Modelos Predictivos de la Función Cognitiva: Aplicación de Análisis de Datos a Factores de Estilo de Vida y Salud

Propuesta Final de Investigación

Asignatura

CC408 - Ciencia de Datos

Alumnas

Micaela Vollert

Sofía Dillon

Juana Matticoli

Profesores

María Noelia Romero

Ignacio Spiousas

Fecha de entrega

07/12/2024

1. INTRODUCCIÓN

En la sociedad actual, la salud y el bienestar se han consolidado como prioridades fundamentales, reflejando una creciente valoración de los estilos de vida saludables. Esta preocupación ha aumentado significativamente debido a una mayor conciencia sobre los riesgos asociados a las enfermedades, el deterioro cognitivo y el envejecimiento. Es en este contexto que se vuelve relevante buscar estrategias para mantener y promover las funciones cognitivas saludables. Estas abarcan habilidades como la memoria, el razonamiento, la atención y la resolución de problemas y son esenciales para la calidad de vida y el bienestar individual (Oosterhuis, et al., 2023). Sin embargo, estas capacidades pueden verse afectadas por diversos factores relacionados con la salud física, el estado mental y el estilo de vida, lo que plantea la necesidad de entender cómo estas variables interactúan y cuáles tienen mayor relevancia en la predicción de un envejecimiento cognitivo saludable.

La motivación de este estudio radica en la oportunidad de predecir la función cognitiva a partir de factores de salud, tales como la presión arterial, el índice de masa corporal (IMC), el nivel de glucosa en sangre, la agudeza visual, entre otros, y evaluar cómo estos pueden ser utilizados para diseñar estrategias personalizadas de prevención y promoción del bienestar. Comprender estas interacciones podría no solo optimizar programas de alimentación, ejercicio y cuidados de la salud adaptados a cada individuo, sino también convertirse en una herramienta para detectar riesgos asociados con el envejecimiento prematuro y prevenir el desarrollo de enfermedades neurodegenerativas. En este sentido, el trabajo busca contribuir a la creación de políticas de salud pública basadas en datos concretos que permitan mejorar la salud cognitiva de la población, con un enfoque preventivo y proactivo.

La pregunta de investigación que guía este trabajo es: *¿Cómo se puede predecir el nivel de función cognitiva de una persona a partir de variables relacionadas con la salud*

física, salud mental y estilo de vida? Responder esta pregunta proporcionará contribuciones valiosas para el diseño de políticas de salud pública y programas de bienestar, y a su vez, enriquecerá la literatura existente sobre los determinantes de la función cognitiva en la población general.

2. ANTECEDENTES

La relación entre las funciones cognitivas y factores como el estilo de vida y la salud física ha sido ampliamente estudiada en la literatura de la salud y las neurociencias. Diversas investigaciones han demostrado que factores como el ejercicio, la nutrición, la salud mental y el estrés influyen significativamente en el rendimiento cognitivo. A continuación, se presentan algunos trabajos relevantes que sirven de base para la propuesta de investigación desarrollada en este informe.

Un ejemplo es el estudio de Byeon (2015) que aplicó Random Forests, un algoritmo de aprendizaje supervisado basado en la combinación de múltiples árboles de decisión, para predecir el deterioro cognitivo leve (MCI). Este modelo se destacó por manejar conjuntos de datos complejos y no lineales, identificando factores relevantes como la edad, el nivel educativo y hábitos relacionados con la salud. El modelo de Random Forests mostró ser más preciso que los modelos tradicionales como la regresión logística y los árboles de decisión, alcanzando una precisión de predicción del 72.1%. El uso de esta herramienta es relevante para la investigación de este informe, ya que puede ayudar a desarrollar un modelo robusto para predecir la función cognitiva a partir de múltiples factores de salud y estilo de vida, manejando grandes volúmenes de datos y evitando el sobreajuste.

