machine learning project

objetivo y contexto

El objetivo del proyecto fue migrar la predicción del Nivel de Ansiedad de una escala ordinal de 10 clases (columna Anxiety Level (1-10)) a un sistema de Clasificación Multiclase de 3 grupos funcionales, dada la ineficacia del modelado inicial.

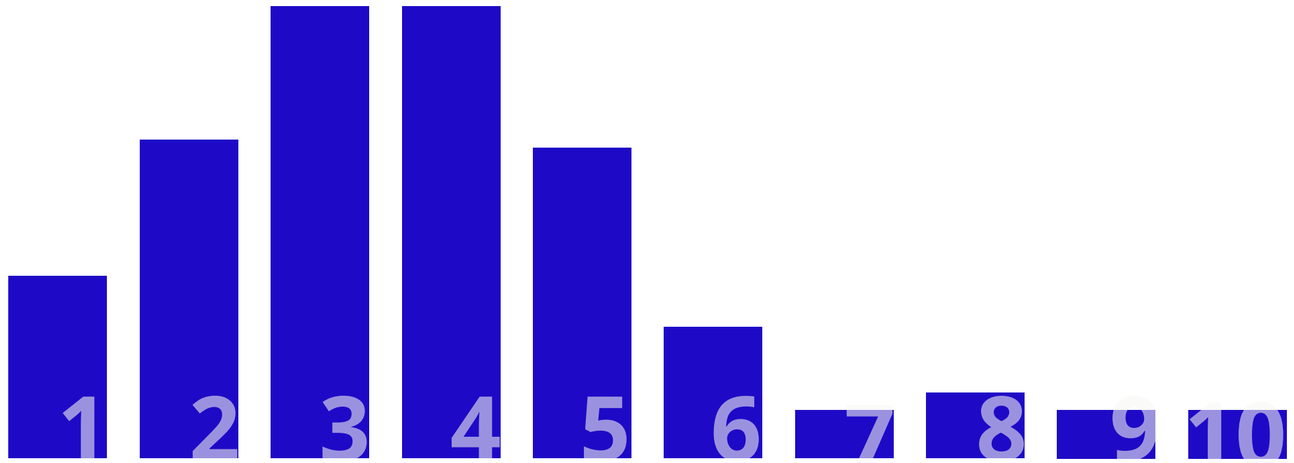
Target Inicial (Y): Anxiety Level (1-10) (Clasificación 10-clases).

Variables de (X): . Este conjunto abarca tanto variables continuas y ordinales como el Nivel de Estrés, Horas de Sueño, Sesiones de Terapia, y el Consumo de Cafeína como variables categóricas previamente codificadas como Smoking, Dizziness, Family History of Anxiety, y las variables One-Hot Encoded de Occupation y Gender. Se observa que las variables con mayor peso o importancia en el análisis son el Nivel de Estrés, las Horas de Sueño y Sesiones de Terapia.

limpieza de los datos

Target

La métrica inicial fue pobre debido a la complejidad de las 10 clases y su inherente desequilibrio. Se realizó un Mapeo a 2 clases:



Resultado: No tiene ansiedad / Si tiene ansiedad

Engineering

Se aplicaron dos técnicas de codificación distintas a las variables categóricas:

Label Encoding transformar variables ordinales o binarias a valores numéricos enteros.

One-Hot Encoding para transformar variables nominales sin orden inherente.

Desequilibrio

Se aplicó **SMOTE** al conjunto de entrenamiento para equilibrar la distribución de la clase objetivo, creando instancias sintéticas de la clase minoritaria.

modelos entrenados

Modelos Lineales

Se utilizan para modelar la relación entre variables mediante una función lineal, siendo la **Regresión Logística** el principal exponente para problemas de clasificación.

