UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE CIENCIA E INGENIERÍA EN ALIMENTOS Y BIOTECNOLOGÍA

CARRERA DE BIOTECNOLOGÍA

Análisis de microarrays para la expresión diferencial del gen LRRK2 asociado a la enfermedad de Parkinson

Informe Final del Trabajo de Titulación, Opción Proyecto de Investigación, previo a la obtención del título de Ingeniero Biotecnólogo, otorgado por la Universidad Técnica de Ambato, a través de la Facultad de Ciencia e Ingeniería en Alimentos y Biotecnología

Autor: David Sebastián Santamaría Jácome

Tutor: M.Sc. Cristian Fernando Galarza Galarza

Ambato – Ecuador

Febrero – 2024

APROBACIÓN DEL TUTOR

Mg. Cristian Fernando Galarza Galarza

CERTIFICA:

Que el presente Informe Final del Trabajo de Titulación ha sido prolijamente revisado. Por lo tanto, autoriza la presentación de este Informe Final del Trabajo de Titulación, opción Proyecto de Investigación, el mismo que responde a las normas establecidas en el Reglamento de Títulos y Grados de la Facultad de Ciencia e Ingeniería en Alimentos y Biotecnología.

Ambato, 12 de enero de 2024

Mg. Cristian Fernando Galarza Galarza
C.I.: 1803160272
TUTOR

AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, David Sebastián Santamaría Jácome, manifiesto que los resultados obtenidos en el presente Informe Final del Trabajo de Titulación, opción Proyecto de Investigación, previo a la obtención del título de Ingeniero Biotecnólogo, son absolutamente originales, auténticos y personales, a excepción de las referencias bibliográficas.

David Sebastián Santamaría Jácome

C.I.: 1850150036

AUTOR

DERECHOS DE AUTOR

Autorizo a la Universidad Técnica de Ambato, para que haga de este Informe Final del Trabajo de Titulación o parte de él un documento disponible para su lectura, consulta y proceso de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales de mi Informe Final del Trabajo de Titulación, con fines de difusión pública, además apruebo la reproducción de este, dentro de las regulaciones a la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor.

David Sebastián Santamaría Jácome

C.I.: 1803983327

AUTOR

APROBACIÓN DEL TRIBUNAL DE GRADO

Los suscritos profesores calificadores, aprueban el presente Informe Final del Trabajo de Titulación, opción Proyecto de Investigación, el mismo que ha sido elaborado de conformidad con las disposiciones emitidas por la Facultad de Ciencia e Ingeniería en Alimentos y Biotecnología de la Universidad Técnica de Ambato.

Para Constancia firman:

Firmado electrónicamente por:
LILIANA PAULINA LALALEO CORDOVA

Presidente de Tribunal

Firmado electrónicamente por:
DANAE FERNANDEZ RIVERO

Msc. Danae Fernández Rivero

1757181209

Firmado electrónicamente por:
PABLO VINICIO TUZA ALVARADO

Dr. Pablo Vinicio Tuza Alvarado

1104063241

Ambato, 05 de febrero de 2024

DEDICATORIA

Principalmente dedico este trabajo a mis padres y hermano,
esperando que lo sientan como suyo, puesto que,
cada charla, pregunta o sugerencia ha sido
de gran ayuda para el desarrollo de este trabajo.
Nunca olviden que sin importar nada les amo demasiado
y sin sus cuidados, paciencia y amor, no me
sentiría tan completo como lo hago en este momento.
A mi "Huesos" que siempre cumplió el rol de segunda madre
y con su mirada vigilante y cariño formó parte de esta travesía
A mis ángeles "Papi Lucho" y "Mami Fanny" por
siempre brindarme su cariño sé que en donde sea que
estén, me seguirán cuidando y amando.
A "Scrapy", gracias por ver en mi un ejemplo a seguir
sin importar mis errores, espero que esto te motive
a seguir adelante y nunca darte por vencido.

- David Sebastián Santamaría Jácome-

¡Del fracaso se aprende!
Del éxito, mmm…No mucho"

- La Familia del Futuro –

AGRADECIMIENTO

Agradezco a la vida por permitirme estar aquí el día de hoy a pesar de la adversidad, por lo bueno y por lo malo

Agradezco infinitamente a mis padres Edgar y Mónica, por nunca dejarme solo en este camino, sus consejos y apoyo incondicional me han permitido forjar un carácter fuerte y perseverante, estoy seguro que sin su motivación y confianza mi historia fuera completamente distinta.

A mi hermano, mi mejor amigo de toda la vida, gracias por creer en mí, aun cuando ni siquiera yo lo hacía, depositando una fe ciega que me motivaba a superarme cada día, gracias por ser mi maestro de vida y apoyo cuando sentía que no podía más.

A mi tutor, Mg. Cristian Galarza por no ser únicamente un docente, sino también un amigo y de manera genuina me brindó consejos y enseñanzas que sirvieron de guía durante la elaboración de este trabajo.

A todos los amigos que he hecho durante esta etapa académica, gracias por compartir conmigo momentos de alegría, tristeza, ira e impotencia; en especial a mis amigos Víctor, Lucho y Paez, que siempre han confiado en mis capacidades, me han brindado ayuda y han valorado mi apoyo y amistad.

A todas las personas que he conocido a lo largo de mi etapa educativa y que, con su afecto y estima me han motivado a seguir adelante. A mi familia, y en especial a mis abuelos, por su paciencia, sus sonrisas y sus ganas de verme triunfar.

Finalmente agradezco a la Universidad Técnica de Ambato, directivos y profesores por formarme como profesional con paciencia y dedicación, brindándome los conocimientos necesarios para esta nueva etapa de mi vida.

ÍNDICE DE GENERAL DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN EJECUTIVO

Las enfermedades neurodegenerativas se encuentran en un estado de evolución constante, a pesar de esto, muchas no han sido comprendidas en su totalidad lo que implica un conocimiento limitado acerca de sus repercusiones en el organismo, este es el caso de la enfermedad del Parkinson. En este contexto las herramientas bioinformáticas emergen como una esperanza para la comprensión de las funciones biológicas afectadas por la enfermedad; mediante la aplicación de algoritmos como el análisis de expresión diferencial o aprendizaje automático para telemonitorización temprana.

Se utilizaron datos extraídos de células madre neuronales (NSC) derivadas de células madre embrionarias (hESC) que albergan una mutación patógena LRRK2 (G2019S) para el análisis de datos Ómicos. Gracias al enriquecimiento funcional de las muestras, es posible determinar la transcriptómica de la mutación. Para el caso del método de telemonitoreo, se emplearon 3 algoritmos de aprendizaje automático, utilizando un dataset de 52 variables con 252 observaciones; correspondientes a muestras de voz de pacientes con y sin Parkinson, puesto que esta empieza a presentar síntomas de disfonía al poco tiempo de contraer la afección.

Se localizaron un total de 139 genes ligados a esta mutación y por ende a la enfermedad del Parkinson y su desarrollo, la mayoría de estos se encuentran presentes en procesos cerebrales, principalmente en los impulsos nerviosos y componentes estructurales del sistema nervioso central. El método predictivo desarrollado cuenta con un nivel de exactitud del 0.81 mediante la aplicación de un algoritmo de "Machine learning" robusto y de alto impacto científico denominado "Random forest"

Palabras clave: Biotecnología médica, trastornos neurodegenerativos, análisis de microarrays, LRRK2, enfermedad de Parkinson (PD).

ABSTRACT

Neurodegenerative diseases are in a constant state of evolution; however, many of them have not been fully understood, leading to limited knowledge about their implications on the organism. This is notably observed in the case of Parkinson's disease. In this context, bioinformatics tools emerge as a hopeful avenue for understanding the biological functions affected by the disease. This is achieved through the application of algorithms such as differential expression analysis or machine learning for early telemonitoring.

Data extracted from neural stem cells (NSCs) derived from embryonic stem cells (hESCs) carrying a pathogenic LRRK2 (G2019S) mutation were used for Omics data analysis. Functional enrichment of the samples allows for the determination of the transcriptomics of the mutation. Regarding the telemonitoring method, three machine learning algorithms were employed, utilizing a dataset with 52 variables and 252 observations, corresponding to voice samples from patients with and without Parkinson's disease. This is crucial as the disease tends to manifest dysphonia symptoms shortly after contraction.

A total of 139 genes linked to this mutation and consequently to Parkinson's disease and its development were identified. The majority of these genes are present in cerebral processes, particularly in nerve impulses and structural components of the central nervous system. The developed predictive method boasts an accuracy level of 0.81 through the application of a robust and scientifically impactful machine learning algorithm known as "Random Forest."

Keywords: Medical biotechnology, neurodegenerative disorders, microarray analysis, LRRK2, Parkinson's disease (PD).

CAPÍTULO I

MARCO TEÓRICO

1.1 Antecedentes investigativos

1.1.1 Justificación

La cantidad de personas diagnosticadas con trastornos neurodegenerativos aumenta considerablemente con el paso del tiempo, esto representa una amenaza para la salud pública. Tan solo en 2019, aproximadamente 50 millones de personas padecían este tipo de enfermedades en todo el mundo, estas a menudo provocan demencia. Se espera que para el año 2060 esta cifra aumente a 152 millones **(Armstrong, 2020)**. Ejemplos de enfermedades neurodegenerativas son la enfermedad de Alzheimer, la enfermedad de Parkinson, la enfermedad de Huntington, la esclerosis lateral amiotrófica, la demencia frontotemporal y las ataxias espinocerebelosas **(Gitler et al., 2017)**, la mayoría de las veces estas patologías se encuentran asociadas con la edad. Aun así, existen diversos factores externos que pueden verse involucrados al momento de contraer algún tipo de esta enfermedad, de entre los cuales se puede destacar factores de riesgo asociados con la dieta, infección patógena, traumatismo craneoencefálico, función mitocondrial, productos farmacéuticos, exposición a metales, entre otros **(Armstrong, 2020; B. R. De Miranda et al., 2022)**.

La enfermedad del Parkinson (PD; Parkinson's disease), alguna vez considerada extraña por la poca relevancia que se le daba a su estudio, **(B. R. De Miranda et al., 2022)** es hoy en día la segunda enfermedad neurodegenerativa más común en el mundo después del Alzheimer **(Reich & Savitt, 2019)**. Se caracteriza por el movimiento involuntario de los músculos, temblor corporal y debilidad muscular en partes que no están en acción e incluso cuando se encuentran en reposo **(Mehanna & Jankovic, 2019).** Es causada por la degeneración de las neuronas dopaminérgicas en el cuerpo estriado **(Boulos et al., 2019)**, estudios sugieren que la PD se desarrolla a través de seis etapas, que se inician en una ubicación periférica y va avanzando hacia el sistema

nervioso central (CNS) a través del olfato o nervios vagales en fases presintomáticas, afectando así a la sustancia negra, el tálamo y la amígdala **(Bloomingdale et al., 2022; Boulos et al., 2019)**. De hecho, se cree que esta patología se desarrolla durante una década o un poco antes de que se manifiesten los síntomas motores clínicos **(Bloomingdale et al., 2022)**.

En nuestro país, este trastorno se encuentra presente con una prevalencia de 243 casos por cada 100 000 habitantes **(Montalvo et al., 2017)**. Según **Cerri et al. (2019)**, el riesgo de desarrollar Parkinson es el doble en hombres que en mujeres, sin embargo, en las mujeres la tasa de mortalidad y progresión es más alta y rápida. El gen *LRRK2* presenta un papel importante en el tráfico intracelular y el mantenimiento de órganos subcelulares **(Kessler et al., 2018)**, trabaja en conjunto con múltiples socios funcionales, entre los que se destaca el gen *Rab7L1*, regulando de manera coordinada el crecimiento de neuritas y el tráfico intracelular **(Fujimoto et al., 2018)**.

Existen varias tecnologías que permiten estudiar la expresión de varios genes a la vez, entre ellas se encuentran los microarrays. La detección en esta tecnología se logra gracias a la hibridación de las moléculas de ADN o ARN, las cuales producen fluorescencia **(Moreno & Solé, 2004).** Esta información se encuentra almacenada en la base de datos GEO del Centro Nacional de Información Biotecnológica NCBI, donde se encuentran bases de datos de ensayos experimentales cuyas muestras se extraen del cerebro completo de ratón, incluido bulbo olfatorio, cerebelo y tronco encefálico, específicamente, neuronas corticales **(Bonin et al., 2017; Jo et al., 2021)**.

Finalmente, con este proyecto se provee de un punto de partida para la implementación de un método predictivo y preventivo para el diagnóstico de la enfermedad basado en técnicas de Machine Learning, además, se verificará la importancia de la mutación del gen mediante un análisis de expresión diferencial, con lo cual, se corrobora la tendencia que tiene el paciente a padecer la enfermedad.

