



Revista EIA
ISSN 1794-1237
e-ISSN 2463-0950
Año XIX/ Volumen 22/ Edición N.43
Enero - junio 2025
Reia4304 pp. 1-17

Publicación científica semestral
Universidad EIA, Envigado, Colombia

PARA CITAR ESTE ARTÍCULO / TO REFERENCE THIS ARTICLE /

Rincón Velásquez, A. C.
Identificación de riesgo de progresión
de la enfermedad de Parkinson
utilizando Machine Learning.

Revista EIA, 22(43), Reia4304
pp. 1-17.
<https://doi.org/10.24050/reia.v22i43.1841>

✉ *Autor de correspondencia:*

Rincón Velásquez, A. C.
Correo electrónico:
acrinconvelasquez@gmail.com

Recibido: 21-10-2024
Aceptado: 10-12-2024
Disponible online: 01-01-2025

Identificación de riesgo de progresión de la enfermedad de Parkinson utilizando Machine Learning.

✉ ANA CAROLINA RINCÓN VELÁSQUEZ¹

1. Escuela Colombiana de Ingeniería Julio Garavito, Colombia

Resumen

Antecedentes: La Enfermedad de Parkinson (EP) afecta de diferentes formas la vida de las personas que la padecen, quienes inicialmente desconocen cómo será la progresión de la enfermedad dentro de los próximos años y los cambios que implicará en su proyecto de vida. El objetivo de este estudio es implementar un modelo de predicción de riesgo de progresión de Parkinson usando técnicas de aprendizaje autónomo.

Metodología: Se seleccionan 11 variables que son usadas en otros estudios relacionados y son tomadas de datos clínicos de pacientes con EP capturados mediante evaluaciones y cuestionarios. Se implementan modelos de predicción empleando las técnicas de Machine Learning: Regresión Logística, Random Forest y Gradient Boosting Machine (GBM). Los rendimientos son evaluados a través del área bajo la curva Receiver Operating Characteristic (ROC).

Resultados: Se obtiene un conjunto de datos con calidad y según las necesidades del caso de estudio haciendo uso de la base de datos Parkinson Progression Markers Initiative (PPMI) con el cuál se implementa un modelo robusto de riesgo de progresión de EP en un tiempo de 10 años, con un área bajo la curva (AUC) de 0.9, interpretable, confiable y con alta usabilidad para pacientes, cuidadores y personal calificado quienes podrán tomar decisiones más acertadas desde los datos.

Palabras Clave: Parkinson; Predicción; Progresión; Modelo; Aprendizaje autónomo; Datos; Pacientes; Biomarcador; Variables; Desempeño; Usabilidad; Escala; Puntaje.

Identification of Parkinson's disease progression risk using Machine Learning

Abstract:

Background: Parkinson's Disease (PD) affects the lives of patients in different ways, who initially do not know what the progression of the disease will be like in the coming years and the changes it will imply in their life plans. The objective of this study is to implement a Parkinson's progression risk prediction model using autonomous learning techniques.

Methodology: 11 variables are selected that are used in other related studies and are taken from clinical data of patients with PD captured through evaluations and questionnaires. Prediction models are implemented using Machine Learning techniques: Logistic Regression, Random Forest and Gradient Boosting Machine (GBM). Performances are evaluated through the area under the Receiver Operating Characteristic (ROC) curve.

Results: A set of data with quality and according to the needs of the case study is obtained using the PPMI database with which a robust model of risk of PD progression is implemented over a period of 10 years, with an area under the curve (AUC) of 0.9, interpretable, reliable and with high usability for patients, caregivers and qualified personnel who will be able to make more accurate decisions from the data.

Keywords: Parkinson's; prediction; Progression; Model; autonomous learning; Data; Patients; Biomarker; variables; Performance; Usability; scale; Score.

1. Introducción

El Parkinson es una enfermedad neurodegenerativa que no tiene cura, no se comporta de la misma forma en todos los pacientes y, aunque ha sido objeto de estudio en los últimos años, aún no se puede prevenir ni controlar por completo. Cerca de seis millones de personas en el mundo han sido diagnosticadas con esta enfermedad (Fundación Michael J. Fox., 2024), y en Colombia hay una prevalencia estimada de 4,7 (IC95%: 2,2 a 8,9) por 1,000 habitantes (Castro et al., 2014).

Es causado por el déficit de dopamina en el cerebro y se manifiesta inicialmente a través de síntomas como movimientos involuntarios en manos, brazos o piernas, lentitud y/o rigidez muscular. A medida que pasa el tiempo, la enfermedad se va tornando más invasiva, incrementando la severidad de los síntomas y generando otros; impactando significativamente la calidad de vida de quien la padece y de sus familiares, acompañantes y/o cuidadores.

Existen imágenes, pruebas y herramientas que apoyan el diagnóstico y tratamiento de la enfermedad de Parkinson EP, sin embargo, no se cuenta con una prueba que diagnostique EP de manera precisa ni con una herramienta que le indique al paciente cómo será la progresión de su enfermedad en los siguientes años y esto puede ser muy útil para su proyección y calidad de vida. Por ejemplo, saber que en 10 años el paciente ya no tendrá el 80% de independencia en sus actividades y movilidad, sino el 40%. Este estudio tiene como objetivo brindar un modelo de aprendizaje autónomo que pueda ser usado para informar al paciente de EP sobre la progresión de la enfermedad.