Asimismo, Liu et al. (2024) realizaron una comparación entre métodos tradicionales, como la regresión lineal múltiple, y técnicas avanzadas de machine learning, como XGBoost y Naive Bayes. Su estudio, centrado en adultos mayores con diabetes de tipo 2, demostró que, en líneas generales, los modelos de machine learning superan a la regresión lineal en

términos de precisión. En particular, el modelo XGBoost se destacó como el más eficaz para predecir la función cognitiva, demostrando su capacidad para manejar datos complejos y captar relaciones no lineales. La aplicación de estos métodos de análisis permitió identificar diversos factores relevantes que afectan la función cognitiva, como el nivel educativo, la edad, el índice de fragilidad, la glucosa en ayuno y el índice de masa corporal. Por ende, aunque el estudio se enfoca en pacientes que padecen diabetes, lo cual no se vincula directamente con el enfoque de este informe, los resultados evidenciaron la habilidad de estas técnicas de machine learning para conocer factores determinantes, subrayando su potencial aplicabilidad en otros contextos vinculados con las funciones cognitivas.

Finalmente, el estudio de Kimura et al. (2023) complementa los trabajos anteriores al abordar la relación entre los factores de estilo de vida y la función cognitiva mediante el uso de sensores portátiles. Nuevamente, a través de un análisis de regresión de Random Forest (RF) se identificaron variables como el número de pasos caminados, el tiempo de conversación, la duración del sueño y la frecuencia cardíaca como factores protectores de la función cognitiva. Al igual que otros estudios, este enfoque destaca la capacidad de los modelos de machine learning para manejar grandes volúmenes de datos y seleccionar variables predictivas determinantes de manera precisa.

En resumen, la literatura existente destaca cómo el uso de técnicas avanzadas de machine learning y modelos de análisis estadístico permiten abordar problemas complejos, como la predicción de la función cognitiva, mediante el manejo eficiente de grandes volúmenes de datos y la identificación de patrones no evidentes. De esta manera, la propuesta presentada en este informe aprovecha estos avances para desarrollar un enfoque preventivo que amplíe la aplicabilidad de estos modelos a una población más diversa, maximizando su impacto en contextos más amplios.

3. BASE DE DATOS

Para llevar a cabo este proyecto se seleccionó el conjunto de datos titulado "Función cognitiva en base a distintos factores de salud y estilo de vida: Un Modelo Predictivo", disponible públicamente en la plataforma Kaggle (Abdullah, 2024). Este conjunto de datos, de naturaleza sintética, fue diseñado originalmente para predecir la edad de individuos basándose en diversos factores relacionados con la salud y el estilo de vida. Sin embargo, para este trabajo, las variables se redefinieron de acuerdo con nuestra pregunta de investigación y el objetivo del estudio, centrado en analizar los determinantes de la función cognitiva.

El dataset está dividido en dos partes: un conjunto de entrenamiento, que contiene 3000 observaciones con 26 variables, y un conjunto de prueba, que incluye 1000 observaciones con las mismas características. Cada variable representa factores relacionados con la salud, el estilo de vida y las características sociodemográficas. De estos factores, doce son numéricos continuos, cómo *Height* (cm) y *BMI*, una es numérica discreta y trece son categóricas, como *Gender*, *Smoking Status* y *Diet*.

En este contexto, las variables independientes incluyen altura, peso, presión arterial, índice de masa corporal (IMC), niveles de glucosa en sangre, niveles de colesterol, densidad ósea, agudeza visual, fuerza muscular, frecuencia cardíaca, consumo calórico diario, nivel de actividad física, consumo de tabaco, consumo de alcohol, tipo de dieta, enfermedades crónicas, uso de medicamentos, antecedentes familiares, salud mental, patrones de sueño, niveles de estrés, exposición a la contaminación, exposición al sol, nivel educativo y nivel de ingresos. Por otro lado, se definió a la variable dependiente como la función cognitiva, medida en una escala de 0 (deficiente) a 100 (excelente), la cual constituye el objetivo principal de predicción.

Con el fin de conocer en profundidad las características del conjunto de datos seleccionado se llevaron a cabo análisis estadísticos descriptivos iniciales utilizando Python 3.12.4 en el entorno de desarrollo Anaconda (2024). Por un lado, se realizaron las estadísticas descriptivas de las variables numéricas (**Figura 1**). Entre los resultados más destacados, se observa que la edad promedio de los participantes es de 53.49 años, con un rango amplio que va desde los 18 hasta los 89 años. En cuanto a las medidas físicas, la altura promedio es de 168.59 cm y el peso promedio es de 72.54 kg. Además, los niveles promedio de colesterol y glucosa en sangre (234.03 mg/dL y 126.65 mg/dL, respectivamente) sugieren la presencia de ciertos riesgos metabólicos en la población analizada.