1.1.2 Enfermedad de Parkinson y sus causas comunes

La enfermedad de Parkinson es un síndrome clínico reconocible con una variedad de causas y presentaciones clínicas **(Tan et al., 2018)**. Esta representa una condición neurodegenerativa de rápido crecimiento; cuenta con una prevalencia ascendente en todo el mundo, lo que le brinda características similares a las de una pandemia, excepto por una causa infecciosa **(Bhat et al., 2018)**. Entre el 3% y el 5% de la enfermedad de Parkinson se explica por causas genéticas vinculadas a genes conocidos de la enfermedad **(Williams-Gray & Worth, 2023)**. Mientras que 90 variantes genéticas explicarían el riesgo hereditario de la enfermedad de Parkinson no monogénica (entre 16% y 36%) **(Reich & Savitt, 2019)**. Esto dificulta aún más la situación, puesto que, debido a la variabilidad de los sistemas humanos no se puede determinar con exactitud la probabilidad de contraer PD en base a los niveles de exposición a factores exógenos, sin mencionar que estos factores no genéticos siguen siendo subestimados y poco estudiados **(B. R. De Miranda et al., 2022; Simon et al., 2020)**.

1.1.2.1 Factores Genéticos

La PD autosómica de aparición temprana se ha visto fuertemente influenciada por mutaciones en los genes PARKIN y PINK1 **(Kitada et al., 1998; Valente et al., 2004)**, ya que están asociados con una vía celular que involucra la eliminación selectiva de mitocondrias disfuncionales en los lisosomas a través de un proceso conocido como "mitofagia" **(Simon et al., 2020)**. Esto quiere decir que la perdida de funciones de estos genes resultaría en una mitofagia deteriorada y por ende en la acumulación de estas mitocondrias disfuncionales **(Shin et al., 2011)**. PARKIN regula indirectamente los niveles de PGC-1alfa, un regulador transcripcional, que controla de manera coordinada la expresión de genes necesarios para la biogénesis mitocondrial **(Shin et al., 2011)**, se ha descubierto que los niveles de PGC-1alfa también son bajos en la PD esporádica **(Zheng et al., 2011)**.

El Parkinsonismo se observa comúnmente en pacientes con mutaciones en varios genes DYT, incluyendo los que están involucrados en la vía de síntesis de dopamina. La mutación en el gen *LRRK2*, es la causa más frecuente de la enfermedad de

Parkinson monogénico o familiar **(Watanabe et al., 2020)**. **Tolosa et al. (2020)** afirman que "estas mutaciones y en concreto la Gly2019Ser, se observan en pacientes con Parkinson autosómico dominante y en aquellos con Parkinson esporádico aparente". Sin embargo, esta no es la única "falla" genética asociada a la enfermedad, también es posible encontrar mutaciones patogénicas que residen en los dominios ROC-COR GTPasa (R1441G/C/H, Y1669C) y quinasa (2020T). De igual forma, el bloqueo farmacológico de los receptores de dopamina puede llegar a ser un causante válido **(Rocha et al., 2022)**.

Adicionalmente, según **Genis et al. (2018)** la presencia del alelo ε4 del gen que codifica la apolipoproteína E (ApoE), se ha asociado como factor de susceptibilidad para desarrollar Parkinson, lo cual se encuentra relacionado con la enfermedad de Alzheimer, al ser ambos padecimientos neurodegenerativos **(Elizondo-Cárdenas et al., 2011)**. Estos vínculos genéticos implican una disfunción del recambio mitocondrial en la PD.

1.1.2.2 Factores exógenos

Exposición a sustancias químicas tóxicas: Gracias a estudios concretos se sabe que las posibilidades de contraer Parkinson son mayores en aquellos individuos que se encuentren expuestos a pesticidas tanto de manera directa como "pasiva". Los químicos agrícolas asociados con la PD son el paraquat, la rotenona, el 2,4-D y varios productos ditiocarbamatos y organoclorados, sin mencionar otro tipo de sustancias químicas como los disolventes clorados **(Y. Chen et al., 2021)**. A pesar de que estos productos hoy en día ya no se utilizan con la misma frecuencia por su impacto en el ambiente y la salud, sus efectos persisten en el ambiente, sin mencionar que de igual forma siguen siendo contaminantes comunes de suelos y aguas aledañas a los sectores agrícolas en donde eran aplicados **(Simon et al., 2020)**.

Lesión craneal: Las lesiones a nivel de cráneo, fuertes o moderadas en individuos pueden ser capaces de aumentar el riesgo de padecer PD, aunque esto no es aplicable en todos los casos, el riesgo aumenta de 2 a 5 veces con el número de lesiones, y los factores de susceptibilidad genética **(Bloem et al., 2021; Simon et al., 2020)**.

Estilo de vida: Los hábitos alimenticios son aquellos que tienen una mayor influencia en el desarrollo de la PD, de entre las causas más comunes se tiene el consumo de alcohol, café, tabaco o incluso té **(Y. Chen et al., 2021)**, aun así, estos suponen un riesgo reducido para contraer la enfermedad. El elemento dietético cuya ingesta descontrolada simboliza un mayor riesgo de padecer la enfermedad son los productos lácteos, según **Boulos (2019),** "Esta relación puede deberse posiblemente a la concentración de sustancias tóxicas presentes en la leche".

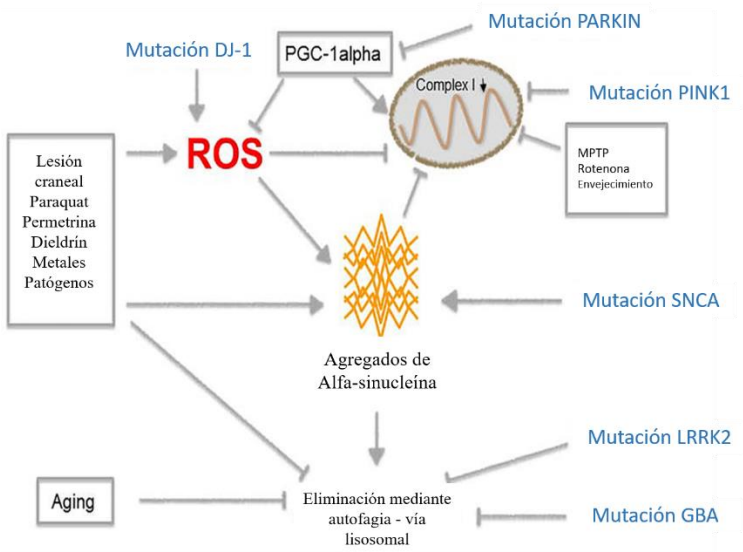

Figura 1. *Factores ambientales y genéticos que influyen en la patogénesis de la EP al afectar vías similares.*

Fuente: **(Simon et al., 2020)**

1.1.3 Factores de riesgo del Parkinson

La PD es considerada altamente compleja, por ende, su estudio requiere de tiempo y técnicas especializadas **(B. R. De Miranda et al., 2022)**, sin mencionar que puede ser causante de otro tipo de enfermedades, que a futuro empeoran la situación del paciente.

1.1.3.1 Demencia con cuerpos de Lewy (DLB)

Patológicamente, la PD se define por la pérdida de neuronas dopaminérgicas en la sustancia negra pars compacta (SN) ubicada en el mesencéfalo y la acumulación de

cuerpos de Lewy **(Bloem et al., 2021)**, estos últimos son inclusiones citoplasmáticas que incluyen agregados anormales de proteínas (α-sinucleína), que se acumulan en las células nerviosas de ciertas áreas del cerebro **(Lewis & Spillane, 2019)**. De la misma forma, la acumulación de cuerpos de Lewy en extensiones neuronales puede denominarse como neuritas de Lewy **(Harvey et al., 2023)**.

Consecuentemente, la demencia con cuerpos de Lewy se encuentra completamente asociada a la PD, esto quiere decir que cuando un paciente que padece PD, muy probablemente contraerá DLB en un futuro, de hecho, se establece que el deterioro cognitivo de PD es frecuente a medida que la enfermedad avanza y aproximadamente en el 50 % de estos casos presentan demencia dentro de los 10 años posteriores al diagnóstico **(Harvey et al., 2023; Hayes, 2019)**.

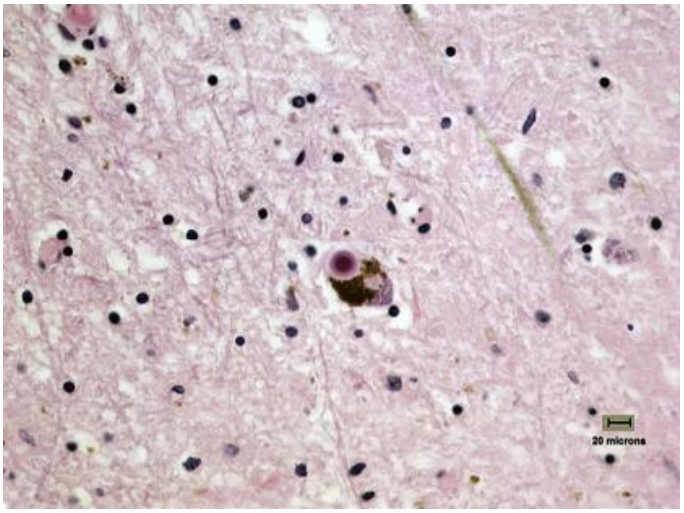

Figura 2. *Vista de un cuerpo de Lewy*
Fuente: **(Hayes, 2019)**

1.1.4 Tratamientos contra el Parkinson

Los efectos resultantes del padecimiento de la enfermedad no se limitan únicamente a los síntomas motores, sin embargo, estos son los que poseen mayor impacto dentro de la calidad de vida del paciente. Dentro de los síntomas no motores se pueden encontrar el estreñimiento, dismotilidad gástrica, alteraciones del sueño y depresión; los cuales aparecen mucho antes que los ya conocidos "temblores" **(Bloomingdale et al., 2022)**.

La decisión de empezar con un tratamiento para la PD se toma en conjunto con el paciente, debido a que esta se basa en el impacto de los síntomas, es decir, en cómo estos afectan su capacidad de realizar actividades diarias **(Hayes, 2019)**.

1.1.4.1 Tratamiento farmacológico

La farmacoterapia dopaminérgica es una de las cuatro estrategias principales para combatir el Parkinson **(Santos-García et al., 2023)**. Levodopa es el primer medicamento de mayor eficacia en el tratamiento de la enfermedad, sin embargo, su uso de esta se encuentra limitado debido a sus características toxicológicas y la probabilidad de que esta acelere la progresión de la patología al promover el estrés oxidativo **(Bloem et al., 2021; Hayes, 2019)**. Es ahí en donde entran en acción los agonistas de la dopamina (pramipexol, ropinirol y rotigotina), los cuales estimulan los receptores dopaminérgicos en el sistema nervioso central y por ende, se emplean mayoritariamente para aliviar efectos secundarios del Parkinson **(Bloem et al., 2021; Hayes, 2019)**.

Medicamentos anticolinérgicos (trihexifenidilo y benzotropina) no son eficaces en el tratamiento de la bradicinesia, sin embargo, pueden tener impactos favorables en la disminución de la rigidez, distonía y temblores, de igual forma, los inhibidores de la monoaminooxidasa aldehído deshidrogenasa B (MAO-B) (rasagilina y selegilina) inhiben las enzimas implicadas en la degradación de levodopa y dopamina **(Bloem et al., 2021; Jankovic & Tan, 2020)**.

Así mismo, se ha demostrado que la nobiletina tiene capacidad antineuroinflamatoria al inhibir la producción y secreción de mediadores proinflamatorios inducida por lipopolisacáridos, además, su uso en modelos animales demuestra que es capaz de proteger las neuronas dopaminérgicas en la SN **(Nakajima & Ohizumi, 2019)**. Adicionalmente y como método alternativo se pretende emplear curcumina para reducir el impacto de los síntomas de la PD, según **Abrahams (2021),** "aunque sea especulativo, es probable que la curcumina resulte beneficiosa antes de que se produzcan síntomas de PD o daño celular excesivo y, por lo tanto, puede ser útil como

suplemento dietético antes de la aparición de la enfermedad.” Esto al hablar específicamente de la enfermedad de Parkinson mutante LRRK2.