2. Metodología:

Pacientes:

Se toman datos de pacientes participantes del estudio PPMI (Parkinson's Progression Markers Initiative) y se adecúan para hallar progresión de estado leve, medido desde el puntaje total de la escala MDS-UPDRS, a un estado moderado/severo en un tiempo de 10 años. Se obtiene un total de 298 participantes. Haciendo uso de los puntos de corte determinados por Martínez-Martín et al. (2015).

Variables:

Tomando como referencia los estudios que utilizan datos de PPMI con ML citados en Gerraty et al. (2023), específicamente el estudio realizado por de Jong et al. (2019) por ser un estudio que logró estratificar a pacientes con enfermedad de Parkinson en subgrupos mediante un método novedoso; se seleccionan:

- Temblor
- Estabilidad Postural
- El total de escala MDS-UPDRS
- Trastorno de conducta del sueño (REM sleep behavior disorder)
- Escala de somnolencia de Epworth (Epworth Sleepiness Scale)
- Actividades de la vida diaria Schwab & England. (Modified Schwab & England Activities of Daily Living)

De otros estudios que usan la misma base de datos (Gramotnev et al., 2019; Nguyen et al., 2020; Salmanpour et al., 2019; Schrag et al., 2017; Tang et al., 2021), se elige la puntuación del cuestionario que mide deterioro cognitivo:

- Evaluación cognitiva de Montreal (MoCA)

Así mismo, se contempla el cuestionario sobre la existencia de comportamientos impulsivo-compulsivos, que son considerados relevantes en EP (Drew et al., 2020; Evans et al., 2019).

- Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease–Rating Scale (QUIP)

Más las siguientes variables demográficas:

- Edad
- Sexo
- Mano dominante

3. Modelos y resultados:

Se implementan algoritmos de machine learning. Para ello, se dividen los datos con el objeto de tener un set de entrenamiento del 80% y otro del 20% para prueba del modelo y, para estandarizar los datos se usa la resta de la media de todos los pacientes y se divide por la desviación estándar de todos los pacientes. Posterior a la prueba, se evalúan los resultados y así se conoce si el modelo es acertado con sus pronósticos.

- Regresión Logística

En este tipo de modelo es importante revisar los coeficientes aprendidos, también llamados Betas.

Tabla 1. Coeficientes aprendidos – Regresión Logística.

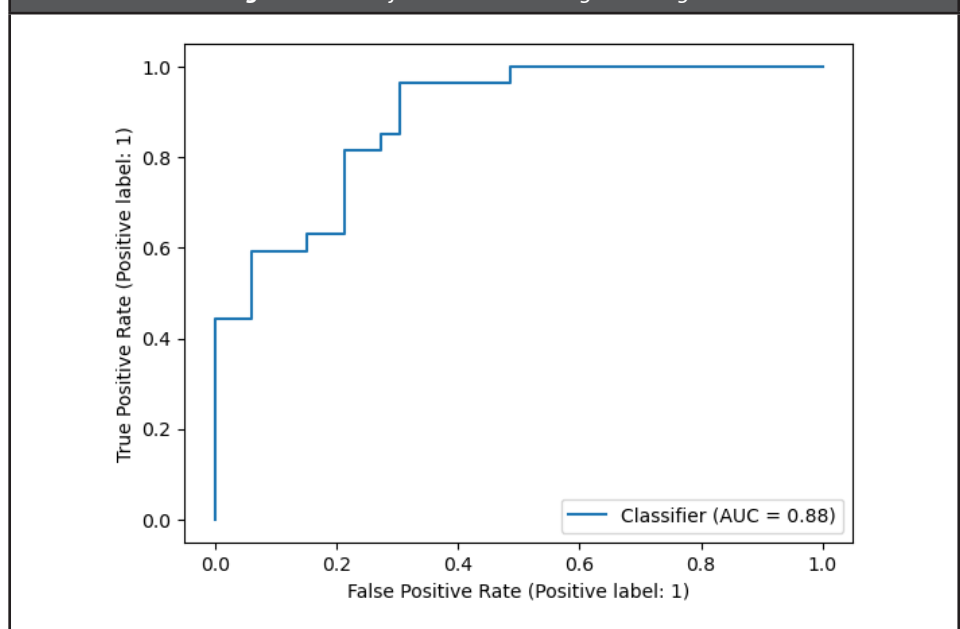
Temblor	Estabilidad Postural	Total MDS_UPDRS	Desorden comportamiento del sueño	Somnolencia	Actividades Diarias	Nivel Cognitivo	Comportamientos compulsivos	Edad	Sexo	Mano Dominante
0.37854715	-0.0806843	1,24424258	0.42983521	0.13894656	-0.1670678	-0.16822	0.1682157	0.13164023	0.10966812	-0.04228

Los anteriores (Tabla 1), son los coeficientes de cada una de las 11 variables seleccionadas.

