Propuesta final Ciencia de Datos - Ignacio Spiousas, Noelia Romero
Esperanza Pereyra Iraola, Daniela Chejfec, Violeta Juliá
Fecha de entrega: 7/12/2024

Introducción

La migraña es una condición neurológica que afecta a más de mil millones de personas en todo el mundo y se encuentra entre las principales causas de discapacidad, según el *Global Burden of Disease Study 2019*. Además, para las mujeres y adultos jóvenes, es una de las principales causas de años vividos con discapacidad (DALYs), una medida que expresa el total de años perdidos debido a enfermedad, discapacidad o muerte prematura. Además, su naturaleza multifacética, que incluye una variedad de síntomas, como dolores de cabeza severos y síntomas gastrointestinales, y desencadenantes, como el estrés y la falta de sueño, dificulta tanto el diagnóstico como el manejo clínico (Cen, *et al.,* 2024).

En este contexto, la importancia de comprender los patrones de los episodios de migraña se ha vuelto fundamental. Identificar los factores que influyen en la prevalencia de la migraña, como los factores genéticos y ambientales, resalta la necesidad urgente de desarrollar herramientas predictivas basadas en datos. Estas herramientas podrían ser cruciales no solo para mejorar el diagnóstico temprano, sino también para personalizar los tratamientos y reducir el impacto de los episodios más severos, que afectan profundamente la calidad de vida de los pacientes. (Puledda, *et al.,* 2017)

Dado el impacto significativo de las migrañas, surge la siguiente pregunta de investigación: ¿Es posible predecir la intensidad de una migraña utilizando datos sobre síntomas previos, características demográficas y patrones del dolor? Este proyecto busca tratar una de las principales restricciones en el manejo de la migraña: la ausencia de instrumentos precisos y accesibles que permitan prever la severidad de los episodios. El desarrollo de modelos predictivos no sólo constituye una aportación innovadora a la disciplina, sino que también posee la capacidad de transformarse en un instrumento esencial para los expertos en salud y los pacientes. Al proporcionar insights personalizados basados en datos, esta investigación podría establecer las bases para una administración más eficaz de la

migraña, añadiendo un valor significativo a la literatura existente y proporcionando una perspectiva más preventiva ante una condición de gran repercusión a nivel mundial.

Literatura previa

El uso de modelos de aprendizaje supervisado, como la regresión logística y los árboles de decisión, ha demostrado ser efectivo para predecir la intensidad y la duración de los episodios de migraña. Estos modelos utilizan características clínicas y demográficas, como la frecuencia de migrañas previas, los síntomas comórbidos y las características individuales de los pacientes, para predecir episodios futuros. Este enfoque metodológico ha mostrado utilidad en la predicción y tratamiento de la migraña, lo que destaca la importancia del uso de machine learning en el campo de la salud (Özdemir, 2023).

Por otro lado, el uso de técnicas de clasificación multiclase, como árboles de decisión y random forest, ha llevado a identificar diferentes tipos de migraña según sus características, como la localización o los factores desencadenantes. Estas técnicas ayudan a segmentar a los pacientes en grupos más específicos, lo que facilita la personalización del tratamiento. De manera complementaria, técnicas de boosting como el gradient boosting han mostrado ser igualmente efectivas, permitiendo combinar datos de cuestionarios clínicos y resonancia magnética, reduciendo características redundantes y mejorando la precisión de los modelos, lo que refuerza su utilidad para personalizar diagnósticos y tratamientos (Cen *et al*., 2024).

Además, la reducción de dimensionalidad mediante técnicas como el Análisis de Componentes Principales (PCA) ha demostrado ser eficaz para explorar grandes conjuntos de datos, permitiendo identificar patrones subyacentes entre múltiples variables, como los síntomas y las características del dolor. Este enfoque ha sido clave para detectar subgrupos de pacientes con características comunes, lo cual resulta fundamental para personalizar las intervenciones médicas y mejorar la efectividad de los tratamientos (Schürks *et al.*, 2011).

Estas metodologías establecen un marco sólido para la exploración de datos en el ámbito de la migraña y resaltan el potencial de los enfoques basados en machine learning para avanzar en la personalización del diagnóstico y tratamiento. A partir de estos avances, la propuesta busca integrar estas técnicas con un enfoque específico en la predicción de la severidad de los episodios de migraña, contribuyendo a un campo donde aún existen oportunidades significativas para la innovación y la mejora clínica.