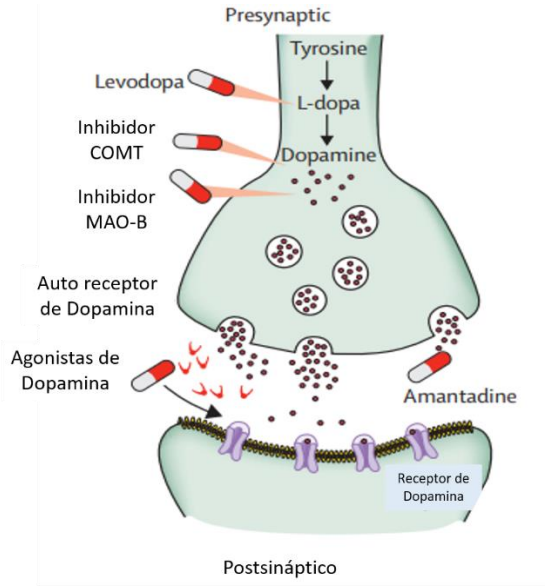

Figura 3. *Sitios de acción de los distintos fármacos antiparkinsonianos*
Fuente: **(Bloem et al., 2021)**

1.1.4.2 Tratamiento microquirúrgico funcional

A pesar de que este método no ha sido utilizado desde que la levodopa se convirtiera en el tratamiento de mayor impacto (a principios de 1960), este procedimiento tuvo un resurgimiento como cirugía ablativa estereotáxica en el tálamo, el núcleo subtalámico y el globo pálido interno, empleando electrodos estimulantes en esos núcleos. Este procedimiento, conocido como *estimulación cerebral profunda* (DBS) se ha convertido en un elemento básico del tratamiento en pacientes con complicaciones en el tratamiento médico convencional, que incluyen fluctuaciones motoras precipitadas e impredecibles, discinesia incapacitante o la presencia de temblores intratables **(Hayes, 2019)**.

1.1.5 Microarrays en la lucha contra el Parkinson

1.1.5.1 Microarrays de expresión

Los avances en la investigación sobre la relación gen-enfermedad han sufrido una revolución técnica significativa gracias a las nuevas tecnologías de análisis genético, mediante la introducción, uso y manejo de herramientas de base genómica **(Tell-Marti et al., 2021)**. Por esta razón, los días en los que se analizaba un gen único y sus efectos quedaron en el pasado **(Angelescu & Dobrescu, 2021)**. En la actualidad, los avances tecnológicos brindan la capacidad de analizar el comportamiento de miles de genes de forma simultánea. Uno de estos sistemas, son los Microarrays, este enfoque cambia la forma en la que se plantean problemas y se obtienen soluciones de los experimentos en enfermedades crónicas degenerativas, esto gracias a que brindan una visión completa del genoma en conjunto **(Yu et al., 2023)**.

Un microarray es un soporte sólido, al que gracias a la ayuda de máquinas especializadas se le han agregado cadenas de material genético mayoritariamente ADN complementario (ADNc) en forma de una matriz ordenada de miles de puntos (hasta 40,000 fragmentos distintos por cada centímetro cuadrado de espacio) perfectamente espaciados entre sí **(Marjit et al., 2023).** Cada punto contiene millones de secuencias clonadas "idénticas" y se podría decir que disponen prácticamente de todo el genoma en estudio **(Lacombe & Rooryck-Thambo, 2018)**, es por esto que, permiten medir simultáneamente la actividad e interacción de cientos de genes.

1.1.5.2 Tipos de microarrays

Microarrays de dos canales: En este tipo de microarray, las pruebas son oligonucleótidos, ADNc o pequeños fragmentos de PCR, que corresponden con ARN mensajero (ARNm). Su funcionamiento se centra en la utilización de preparaciones obtenidas a partir de dos muestras biológicas distintas, en donde cada una se marca con un fluoróforo diferente. Finalmente, las dos preparaciones marcadas se mezclan e hibridan sobre el mismo chip de ADN. De esta forma, se pueden observar genes que se activan o se reprimen en distintas condiciones, sin embargo, no se pueden observar

niveles absolutos en la expresión y requieren PCR cuantitativa (qPCR) para un análisis absoluto **(Hung & Weng, 2017)**.

Chips de ADN de oligonucleótidos: También llamados microarrays de canal único, las pruebas son designadas a partir de una secuencia conocida o un ARNm predicho. Estas secuencias se construyen mediante la elongación secuencial de una cadena en crecimiento con un solo nucleótido empleando fotolitografía, un proceso por el cual se transfiere un patrón desde una máscara fotográfica a una superficie sólida **(Betanzos et al., 2019)**. Generando estimaciones del nivel de expresión, pero en caso de que se requieran distintas condiciones, no pueden ser observadas en una misma matriz, debido a que no están basadas en la hibridación competitiva, en otras palabras, un chip para una muestra **(Hung & Weng, 2017)**.

1.1.5.3 Análisis de datos de microarrays

La imagen se obtiene tras la excitación del tinte fluorescente de cada target mediante una luz monocromática producida por un láser o luz blanca y colectando la luz de emisión (fluorescencia) convirtiendo la corriente de fotones en valores digitales que pueden ser almacenados en una computadora como un archivo de imagen (.TIFF). Los tintes más usados son los de cianina, Cy3 (verde) y Cy5 (rojo), los cuales tienen emisiones en los rangos de 510-550 nm y 630-660 nm, respectivamente **(Jin et al., 2022)**. Tras el análisis de la imagen obtenemos archivos del tipo ".gpr" o ".spot", dependiendo del software utilizado. Estos contienen, las intensidades de cada canal correspondientes a los spots Foreground, background, medidas cualitativas (medidas de variabilidad, tamaño del spot, medidas de circularidad, entre otras) **(Scionti et al., 2022).**

La información resultante de los análisis de microarrays cuenta con varios niveles de complejidad. El nivel más bajo corresponde a datos simples sobre genes concretos, es decir, identifica el nivel de expresión de un gen individual asociado a un determinado fenotipo, simulando a experimentos tradicionales. Ahora bien, el nivel de alta complejidad, facilita la exploración de la figura completa del conjunto del transcriptoma, es decir, de todos los genes que se expresan de forma asociada a ese

fenotipo, magnitud que es imposible lograr con otros sistemas distintos **(Agapito & Arbitrio, 2022)**.

Los datos que se generan con microarrays, aparte de tener un gran volumen, se caracterizan por ser altamente variables, por tal motivo es básico un análisis estadístico y en ocasiones el diseño experimental que se plantee para solucionar las diferentes cuestiones biológicas propuestas **(Agapito & Arbitrio, 2022)**, En resumen, una cantidad significativa de datos es redundante y es esencial filtrar cuidadosamente los genes importantes. Para el análisis estadístico de estos datos se utiliza R (versión 3.3.2, free software environment for statistical computing and graphics) debido a su versatilidad y capacidad de utilizar "Librerías" que facilitan el análisis de estos **(de Lima et al., 2022)**. Se apoya en los paquetes de "Bioconductor", un proyecto de código abierto con acceso a potentes métodos estadísticos y gráficos para el análisis de datos genómicos.

1.1.5.4 Microarrays en el estudio de la enfermedad de Parkinson

Se han identificado mutaciones en cuatro genes, α-sinucleína, parkina, DJ 1 y UCH-L1, en formas de enfermedad de Parkinson hereditarias autosómicas, sin embargo, el gen *LRRK2* no ha sido fuente de estudio, a pesar de ser una causa directa de la enfermedad, en su mayoría, puede deberse a que este se encuentra presente en ambos tipos de enfermedad (Genética y adquirida por naturaleza), así que no resulta tan llamativa, situación errónea puesto que analizando las mutaciones de *LRRK2* es posible ampliar el método de diagnóstico e implementar un método predictivo de la enfermedad sin importar su origen **(Fujimoto et al., 2018)**. UCH-L1 ha sido el gen que mayor impacto ha tenido dentro de los análisis de microarrays, en su mayoría las mutaciones de este ya se encuentran cubiertas en su totalidad y tienen una relación ya establecida con la enfermedad. Lo que normalmente se emplea para este tipo de estudios es una serie de muestras de pacientes que presenten la enfermedad, y un grupo control en donde el trastorno se encuentre en fase latente o inexistente **(Jo et al., 2021)**.

Los arrays de expresión se han convertido en una herramienta de gran utilidad en la medicina clínica para el descubrimiento de marcadores moleculares, revelando así

cada vez más principios fundamentales que ayudan a clasificar una enfermedad de acuerdo con su perfil de expresión específico. Los datos obtenidos del análisis de microarrays son reproducibles, claros, y esporádicamente revelan hallazgos inesperados. **(Agapito & Arbitrio, 2022; Watanabe et al., 2020)**.

1.2 Objetivos

1.2.1 Objetivo general

- Analizar la expresión diferencial del gen LRRK2 asociado a la enfermedad de Parkinson mediante la técnica de microarrays.

1.2.2 Objetivos específicos

- Emplear la base de datos GEO del NCBI para descargar un conjunto de datos experimentales de microarrays enfocados en la enfermedad del Parkinson.
- Analizar la expresión diferencial del gen LRRK2 buscando la identificación de fenotipos relacionados con la enfermedad de Parkinson.
- Proponer un modelo de predicción para el diagnóstico temprano de la enfermedad de Parkinson en fenotipos relacionados a la enfermedad.

CAPITULO II
METODOLOGÍA

2.1 Materiales

Tabla 1. Recursos bioinformáticos

Recursos	**Enlace**
GEO (Gene Expression Omnibus)	https://www.ncbi.nlm.nih.gov/geo/
R	https://cran.r-project.org/bin/windows/base/
R-studio	https://posit.co/download/rstudio-desktop/
GeneCards	https://www.genecards.org/
Bioconductor	https://www.bioconductor.org/
PubMed	https://pubmed.ncbi.nlm.nih.gov/
UC Irvine Machine Learning Repository	https://archive.ics.uci.edu/
Gene Ontology and GO Annotations	https://www.ebi.ac.uk/QuickGO/

Tabla 2. Equipos

Recursos	**Cantidad**
Computadora personal	1
Router y servicio de internet	1

2.2 Métodos

2.2.1 Búsqueda de datos referentes a la enfermedad

GEO (Gene Expression Omnibus), es un repositorio público internacional desarrollado por el NCBI (National Center for Biotechnology Information) en donde se pueden encontrar datasets provenientes de diferentes tecnologías, entre ellos microarrays, secuenciación de próxima generación (NGS) y otras formas de datos genómicos funcionales de alto rendimiento presentados por la comunidad de investigación **(Clough & Barrett, 2016)**. Dentro de la plataforma se seleccionó la opción de "GEO Datasets", empleando como criterios de búsqueda: ("Parkinson disease"[MeSH Terms] OR Parkinson's disease [All Fields]) AND "Expression profiling by array" [DataSet Type] AND (LRRK2*) para la identificación del conjunto de datos (Datasets).

Para la selección de los mismos se emplearon como criterios de inclusión/exclusión los brindados por **Oerton & Bender (2017)**, que son: 1) Los estudios debieron diseñarse específicamente para la investigación de la PD; 2) Tipo de muestra (sustancia negra, cuerpo estriado, sangre, corteza frontal, línea celular, sección de todo el cerebro o cerebelo); 3) Los contrastes de la PD versus control sano deben estar disponibles con al menos dos muestras para cada condición; 4) La expresión génica debió medirse utilizando tecnología de microarrays; 5) Los experimentos debieron ser realizados en la plataforma GPL571 [HG-U133A_2] Matriz Affymetrix Human Genome U133A 2.0.

2.2.2 Validación de la información (comprobación de calidad de conjuntos de datos)

Todos los análisis bioinformáticos se realizaron utilizando el lenguaje estadístico R versión 3.3.2 mediante el uso de RStudio 2023.12.0, librerías de Bioconductor v3.17 desarrolladas para la expresión diferencial de los genes. Un método eficiente para leer los datos y asignar a cada muestra los valores de las covariables (grupo para el análisis, ID, color distintivo, etc.), consiste en la creación de un archivo de texto, al que le hemos denominado "target", el cual contiene la identificación (diseño del

experimento) de cada archivo de datos, lo que permite realizar la asignación de cada muestra a cada condición experimental **(Chung et al., 2021)**.

Las sondas mapean los nombres de los genes correspondientes; para esto fue necesario la anotación de los ID de sonda en sus genes asociados utilizando el archivo GPL de anotación pertinente (obtenido de GEO) **(Huang et al., 2022)**. Mediante la exploración de los datos, basada en técnicas univariantes (diagramas de caja) y técnicas multivariantes (cluster jerárquico y análisis de componentes principales o PCA) se determinó que el array del conjunto etiquetado con 90_NSC_WT (GSM888890_NSC_H9_WT_3.CEL.gz) correspondía a una pluralidad de genes, razón por la cual no fue considerado en el estudio, esperando maximizar la cantidad de genes disponibles para las comparaciones entre diferentes plataformas. Para aquellos casos en el que un símbolo de gen coincidía con múltiples ID de sonda, el valor de expresión medio fue elegido del nivel de expresión de este **(Oerton & Bender, 2017; Y. Sun et al., 2018)**.

2.2.3 Análisis de la expresión diferencial del gen LRRK2

Los archivos se preprocesaron a través de la corrección de fondo, la normalización del cuantil y el cálculo de la expresión utilizando el algoritmo robusto de promedio de matriz múltiple (RMA) con paquete Oligo **(Oerton & Bender, 2017; Y. Sun et al., 2018)**. Mediante la utilización de la función “rma” se obtuvo el objeto “ExpressionSet” que almacena la información del dataset. Se observó que los datos del ensayo constan de 22277 genes (features) en las 8 muestras que se manejan mediante los nombres de los elementos de la expresión. Se utiliza el paquete "pd.hg.u133a.2" para efectuar la anotación de los genes (que cada posición pueda asociarse con cada sonda). Como resultado final se obtuvo una matriz con los datos normalizados y que ya podían ser filtrados.