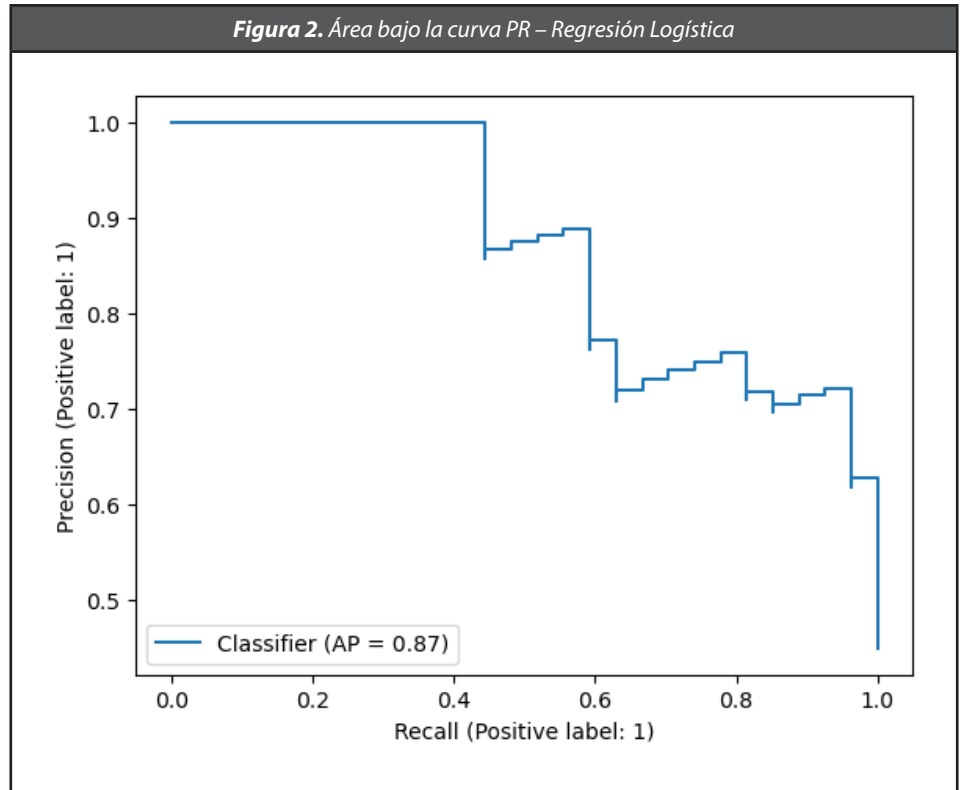
Los betas están mostrando que las variables: Temblor, Total MDS_UPDRS, Desorden Comportamiento del sueño, Somnolencia, Comportamientos Compulsivos, Edad y Sexo tienen una relación positiva con la probabilidad de que haya progresión de EP entonces si aumentan, la probabilidad también. Pasa al contrario con Actividades Diarias y Nivel Cognitivo que a medida que disminuyen, aumentan la probabilidad de progresión. Para el caso de estabilidad postural, la relación es muy pequeña.

Para conocer el rendimiento de los modelos, se grafica el área bajo la curva ROC (AUC), que para este caso es 0.88. Este es un buen rendimiento porque se acerca a 1.

Figura 1 Área bajo la curva ROC – Regresión Logística



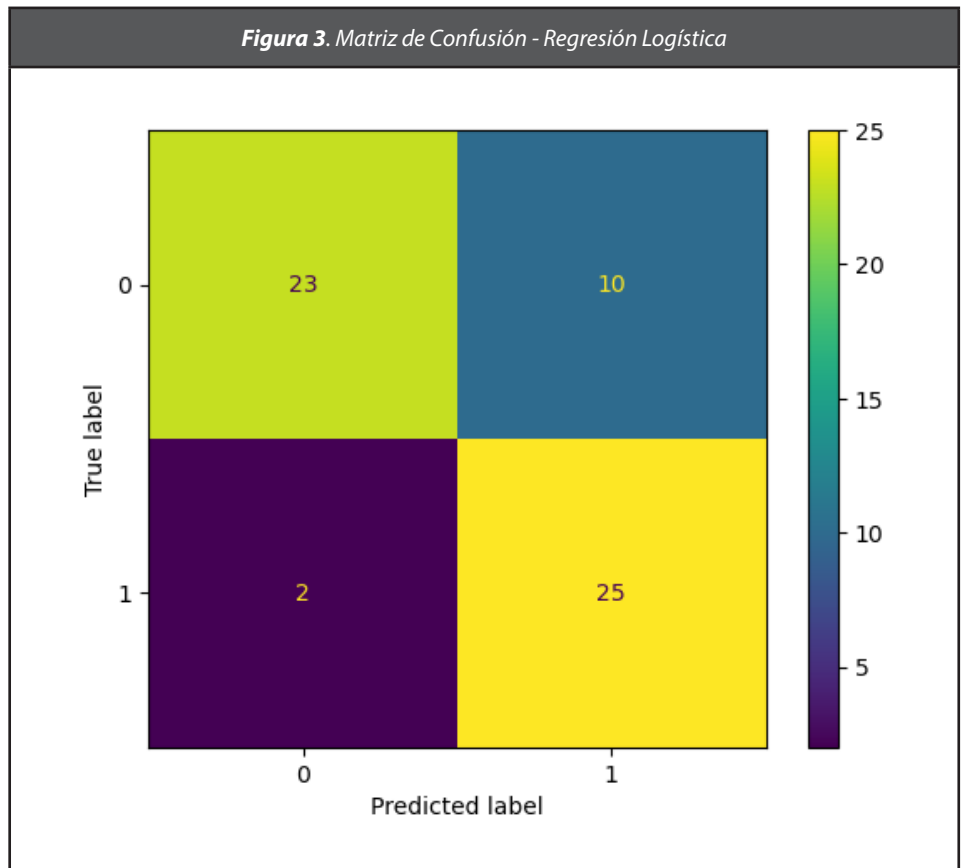
Se dibuja la curva de Precisión – Sensibilidad (PR). Tiene un AUC de 0.87.



