Predicción del diagnóstico de Parkinson: Enfoque basado en modelos de regresión logística

Asignatura: Ciencia de Datos

Profesores: Noelia Romero e Ignacio Spiousas

Alumnas: Justina Reinke, Luciana Crupnik y Josefina Laborda

1. INTRODUCCIÓN

Las enfermedades neurodegenerativas comparten como rasgo común un proceso progresivo de deterioro de las neuronas, en el cerebro y/o otras partes del sistema nervioso central o periférico. Estas enfermedades pueden compartir síntomas y alteraciones neuropatológicas, lo que dificulta un diagnóstico preciso, ya que no existen herramientas definitivas para identificarlas. Los médicos suelen basarse en la evolución gradual de los síntomas, lo que puede llevar a ajustes en el diagnóstico inicial.

El parkinson es una enfermedad neurodegenerativa asociada con síntomas motores (movimientos lentos, temblores, y desequilibrio) y una amplia variedad de complicaciones no motoras (síntomas neuropsiquiátricos y autonómicos, trastornos del sueño, dolor y otras alteraciones sensoriales). A nivel mundial, la discapacidad y la muerte debido a esta enfermedad están aumentando más rápido que cualquier otro trastorno neurológico. En los últimos 25 años, su prevalencia se ha duplicado. *(Organización Mundial de la Salud, 2022)*

El principal factor de riesgo es la edad, aunque también puede afectar a personas más jóvenes. Dado que los síntomas iniciales suelen ser difíciles de detectar, un diagnóstico temprano es crucial no solo para mejorar la calidad de vida de los pacientes, sino también para contribuir a la investigación. En este contexto, estudiar cómo los factores demográficos, de estilo de vida y clínicos influyen en el riesgo de desarrollar esta enfermedad resulta fundamental para comprender mejor su etiología y progresión. *(Organización Mundial de la Salud, 2022)*

La pregunta central que dirige esta investigación es: *¿Cómo influyen los factores demográficos, de estilo de vida y clínicos para predecir el riesgo de desarrollar la enfermedad de Parkinson?* Este conocimiento no solo podría mejorar las estrategias de prevención y detección temprana, sino también optimizar los recursos de salud pública,

permitiendo intervenir antes de que los síntomas afecten gravemente la calidad de vida de las personas.

2. ANTECEDENTES

La enfermedad de Parkinson ha sido ampliamente estudiada en los últimos años, analizando qué factores están involucrados y utilizando distintas herramientas para poder predecir su desarrollo. Resulta pertinente utilizar estos hallazgos para poder guiar nuestra investigación.

En primer lugar, el estudio de Belvisi et al. (2020) evaluó factores de riesgo y protección para la enfermedad de Parkinson con 694 pacientes y 640 controles. Se analizaron 31 factores mediante modelos de regresión logística y análisis de clúster de k-medias. Los resultados identificaron nueve factores asociados independientemente con la EP: historia familiar de la enfermedad, dispepsia, exposición a pesticidas, aceites, metales, anestesia general (factores de riesgo), y consumo de café, tabaquismo y actividad física (factores protectores). El análisis de clúster reveló cuatro subtipos de pacientes en la población: pacientes con antecedentes familiares de EP (Grupo 1), con exposición a agentes tóxicos (Grupo 2), con dispepsia predominante (Grupo 3), y sin factores de riesgo ni protección asociados (Grupo 4). Esto es relevante para comprender la cantidad de combinación de factores de riesgo de la enfermedad de Parkinson.

Por otro lado, el estudio de Challa et al. (2016) evaluó diferentes modelos de machine learning para predecir la enfermedad de Parkinson utilizando síntomas no motores como el REM (rapid eye movement), trastorno de la conducta del sueño (RBD) y la pérdida olfativa, y biomarcadores de líquido cefalorraquídeo y neuroimágenes. Utilizaron Multylayer Perceptron, BayesNet, Random Forest y Boosted Logistic Regression, destacando esta última con una precisión del 97.16% y un AUC del 98.9%, sugiriendo su efectividad para la

detección temprana de Parkinson. Los resultados evidencian la capacidad de estas técnicas de aprendizaje automático para predecir la enfermedad.