Por otro lado, se llevaron a cabo las estadísticas descriptivas pertenecientes a las variables categóricas (**Figura 2**). Se destaca un equilibrio en la distribución de género, con un leve predominio femenino. En relación con los estilos de vida, el nivel de actividad física más reportado fue "Moderado", mientras que la dieta más común fue la "Equilibrada". En términos de salud, los estados de sueño "Normal" y salud mental "Buena" fueron predominantes. Asimismo, la hipertensión fue la enfermedad crónica más frecuente, mientras que en términos socioeconómicos, la mayoría de los participantes reportaron niveles educativos de grado universitario y un nivel de ingresos medio.

	count	mean	std	min	25 %	50 %	75 %	max
Height (cm)	3000.0	168.591	9.293	141.131	161.63	168.216	175.523	198.112
Weight (kg)	3000.0	72.537	13.191	32.538	63.223	71.449	81.703	123.599
Cholesterol Level (mg/dL)	3000.0	234.03	24.521	148.812	216.757	234.377	250.647	331.301
BMI	3000.0	25.55	4.367	12.05	22.454	25.352	28.404	43.33
Blood Glucose Level (mg/dL)	3000.0	126.654	18.226	69.867	114.393	126.802	139.377	185.736
Bone Density (g/cm²)	3000.0	932	444	-0.22	561	0.94	1.295	2.0
Vision Sharpness	3000.0	475	0.21	0.2	282	462	639	1.063
Hearing Ability (dB)	3000.0	47.016	14.336	0.0	36.735	46.964	56.829	94.004
Cognitive Function	3000.0	63.868	11.756	30.382	55.648	64.015	72.087	106.48
Stress Levels	3000.0	5.477	2.585	1.0	3.222	5.497	7.68	9.996
Pollution Exposure	3000.0	5.029	2.871	6	2.607	5.096	7.476	9.998
Sun Exposure	3000.0	5.956	3.475	2	2.873	5.957	8.991	11.993
Age (years)	3000.0	53.486	20.57	18.0	36.0	53.0	72.0	89.0

Figura 1. Estadísticas Descriptivas de las Variables Numéricas

	count	unique	top	freq
Gender	3000	2	Female	1511
Blood Pressure (s/d)	3000	1606	135/93	9
Physical Activity Level	3000	3	Moderate	1407
Smoking Status	3000	3	Former	1181
Alcohol Consumption	1799	2	Occasional	1057
Diet	3000	4	Balanced	1183
Chronic Diseases	1701	3	Hypertension	676
Medication Use	1802	2	Regular	1063
Family History	1549	3	Diabetes	645
Mental Health Status	3000	4	Good	1073
Sleep Patterns	3000	3	Normal	1519
Education Level	2373	3	Undergraduate	884
Income Level	3000	3	Medium	1223

Figura 2. Estadísticas Descriptivas de las Variables Categórica

A su vez, se evaluó la presencia de valores faltantes en las columnas, siendo las más afectadas *Alcohol Consumption* con 1201 valores faltantes, *Chronic Diseases* con 1299, *Medication Use* con 1198, *Family History* con 1451 y, por último, *Education Level* con 627. Este hallazgo subraya la importancia de abordar estos valores ausentes de manera adecuada, ya sea a través de técnicas de imputación o mediante su exclusión, antes de proceder con análisis posteriores.

4. METODOLOGÍA

En cuanto a la metodología de la presente investigación, se procede con una primera etapa de análisis exploratorio de los datos, para luego desarrollar una etapa predictiva y una de validación. Estas etapas permitirán primero explorar los datos, luego construir un modelo predictivo de la función cognitiva y, finalmente, evaluar y mejorar la calidad de ese modelo.

En primer lugar, se realizará la limpieza de la base de datos. Se identificaron las variables con valores faltantes como *Alcohol Consumption* y *Medication Use*. Para ello, se emplearán técnicas de imputación: se reemplazará el valor faltante por la media o la mediana de la variable numérica en cuestión. Si se tratase de una variable categórica, se reemplazaría el valor faltante por el valor más frecuente (la moda). En caso de que una variable tenga un porcentaje muy alto de valores faltantes (mayor al 60%), se considerará su exclusión del análisis para evitar que distorsione los resultados. En cuanto a los valores atípicos (outliers), se decidirá si es necesario transformarlos o excluirlos dependiendo de su impacto y relevancia en el análisis.