Indica que tiene una alta precisión y sensibilidad y se confirma que las clases son balanceadas porque tanto AUC como AUC-PR son cercanas a 0.9.

Para este modelo la exactitud o accuracy es 0.8.

Se grafica la matriz de confusión para ver los casos que acierta correctamente y los que no.



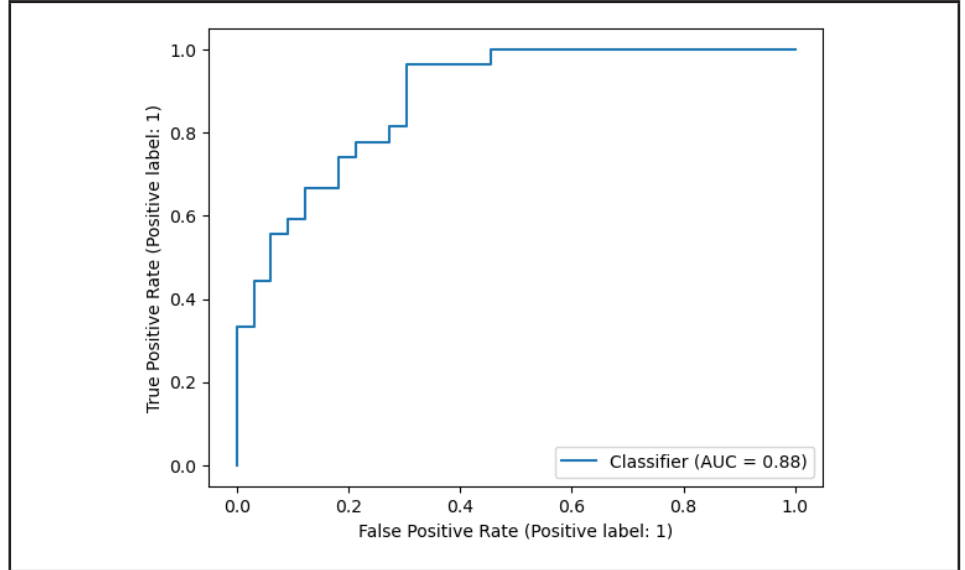
El modelo predijo correctamente 23 casos positivos con EP y 25 casos negativos. Predijo 10 casos como negativos, siendo positivos y 2 como positivos, siendo negativos. En resumen, es un buen modelo porque la tasa de predicciones adecuadamente es más alta que la de predicciones incorrectamente y por los resultados vistos anteriormente.

- Regresión Logística con validación cruzada.

Este modelo se implementa y se hace ajuste de los parámetros relacionados, sin embargo, los resultados son los mismos que en el modelo anterior como se puede ver en la figura 4.

De hecho, el parámetro de regularización óptimo es 0.0001, es muy pequeño lo cual significa que el modelo no tiene sobreajuste u overfitting.