Por otro lado, el estudio de Chairta et al. (2021) evaluó la capacidad predictiva de un Polygenic Risk Score (PRS)[1] basado en 12 polimorfismos de un solo nucleótido (SNP) combinados con factores de riesgo ambientales y demográficos (edad, género, antecedentes familiares, lesiones en la cabeza, índice de masa corporal, depresión, y tabaquismo) en una población griega-chipriota. Se aplicaron métodos como regresión logística univariada para analizar las asociaciones individuales, regresión logística múltiple para evaluar el modelo combinado, y selección de variables mediante *stepwise regression*. El mejor modelo predictivo incluyó el PRS y siete factores adicionales, alcanzando un AUC de 0.79. Los autores concluyen que combinar factores genéticos y ambientales mejora significativamente la predicción del riesgo de Parkinson, lo que es resulta relevante para nuestro análisis. Además, resaltan la importancia de hacerlo con muestras más grandes, y nuestra base de datos cuenta con un número de pacientes mucho mayor.

El estudio de Dadu et al. (2022) utilizó datos longitudinales de Parkinson Disease Biomarker Program (PDBM) y Parkinson's Progression Marker Initiative (PPMI) para identificar subtipos de Parkinson y predecir su progresión mediante aprendizaje automático. Aplicaron reducción de dimensionalidad con factorización de matrices no negativas (NMF), modelos de mezcla gaussiana (GMM) para subtipos según velocidad de progresión, y algoritmos supervisados como Random Forest. Los resultados revelaron tres subtipos clínicos (progresión lenta, moderada y rápida) con diferencias significativas en dimensiones motoras, cognitivas y de sueño. Los modelos alcanzaron un AUC promedio de 0.92 en la predicción a largo plazo, destacando biomarcadores como el neurofilamento de cadena ligera (NfL) como

[1] Una puntuación de riesgo poligénico (PRS) utiliza solamente información genómica para evaluar las probabilidades de que una persona tenga o desarrolle una afección médica particular. (Instituto Nacional de Investigación del Genoma Humano, 2024)

indicador de progresión rápida. Estos hallazgos son relevantes para tener en cuenta los factores que influyen en la progresión de la enfermedad, para poder hacer predicciones más precisas.

3. BASE DE DATOS

Se utilizará la base de datos “Parkinson 's Disease Dataset Analysis”, publicada en Kaggle por Rabie El Kharoua, actualizada por última vez hace seis meses (Ver Anexo 1.1). La base de datos incluye información de 2,105 pacientes diagnosticados con la enfermedad de Parkinson. Este conjunto de datos es sintético y se generó con fines educativos.

En el siguiente análisis se utilizará una serie de variables relacionadas a distintos factores. Por un lado, las variables demográficas incluyen la edad, el género, la etnia, y el nivel educativo. En cuanto a las variables de estilo de vida, se incluyen el índice de masa corporal (BMI), si el paciente fuma, el consumo semanal de alcohol (0-20), la actividad deportiva por semana, calidad de dieta y calidad de sueño. La base de datos también proporciona medidas clínicas como lo son la presión arterial sistólica y diastólica, el colesterol total, niveles de colesterol de lipoproteínas de alta y baja densidad, y niveles de triglicéridos. Las evaluaciones cognitivas y funcionales incluyen la Puntuación de la Escala Unificada de Calificación de la Enfermedad de Parkinson (UPDRS), Puntuación de la Evaluación Cognitiva de Montreal y el Puntuación de la evaluación funcional. Por otro lado, la base de datos incluye la presencia de los siguientes síntomas: temblor, rigidez muscular, bradicinesia (lentitud de movimientos), inestabilidad postural, problemas de discurso, desórdenes del sueño y constipación. También se proporciona información sobre historial médico como la presencia de diabetes, depresión, paros cardíacos, hipertensión, lesiones cerebrales traumáticas, y el historial familiar de Parkinson. Las variables son categóricas y continuas, y se buscará predecir la variable *diagnosis*, que representa el estado del diagnóstico de la enfermedad de Parkinson, donde 0 indica no y 1 indica sí.

Para explorar en detalle las características del conjunto de datos seleccionado, se realizaron análisis estadísticos descriptivos iniciales utilizando Python.

