metodología

### Limpieza de datos

Previo a la implementación del modelo se hizo el análisis de los datos. Lo primero que se observó fue un problema en la columna de Sleep Disorder la cual es de suma importancia pues en base a esta variable es que se realizarán las predicciones. Se determinó que el problema se daba porque la librería de pandas identificaba el valor de none como vacío. Por lo que se hizo el ajuste a esta columna para que en vez de decir none, dijera ¨No¨.

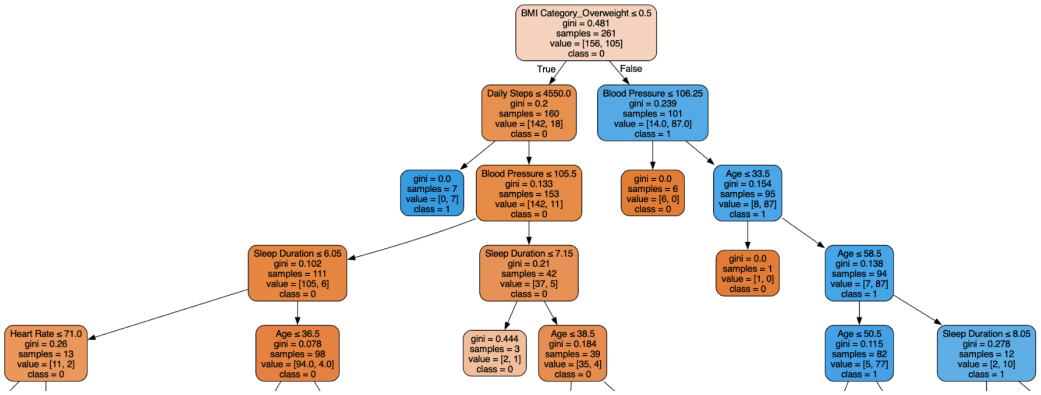
Ya que se pretende realizar predicciones sobre si una persona presenta o no enfermedades del sueño se hizo un ajuste a la columna Sleep Disorder. Por lo que se generó la nueva columna objetivo llamada Has Sleep Disorder. Esta columna indica un 1 si la persona presenta enfermedades del sueño, indicadas previamente en la columna Sleep Disorder como Insomnia o Sleep Apnea. Al igual que indicando un 0 si el valor de la columna es No.

### Transformación de datos

Para tener un mejor manejo de los datos se le aplicó one hot encoding a las variables de BMI Category y Gender. Con lo cual se obtuvo 5 columnas nuevas, 2 correspondientes al bmi dividiéndolo en peso normal, sobrepeso y obeso. Y las últimas 2 correspondientes a sexo masculino y femenino. Se pensó incluir en el one hot encoding a la variable de ocupación. Sin embargo, debido a que esta variable cuenta con muchos posibles campos se tomó la decisión de despreciarla para evitar confusiones en el modelo incluso overfitting.

### Exploración de los datos

Para tomar una decisión sobre las variables a utilizar para el entrenamiento del modelo se hizo uso de la herramienta árbol de decisión. Con lo que se confirmó la eliminación de la variable ocupación. Así como el entendimiento de la importancia del índice de masa corporal. Además se identificó que los principales factores para la aparición de una posible enfermedad de sueño son la edad, el índice de masa corporal, pasos diarios, presión arterial y duración del sueño.




## 3. Modelo

### Selección del modelo

Debido a que era necesario que el modelo tuviera la habilidad de predecir una variable categórica se tomó en cuenta la posibilidad de uso de modelos específicamente categóricos. Entre ellos regresión logística, random forest e incluso XGBoost. Sin embargo, debido a que regresión logística es un modelo simple y efectivo en especial para predicciones binarias se tomó la decisión de utilizarlo. (Marron, 2020)

### Separación de Datos

**Entrenamiento y Validación:** El conjunto de datos se dividió en un 70% para entrenamiento más validación y un 15% para prueba. Dentro del conjunto de entrenamiento, se realizó una segunda división para crear un conjunto de validación que representa el 15% de los datos de entrenamiento.