Figura 4. Área bajo la curva ROC - Regresión Logística Validación Cruzada

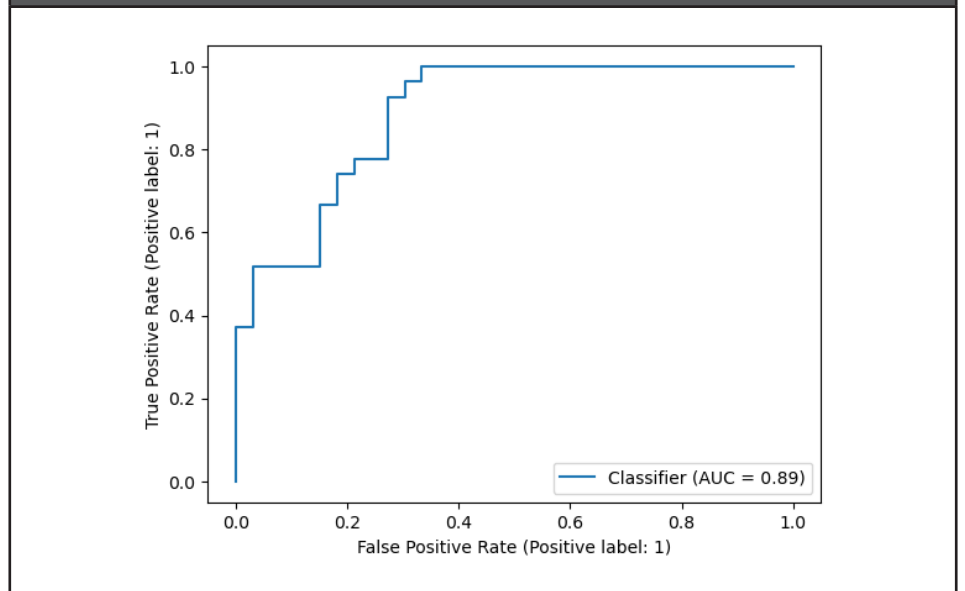


El AUC de la curva ROC es la misma del modelo anterior, 0.88.

- Random Forest.

Al aplicar este modelo, aunque es leve la diferencia, se obtiene un mejor rendimiento con un AUC = 0.89

Figura 5. Área bajo la curva ROC – Random Forest

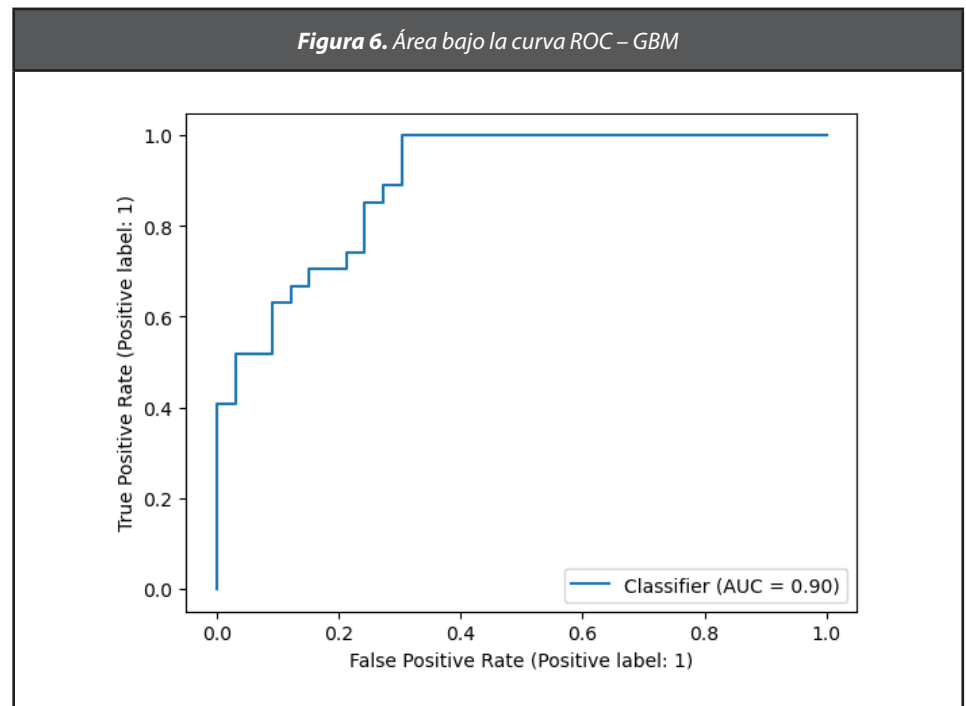


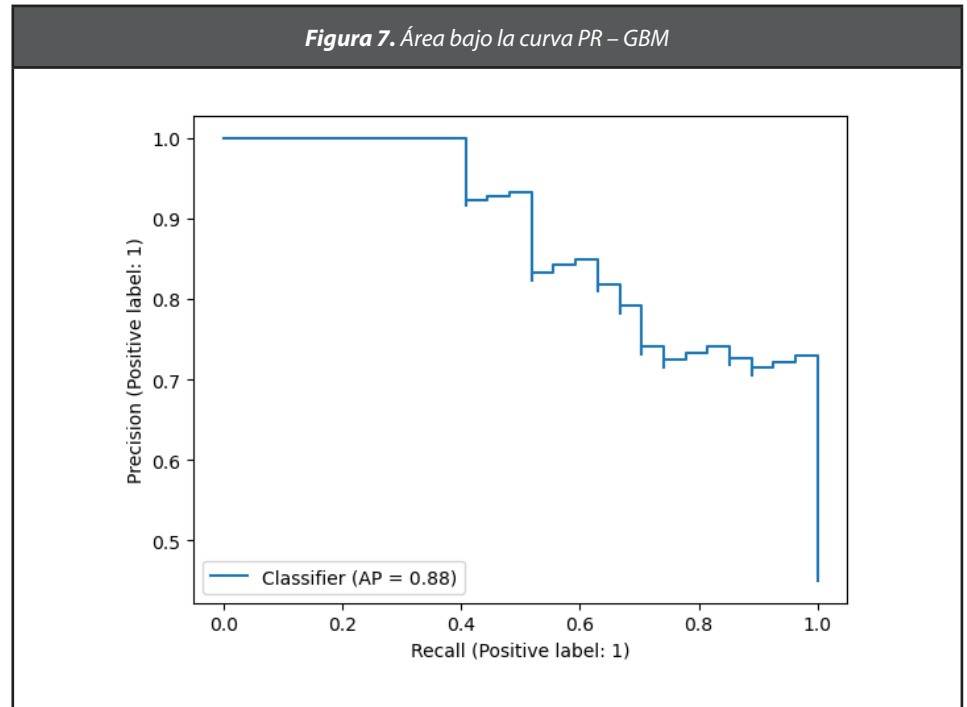
Se ve un leve incremento en AUC de este modelo, significa que discrimina mejor entre una clase y otra. Esto también se debe al tipo de modelo que es más robusto y adecuado para un problema como este de clasificación.