Modelos Basados en Árboles

Son poderosos métodos no lineales que dividen el espacio de características en regiones. Se incluyen **Árbol de Decisión, Random Forest** y **Gradient Boosting.**

Modelos Basados en Distancia

Clasifican las observaciones midiendo la proximidad a otros puntos de datos o fronteras. Se evalúan **KNN** y **SVC.**

Exploración Adicional

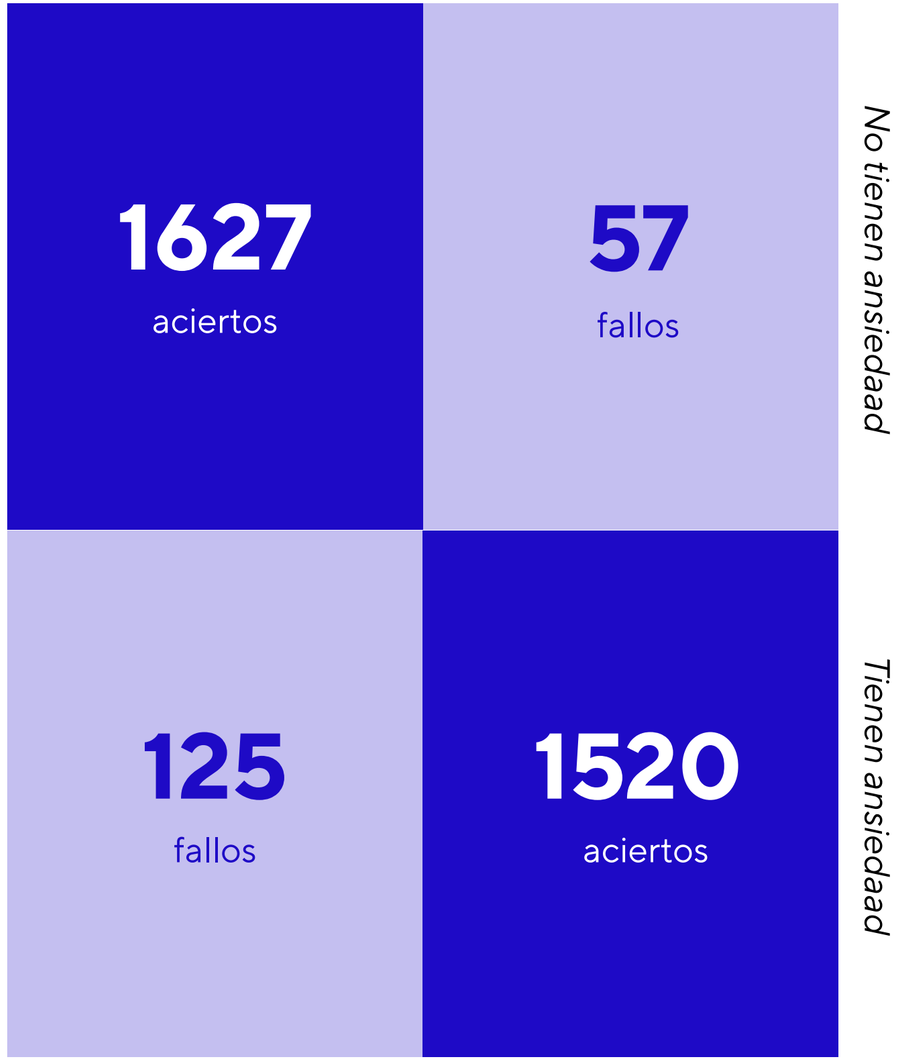
Se empleó el agrupamiento no supervisado **K-Means** para entender la estructura intrínseca de los datos y su potencial como técnica exploratoria o de feature engineering.