En primer lugar, obtuvimos la media, el valor mínimo y máximo, y la desviación estándar para las variables continuas. Observamos que la edad media de los pacientes es de 70 años. A su vez, para la variable de consumo de alcohol por semana que varía entre 0-20, el valor de la media se encontró en el punto medio, promediando un valor de 10. Lo mismo sucede para actividad física por semana y calidad de dieta, que varía de 0 a 10, y promediaron un valor aproximado de 5. En cuanto a las medidas clínicas, los valores se distribuyeron en un rango normal, excepto por las medidas de colesterol que se encontraron ligeramente elevadas. La puntuación promedio de UPDRS (Unified Parkinson 's Disease Rating Scale) indica una severidad moderada de Parkinson en la muestra, pero hay individuos tanto con síntomas leves como con severos. Esto coincide con la media del MoCA (Montreal Cognitive Assessment), que sugiere un nivel de deterioro cognitivo moderado en la muestra, con algunos pacientes en rangos normales y otros con deterioro severo. Por último, la puntuación promedio de la evaluación funcional indica un nivel moderado, con algunos en los niveles más altos de independencia.

Por otro lado, se analizaron las variables categóricas, mediante el cálculo de la moda de los valores. Se observa que la mayoría de los pacientes son hombres, de etnia caucásica y con un nivel educativo secundario alcanzado. Se observa que la mayoría no fuma, y tampoco presenta síntomas asociados con la enfermedad, probablemente debido a que algunos pacientes se encuentran dentro rangos normales, mientras que otros presentan un deterioro significativo. Asimismo, también se evidencio que la mayoría no presentaba un historial de diabetes, depresión, paro cardiacos, hipertensión, lesión traumática del cerebro, ni historial familiar con parkinson.

Variable	count	mean	std	min	max
Age	2105.0	69.60190023752969	11.594511460477861	50.0	89.0
BMI	2105.0	27.209492765200114	7.2080989337688095	15.008333144286365	39.99988683753807
AlcoholConsumption	2105.0	10.040413032423082	5.687014060970639	0.0022279553989323	19.988865972822232
PhysicalActivity	2105.0	5.016674447977553	2.890919008035536	0.0041566877293075	9.995254857028534
DietQuality	2105.0	4.912900770128542	2.872114829177342	1.053817508700483e-05	9.99586431335674
SleepQuality	2105.0	6.996638843435026	1.7530650155281342	4.0004973021486965	9.999820998517183
SystolicBP	2105.0	133.71971496437055	26.502355084289594	90.0	179.0
DiastolicBP	2105.0	90.24988123515439	17.061488029995395	60.0	119.0
CholesterolTotal	2105.0	226.86083976879297	43.589406479550384	150.06269820288267	299.9630739305192
CholesterolLDL	2105.0	126.1478578621359	43.40703598260851	50.02282828765501	199.98598067436163
CholesterolHDL	2105.0	59.6703515412509	23.37091987799962	20.02798097657361	99.98226534562758
CholesterolTriglycerides	2105.0	222.9404998224325	101.89582169009404	50.11360418775354	399.9750224753191
UPDRS	2105.0	101.41531827710152	56.59144820532081	0.0284410965338546	198.9536041801936
MoCA	2105.0	15.094313546161898	8.643013907611003	0.0211912121245616	29.97010688479763
FunctionalAssessment	2105.0	4.989694186120927	2.93387669403383	0.0015051231673135	9.992697334523715

Variable	Moda
Gender	0
Ethnicity	0
Smoking	0
EducationLevel	1
FamilyHistoryParkinsons	0
TraumaticBrainInjury	0
Hypertension	0
Diabetes	0
Depression	0
Stroke	0
Tremor	0
Rigidity	0
Bradykinesia	0
PosturalInstability	0
SpeechProblems	0
SleepDisorders	0
Constipation	0

Figura 1(izquierda): Estadísticas descriptivas de las variables numéricas y figura 2(derecha): Estadísticas descriptivas de las variables categóricas

4. METODOLOGÍA

El objetivo principal de este análisis es predecir la presencia de Parkinson (Y = {0,1}) utilizando un conjunto de variables predictoras nombradas previamente. Este enfoque busca identificar a los pacientes que padecen la enfermedad, priorizando la minimización de los falsos negativos, ya que no diagnosticar correctamente a un paciente con Parkinson puede tener consecuencias críticas para su salud, retrasando el tratamiento necesario y potencialmente agravando su condición.