El filtraje no específico permite eliminar los genes que varían poco entre condiciones de estudio, o, que se desea quitar por otras razones, como por ejemplo que no se dispone de anotación para ellos **(Irizarry et al., 2012)**. Para la selección de genes se realizó una distribución de los niveles de expresión de cada uno a través de las

muestras, para esto fue necesario graficar el coeficiente intelectual (CI) de cada gen y trazar estos valores. La función "nsFilter" permite eliminar los genes que presenta baja variabilidad o bien no se dispone de anotación para ellos **(Gentleman et al., 2023)**. Para este análisis, se utilizaron las anotaciones como criterio de filtraje, por ende, se dispuso del paquete de anotaciones "hgu133a2.db", ya que considera todos los genes que están presentes en los arrays, filtrando aquellos que se emplean en la base de datos (BDs). Tras haber realizado el filtrado de genes, se determinó que estos se han reducido a 3161 genes (features) que son los que tienen mayor variabilidad, con los que se trabaja a partir de este punto, los demás componentes se mantienen iguales.

Para seguir con el análisis basado en modelos lineales se procedió a la creación de la matriz de diseño. Básicamente, describe la asignación de cada muestra a un grupo a manera de tabla. Tiene tantas filas como muestras y tantas columnas como grupos (LRRK2_GS y WT) **(Smyth, 2004)**. En las cuales se van intercalando entre 0 y 1 dependiendo del grupo al que pertenece la muestra.

Ahora bien, la matriz de contrastes se utilizó para describir las comparaciones entre grupos, en las columnas se muestran las comparaciones y en las filas los grupos que utiliza el estudio. Una comparación entre grupos (contraste) se representa con un 1 y un -1 en las filas de los grupos a comparar. En caso de que fueran varios grupos los que intervinieran en la comparación se tendría tantos coeficientes como grupos, simplemente que su suma sería cero **(Salicrú et al., 2011)**.

Cada grupo de chips fue comparado con PD y Control, respectivamente (G. Ma et al., 2019). Se usó la función "makeContrasts", en donde se observó los contrastes que se formaron en cada una de las interacciones de los grupos. Estos fueron integrados al modelo ajustado con "contrasts.fit" y mediante estadística bayesiana (modelos de Bayes empíricos) presente en la función "ebayes" se obtuvieron los genes de expresión diferencial.

2.2.4 Ajuste de modelo experimental

Gracias al método implementado en el paquete "limma" el ajuste se efectuó por mínimos cuadrados ponderados o generalizados. Posteriormente, se utilizó la

instrucción “topTable”, la cual permite extraer una tabla con los genes mejor clasificados de un ajuste lineal. Su utilidad se basa en que este aplica un filtro basado en dos criterios, “log fold change” que permite cuantificar la proporción de dos valores (cuantas veces más esta expresado un gen en una condición que en otra) y “p.value” que permite verificar su significancia estadística **(Smyth, 2004)**. Para el análisis de expresión se utilizó un valor de “log-fold-change” mayor a 3, un p-valor ajustado (probabilidad de obtener una diferencia significativa entre factores al azar) inferior a 0.05 para acceder a una estadística del 95% de significancia; normalmente se trabaja con este p-valor debido a que es estadísticamente improbable que las diferencias entre una familia de comparaciones se deban al azar, afirmando de esta forma que la diferencia es real **(Arias-Molina, 2017; Diaz & Rios, 2018; Kain & MacLaren, 2007)**. De igual forma se aplica un ajuste FDR (false discovery rate) que es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por el test estadístico.

Los datos de expresión se identifican por el nombre de la muestra (que además identifica el grupo) y por el identificador del gen. El proceso de anotación fue aplicado a cada uno de los genes que están contenidos en las muestras y añadido a la tabla de expresión los identificadores de los genes, mediante el ID de ENTREZID y SYMBOL, identificadores universales con los cuales se analizó la significancia biológica de la expresión.

Se empleó la función “AnnotationDbi” encargada de seleccionar los genes down-regulated y up-regulated para poder ser mostrados en un gráfico denominado volcano_plot, estos se ubican en la zona superior y hacia la izquierda y derecha del gráfico respectivamente **(Pagès et al., 2023)**. Al ser un gráfico con limitaciones de espacio, todos los genes altamente expresados no fueron fácilmente diferenciables dentro del mismo, debido a esto, se realizó un análisis de componentes principales (PCA) en donde se detalló el símbolo, ID, descripción y categoría correspondiente a cada uno de los genes (Anexo 5). De igual forma se generó un mapa de calor comparando las 8 muestras trabajadas con los resultados obtenidos. Estos gráficos resultaron de mucha utilidad al momento de interpretar los resultados obtenidos en

tablas puesto que su función es brindar un análisis visual de los genes de baja y alta expresión.

2.2.5 Cuantificación de expresión de genes

El análisis de significación biológica de las listas mediante análisis de enriquecimiento, (OverRepresentation Analysis) permite detectar si las listas de genes diferencialmente expresadas presentan una cantidad superior a la esperada (se encuentran enriquecidas), de genes asociados a funciones o procesos biológicos determinados, es decir, anotados en categorías biológicamente relevantes para el problema que se estudia **(Khatri & Drăghici, 2005)**.

Se trabajó con dos colecciones de genes: la lista seleccionada y el universo de genes, es decir, todos los genes que se han incluido en el análisis. Se utilizó el paquete "GOstats" para realizar el análisis de enriquecimiento genético. Se creó un objeto denominado "Hiperparámetro" de clase "GOHyperGParams", el cual es una clase de parámetro general (en lugar de argumentos individuales) que incluye los elementos necesarios para ejecutar una serie de análisis relacionados de forma organizada; en este caso, el método hyperGTest para hacer un análisis basado en la Gene Ontology (GO) como categoría **(Álvarez, 2020; Ayala, 2018; Falcon & Gentleman, 2007; Kruschke, 2015)**. Una vez creados los parámetros, el análisis se llevó a cabo invocando la función "hyperGTest" que realiza un test de Fisher (pruebas hipergeométricas) para todas las categorías evaluando la representación excesiva de identificadores de categoría en el conjunto de genes. Se mostró finalmente los genes que se han anotado y el análisis de significancia biológica efectuado sobre los mismos **(Falcon & Gentleman, 2007)**.

2.2.6 Búsqueda datos biomédicos referentes al Parkinson para propuesta de un trabajo sistematizado para el diagnóstico temprano

UC Irvine Machine Learning Repository es una colección de bases de datos, teorías de dominio y datasets que utiliza la comunidad científica para el análisis empírico de algoritmos de "Machine Learning", el repositorio fue creado en 1987 y actualmente

mantienen 653 conjuntos de datos como servicio para la comunidad de aprendizaje automático **(Amarnath et al., 2016)**. Los datasets se encuentran disponibles en función de la tarea objetivo, el tipo de los atributos, la naturaleza de los datos, el área de conocimiento y, finalmente, los aspectos relativos al propio dataset (tamaño y formato).

Dentro del apartado "Search" se ingresaron las palabras clave: "Parkinson's Disease", arrojando así 6 resultados; todos relacionados con disfonía. Considerando este último parámetro de realizó una segunda búsqueda para la depuración de resultados, ahora empleando los términos: "Parkinson's Disease" AND "Classification" ADN "Sound". Se definieron como criterios de selección: 1) Los datos deben provenir de un estudio en el que se trabajen 70 variables o más; 2) Los grupos de pacientes tanto enfermos, como de control deben estar conformados por hombres y mujeres, encontrándose en una misma proporción en el conjunto de datos final; 3) Las edades de los pacientes deben tener un amplio rango de separación, mínimo 30 años de diferencia. Criterios que fueron adaptados de los propuestos por **Sakar et al. (2019)** con influencia del trabajo planteado por **Lahmiri (2017),** con el objetivo de contribuir a la variabilidad de los datos y asegurar resultados mucho más realistas. Con esto, se estimó la probabilidad de concordancia con características que presenten los individuos y el riesgo de desarrollar la patología, proponiendo así una herramienta de detección temprana de la enfermedad.

2.2.7 Pruebas con los datos para la predicción

Al tratarse de disfonía en pacientes, fue posible considerar una amplia serie de variables antes de pasar al apartado de aprendizaje.

Funciones de línea de base, estas fueron utilizadas como punto de referencia para el proceso de entrenamiento del modelo, es posible considerar: Variantes de fluctuación, Variantes de brillo, Parámetros de frecuencia fundamental, Parámetros de armonía, entropía del período de tono, etc.

Funciones de frecuencia de tiempo, estas fueron empleadas para el modelamiento de patrones temporales en los datos durante el entrenamiento, es posible considerar: Parámetros de intensidad, Frecuencias formantes y Banda ancha.

Coeficientes cepstrales de frecuencia Mel (MFCC), se los empleó para captar los efectos de la EP en el tracto vocal por separado de las cuerdas vocales **(Sakar et al., 2019)**.

Características de las cuerdas vocales, se tomaron en cuenta ya que de esta forma se pueden identificar aspectos funcionales, que son relevantes para la producción de la voz y por ende del diagnóstico de la enfermedad, es posible considerar: Cociente de glotis, Excitación de las cuerdas vocales, Excitación glotal a ruido, etc.

Debido al hecho de que el dataset contenía 755 variables con 756 observaciones cada una, fue necesario emplear un modelo de reducción de datos. El algorithmo Minimum Redundancy, Maximum Relevance (mRMR) minimiza la redundancia de un conjunto de características y maximiza la relevancia de las mismas, esto es posible gracias a la información mutua de variables, básicamente, si dos características son similares, sólo la más relevante se considerará importante **(Ding & Peng, 2003)**.

De este modo se obtuvieron 50 variables (Sin considerar el "Género" y la "Clase"), cuyas observaciones al encontrarse por triplicado fueron reducidas calculando la media de las 3. Construyendo así, un dataset final que consta de 52 variables con 252 observaciones, finalmente, se curaron los datos de acuerdo con los requisitos técnicos del algoritmo, es decir que estos se normalizaron empleando la técnica de normalización "*Escalar de robusto*".

2.2.8 Selección del mejor algoritmo de predicción

Existen una serie de algoritmos de predicción, algunos de los más utilizados son: Naive-Bayes, Artificial Neural Network, Support Vector Machine, Árbol de decisión, etc. En este apartado, lo que se buscó fue el seleccionar el algoritmo que mejor se ajuste a las necesidades en base a la naturaleza y tipo de muestra seleccionado. Al tratarse de muestras sonoras para detección de disfonía, se consideraron 3 tipos de algoritmos de aprendizaje. *Support vector machine*, que busca un hiperplano que separa de manera óptima las muestras de diferentes clases en un espacio de características **(Pisner & Schnyer, 2019)**; *Random Forest*, que se centra en la creación de varios árboles de decisión con el objetivo de obtener un modelo único robusto en comparación a los individuales **(Espinosa Zúñiga, 2020)** y *Árbol de clasificación o*

decisión, el cual es una representación en forma de árbol que organiza los datos en una serie de decisiones basadas en características o atributos de la muestra **(Origel-Rivas et al., 2020)**.

Para emplear estos tres algoritmos de predicción fue necesario establecer previamente individuos de train y test (entrenamiento y prueba), estos se han dividido en una proporción aproximada de 70 y 30 % respectivamente, porcentaje optimo establecido para ejecutar un correcto análisis de "machine learning" **(Wang et al., 2020)**. En base a esto, el data set denominado "class_train" estaba formado por 168 observaciones, mientras que el "class_test" 84 observaciones, utilizando para ambos casos la semilla "2018" para garantizar resultados repetibles.

Support vector machine

Para este caso en específico, se trabajó usando los datos con las variables cuantitativas estandarizadas. Se realizó utilizando la librería "kernlab" junto con la función "ksvm" para el entrenamiento de los modelos que se emplearon. Posteriormente, se probaron los modelos con los datos "class_test", y se procedió a la evaluación correspondiente.

Random Forest

Se trabajó el algoritmo con los datos originales, mediante la librería "randomForest" y la función del mismo nombre, se entrenaron dos modelos con 100 y 200 árboles cada uno, los mismos que son candidatos ideales para el *bagging*, una técnica que corresponde a un método de aprendizaje por conjuntos que se usa comúnmente para reducir la varianza dentro de un conjunto de datos ruidoso, pero aproximadamente imparciales **(Lujan-Moreno et al., 2018)**.

Árbol de decisión

Al igual que el anterior, este algoritmo de predicción se trabajó con los datos originales. Se utilizó la librería "C50" y empleando la función "C5.0" se entrenaron dos modelos, el C5.0 simple y el C5.0 haciendo boosting con 10 ensayos, finalmente, se probaron ambos modelos con los datos "class_test", y se obtuvo la evaluación correspondiente.