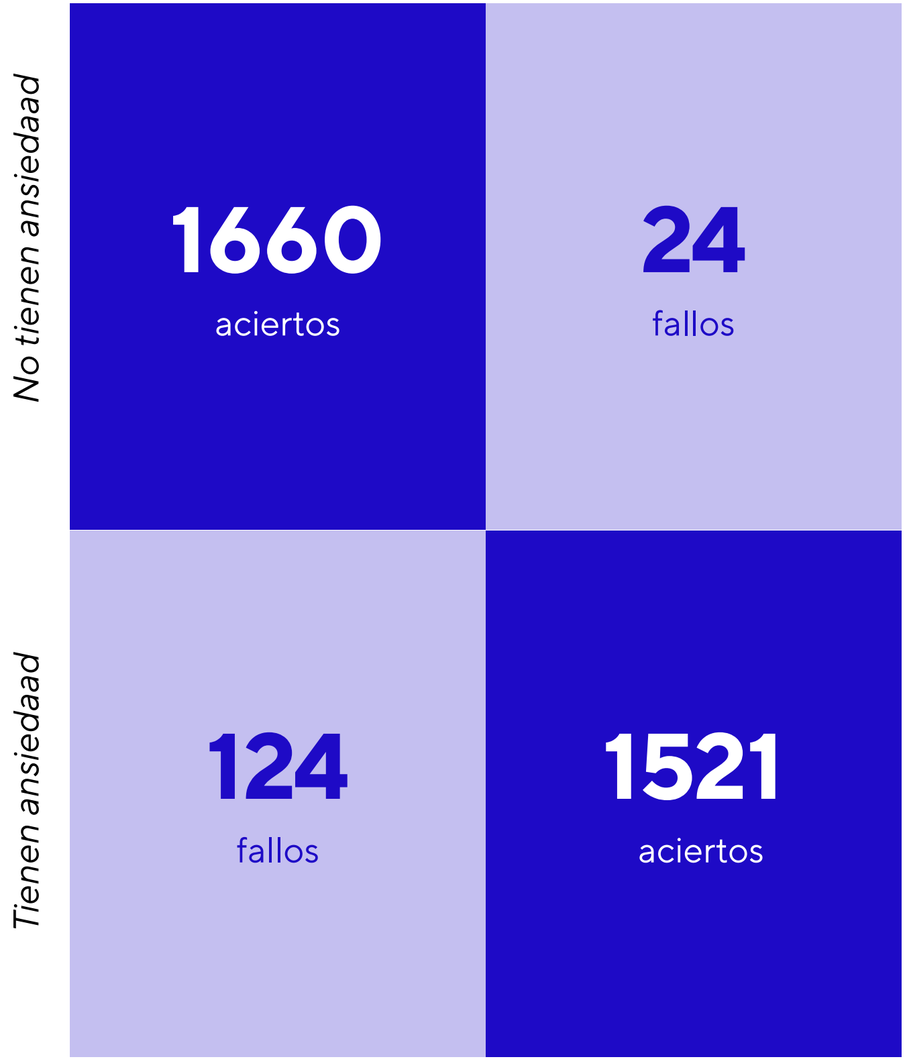
Regresión logística

La Regresión Logística ha sido evaluada como nuestro modelo de referencia, logrando una Precisión del 94.99%. Para entender su rendimiento predictivo, analizamos la matriz de confusión, que detalla sus aciertos y fallos:

- Aciertos en 'No tiene ansiedad': El modelo clasificó correctamente 1627 casos (Verdaderos Negativos).
- Aciertos en 'Sí tiene ansiedad': El modelo clasificó correctamente 1520 casos (Verdaderos Positivos)

Sin embargo, se registraron 57 fallos en 'No tiene ansiedad' y 125 fallos en 'Sí tiene ansiedad'. Estos errores demuestran que la Regresión Logística, aunque sólida, presenta una dificultad notable en la identificación de pacientes que realmente tienen ansiedad .





Random Forest

El algoritmo Random Forest fue evaluado como uno de nuestros principales modelos debido a su capacidad para manejar relaciones no lineales en los datos. Este modelo logró la Precisión más alta de 95.59%. La matriz de confusión del modelo reveló los siguientes resultados específicos:

- Aciertos en 'No tiene ansiedad': El modelo clasificó correctamente 1660 casos.
- Aciertos en 'Sí tiene ansiedad': 1521 casos fueron clasificados correctamente.