Base de datos

La base de datos utilizada en este proyecto, “Migraine Dataset”, fue extraída de la plataforma Kaggle, publicada originalmente por Ranzeet Raut (Raut, R., n.d.). Además, la limpieza y preprocesamiento inicial del dataset fueron realizados por un usuario de Kaggle (Blas Rimac, B., n.d.), quien eliminó seis filas duplicadas, asegurando así la calidad de los datos para su análisis. Este conjunto de datos, compuesto por 400 registros y 24 columnas, contiene información detallada y diversa sobre las características de los episodios de migraña. Incluye datos demográficos, la edad de los pacientes, aspectos específicos de cada episodio, y su duración, frecuencia, intensidad y localización del dolor. También documenta síntomas acompañantes, como náuseas, vómitos, fonofobia, fotofobia y disturbios visuales, además de otros factores sensoriales y neurológicos, como vértigo, tinitus, disartria, parestesias y ataxia, que proporcionan una visión integral de esta compleja condición.

La base de datos también incluye una variable de clasificación que permite identificar diferentes tipos de migraña, como episodios con aura típica, sin aura, y otros subtipos específicos, lo que facilita el análisis comparativo y la segmentación. Los gráficos descriptivos generados a partir de este dataset (Ver Anexo) muestran la distribución de las variables principales, proporcionando una primera visión de los patrones observados. Por ejemplo, se observa que la mayoría de los pacientes reportan migrañas de intensidad moderada y síntomas recurrentes como fotofobia y fonofobia. Asimismo, la duración de los episodios se concentra mayoritariamente en rangos específicos, mientras que el tipo de migraña predominante en la muestra es “Typical aura with migraine”. Estos gráficos, extraídos también de la plataforma Kaggle, ayudan a identificar tendencias que resultarán clave para el desarrollo de los análisis posteriores.

Metodología

El presente proyecto utiliza una combinación de técnicas de análisis de datos y aprendizaje supervisado para predecir la intensidad de los episodios de migraña y explorar los factores asociados a su manifestación. Las metodologías aplicadas incluyen el análisis de componentes principales (PCA), árboles de decisión y modelos de predicción como Random Forest y Bagging, lo que asegura un enfoque robusto y multidimensional.

En primer lugar, se llevará a cabo un análisis exploratorio de datos para comprender las relaciones entre las variables, identificar valores atípicos y descubrir patrones iniciales en el dataset. Se utilizarán visualizaciones como gráficos de dispersión, histogramas y matrices de correlación para estudiar características como la frecuencia, intensidad y síntomas asociados a las migrañas, lo cual permitirá guiar las siguientes etapas del análisis.

El PCA es una técnica de reducción de dimensionalidad que transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales (James, et al., 2023). Este método conserva la mayor cantidad posible de variabilidad presente en los datos originales, ayudando a simplificar el análisis y mejorar la eficiencia computacional. En este proyecto, el PCA se aplicará para identificar patrones en las múltiples variables relacionadas con los episodios de migraña, como síntomas sensoriales, neurológicos y características del dolor. Por ejemplo, si variables como fotofobia, fonofobia y vértigo están correlacionadas, el PCA puede agruparlas en una o dos dimensiones que representen un "perfil sensorial" de las migrañas. Esta reducción permitirá priorizar las características más relevantes para los modelos predictivos y eliminar redundancias en los datos.

Los árboles de decisión son modelos supervisados que utilizan una estructura jerárquica para dividir iterativamente los datos en subconjuntos más pequeños basados en valores de sus variables predictoras (James, et al., 2023). Cada nodo del árbol representa una decisión basada en una variable, mientras que las hojas representan las predicciones finales. Este método es especialmente útil por su capacidad de manejar tanto datos categóricos como valores numéricos continuos y su facilidad para interpretar las relaciones no lineales entre variables. En este proyecto, los árboles de decisión se emplearán para identificar qué características, como intensidad, localización del dolor o síntomas específicos, son más influyentes en la predicción de la severidad de un episodio de migraña.