Se evaluó la eficiencia y precisión del algoritmo utilizando conjuntos de datos independientes (en donde se incluía un grupo que posee la enfermedad y otro control) para después analizar el valor de precisión que arrojaron mediante la creación de una tabla resumen de resultados empleando la función “kable” de la librería “knitr”. Esto ayudó en la selección del algoritmo que mejor se ajustó a estos datos, lo cual permitió decidir qué tan bien se ha desarrollado el modelo y su capacidad de predecir la enfermedad en nuevos individuos.

2.2.9 Predicción de PD basado en evidencia

Una vez se obtuvieron los resultados, los tres modelos fueron analizados para determinar al mejor en base a la puntuación obtenida en el parámetro “Accuracy”, es decir, aquel modelo que tuvo un nivel de precisión y exactitud mayor al resto. Cuando este fue escogido, se determinó que puede ser utilizado para predecir el padecimiento de la EP en posibles pacientes.

CAPITULO III

RESULTADOS Y DISCUSIÓN

3.1 Análisis y discusión de resultados

3.1.1 Análisis DEGs del conjunto de datos GSE36321.

La prevalencia de la EP aumenta con la edad y puede alcanzar altos niveles en personas mayores de 80 años **(Benamer & De Silva, 2010)**. A pesar de los avances en la ciencia la EP sigue representando un reto dentro de la medicina, esto debido a la gran cantidad de variantes que puede presentar la enfermedad y el sinnúmero de posibles causas que posee. Recientemente se ha descubierto que las mutaciones del gen LRRK2 (proteína multidominio con actividades GTPasa y quinasa) poseen una frecuencia mundial del 1 % en la PD esporádica y del 4 % en la PD familiar **(Schootemeijer et al., 2023)**.

La mutación LRRK2 asociada a la PD mejor caracterizada es la mutación genómica autosómica dominante c.6055G>A, una sustitución en la posición 6055 del exón 41 del gen que da como resultado el cambio de una glicina a serina en el codón 2019 de LRRK2 **(Cookson, 2010; Ren et al., 2019)** caracterizada de mejor manera como "LRRK2-G2019S".

Esta mutación afectaría la actividad quinasa de LRRK2 con un aumento de su función, a través de la actividad de hiperfosforilación, lo que conduciría a niveles elevados de proteínas αSyn (α-sinucleína) y tau, lo que puede causar disfunción mitocondrial, interrupción del transporte de vesículas y autofagia, y crecimiento neuronal anormal; acercando al individuo a la muerte neuronal **(Arbez et al., 2020; X. Chen & Le, 2021; M. Miranda, 2007; Ren et al., 2019; Volta et al., 2017)**. Cabe mencionar que la penetrancia de LRRK2-G2019S es incompleta y variable, lo que resulta en una comprensión incompleta de la misma **(Pischedda et al., 2021)**.

Ahora bien, se ha demostrado que mutaciones genéticas pueden interferir en el desarrollo de la enfermedad, dando lugar a manifestaciones clínicas diversas. Por tal

motivo el dataset con el que se trabajó en el presente estudió contaba con 9 muestras originales donde se extraía células madre neuronales (NSC) que albergan la mutación LRRK2 (G2019S), muestras que mantienen concordancia con lo dicho por **Walter (2021),** que establece en sus estudios que la desregulación de las células madre neuronales y el desarrollo neurológico comprometido resultante de esta, desempeñan un papel de impacto en la PD.

Otro de los aspectos de importancia a destacar sobre las muestras presentes en el estudio es que se encuentran derivadas de células madre embrionarias humanas (hESC) que albergan una mutación patógena, un origen alentador puesto que el principal objetivo de este tipo de análisis referentes al PD espera un estudio exhaustivo del impacto de la mutación LRRK2-G2019S en la dinámica de la especificación del destino celular a lo largo de la diferenciación neuronal de células madre específicas del paciente **(Sanders et al., 2014; Walter et al., 2021)**.

Por ende, la mutación LRRK2:G2019S resulta ser un modelo de enfermedad efectivo y un avance potencial en la investigación de la patología al momento de determinar los mecanismos de esta y desarrollar nuevos métodos de tratamiento mediante el estudio de pacientes con PD **(Ren et al., 2019)**.

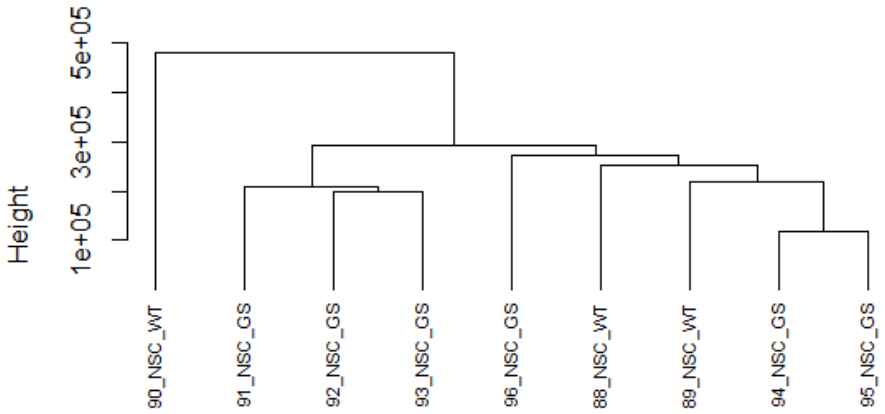

Figura 4. *Agrupación jerárquica (Gráfico de cluster) de las muestras originales del dataset GSE36321*

Multiples estudios han encontrado que dentro de experimentos de microarrays, los tamaños de muestra juegan un papel determinante en la confiabilidad de los mismos, es decir, que tamaños de muestra más robustos brindan una mayor confianza al llamar a genes expresados diferencialmente, lo que resulta en la obtención de listas concretas de genes expresados diferencialmente **(Oerton & Bender, 2017)**, sin embargo, en aspectos de concordancia promedio entre los diferentes conjuntos de datos, el efecto del tamaño de la muestra no ha sido examinado directamente **(Q. Sun et al., 2020)**; por tal motivo, el análisis visual de las muestras sigue siendo necesario para comprobar la calidad de las mismas.

Una de las formas más comunes que se emplean para observar si las muestras se agrupan según los grupos experimentales, es utilizando gráficos de cluster jerárquico; estos se emplean con frecuencia ya que realizan una agrupación básica de las muestras por grado de similitud **(McLachlan et al., 2017)**. Visualizando el panorama de la expresión genética de los estudios de PD, se revela un subconjunto distinto al resto de muestras empleadas, este es el caso de 90_NSC_WT cuyo grado de similitud entre muestras que conforman el dataset original no fue significativo (Ver figura 4), por esta razón se ha decidido separar esta muestra del dataset original, con el cual se efectúa el procesamiento de los datos.

La razón por la cual esta muestra presenta esa variación en comparación con las demás es una variación de las condiciones en el protocolo de tratamiento de las NSC derivadas de hESC, puesto que a diferencia del resto de muestras "WT" que se trataron con LRRK2-IN-1 3 μM durante 5 días, esta únicamente se trató de forma simulada, lo que indica que posiblemente esta variante capture los patrones de afinidad observados con el resto de muestras del dataset **(G. Ma et al., 2019; Oerton & Bender, 2017)**.

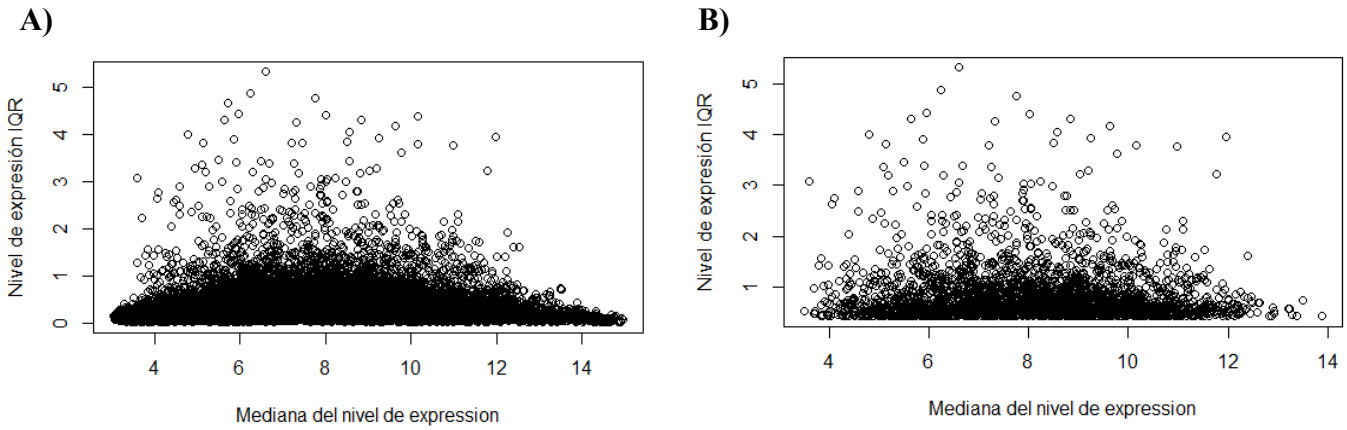

Figura 5. *Características principales en niveles de genes expresados diferencialmente (DEG)*

A) Gráfico de características principales en niveles DEG original B) Gráfico de características principales en niveles DEG tras efectuar filtrado de genes innecesarios

La variabilidad genética se refiere a la diversidad en las frecuencias de los genes. La variabilidad genética puede referirse a las diferencias entre individuos o las diferencias entre poblaciones. Las mutaciones son la causa fundamental de la variabilidad genética, pero mecanismos tales como la reproducción sexual y la deriva genética también contribuyen a la misma **(Miller & Federoff, 2010)**.

Al empezar a trabajar con los genes presentes en el dataset es recomendable desarrollar un gráfico correspondiente a las características principales de la muestra con la que se trabaja, de esta forma es muy sencillo identificar si esta posee elementos innecesarios para el análisis **(Kong et al., 2018)**. En primera instancia, se observa que un gran número de los genes presentan baja variabilidad (Ver Figura 5A), por ende, es poco probable que sean útiles para descifrar diferencias entre los tipos de tejido con los que se está trabajando.

Debido a esto, se ejecuta un filtraje de genes con baja variabilidad ya que no aportarán información relevante al análisis de expresión diferencial; manteniendo únicamente a aquellos cuya presencia implique significancia en el estudio (Ver Figura 5B). El objetivo de esto es reducir el ruido y el impacto de múltiples ajustes de prueba, haciendo énfasis en la eliminación de aquellos genes o spots cuyas señales medias no

superaron un umbral mínimo en todos los grupos **(Botta-Orfila et al., 2012; Lin et al., 2018)**, lo que ha resultado en la eliminación de 9572 genes repetidos, 9482 que tienen una baja variabilidad y 62 que han sido excluidos directamente.

Ajuste modelo experimental

El DEG entre grupos de pacientes realiza un análisis de modelo lineal, con moderación bayesiana empírica de la varianza, buscando ajustar el tamaño de muestra de experimentos de microarrays (**Botta-Orfila et al., 2012**). Básicamente, se aplica un filtro en base a "log fold change (logFC)" un parámetro que cuantifica la proporción de dos valores, es decir, cuanto más expresado se encuentra un gen en una condición que en otra, y el estadístico "p.value".

Dentro del análisis existen varias posibilidades que varían en cuanto al nivel de confianza que se pretende alcanzar; tradicionalmente, un "p.value" ≤ 0,25 y un "logFC" de al menos 2 tienen la capacidad de mostrar genes diferencialmente expresados de manera confiable **(Botta-Orfila et al., 2012)**, sin embargo, **Gao (2015)** propone que para que un DEG pueda considerarse efectivo, los parámetros de selección deben ajustarse a un "logFC" ≥ 1,2 y un "p.value" ≤ 0,05. Con fines académicos y buscando el mayor nivel de confianza (95 %), se utilizó un valor de "logFC" ≥ 3 y un "p.value" ≤ 0,05.

Tabla 3. Datos de la expresión de forma tabular junto con identificadores de los genes (Entrezid y Symbol)

PROBEID	ENTREZID	SYMBOL	logFC	AveExpr	P.Value	adj.P.Val
206915_at	4821	NKX2-2	3.827	9.22	1.810844e-09	5.724078e-06
201427_s_at	6414	SELENOP	3.119	7.54	4.012864e-08	4.289519e-05
40284_at	3170	FOXA2	3.209	8.49	4.071039e-08	4.289519e-05
205646_s_at	5080	PAX6	-3.753	8.58	2.464973e-07	1.947945e-04
203697_at	2487	FRZB	-3.181	6.31	2.004203e-06	9.301202e-04
214451_at	7021	TFAP2B	-3.715	6.72	3.463923e-06	1.094946e-03
202404_s_at	1278	COL1A2	-3.603	8.96	5.053777e-06	1.452272e-03
205206_at	3730	ANOS1	-3.825	7.50	6.719416e-05	1.062004e-02

En la Tabla 3 DEG se observan campos como el logFC, AveExpr, P.Value, adj.P.Val, parámetros representativos de un análisis apropiado, puesto que **Zhao et al. (2022)** manifiestan que estos se utilizan para la clasificación de genes y permiten valorar sus niveles de expresión.