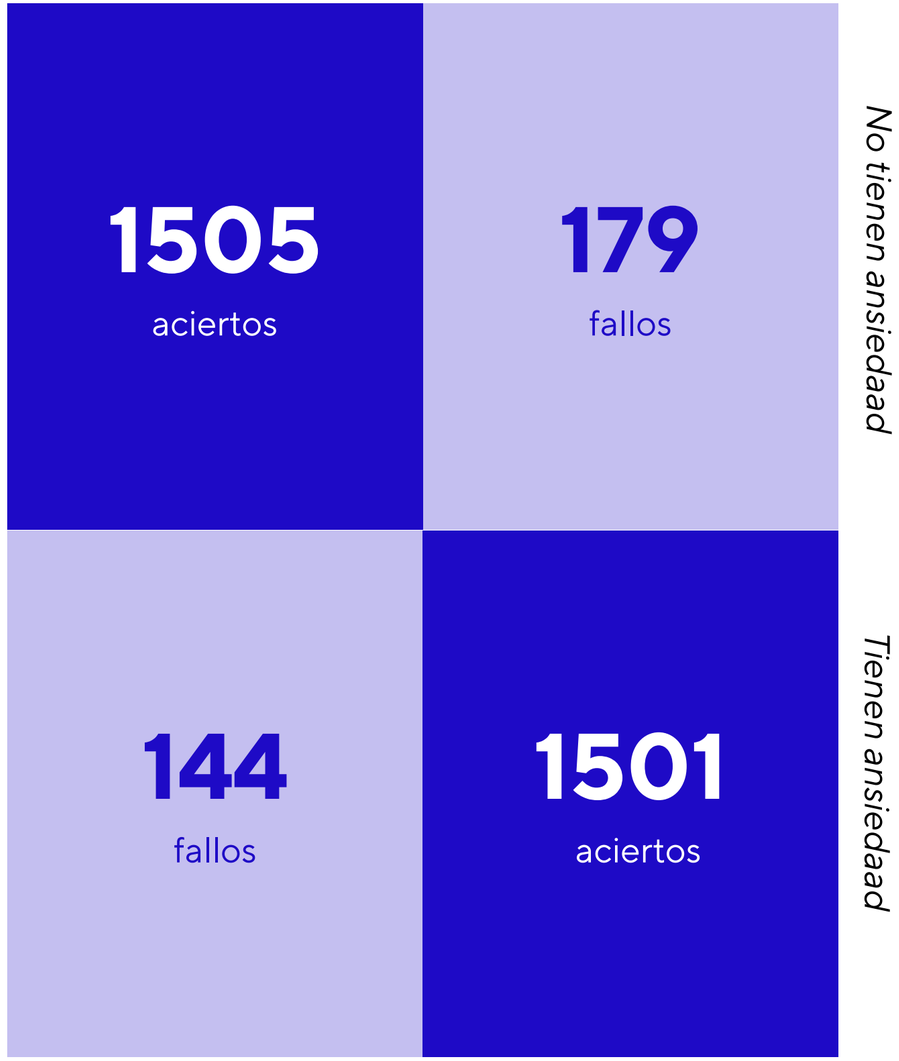
En cuanto a los errores, el modelo solo falló en 24 casos de 'No tiene ansiedad' y 124 casos de 'Sí tiene ansiedad'. La capacidad del Random Forest para minimizar los errores de clasificación, especialmente los falsos negativos, lo convierte en un candidato extremadamente fuerte para una aplicación donde la detección es precisa.