**Prueba:** El 15% restante de los datos se reservó para la evaluación final del modelo.

### Evaluación del Modelo

**Conjunto de Prueba:** El modelo final se evaluó en un conjunto de prueba independiente. La precisión alcanzó 0.965, con informes de clasificación que mostraron un buen equilibrio entre precisión y recall para ambas clases, indicando una capacidad sólida para generalizar datos no vistos.

**Conjunto de Validación:** El modelo también se evaluó en un conjunto de validación. La precisión fue de 0.946, y el informe de clasificación reveló que el modelo mantuvo un rendimiento consistente, con métricas de precisión y recall que indicaron una buena generalización en datos no utilizados durante el entrenamiento.

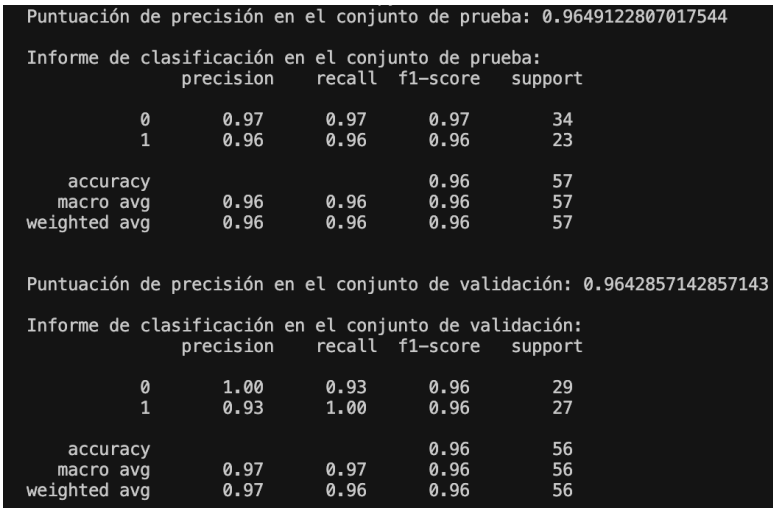
## Diagnóstico del Modelo

## Diagnóstico y Explicación del Grado de Bias (Sesgo)

**Bias (Sesgo): Bajo**
El modelo mostró una alta precisión en ambos conjuntos, prueba y validación, y un buen equilibrio entre precisión y recall. La falta de un sesgo significativo se debe a la capacidad del modelo para generalizar bien a datos no vistos, indicando que no está ajustado de manera inadecuada a los datos de entrenamiento.

## Diagnóstico y Explicación del Grado de Varianza

**Figura No. 1**: *Resultados predicción sin grid search*



```
Puntuación de precisión en el conjunto de prueba: 0.9649122807017544

Informe de clasificación en el conjunto de prueba:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        34
           1       0.96      0.96      0.96        23

    accuracy                           0.96        57
   macro avg       0.96      0.96      0.96        57
weighted avg       0.96      0.96      0.96        57


Puntuación de precisión en el conjunto de validación: 0.9642857142857143

Informe de clasificación en el conjunto de validación:
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        29
           1       0.93      1.00      0.96        27

    accuracy                           0.96        56
   macro avg       0.97      0.97      0.96        56
weighted avg       0.97      0.96      0.96        56
```

**Figura No. 2**: *Resultados predicción con grid search*

```
Accuracy Score on Test Set: 0.9649122807017544

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        34
           1       0.96      0.96      0.96        23

    accuracy                           0.96        57
   macro avg       0.96      0.96      0.96        57
weighted avg       0.96      0.96      0.96        57


Accuracy Score on Validation Set: 0.9464285714285714

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.96      0.93      0.95        29
           1       0.93      0.96      0.95        27

    accuracy                           0.95        56
   macro avg       0.95      0.95      0.95        56
weighted avg       0.95      0.95      0.95        56
```

**Varianza: Bajo**
**Explicación:** La precisión y los valores de F1-score en los conjuntos de prueba y validación fueron consistentes, lo que sugiere que el modelo no presenta una varianza alta. Esto indica que el modelo no está sobreajustado y mantiene un rendimiento estable en datos nuevos.