La columna "PROBEID" corresponde al ID de los genes analizados, aquellos que se encuentran presentes en el estudio y cuyo identificador común se muestra en la columna "SYMBOL". La columna "AveExpr" da el nivel medio de las expresiones asociadas al gen independientemente del grupo al que pertenezcan (LRRK2_GS y WT) en función del $\log_2$ **(Chamorro-Poyo, 2019)**. "P.Value" es el menor nivel para el cual el test resultaría en rechazo para los datos observados, en este caso, un p-value del 0,05 significa que el 5% de los genes que se listan no se expresan diferencialmente, es decir, corresponden a un falso positivo **(Fan et al., 2018; Fu et al., 2022)**. El objetivo de emplear este valor es reducir el número de falsos positivos (FDR, false discovery rate), a pesar de que haya la posibilidad de perder genes diferenciales, no se trata de tener más genes, más bien, aumentar la seguridad del análisis.

Por otro lado, la columna "logFC", correspondiente al $\log_2$-fold-change, muestra el cambio en la proporción de "reads" para ambas condiciones (LRRK2_GS y WT) en función del $\log_2$ **(Miyashita et al., 2023)**. Los genes para los cuales el valor logFC es negativo, constituyen aquellos que se expresan en menor medida, a diferencia de aquellos cuyo valor logFC es positivo, pues estos se expresan en mayor medida **(Sánchez, 2015)**.

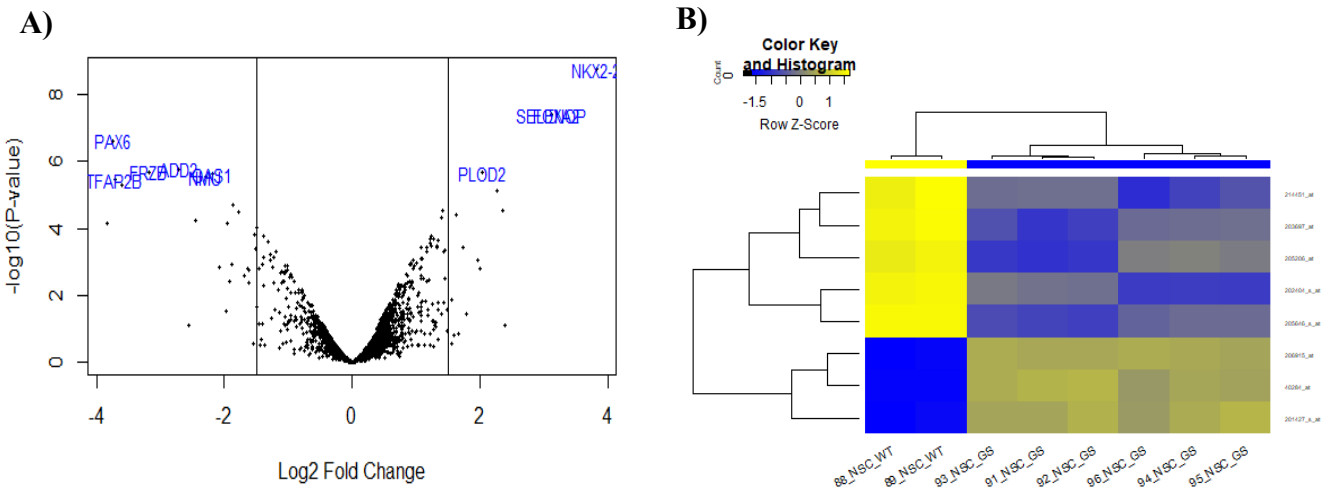

Figura 6. *Análisis de expresión diferencial del GEO dataset: GSE36321*

A) "Volcano plot", en donde se muestran los genes que presentan alta y baja expresión. B) Mapa de calor en donde se representa la co-expresión de los genes en cada uno de los grupos cuya leyenda corresponde al valor de "logFC".

La Figura 6A corresponde a una de las representaciones gráficas del análisis de enriquecimiento funcional de los DEG, se puede observar que se han marcado 10 genes tanto *up* como *down regulated*, aquellos genes que se encuentran por encima del umbral de p-value presentan cambios significativos de expresión, contrariamente a aquellos que se encuentran cerca del origen, puesto que no muestran cambios significativos en la expresión **(Q. Sun et al., 2020)**. Gracias al resultado expuesto en el heatmap (Ver Figura 6B, mapa de calor) se demuestra que los genes no presentan una tendencia a mostrarse sobreexpresados o inhibidos, debido a que la distribución de colores no es prolija (completamente azul o amarillo), mostrando únicamente irregularidades en comparaciones con las muestras tipo WT, lo cual no afecta al estudio.

Botta-Orfila et al. (2012) manifiesta que los genes obtenidos mediante el uso de un p-value $\leq 0{,}25$ y un logFC de al menos 2 se consideran expresados diferencialmente; considerando las características empleadas en este análisis, se determina que los genes mostrados en el Figura 6 A y B cumplen con las condiciones necesarias para ser considerados y por ende el estudio proporciona resultados reales y confiables, es decir, genes relacionados con la función biológica afectada por la enfermedad.

De igual forma, **Lin et al. (2018)** presentan en su estudio 70 grupos expresados diferencialmente en el grupo modelo de la PD, cuyas características estadísticas fueron un logFC igual a 1,5 y un p-value ≤ 0,05, parámetros que se encuentran acorde con lo antes mencionado. Sin embargo, la razón por la cual los resultados de expresión varían significativamente con los que se han obtenido en esta ocasión (8 grupos) (Ver Figura 6B) es debido a las mismas variables estadísticas, y que trabajar con un nivel de logFC de 3 es sinónimo de una clasificación de genes más selectiva, robusta y confiable que un nivel logFC de 1.

Tabla 4. Pruebas hipergeométricas para evaluar la representación de identificadores de categoría en el conjunto de genes

GOBPID	OddsRatio	ExpCount	Count	Size	Term
GO:0031018	118.28	0.0483	3	18	Endocrine pancreas development
GO:0021778	990.33	0.0080	2	3	Oligodendrocyte cell fate specification
GO:0021779	990.33	0.0080	2	3	Oligodendrocyte cell fate commitment
GO:0031016	84.31	0.0644	3	24	Pancreas development
GO:0001708	70.72	0.0751	3	28	Cell fate specification
GO:0021780	329.88	0.0134	2	5	Glial cell fate specification

Finalmente, se muestran 139 genes que se han anotado tras efectuar el análisis de significancia biológica a través del estudio de DEG. Cada uno de estos presenta información importante para la estimación de la relevancia de cada gen en la investigación. (Ver tabla 4).

" *OddsRatio*", también conocida como "razón de exceso" o "razón de probabilidades", cuantifica la representatividad de la expresión del gen, es decir, que mientras este valor sea mayor, el gen se encuentra más expresado **(Mohapatra & Krishnan, 2011)**. Por otro lado "*Count*" corresponde a la cantidad de veces que se encontró el termino ("GOBPID"), "*ExpCount*" la cantidad esperada, "*Size*" es el número de genes asociado a este término en el dataset **(Guri et al., 2010; Mohapatra & Krishnan, 2011)**. Finalmente, mediante los términos de Gene Ontology, "*Term*" se muestra la relación con la función biológica de los genes resultantes del estudio **(Dimmer et al., 2009)**

LRRK2 se expresa en tejidos cerebrales como el tronco del encéfalo (mesencéfalo), el cuerpo estriado, el bulbo olfatorio, la corteza, el hipocampo y el cerebelo **(Gilsbach et al., 2018)**. Participa principalmente en la regulación de la traducción de proteínas, el crecimiento de los axones y el envejecimiento en el sistema nervioso; destacando que su mutación fue descubierta en 2004 **(Ait Wahmane et al., 2021)**.

El DEG muestra funciones biológicas que pueden resultar contrarias a lo que se ha mencionado hasta este momento referente al gen LRRK2, específicamente las anotaciones "GO:0031018" y "GO:0031016" que corresponden a "Desarrollo del páncreas endocrino" y "Desarrollo del páncreas" respectivamente. Sin embargo, y aunque resulte difícil de entender, estas funciones si se encuentran relacionadas con la actividad cerebral de manera indirecta y por ende con el gen LRRK2. **Beumer & Clevers (2021)** establecen que en el páncreas se encuentran células enteroendocrinas productoras de hormonas (EECs), las cuales presentan características fenotípicas de las neuronas, incluidas las sinapsis y la producción de neurotransmisores; de igual forma, estas células son capaces de expresar neurogenina 3 (NEUROG3), el cual es un factor de transcripción que se requiere para el desarrollo de los islotes pancreáticos (grupos de células en el páncreas) **(Olvera-Granados et al., 2008)**, lo que explicaría el porqué de estos resultados en las anotaciones. No obstante, **Ruiz (2017)** menciona que, "estudios previos han mostrado que la Neurogenina 3 actúa como un factor de neurogénesis (generación de nuevas neuronas) en el sistema nervioso" lo cual crea una relación entre el páncreas y el cerebro.

Ahora bien, se encontró que NEUROG3 es fosforilado por quinasas dependientes de ciclina **(Beumer & Clevers, 2021)**, considerando que la mutación LRRK2-G2019S altera la actividad quinasa del gen provocando hiperfosforilación, la muestra del dataset utilizado se extrajo de pacientes que presentaban esta alteración genética en específico; sería una forma más que permite comprobar que estas funciones biológicas se encuentren presentes y en verdad tienen un papel significativo dentro del análisis. De igual forma, **Zeve et al. (2022)** estipulan que las EECs ofrecen una puerta de entrada al sistema nervioso central a través de su papel en el eje intestino-cerebro y son objetivos potenciales para influir en el apetito, la liberación de insulina y la

motilidad intestinal, conclusión que concuerda con lo dicho por **Mastracci & Sussel (2013)**, un factor de transcripción neuronal que está implicado en la transcripción del gen del glucagón puede explicar la presencia de proglucagón en determinadas zonas del cerebro, así como en las células alfa del páncreas.

Como otra posible conexión, cabe mencionar que el desarrollo de los islotes pancreáticos es posible gracias a la ayuda de "laminina"; **Gittes (2009)** menciona que estas glucoproteínas regulan la guía para ciertos axones y, por ende, favorecen el crecimiento axonal, proceso en el cual también se encuentra involucrado LRRK2.

"GO:0001708" corresponde a aquellos genes que regulan la "Especificación del destino celular", la presencia de este término en el DEG se encuentra justificada ya que la región N-terminal de LRRK2 está encargada de regular el correcto desarrollo del ciclo celular, el cual se encuentra involucrado en el compromiso del destino celular en el que la célula está designada para seguir un camino de desarrollo **(Gilsbach et al., 2018)**.

"GO:0021778" y "GO:0021779" corresponden a la "especificación del destino de las células de oligodendrocitos" y "compromiso del destino de las células oligodendrocitos" respectivamente, su presencia en el análisis está justificado ya que, la función principal de los oligodendrocitos es la mielinización de los axones nerviosos en el sistema nervioso central **(Paredes et al., 2021)**, ambas funciones biológicas se encuentran relacionadas con la diferenciación celular en oligodendrocitos; "GO:0021778" mediante un proceso autónomo, mientras que "GO:0021779" correspondiente a un proceso que restringe el destino del desarrollo celular **(Carbon & Mungall, 2023; Clayton & Tesar, 2021; Frankish et al., 2023)**.

3.1.2 Machine Learning para el diagnóstico temprano de la PD

Las herramientas bioinformáticas evolucionan a pasos agigantados y por ende sus aplicaciones también, el aprendizaje automático o también llamado "Machine learning" es una técnica de inteligencia artificial, que surge por la necesidad de

recopilar y procesar grandes volúmenes de datos; ha ido evolucionando constantemente para abordar problemas de diversas áreas con el objetivo de predecir datos desconocidos **(El Mestari et al., 2024; C.-X. Liu et al., 2022)**. Una de sus aplicaciones más significativas seria la detección temprana de enfermedades en base a ciertos parámetros relacionados con la misma, brindando un diagnóstico por tele monitoreo.