La curva PR se mantiene igual que en los modelos anteriores, con AUC = 0.87.

- Gradient Boosting Machines (GBM).

Implementando este modelo se obtiene un mejor rendimiento por ser de ensamble e ir optimizando su predicción. El área bajo la curva ROC es 0.9. Con respecto a los anteriores, este es el más alto.



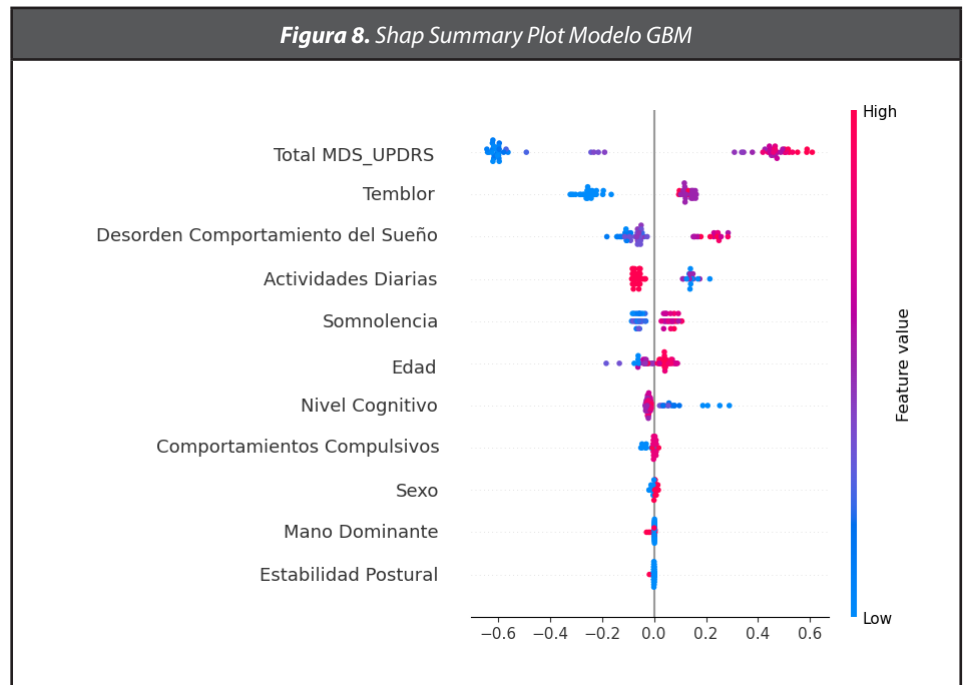


El AUC de la curva PR es 0.88.

En general el rendimiento de los modelos es bueno. Las AUC son muy cercanos al aplicar los diferentes algoritmos, esto se debe a la calidad en el set de datos y al tipo de problema. Sin embargo, se prefiere este último por la precisión y la naturaleza del modelo.

- **Biomarcadores versus Progresión**

Mediante Shap Plot, se interpretan los resultados del modelo GBM que fue el que obtuvo un mejor rendimiento. La figura 8 es un summary plot que muestra cómo cada biomarcador o variable influye en las predicciones.



Es una herramienta muy útil porque permite brindar interpretación al modelo que, aunque tiene un rendimiento muy bueno, no es fácilmente explicable por sí solo, como lo fue la regresión logística brindando muy buena información en el momento de su implementación.

En este caso, se puede ver que en sentido vertical de la gráfica 8 se muestran los biomarcadores en orden de relevancia, siendo los que más influyen los que se encuentran en la parte superior y los de menor impacto en la parte inferior. Así las cosas, el Total de la escala MDS_UPDRS es el más representativo para el modelo y Estabilidad postural es la menos influyente.

En el sentido horizontal, se tiene la relación de la variable con la predicción, hacia la izquierda representa una relación menor o negativa y va incrementando hacia la derecha. Y los colores de los puntos indican el valor del biomarcador, según el mapa del lado derecho. Siendo rojo un valor alto, morado medio y azul bajo.

En este punto del estudio, las variables son consideradas Biomarcadores porque ya hacen parte del modelo construido y que, mediante datos clínicos reales, pronostican progresión de EP en un tiempo de 10 años. Son características del paciente.