En la etapa de limpieza de los datos, comenzaremos por el manejo de valores faltantes. De acuerdo con el tipo de datos, procederemos a hacer una imputación múltiple, sustitución con la mediana o con la moda. Asimismo, se considerará la normalización y estandarización de las variables continuas, como la edad o los niveles de colesterol, asegurando que estén en escalas comparables para el modelo. Por otro lado, las variables categóricas, como el género, la etnicidad o los síntomas binarios (e.g., presencia de temblores o rigidez), serán transformadas a variables dummies, lo que permitirá su adecuada interpretación por el modelo predictivo.

En el análisis exploratorio de datos, se buscará obtener una comprensión más profunda de las características del conjunto de datos mediante técnicas visuales y estadísticas. Se analizará la distribución de las variables, utilizando herramientas gráficas como

histogramas para las variables continuas (e.g., edad, IMC, niveles de colesterol) y gráficos de barras para las categóricas (e.g., género, etnicidad, síntomas), lo que permitirá identificar patrones, anomalías o posibles valores atípicos. En concreto, se utilizará un histograma para estimar la función de densidad de la variable predictora. Además, se empleará una matriz de correlación para analizar la correlación entre las variables, y poder detectar potencial colinealidad entre los predictores. También se examinará la distribución de la variable objetivo para identificar desequilibrios en las clases (por ejemplo, si hay más casos negativos que positivos de Parkinson), lo que es crucial para garantizar un análisis robusto y representativo.

En la etapa de selección de variables predictoras, se identificarán aquellas que resulten más relevantes para el modelo predictivo. Este proceso permitirá determinar qué columnas tienen un impacto significativo en el diagnóstico de Parkinson. Para la elección de variables predictoras se considerará la aplicación de técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), para simplificar el modelo y eliminar redundancias sin perder información clave. Esto garantizará que el modelo sea eficiente y centrado en las variables más importantes para la predicción.

En la etapa de modelado y predicción, utilizaremos un modelo de regresión logística para estimar la probabilidad de "diagnóstico" positivo de Parkinson (Y = 1) en función de un conjunto de variables predictoras. Este enfoque permitirá capturar de manera integral los factores asociados con el diagnóstico de Parkinson para poder predecir la probabilidad de que una nueva observación pertenezca a la categoría de diagnóstico positivo condicional al vector de predictores.

El modelo se entrenará utilizando el 70% de los datos, mientras que el 30% restante se destinará a la fase de prueba. Para estimar los coeficientes del modelo estadístico, logit utiliza el método de máxima verosimilitud; el objetivo es encontrar los valores de los parámetros

que hacen que los datos observados sean lo más probable posible. La regresión logística modelará la relación entre las variables predictoras y la probabilidad de diagnóstico positivo mediante una función sigmoide, lo que hace posible una clasificación binaria efectiva (en probabilidades acotadas en el rango [0,1]). Para clasificar a los pacientes, utilizaremos un umbral basado en el clasificador de Bayes, establecido previamente. Este umbral, inicialmente definido en 0.5, permitirá clasificar como positivo (Y = 1) a aquellos pacientes cuya probabilidad estimada sea superior a este valor. Sin embargo, dado el contexto clínico, donde los falsos negativos (no diagnosticar a alguien con Parkinson) tienen consecuencias graves, ajustaremos el umbral para priorizar la sensibilidad del modelo, reduciendo este tipo de error a expensas de un posible incremento en los falsos positivos.

Para evaluar el desempeño del modelo, generaremos la curva ROC, que mostrará el trade-off entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos para diversos umbrales de decisión. Calcularemos el área bajo la curva (AUC) como métrica de la capacidad predictiva del modelo, donde valores más cercanos a 1 indican un mejor desempeño. Este análisis nos permitirá identificar el umbral óptimo para nuestro contexto clínico y validar la efectividad del modelo en clasificar correctamente a los pacientes.