Árbol de de decisión

Nuestro tercer modelo evaluado fue el Árbol de Decisión, que sirve como un modelo base interpretable para la clasificación. Tras la optimización, este modelo alcanzó una Precisión de CV del 90.80%. La matriz de confusión del Árbol de Decisión mostró el siguiente detalle en sus predicciones:

- Aciertos en 'No tiene ansiedad': 1505 casos clasificados correctamente.
- Aciertos en 'Sí tiene ansiedad': 1501 casos clasificados correctamente.

Sin embargo, el modelo incurrió en 179 fallos en la clase 'No tiene ansiedad' y 144 fallos en la clase 'Sí tiene ansiedad'. Este incremento en los errores, especialmente en la clase de 'No tiene ansiedad', subraya la limitación de un único árbol para capturar la complejidad total de los datos en comparación con métodos de ensemble como Random Forest.



Gradient boosting

El algoritmo Gradient Boosting conocido por su alta capacidad predictiva. Este enfoque iterativo alcanzó una Precisión de CV del 95.22%.Al examinar la matriz de confusión, el modelo mostró las siguientes métricas de clasificación:

- Aciertos en 'No tiene ansiedad': El modelo clasificó correctamente 1640 casos.
- Aciertos en 'Sí tiene ansiedad': 1521 casos fueron clasificados correctamente.

En cuanto a los fallos, el modelo registró 46 errores en la clase 'No tiene ansiedad' y 126 errores en la clase 'Sí tiene ansiedad'. Aunque su precisión global es muy alta, es notable que Gradient Boosting tiene una tasa de Falsos Negativos (fallos en la clase 'Sí tiene ansiedad') ligeramente superior a nuestro mejor modelo, lo cual es una consideración clave para la decisión final.

knn

K-Nearest Neighbors , que clasifica los puntos basándose en la proximidad de sus vecinos. Este modelo alcanzó una Precisión de CV del 91.30%, demostrando una buena capacidad de generalización en los datos.

Al analizar el desempeño específico, la matriz de confusión de KNN arrojó los siguientes resultados:

- Aciertos en 'No tiene ansiedad': 1474 casos clasificados correctamente.
- Aciertos en 'Sí tiene ansiedad': 1587 casos clasificados correctamente.

Un aspecto importante a destacar de KNN es la distribución de sus errores. Aunque el modelo tuvo 210 fallos en la clase 'No tiene ansiedad', solo registró 58 fallos en la clase 'Sí tiene ansiedad' (Falsos Negativos). Esta baja tasa de falsos negativos es notable, ya que implica que es muy efectivo identificando a los pacientes que verdaderamente tienen ansiedad, un punto fuerte de este clasificador.

SVC

El modelo Support Vector Classifier (SVC) fue evaluado como una técnica de clasificación avanzada que busca el hiperplano óptimo para separar las clases. SVC alcanzó una impresionante Precisión de CV del 95.16%, confirmando la efectividad de los modelos complejos en nuestro conjunto de datos. Al analizar el desempeño específico, la matriz de confusión de SVC mostró lo siguiente:

- Aciertos en 'No tiene ansiedad': 1669 casos clasificados correctamente.
- Aciertos en 'Sí tiene ansiedad': 1499 casos clasificados correctamente.

SVC es notable por su baja tasa de errores en la clase negativa, con solo 15 fallos en la predicción 'No tiene ansiedad'. Sin embargo, registró 146 fallos en la clase 'Sí tiene ansiedad'. Este desequilibrio indica que SVC priorizó la clasificación correcta de la clase 'No tiene ansiedad', un aspecto a considerar al compararlo con otros modelos que minimizan más los Falsos Negativos.

Se evaluó un amplio espectro de algoritmos de ML para cribado de ansiedad incluyendo modelos lineales, KNN, Árbol de Decisión y Ensemble. Los modelos con mayor precisión fueron: Random Forest: 95.55%, Gradient Boosting: 95.22% y SVC: 95.16%

Regresión Logística: 57 Falsos Negativos

Random Forest: 24 Falsos Negativos

Gradient Boosting: 44 Falsos Negativos