La PD es bastante invasiva en el ser humano, y el movimiento involuntario de articulaciones no son el único problema, puesto que aproximadamente entre el 70 al 90% de las personas afectadas presentan también deterioro vocal o mejor conocido como disfonía **(A. Ma et al., 2020)**, un trastorno que se presenta cuando existen alteraciones en la calidad de la voz sin ninguna explicación biológica (dificultad neurológica, anatómica u otra dificultad orgánica que presente la laringe) sino más bien por los trastornos del movimiento que subsecuentemente afectan a la laringe **(Mittapalle et al., 2023; Snow & Guardiani, 2019)**. Este "efecto secundario" puede manifestarse en dos formas principales; disfonía hiperfuncional (hipercinética), caracterizada por contracciones involuntarias de la musculatura laríngea; y la disfonía hipofuncional (hipocinética) asociada con el cierre incompleto de las cuerdas vocales debido a la falta de tensión muscular en la laringe **(Mittapalle et al., 2023; Rodríguez-Martín, 2023; Snow & Guardiani, 2019)**.

Ma et al. (2020) mencionan que, en personas con PD el análisis acústico presenta una relación de armónicos-ruido (HNR) decreciente; mientras que la fluctuación, que es una medida de la perturbación de la frecuencia y representa la variabilidad de esta durante la fonación constante, presenta aumentos significativos a medida que avanza la enfermedad; todo esto, desencadenando en una voz áspera y ronca. Siendo esta una de las principales razones por la cual se ha escogido como criterio de selección del dataset al parámetro que establece, que las edades de los pacientes deben tener un rango de separación de mínimo 30 años de diferencia; lo cual guarda relación con lo mencionado por Mittapalle **(2023)** en cuyo estudio menciona que la disfonía se encuentra bastante marcada en adultos de entre 19 y 60 años, brindando así un rango de 40 años en el que es bastante fácil y acertado diferenciar cambios significativos en los niveles de voz **(Morello-Da Cruz et al., 2020)**.

Ahora bien, las herramientas de aprendizaje automático y la disfonía podrían trabajar en conjunto debido a que la medición de la voz resulta ser simple y nada invasiva para ser considerada en la detección de sujetes enfermos **(Lahmiri, 2017)**. A pesar de que los recursos necesarios para un "machine learning" no son numerosos, esto no quiere decir que se puedan trabajar los datos sin ningún tipo de cuidado o preparación previa, en este caso, el dataset contenía una serie de variables que no resultaban útiles para este estudio, estas correspondían a la transformada wavelet de factor Q sintonizable (432 parámetros) que no representan información fundamental, por lo cual fueron eliminadas.

A pesar de este cambio, las variables de análisis seguían siendo numerosas (321), razón por la cual fue necesario realizar un segundo filtraje de variables; **Sakar et al. (2019)** mencionan que la mejor manera de determinar las mejores características y al mismo tiempo obtener un modelo de clasificación de PD robusto y preciso es efectuando un filtrado basado en redundancia mínima-relevancia máxima (mRMR), cuyo objetivo es reducir de manera directa y explicita la redundancia en la selección de funciones por medio de un enfoque de filtro; el cual viene dado gracias a que cada característica se clasifica en función de su relevancia para la variable objetivo ("Class") y también de su redundancia en el conjunto de características **(Ju & He, 2018)**. Básicamente este método de filtrado se encuentra basado en información mutua (MI), que viene siendo la relación entre dos variables aleatorias (X e Y) **(L. Chen et al., 2017; Saffari et al., 2021)**; sin embargo, **Ding & Peng (2003)** afirman que también puede lograrse un análisis Mrmr exitoso gracias al uso del coeficiente de correlación de Pearson combinando con distancia euclidiana, en donde, las correlaciones positivas y negativas altas son indicadores de redundancia; tomándose como estimación final el valor absoluto de las mismas.

Liu et al. (2017) indican que las características elegidas por este método desembocan en una mayor precisión que aquellas seleccionadas únicamente mediante la máxima relevancia. Gracias a esto es posible reducir el problema de alta dimensión a un conjunto mínimo con máxima dependencia conjunta, es decir, extraer las variables que presente poca redundancia entre sí, favoreciendo a la relevancia de las mismas

(Radovic et al., 2017). El algoritmo mRMR, fue propuesto por **Ding & Peng (2003)** como método experimental para filtrado de genes repetidos o sin significancia dentro del análisis de microarrays; la técnica fue tan exitosa que a partir de ese momento se ha aplicado a una amplia variedad de problemas (predicción de estructuras de proteínas, sistemas de apoyo a decisiones biomédicas, clasificación de datos hiperespectrales, etc.) de aprendizaje automático como paso de preprocesamiento **(Buś et al., 2022)**.

Las 50 variables consideradas para este estudio tras el segundo filtraje, concuerdan con las obtenidas por **Sakar et al. (2019)**, con la diferencia de que este último, mantiene a las características correspondientes a la transformada wavelet de factor Q sintonizable (TQWT), a diferencia de este análisis en donde no se tomaron en cuenta; esto debido a que, **Sakar et al. (2019)** pretenden aplicar por primera vez la "TQWT" a las señales de voz de pacientes con PD para la extracción de características, puesto que su hipótesis se centra en que estas presentan una mejor resolución de frecuencia que la transformada wavelet discreta, sin embargo, este sigue siendo un método experimental y por lo tanto no brinda los niveles de confianza necesarios para este caso. En la Figura 7 es posible observar la variabilidad que poseen las características utilizadas para el aprendizaje en este estudio. Mostrando de manera representativa únicamente las primeras 19 propiedades.

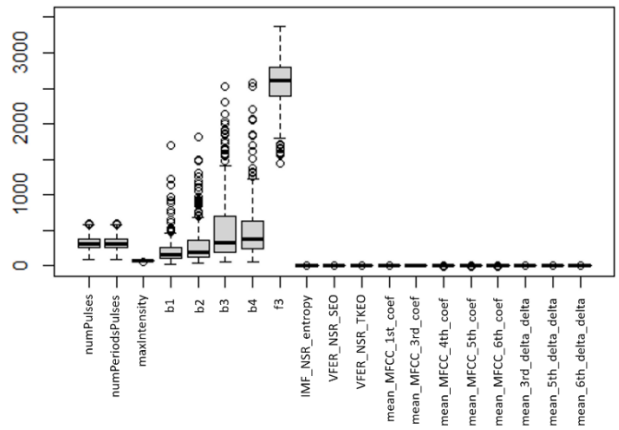

Figura 7. *Boxplot de datos obtenidos a partir de pacientes que padecen disfonía por Parkinson (19 primeras variables)*

Los datos “en crudo” deben ser curados para su procesamiento, al contar con una dispersión significativa entre datos que componen las variables (Ver figura 7), se debe efectuar un proceso de “normalización”, el cual elimina los posibles causales de “ruido” en el dataset, este procesamiento nos permite definir bases de datos más naturales y limpias, disminuyendo su volumen y simplificando su estructura para que la información resultante sea fácilmente localizable, comparable y recuperable **(Jain et al., 2018)**.

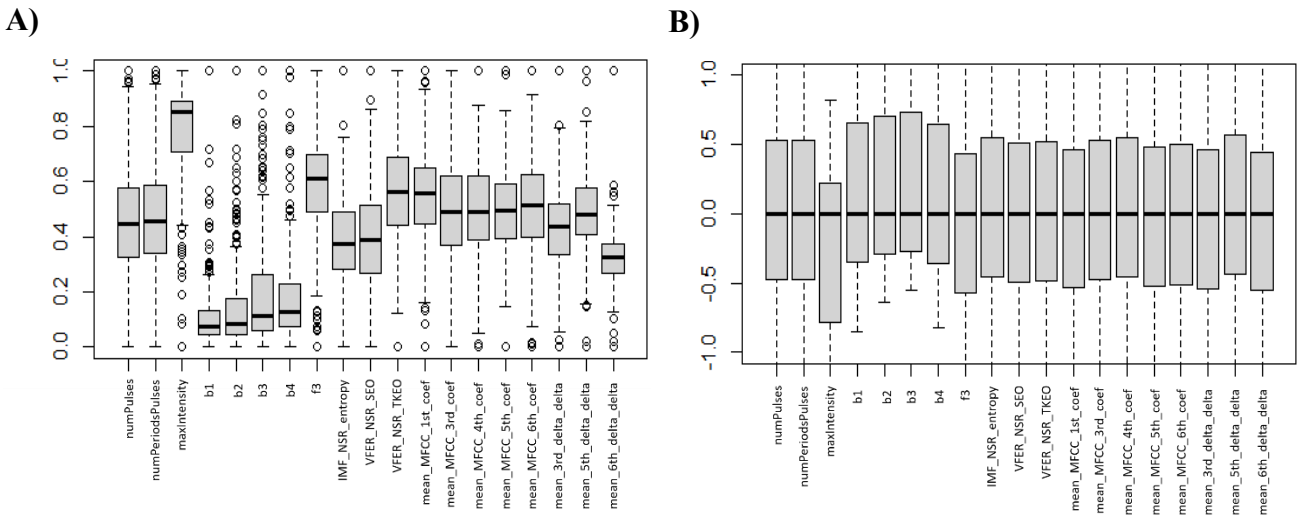

Figura 8. *Boxplot de datos tras efectos de normalización*

A) Método de normalización máximos y mínimos. B) Método de normalización escalar de robustos

La Figura 8 presenta los dos tipos de normalización utilizados en el análisis con el fin de encontrar el proceso óptimo para el trabajo. La estandarización mín-máx lleva los datos a la escala de 0 a 1, el método resuelve el problema de rangos desiguales empatando los valores máximo y mínimo, pero no corrige asimetría en la distribución **(Ganz et al., 2019; Treviño-Cantú, 2022)** (Ver Figura 8A). Por otro lado, la normalización robusta empareja las medianas y el tamaño de las cajas en un recorrido de 1 a -1, sin embargo, los extremos máximo y mínimo quedan indefinidos, pero la asimetría es controlada gracias a que elimina las diferencias en las amplitudes (Ver Figura 8B) **(Brimicombe, 2000; Radovic et al., 2017; Treviño-Cantú, 2022)**.

Comúnmente la estandarización mín-máx es más que necesaria para poder empezar a trabajar con datos, de hecho, **Ganz (2019)** menciona que este “es el método por

excelencia puesto que permite conservar las relaciones de los valores al no introducir ningún sesgo potencial en los datos". Sin embargo, se tomó como método de normalización válido aquel que de manera visual presentó mayores niveles de "armonía" entre datos, este es el caso de la normalización robusta (Ver Figura 8). Decisión que se encuentra acorde con lo mencionado por **Radovic (2017)**, el cual manifiesta que, para abordar el problema de tratar con un gran número de datos, todos se deben normalizar mediante la normalización robusta (z_{NR}), que normalmente se utiliza como paso de preprocesamiento para datos de amplitud sonora .

El análisis de aprendizaje debe realizarse bajo ciertos parámetros para ser considerado exitoso, el parámetro más importante es la división de las observaciones mediante una proporción entre los subgrupos de entrenamiento ("train") y prueba, ("test"); la proporción aplicada es de 70 y 30 % respectivamente, se los escogió en base a los parámetros propuestos por **Sakr et al. (2018)**, que aplica la misma proporción obteniendo resultados prometedores, sin embargo, esto se explica gracias a **Gholamy (2018)** en cuyo estudio explica que las mejores proporciones para efectuar un análisis de aprendizaje automático son 70-30 u 80-20 debido a que "no sobreestiman la precisión (no subestiman el error de aproximación) y son las más precisas entre las estimaciones válidas, es decir, la sobreestimación del error de aproximación es el menor posible".

Tabla 5. Resultados Support Vector Machine

Modelo	Accuracy	Kappa	AccuracyLower	AccuracyUpper
Lineal	0.75	0.179	0.644	0.838
Gaussiano	0.75	0.179	0.644	0.838

Support Vector Machine (SVM) es un algoritmo de aprendizaje ampliamente utilizado para la clasificación o análisis de regresión, donde se puede utilizar una función de kernel para transformar las características **(Deiss et al., 2020; Mammone et al., 2009)**. Estas funciones asignan los datos a un espacio dimensional diferente (suele ser superior), con el objetivo de determinar una solución lineal óptima separando un hiperplano, simplificando los límites de decisión complejos no lineales para hacerlos

lineales proporcionando la distancia más pequeña entre los puntos del hiperplano y el margen más grande entre las clases **(Lee et al., 2021; Noble, 2006)**.

Los modelos “kernels” empleados fueron el “Gaussiano” y “Lineal”, para aprendizaje de una clase y de dos clases respectivamente. En ambos métodos los parámetros de selección poseen el mismo valor, es decir, que la predicción se realizaría de igual forma sin importar el modelo “kernel” escogido. Esto puede deberse a que la separación entre clases es aproximadamente lineal **(P. H. Chen et al., 2005)**, de ahí que, ambos modelos presenten el nivel de exactitud considerable (0.75) (Ver tabla 5). **Deiss (2020)** menciona que en su estudio la función de base radial gaussiana se seleccionó puesto que representa el mejor rendimiento del modelo en la mayoría de casos, ya que su flexibilidad puede ir desde un clasificador lineal a uno muy complejo, razón por la cual se escoge como mejor modelo de entrenamiento.