Posteriormente a la limpieza de la base, se identificarán grupos naturales en los datos a través de técnicas de clustering, sin todavía clasificarlos según la variable a predecir, la función cognitiva. Por lo tanto, se hará uso del algoritmo de agrupamiento K-means para reconocer subgrupos de sujetos con características similares en cuanto a estilo de vida, salud física y mental. K-means es un método útil especialmente cuando se desea inspeccionar la

estructura interna de los datos y encontrar subpoblaciones en función de sus similitudes. De esta forma, resultaría posible identificar un grupo con características de salud y estilo de vida que se asocian a un buen rendimiento cognitivo, por ejemplo. Es necesario que, previamente a la clusterización, se normalicen las variables continuas como 'IMC' y los niveles de glucosa, para que las diferencias en las escalas no afecten el agrupamiento y permitan obtener resultados consistentes. La cantidad de clusters seleccionados quedará sujeto a cómo se comporten los datos y a su interpretación, pero sería ideal seleccionar 3 clusters: 'alto nivel de salud y lifestyle', 'nivel medio de salud y lifestyle' y 'bajo nivel de salud y lifestyle'. Una vez finalizada la clusterización, y con ella, la etapa de análisis exploratorio de los datos, se obtendrá una idea acerca de cómo se agrupan las personas y qué características identifican a cada grupo, lo que ofrece un buen punto de partida para comenzar con la etapa predictiva.

Se continúa el análisis construyendo un modelo de Random Forest para predecir el nivel de función cognitiva a partir de las variables independientes. Esta técnica de ensamble combina varios árboles de decisión que utilizan diferentes subconjuntos de variables para realizar predicciones sobre los datos de entrenamiento. Dado que los factores de salud y estilo de vida no siempre actúan de manera lineal, sino que suelen relacionarse de forma compleja, este método es adecuado para el presente estudio. Para entrenar este modelo se hará uso del conjunto de datos de entrenamiento ya dado por la base de datos original, que contiene 3000 observaciones con 26 variables. Luego de aprender del conjunto de entrenamiento, se evaluará la precisión predictiva del modelo en el conjunto de prueba, que incluye 1000 observaciones. Random Forest también permitirá ver cuáles son los factores más influyentes en la función cognitiva de las personas, lo cual puede ser información útil para pensar en posibles intervenciones o recomendaciones sobre los determinantes de la salud cognitiva.

Por último, en la etapa de validación, se implementará Cross-validation (validación cruzada) para evaluar y mejorar la calidad del modelo. Se busca asegurar que el modelo

predictivo sea confiable y tenga la capacidad de generalizar adecuadamente en un nuevo conjunto de datos. Para ello, se dividirá el conjunto de datos en 10 particiones (10-fold Cross-validation) para entrenar y validar el modelo en cada una. Se opta por realizar 10 particiones ya que es el estándar más utilizado, y ofrece un equilibrio entre costo computacional y el estimación del desempeño del modelo. El conjunto de entrenamiento, al ser de 3000 observaciones, es demasiado grande para optar por técnicas como leave-one-out cross-validation (LOOCV), donde el costo computacional es demasiado alto y es usualmente utilizado para bases pequeñas. En caso de que se presente una limitación en cuanto a tiempo y costo computacional, se podrían realizar 5 particiones en vez de 10. Esta decisión estará sujeta al momento de trabajar con los datos. En cada iteración, cada una de las particiones será utilizada como conjunto de validación mientras que las otras entrenan al modelo, obteniendo así una medida promedio del rendimiento del modelo. De esta forma, se evita que el modelo esté sobreajustado para que funcione con otros conjuntos de datos. Se utilizarán las siguientes métricas de desempeño para conocer la capacidad de predicción del modelo y su variabilidad: matrices de confusión, accuracy, AUC y curva ROC. Con estos resultados, se podría optimizar el modelo de Random Forest si así se quisiera, ajustando aspectos como la cantidad de árboles o la profundidad máxima de cada árbol. Esto podría ayudar a mejorar la precisión y la capacidad de predicción. Al finalizar con la etapa de validación, se pretende contar con un modelo predictivo que nos guíe a resolver la pregunta de investigación del presente estudio: *predecir el nivel de función cognitiva de una persona a partir de variables relacionadas con la salud física, salud mental y estilo de vida.* A partir de esta predicción, se podrían diseñar estrategias de prevención y/o promoción de la salud cognitiva en poblaciones específicas.