Revisando en orden de influencia:

- Para el Total MDS_UPDRS, su relación es positiva con la predicción. Valores altos llevan a casos en los que EP tendrá progresión.
- El Temblor sigue en la escala de relevancia, también tiene una relación positiva; valores medios y altos conducen a un estado de progresión.
- Los desórdenes en el comportamiento del sueño afectan positivamente la predicción, valores altos indican que habrá progresión de EP.
- La independencia para realizar actividades diarias es inversa. Valores bajos de independencia influyen positivamente en la progresión de EP.
- Somnolencia tiene una relación positiva, aunque es menos influyente que los anteriores biomarcadores.
- Con respecto a la edad, se nota que entre mayor es, el pronóstico tiende a ser positivo.
- Nivel cognitivo tiene relación inversa, un déficit en este biomarcador, impacta positivamente la progresión de EP.
- Los comportamientos compulsivos tienen una relación positiva, aunque es menos influyente.
- En cuanto al género del participante o sexo, se visualiza que los hombres que tienen el valor más alto de este biomarcador, mostrado en color rojo, son los que mayor riesgo de progresión tienen.
- Mano dominante y estabilidad postural tienen menor impacto en el pronóstico.

4. Conclusiones y discusión.

La construcción de un conjunto de datos óptimo y adecuado según las condiciones del problema fue fundamental para este trabajo de grado, pues condujo a obtener modelos con alto desempeño de

clasificación de riesgo de progresión de la enfermedad de Parkinson, permitiendo que el resultado pueda ser usado de manera confiable, ya que los hallazgos han sido validados y se pueden generalizar, a diferencia de otros estudios que usan técnicas de machine learning y la base de datos de PPMI. Y, si bien es cierto que este proceso de extracción tomó más tiempo del esperado, fue conveniente porque se comprendió que tiene todo el sentido para la ciencia de datos garantizar modelos obtenidos a partir de datos con calidad y por ende, resultados verídicos.

Una vez se logró obtener el conjunto de datos de interés y modelarlos, se prefirió el resultado del modelo GBM, teniendo en cuenta que se trata de un problema relacionado con datos clínicos y se requiere la mayor precisión posible. Incluso contando con una leve diferencia de 1% con respecto a los otros modelos implementados, ese porcentaje equivale a 3 pacientes en los que se puede predecir su riesgo de progresión para una mejor toma de decisiones en cuanto a su tratamiento y su calidad de vida en los 10 años siguientes al diagnóstico.

Con respecto a la interpretabilidad, se halló que en los modelos de regresión es mayor con relación al modelo elegido, por esta razón se hace uso de SHAP plots sobre el modelo aplicado para dejar claro cómo cada biomarcador influye en la predicción de progresión de EP.

Hay un plus enorme brindado por este estudio y es que ya se conocen los datos de la iniciativa PPMI que cuenta con información muy relevante de pacientes con EP y sigue capturando nuevas características. Porque, aunque la población a nivel mundial con EP es menor a 1%, viéndolo como proporción. Esto equivale aproximadamente a 8 billones de personas, un número considerable. En Colombia, se estiman 244000 habitantes con EP que también es una cifra importante, pero más allá de las cifras lo trascendental es la complejidad de esta enfermedad y otras que son muy similares, como el Alzheimer causante del 10% de las muertes a nivel mundial. Entonces, su estudio se hace primordial, hallar biomarcadores concretos para EP puede dar claridad en los indicadores de otras patologías que aún tienen causas desconocidas, contribuiría a las personas que lo tienen, evitaría que otras lo desarrollen y también se generarían menos costos para todo el sistema de salud.

Bibliografía

- Castro Toro Aracelly, Buriticá Omar Freddy. (2014). Enfermedad de parkinson: criterios diagnósticos, factores de riesgo y de progresión, y escalas de valoración del estadio clínico. <http://www.scielo.org.co/pdf/anco/v30n4/v30n4a10.pdf>.
- DAMA International. (2017). Data Management Body of Knowledge. Second Edition. Basking Ridge, NJ 07920 USA: Technics Publications.
- De Jong Johann, Emon Mohammad Asif, Wu Ping, Karki Reagon, Sood Meemansa, Godard Patrice, Ahmad Ashar, Vrooman Henri, Hofmann-Apitius Martin, Holger Fröhlich. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience*, Volume 8, Issue 11, November 2019, giz134, <https://doi.org/10.1093/gigascience/giz134>.
- Drew Daniel S, Muhammed Kinan, Baig Fahd, Kelly Mark, Saleh Youssuf, Sarangmat Nagaraja, Okai David, Hu Michele, Manohar Sanjay and Husain Masud. (2020). Dopamine and reward hypersensitivity in Parkinson's disease with impulse control disorder. <https://doi.org/10.1093/brain/awaa198>.
- Espinosa-Zúñiga, Javier Jesús. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, 21(3), 00002. Epub 02 de diciembre de 2020. <https://doi.org/10.22201/fi.25940732e.2020.21.3.022>.
- Evans, A. H., Okai, D., Weintraub, D., Lim, S. -Y., O'Sullivan, S. S., Voon, V., Krack, P., Sampaio, C., Post, B., Leentjens, A. F. G., Martinez-Martin, P., Stebbins, G. T., Goetz, C. G., Schrag, A., Bhidayasiri, R., Brown, R. G., Marinus, J., Mestre, T. A., Violante, M. R. & Skorvanek, M. (2019). Scales to assess impulsive and compulsive behaviors in Parkinson's disease: Critique and recommendations. <https://doi.org/10.1002/mds.27689>.
- Fuentes-Arderiu, X. (2013). What is a biomarker? It's time for a renewed definition. *Clinical Chemistry and Laboratory Medicine*, Vol. 51 (Issue 9), pp. 1689-1690. <https://doi.org/10.1515/cclm-2013-0240>.
- Fundación Michael J. Fox. About Parkinson's. (2024). <https://www.michaeljfox.org/parkinsons-101>.
- Gerraty RT, Provost A, Li L, Wagner E, Haas M and Lancashire L. (2023). Machine learning within the Parkinson's progression markers initiative: Review of the current state of affairs. doi: 10.3389/fnagi.2023.1076657
- Gramotnev, G., Gramotnev, D.K. & Gramotnev, A. (2021). Publisher Correction: Parkinson's disease prognostic scores for progression of cognitive decline. *Sci Rep* **11**, 5915. <https://doi.org/10.1038/s41598-020-76437-z>.
- International Parkinson and Movement Disorder Society (MDS). MDS-UPDRS. (2019).