## Diagnóstico y Explicación del Nivel de Ajuste del Modelo

**Nivel de Ajuste del Modelo: Ajustado (Fitt)**
El modelo está bien ajustado, mostrando un equilibrio adecuado entre sesgo y varianza. Los resultados no indican sobreajuste ni subajuste, ya que, el modelo logró una alta precisión tanto en el conjunto de prueba como en el de validación, y mostró una buena capacidad de generalización.

## Mejora del Desempeño del Modelo

## Técnicas de Regularización y Ajuste de Parámetros

**Uso de Grid Search:** Se utilizó Grid Search para encontrar la mejor combinación de parámetros para el modelo de regresión logística. La búsqueda incluyó la regularización L1 y el ajuste del parámetro C para encontrar el valor óptimo. (Raschka, 2018)

**Resultados:** Los mejores parámetros encontrados fueron C = 100, penalty = '1' y solver = 'liblinear'. La aplicación de estos parámetros mejoró el desempeño del modelo, alcanzando una precisión de 0.965 en el conjunto de prueba y manteniendo un rendimiento sólido en el conjunto de validación con una precisión de 0.946.

Además, se puede observar en la figura No. 1 que al no utilizar gridsearch se tenían problemas en la predicción para el conjunto de datos de verificación. Esto se puede observar específicamente en la precisión de predicción de 0 y en el recall de 1. Demostrando que el

modelo podía estar presentando underfitting, lo cual se arregló con la implementación de gridsearch.

## 4. Conclusiones

Se logró desarrollar un modelo de regresión logística optimizado mediante el uso de Grid Search, lo cual permitió identificar los parámetros más efectivos para el modelo. Este modelo demostró una notable capacidad de generalización, alcanzando una precisión del 96.49% en el conjunto de prueba y del 94.64% en el conjunto de validación. Estos resultados reflejan que el modelo no solo es eficiente con los datos de entrenamiento, sino que también mantiene un alto desempeño en conjuntos de datos nuevos y no vistos.

El proceso de optimización del modelo de regresión logística se llevó a cabo utilizando el conjunto de datos Sleep Health extraído de Kaggle. La metodología incluyó una serie de pasos detallados y rigurosos, tales como la limpieza de datos, la ingeniería de características, y el refinamiento del modelo. Estas etapas contribuyeron significativamente a mejorar tanto la precisión como la capacidad de generalización del modelo.

La limpieza de datos, la transformación de variables categóricas, y la exploración exhaustiva de las características permitieron construir un modelo robusto y fiable para la predicción de trastornos del sueño. La elección de la regresión logística, combinada con técnicas de regularización y ajuste de parámetros, resultó ser efectiva para alcanzar una alta precisión y una capacidad de generalización sólida en los conjuntos de prueba y validación. Este enfoque integral asegura que el modelo no solo sea preciso, sino también que se ajuste adecuadamente a los datos, evitando problemas de sobreajuste y mejorando la capacidad predictiva.

Finalmente, en base a la información proporcionada por el modelo se puede decir que la salud del sueño se ve afectada por diversos factores, sin embargo unos más influyentes que otros. Se puede decir que uno de los principales factores que definen si una persona padece enfermedades del sueño es el índice de masa corporal. Además se destacan como factores influyentes la presión arterial y la actividad física de la persona.

## 5. Referencias


Tharmalingam, L. (s.f.). *Sleep health and lifestyle dataset*. Kaggle. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

Raschka, S. (2018). Aprendizaje automático con Python. Packt Publishing.

Marron, J. S., & McCormick, M. (2020). Selección de modelos en machine learning: Una revisión y direcciones futuras. *Journal of Machine Learning Research*, 21(1), 1-32.