El Random Forest es una técnica de ensamble basada en la construcción de múltiples árboles de decisión independientes, donde cada árbol se entrena con un subconjunto aleatorio del dataset y de las características disponibles (James, et al., 2023). La predicción final se obtiene mediante el promedio (para regresión) o el voto mayoritario (para clasificación) de los resultados de todos los árboles. Este enfoque mejora la precisión y generalización del modelo, reduciendo el riesgo de sobreajuste que podría presentarse en un único árbol de decisión. En este proyecto, Random Forest se utilizará para capturar relaciones complejas entre variables y mejorar la predicción de la intensidad de los episodios. Además, la técnica permitirá evaluar la importancia relativa de cada variable predictora, proporcionando información valiosa sobre los factores que más contribuyen a la severidad de las migrañas.

Boosting es una técnica de ensamble que crea una secuencia de modelos, donde cada modelo posterior se construye para corregir los errores cometidos por los modelos anteriores. A diferencia de métodos como Random Forest, donde los modelos se entrenan de manera independiente, Boosting entrena los modelos de manera secuencial, asignando más peso a las

observaciones mal clasificadas o mal predichas en cada iteración (James, et al., 2023). Uno de los algoritmos más utilizados para Boosting es Gradient Boosting, que optimiza una función objetivo ajustando modelos simples, como árboles de decisión, para minimizar los errores residuales de los modelos anteriores. En este proyecto, Boosting se aplicará para mejorar la precisión en la predicción de la intensidad de los episodios de migraña, capturando interacciones complejas entre variables y proporcionando un modelo más robusto frente a los datos. Al combinar múltiples modelos débiles en un modelo fuerte, Boosting permitirá identificar patrones sutiles en el dataset y contribuirá a la generalización del modelo, reduciendo la posibilidad de errores en nuevas predicciones.

Finalmente, los resultados obtenidos de estos métodos serán interpretados en términos de su aplicabilidad clínica, destacando los factores más relevantes y las interacciones detectadas. Esta metodología permite no solo predecir la intensidad de los episodios de migraña con mayor precisión, sino también generar hipótesis sobre los mecanismos subyacentes que afectan a esta condición, ofreciendo una base sólida para futuros estudios e intervenciones personalizadas.

Conclusiones y limitaciones

Este proyecto busca demostrar cómo el uso de técnicas avanzadas de análisis de datos y machine learning, como PCA, árboles de decisión, Random Forest y Boosting, puede mejorar la capacidad de predecir la intensidad de los episodios de migraña. Se esperan identificar patrones entre las variables del dataset para destacar factores clave como los síntomas previos, las características demográficas y los desencadenantes específicos. Estas herramientas no solo tienen el potencial de aumentar la precisión en las predicciones, sino también de puede dar un enfoque más personalizado y efectivo para el manejo clínico de esta condición.

Aplicando estas metodologías, se espera que los resultados permitan identificar las variables que más influyen en la intensidad de una migraña, haciendo que el acercamiento a futuras investigaciones, estudios e intervenciones sea menos costoso y más acertado, al necesitar tener en cuenta menos factores para medir y tratar una migraña. Por ejemplo, podría ser posible establecer perfiles específicos de pacientes basados en combinaciones de síntomas y desencadenantes, facilitando la personalización de tratamientos y optimizando los recursos clínicos.