5. CONCLUSIONES Y LIMITACIONES

En el presente proyecto, se espera alcanzar un modelo predictivo robusto que responda a nuestra pregunta y objetivo de investigación. En el análisis desarrollado, se contempla una búsqueda de patrones significativos y grupos naturales en los datos, lo que podría aportar una comprensión de los factores comunes entre individuos con perfiles similares. Asimismo, en la etapa predictiva, se sostiene que el modelo de Random Forest logrará identificar las variables más relevantes para la función cognitiva, lo cual proporciona una base sólida para realizar futuras investigaciones. Finalmente, la validación cruzada permitirá garantizar que el modelo sea confiable y generalizable para poder aplicar el modelo en distintos contextos.

En cuanto a las limitaciones, es importante tener en cuenta que el conjunto de datos utilizado es de naturaleza sintética. Aunque los datos sintéticos suelen ser creados a partir de bases de datos reales, puede suceder que no representen de la manera más fiel la salud y los estilos de vida en situaciones reales. Por lo tanto, podría ser necesario validar el modelo y sus hallazgos con datos reales en próximas instancias. En añadidura, la base de datos presentó una gran cantidad de valores faltantes. Si bien se recomiendan técnicas adecuadas para realizar la limpieza de los datos, existe el riesgo de reducir la representatividad del modelo. Otra observación realizada con lo que respecta a la base de datos es que no contempla factores culturales y contextuales que pueden influir sobre el estilo de vida y la salud de los sujetos observados. Sería interesante obtener información sobre estos aspectos para enriquecer el análisis y lograr una mejor predicción aplicable a poblaciones con distintas características demográficas. Por otro lado, se han hallado limitaciones a definir el número de clusters a obtener en el agrupamiento, así como definir el número de particiones adecuado para la validación del modelo Random Forest. Para obtener un número más preciso de clusters y particiones, sería adecuado definirlos en la práctica, al trabajar con los datos. Por

último, desde una perspectiva práctica, se cree que los resultados del modelo dependerán en gran parte de los datos utilizados. Este modelo, para ser implementado en un contexto real, requiere de una base de datos robusta para garantizar precisión y representatividad.

6. REFERENCIAS BIBLIOGRÁFICAS

Abdullah, M. (2024). Human Age Prediction Synthetic Dataset. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9314588

Anaconda, Inc. (2024). Anaconda Navigator (versión 2024). Recuperado de https://www.anaconda.com

Byeon, H. (2015). A Prediction Model for Mild Cognitive Impairment Using Random Forests. International Journal of Advanced Computer Science and Applications, 6.

Kimura, N., Aso, Y., Yabuuchi, K., Ishibashi, M., Hori, D., Sasaki, Y., Nakamichi, A., Uesugi, S., Fujioka, H., Iwao, S., Jikumaru, M., Katayama, T., Sumi, K., Eguchi, A., Nonaka, S., Kakumu, M., & Matsubara, E. (2019). Modifiable Lifestyle Factors and Cognitive Function in Older People: A Cross-Sectional Observational Study. Frontiers in Neurology.

Liu, C. H., Peng, C. H., Huang, L. Y., Chen, F. Y., Kuo, C. H., Wu, C. Z., & Cheng, Y. F. (2024). Comparison of multiple linear regression and machine learning methods in predicting cognitive function in older Chinese type 2 diabetes patients. BMC neurology, 24(1), 11.

Oosterhuis, E. J., Slade, K., May, P. J. C., & Nuttall, H. E. (2023). Toward an understanding of healthy cognitive aging: the importance of lifestyle in cognitive reserve and the scaffolding theory of aging and cognition. The Journals of Gerontology: Series B, 78(5), 777-788.

Yaffe, K., Fiocco, A. J., Lindquist, K., Vittinghoff, E., Simonsick, E. M., Newman, A. B., ... & Harris, T. B. (2009). Predictors of maintaining cognitive function in older adults: the Health ABC study. Neurology, 72(23), 2029-2035.