- International Parkinson and Movement Disorder Society. MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). (2024) <https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm>.
- Johns MW. (1991). A new method for measuring daytime sleepiness: The Epworth Sleepiness Scale. *Sleep*. <https://nasemso.org/wp-content/uploads/neuro-epworthsleepscale.pdf>.
- Martínez-Martín P, Rodríguez-Blázquez C, Mario Alvarez, Arakaki T, Arillo VC, Chaná P, Fernández W, Garretto N, Martínez-Castrillo JC, Rodríguez-Violante M, Serrano-Dueñas M, Ballesteros D, Rojo-Abuin JM, Chaudhuri KR, Merello M. (2015). Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. *Parkinsonism Relat Disord*. doi: 10.1016/j.parkreldis.2014.10.026.
- McEwen Barbara. (2023). Administering the MoCA: What Does it Mean and Why is it Important?. <https://rivcodpss.org/sites/g/files/aldnop276/files/2023-04/MoCA%20Presentation.FULL%20PAGE.pdf>.
- Mohammad R. Salmanpour, Mojtaba Shamsaei, Abdollah Saberi, Saeed Setayeshi, Ivan S. Klyuzhin, Vesna Sossi, Arman Rahmim. (2019). Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. <https://doi.org/10.1016/j.compbimed.2019.103347>.
- Nalls Mike A, Keller Margaux F, Hernandez Dena G, Chen Lan, Stone David J, Singleton Andrew B. (2016). Baseline genetic associations in the Parkinson's Progression Markers Initiative (PPMI). doi: 10.1002/mds.26374.
- Nguyen Alex A., Maia Pedro D., Gao Xiao, Damasceno Pablo F., and Raj Ashish. (2020). "Dynamical Role of Pivotal Brain Regions in Parkinson Symptomatology Uncovered with Deep Learning" *Brain Sciences* 10, no. 2: 73. <https://doi.org/10.3390/brainsci10020073>.
- NIH. Form Structure: PDBP Modified Schwab and England Scale - Legacy Data. (2015). https://fitbir.nih.gov/dictionary/publicData/dataStructureAction!lightboxView.ajax?dataStructureName=LD_ModSchwabEnglandScale&publicArea=true.
- Nomura Takashi, Inoue Yuichi, Kagimura Tatsuo, Uemura Yusuke, Nakashima Kenji. (2011). Utility of the REM sleep behavior disorder screening questionnaire (RBDSQ) in Parkinson's disease patients. doi: 10.1016/j.sleep.2011.01.015.
- Orozco, J. L., Valderrama-Chaparro, J. A., Pinilla-Monsalve, G. D., Molina-Echeverry, M. I., Pérez Castaño, A. M., Ariza-Araújo, Y., Prada, S. I., & Takeuchi, Y. (2020). Parkinson's disease prevalence, age distribution and staging in Colombia. *Neurology international*, 12(1), 8401. <https://doi.org/10.4081/ni.2020.8401>.
- Parkinson's Progressive Markers Initiative (PPMI). (2023). Data User Guide – updated August 22nd, 2023. PPMI Dataset.

Schrag Anette, Siddiqui Uzma Faisal, Anastasiou Zacharias, Weintraub Daniel, Schott Jonathan M. (2017). Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. [https://doi.org/10.1016/S1474-4422\(16\)30328-3](https://doi.org/10.1016/S1474-4422(16)30328-3).

Section. (2024). How to Analyze Machine Learning Models using SHAP. <https://www.section.io/engineering-education/how-to-analyze-machine-learning-models-using-shap/>.

Tang Chunyan, Zhao Xiaoyan, Wu Wei, Zhong Weijia, Wu Xiaojia. (2021). An individualized prediction of time to cognitive impairment in Parkinson's disease: A combined multi-predictor study. <https://doi.org/10.1016/j.neulet.2021.136149>.