Finalmente, analizaremos los coeficientes estimados por el modelo logístico para interpretar cómo cada variable predictora influye en el log-odds del diagnóstico de Parkinson. Esta interpretación detallada permitirá identificar los factores más relevantes y sus efectos marginales, facilitando tanto el diseño de intervenciones preventivas como la priorización de recursos en salud pública.

5. CONCLUSIONES Y LIMITACIONES

El presente proyecto tiene como objetivo predecir la probabilidad de diagnóstico positivo de Parkinson utilizando un modelo de regresión logística basado en un conjunto diverso de variables demográficas, de estilo de vida, clínicas y cognitivas. Esperamos que el

modelo permita identificar patrones significativos en los datos que contribuyan a mejorar la detección temprana de esta enfermedad, priorizando la sensibilidad para minimizar falsos negativos y garantizar que los pacientes con Parkinson sean diagnosticados a tiempo. La implementación del clasificador de Bayes para establecer umbrales óptimos y la evaluación mediante curvas ROC aportarán una perspectiva robusta sobre la precisión del modelo.

Sin embargo, este estudio enfrenta varias limitaciones. En primer lugar, el uso de un conjunto de datos sintético, aunque útil para propósitos educativos y de modelado, plantea serias restricciones en términos de la generalización de los resultados a poblaciones reales. La falta de validación en datos empíricos podría limitar la aplicabilidad del modelo en contextos clínicos prácticos. En el futuro, sería necesario probar el modelo con datos reales para evaluar su desempeño en entornos más variados y representativos.

En cuanto al modelo de regresión logística, aunque es una herramienta poderosa para la clasificación binaria, presenta ciertas limitaciones. La principal es su dependencia en supuestos como la relación lineal entre los predictores y el log-odds, lo que puede ser un problema cuando las relaciones subyacentes son otras. Además, la regresión logística puede ser sensible a multicolinealidad entre variables predictoras y no está diseñada para manejar relaciones no lineales sin transformaciones adicionales o términos de interacción. La inclusión de técnicas de aprendizaje automático, como Random Forest, podría ofrecer modelos más flexibles y robustos en futuras investigaciones.

Finalmente, los resultados del modelo dependen en gran medida de la calidad y representatividad de las variables predictoras seleccionadas. Aunque se ha seguido un enfoque riguroso de limpieza y selección de datos, es posible que existan factores relevantes no contemplados en este conjunto de datos. Estos factores podrían influir significativamente en el riesgo de desarrollar Parkinson y deben considerarse en futuras investigaciones con bases de datos más completas.

REFERENCIAS BIBLIOGRÁFICAS

Chairta, P. P., Hadjisavvas, A., Georgiou, A. N., Loizidou, M. A., Yiangou, K., Demetriou, C. A., ... & Zamba-Papanicolaou, E. (2021). Prediction of Parkinson's disease risk based on genetic profile and established risk factors. *Genes*, *12*(8), 1278.

Dadu, A., Satone, V., Kaur, R., Hashemi, S. H., Leonard, H., Iwaki, H., ... & Faghri, F. (2022). Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinson's Disease*, *8*(1), 172.

Daniele, B., Roberta, P., Andrea, F., Matteo, C., Sara, P., Nicola, M., ... & Giovanni, D. (2020). Risk factors of Parkinson's disease: simultaneous assessment, interactions and etiological subtypes. *NEUROLOGY*, *95*(18), e2500-e2508.

Instituto Nacional de Investigación del Genoma Humano. (2024, 4 de diciembre). *Puntuación de riesgo poligénico*. Recuperado de https://www.genome.gov/es/genetics-glossary/Polygenic-Risk-Score

Nayan Reddy Challa, K., Sasank Pagolu, V., Panda, G., & Majhi, B. (2016). An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques. *arXiv e-prints*, arXiv-1610.

Organización Mundial de la Salud (2022*) Parkinson disease: a public health approach. Technical brief*. Recuperado de https://iris.who.int/bitstream/handle/10665/355973/9789240050983-eng.pdf?sequence=1

ANEXO

1.1 Link de acceso a la base de datos: https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis/data