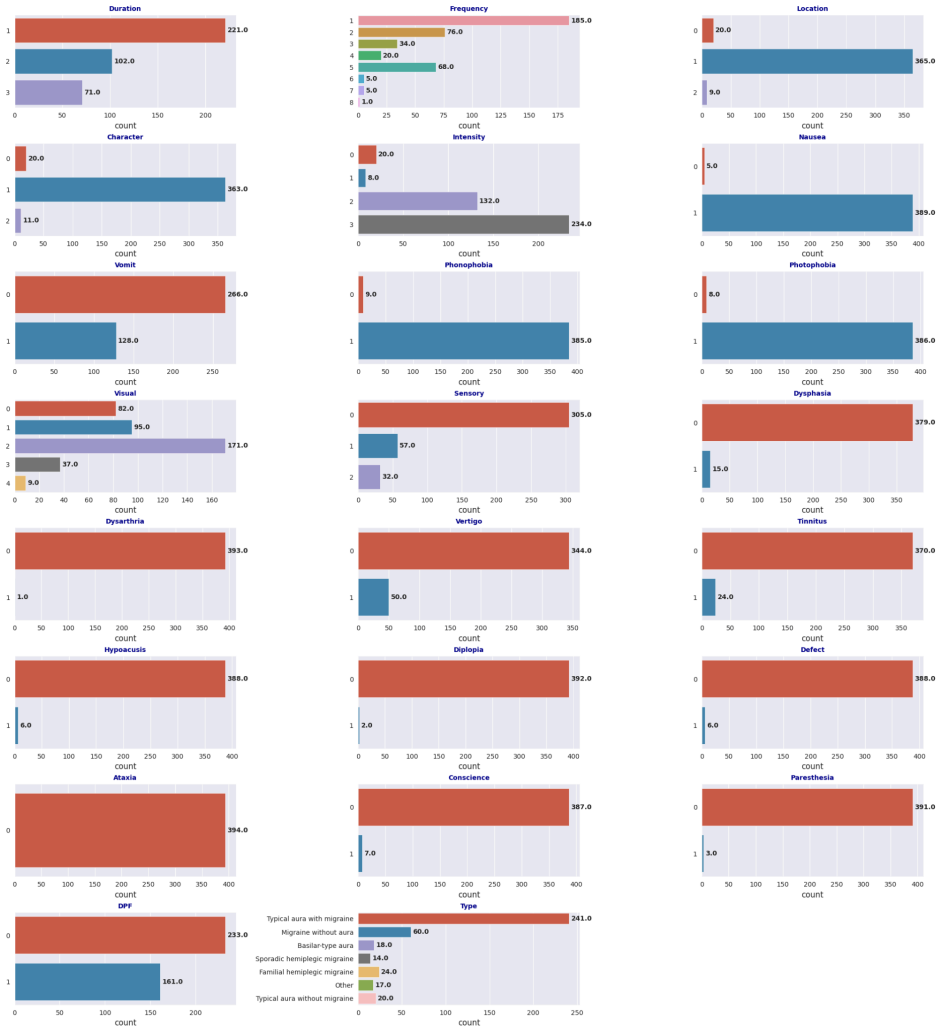
Este proyecto enfrenta algunas limitaciones inherentes. En primer lugar, la calidad y cantidad de los datos disponibles puede influir directamente en la precisión de los modelos.

Aunque el dataset utilizado fue preprocesado, sigue siendo relativamente pequeño, posiblemente limitando la generalización de los resultados. Además, la presencia de sesgos en los datos, como una representación desigual de ciertos subgrupos demográficos o clínicos, podría influir en las predicciones y dificultar la extrapolación a poblaciones más amplias.

Otro desafío importante surge de la interpretación clínica de los resultados. Aunque las técnicas de machine learning permiten identificar patrones complejos, la forma en la que se modelan las interacciones entre variables podría ser difícil de entender para alguien que no esté familiarizado con estos métodos. Esto podría limitar su aplicabilidad en entornos clínicos donde la transparencia del modelo es crucial. A la misma vez, la implementación práctica de estas herramientas requerirá la colaboración de especialistas clínicos y expertos en datos para garantizar que las predicciones se traduzcan en mejoras reales en el manejo de la migraña.

En conclusión, este proyecto busca sentar las bases para un enfoque predictivo y personalizado en el tratamiento de las migrañas. Aunque enfrenta limitaciones metodológicas y prácticas, sus resultados tienen el potencial de contribuir significativamente a la comprensión de esta compleja condición y a la mejora de la calidad de vida de los pacientes.

Anexo



Bibliografía

Blas Rimac, B. (n.d.). *Classification, cross-validation, EDA - accuracy 90%*. Kaggle. Retrieved December 5, 2024, from https://www.kaggle.com/code/bryamblasrimac/classification-crossvalidation-eda-accuracy-90

Cen, J., Wang, Q., Cheng, L., Gao, Q., Wang, H., & Sun, F. (2024). Global, regional, and national burden and trends of migraine among women of childbearing age from 1990 to 2021: insights from the Global Burden of Disease Study 2021. The Journal of Headache and Pain, 25(1), 96.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.*

Özdemir, B. (2023). Improving Migraine Diagnosis by using Binary and Multi-Class Classification of Machine Learning.

Puledda, F., Messina, R., & Goadsby, P. J. (2017). An update on migraine: current understanding and future directions. Journal of neurology, 264, 2031-2039.

Raut, R. (n.d.). *Migraine dataset*. Kaggle. Retrieved December 5, 2024, from https://www.kaggle.com/datasets/ranzeet013/migraine-dataset/data

Schürks, M., Buring, J. E., & Kurth, T. (2011). Migraine features, associated symptoms and triggers: A principal component analysis in the Women's Health Study. Cephalalgia, 31(7), 861-869.