Tabla 6. Resultados Árbol de decisión

Modelo	Accuracy	Kappa	AccuracyLower	AccuracyUpper
Simple	0.655	0.128	0.543	0.755
Boosting	0.750	0.179	0.644	0.838

Los *árboles de decisión* (C5.0) son modelos de clasificación y regresión cuyo objetivo es poder predecir a qué clase pertenece un caso del que conocemos uno o más atributos o mediciones **(Li et al., 2022)**, a diferencia de los modelos lineales, los árboles de decisión son adaptables para resolver cualquier tipo de problema puesto que mapean bastante bien las relaciones no lineales **(Suguiura-Rivero, 2022)**. El algoritmo que se utiliza para crear los árboles es llamado "*partición binaria recursiva*", adquiere este nombre gracias a que, en cada paso de entrenamiento se producen varias divisiones de un subconjunto de datos en regiones simples, para que el proceso se pueda representar mediante un árbol binario y la aplicación de una decisión asociada a una de las variables **(Arana, 2021; Trujillano et al., 2010)**.

La Tabla 6 presenta los dos modelos utilizados para el entrenamiento del "*Árbol de decisión*", antes de comparar resultados, se debe mencionar que el modelo "Boosting" crea un modelo de conjunto mediante la combinación secuencial de varios árboles de decisión débiles **(Khan et al., 2023)**. Este modelo puntúa a las salidas de los árboles individuales y da una ponderación mayor a las clasificaciones incorrectas del primer árbol. Después de numerosos ciclos, el método une todas las reglas débiles en una única regla de predicción bastante fuerte **(Lu & Ma, 2020; Ruiz-Villafranca et al., 2023; Sol et al., 2023)**. Por el contrario, el modelo "Simple" mide la importancia del predictor determinando, el porcentaje de muestras del conjunto de entrenamiento que caen en todos los nodos terminales después de la división **(González-Fernández et al., 2022)**, es decir, que a pesar de que el predictor tenga una medida de importancia de 100 si los nodos terminales comprenden únicamente un par de muestras del conjunto de entrenamiento, la puntuación de importancia baja pudiendo llegar a cero **(Elsayad et al., 2020; Kuhn & Johnson, 2013; Siknun & Sitanggang, 2016)**.

Ambos modelos presentan un nivel de exactitud relativamente alto, sin embargo, el modelo "Boosting" destaca con una diferencia 0.095, con lo cual, es el mejor modelo predictivo para "Árbol de decisión", este resultado concuerda con el de **Ruiz et al. (2023)** los cuales explican que el modelo Boosting logra un mayor rendimiento debido a la complejidad del mismo, es decir, que considera más parámetros que el resto, por lo que es una buena opción para implementar y dar solución al aprendizaje automático en diversos escenarios.

Tabla 7. Resultados Random Forest

ntree	Accuracy	Kappa	AccuracyLower	AccuracyUpper
100	0.798	0.452	0.696	0.877
200	0.810	0.475	0.709	0.887

El algoritmo de *Random Forest* (RF) es un método de aprendizaje que emplea la aleatoriedad de características para crear un bosque de árboles de decisión no correlacionados, cada árbol calcula sus resultados y obtiene el promedio de los

resultados de la predicción, este enfoque permite reducir la varianza en los árboles de decisión **(Jiang et al., 2023; Lee et al., 2021)**. Esto marca la diferencia más significativa entre los árboles de decisión y los bosques aleatorios; los primeros consideran todas las posibles divisiones de características, mientras que el "bosque aleatorio" solo seleccionan un subconjunto de esas características (las más importantes) **(Dinh et al., 2023; F. Zhao et al., 2024)**.

El número de árboles de decisión que componen el "Random forest" ha sido elegido en base a los parámetros estándar del algoritmo **(Jahanshahi & Baydogan, 2022)**. **Dinh et al. (2023)** mencionan que mientras mayor sea el número de árboles que se utilicen, mayor será el tiempo que tome el proceso de entrenamiento, lo que, bajo condiciones de potencia informática limitada, el algoritmo no sea práctico para aplicaciones o sistemas en tiempo real. Conjuntamente **Josso et al. (2023)** establecen que, dentro de un análisis de "Random Forest" la puntuación de pérdida de registros disminuye drásticamente durante los primeros incrementos de la prueba (No significativos) y se estabiliza en valores correspondientes a un bosque de 100 árboles o más. Motivo por el cual, únicamente se han trabajado con 100 y 200 "ntree" (Número de árboles).

Para un "ntree" de 200 el porcentaje de exactitud presenta resultados bastante eficientes (0.810) (Ver Tabla 7), ya que al estar muy cercano al 1 es sinónimo de un exitoso aprendizaje. **Josso et al. (2023)** brindan una explicación a esto, e indica que estos resultados son provocados debido al gran tamaño del conjunto de datos de entrada, por lo cual no resulta sorprendente que el sesgo y la varianza del modelo entrenado sean estables cuando se trabajan más de 100 árboles.

Tabla 8. Comparativa de los mejores resultados "Machine Learning"

Algoritmo	Parametros	Accuracy	Kappa	Accuracy Lower	AccuracyUpper
SVM	modelo = Gaussiano	0.75	0.179	0.644	0.838
C5.0	modelo = Boosting	0.750	0.179	0.644	0.838
RF	ntree = 200	0.810	0.475	0.709	0.887

En la Tabla 8 se muestran los algoritmos de aprendizaje automático evaluados junto con los parámetros de que presentan un mejor rendimiento. Se observa, que todos los algoritmos tienen un comportamiento eficiente; de hecho, dos de los modelos tienen el mismo valor de rendimiento, con una precisión (accuracy) de 0,75. Ya que se ha fijado el rango de precisión entre 0,750 y 0,810, lo que muestra una diferencia mínima entre el valor mínimo y máximo de 0,06. De igual forma, se muestran valores correspondientes a la estadística "Kappa", la cual es utilizada para medir los niveles de concordancia entre modelos predictores, es decir, la posibilidad de que estos adivinen ciertas variables debido a la incertidumbre, para lo que se toman valores entre −1 y 1 **(Kuhn & Johnson, 2013)**, donde el 0 vendría a simbolizar el nulo acuerdo entre las clases observadas y predichas **(McHugh, 2012)**. **Cohen (1960)**, su creador, menciona que, los resultados de niveles "Kappa" se interpretan de la siguiente forma: "valores ≤ 0 indican que no existe acuerdo, 0.01–0.20 ninguno o leve acuerdo, 0.21–0.40 representan acuerdos regulares, 0.41–0.60 acuerdos moderados, 0.61–0.80 acuerdos sustanciales y 0.81–1.00 son casi un acuerdo perfecto". En base a esto, y considerando los valores obtenidos tras efectuar las pruebas de aprendizaje con los tres métodos seleccionados, se selecciona como mejor algoritmo de aprendizaje de los tres evaluados al *Random Forest* (RF), ya que presenta un mayor nivel de "Accuracy" y su valor "Kappa" muestra un grado de concordancia moderado (Ver Tabla 8), por lo que se considera que posee mejor potencia de predicción.

Tabla 9. Resultados obtenidos tras aplicar el algoritmo RF al conjunto de prueba

	Predicho	
Actual	A	B
A	9	11
B	5	59

De entre un total de 252 observaciones con las que se ha trabajado, se ha destinado 168 para el conjunto de entrenamiento denominado "Train" y 84 para el conjunto de prueba denominado "Test", correspondiendo a una proporción aproximada del 70 % y 30 % del total, respectivamente. En el aprendizaje automático se denomina "Train" al conjunto de datos que contiene características objetivo para ajustar los parámetros del

modelo, de modo que un algoritmo pueda hacer predicciones **(Choi et al., 2020; Hastie et al., 2009)**, es decir, aquel conjunto de datos que permite que el algoritmo "aprenda" sobre la enfermedad y sus características con el objetivo de "diagnosticarla" en el futuro. Por otro lado, el conjunto "Test" sirve como herramienta de evaluación imparcial, como su nombre lo indica, probar el ajuste final del modelo **(Deo, 2016; Greener et al., 2022)**.

La Tabla 9 presenta los resultados de la predicción que se ha obtenido aplicando el mejor algoritmo de aprendizaje (*Random Forest*) sobre el conjunto de prueba "Test", donde, "A" representa pacientes sanos y "B" representa pacientes enfermos de disfonía por efecto de la PD. Comparando las relaciones que se presentan en cada nivel de la tabla, se observa que la ejecución del algoritmo arroja resultados que muestran un alto grado de predicción. Para el caso de pacientes sanos, se presentan 9 predicciones correctas y 11 incorrectas; por el contrario, para el caso de pacientes enfermos, para este estudio, el que tiene mayor impacto por la naturaleza del estudio y sus niveles de detección, ha predicho correctamente 59 casos de los 64 presentes en el conjunto; estos resultados nos indican que disponemos de un algoritmo funcional, cuyos resultados son confiables, ya que como se había especificado se alcanzado una precisión de predicción de 81%, puesto que de los 84 pacientes que componen el grupo de prueba se han predicho correctamente 64.

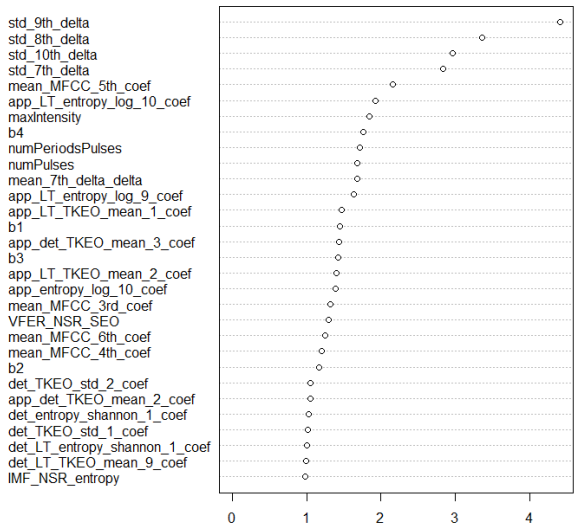

Figura 9. *Variables importantes dentro del algoritmo Random Forest*

Los “MFCC” son un sistema de reconocimiento del habla basados en la percepción auditiva humana, representan la amplitud del espectro del habla de manera compacta, mediante la aplicación de un filtro de pre-énfasis a la señal y posteriormente la división de esta en tramas. Estas últimas deben someterse a un tratamiento mediante el uso de una ventana, que las elimina de los bordes y las acentúa a la parte central de la trama para su análisis **(X. Chen et al., 2023; Kadiri & Alku, 2019; Martínez & Torres, 2013)**.

La Figura 9 muestra la importancia entre los predictores (variables que tienen mayor peso para la predicción de la PD) aplicando el modelo Random Forest. Aquí se observa que “std_9th_delta” correspondiente a los Coeficientes cepstrales de frecuencia Mel (MFCC) es el predictor que mayor peso presenta **(Espinosa Zúñiga, 2020)**.

Se plantea como trabajo futuro, probar nuevos entornos de predicción utilizando únicamente los predictores mejores puntuados que se muestran en la Figura 9, con lo cual se buscaría mejorar la potencia de predicción del algoritmo.

CAPITULO IV

CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- A través del análisis de expresión diferencial, se identificó un total de 139 genes que posiblemente estén involucrados con el desarrollo de la enfermedad de Parkinson. Tras efectuar un análisis de enriquecimiento funcional se determinó que la mayoría de los genes obtenidos están implicados en procesos cerebrales, principalmente en los impulsos nerviosos y componentes estructurales del sistema nervioso central.
- Se evaluó la relación que presenta la mutación LRRK2 G2019S con el aparecimiento de células que promueven el desarrollo pancreático, determinando que de forma indirecta la PD puede influir en los componentes estructurales del páncreas mediante células involucradas en actividades neuroendocrinas
- Se implementó un trabajo sistematizado para el diagnóstico temprano de la enfermedad mediante el uso de datos de disfonía por Parkinson. Utilizando algoritmos de aprendizaje automático computacional se logró una predicción aceptable del 81 %, destacando a los Coeficientes cepstrales de frecuencia Mel como los elementos de mayor importancia para el correcto telemonitoreo de pacientes candidatos a desarrollar la PD.
- Se trabajó con algoritmos que presentan mayor significancia dentro de los análisis para aprendizaje automático. “Random Forest” se destaca de entre los tres estudiados ya que muestra un nivel aceptable de predicción, esto debido al principio de análisis de este, ya que al añadir árboles de decisión que conforman una red estructurada que permite una predicción más significativa. Sin embargo, los otros dos métodos analizados no varían significativamente en sus resultados de predicción.

4.2 Recomendaciones

- Analizar varios tipos de datasets referentes a la PD y sus múltiples mutaciones con el objetivo de encontrar similitudes entre genes presentes en cada una de ellas. Mientras mayor cantidad de datsets provenientes de diferentes tipos de muestra se analicen, la comprensión de la enfermedad será mucho más amplia, pues se cubriría un número mayor de genes y de esta forma seria posible encontrar más factores comunes entre la PD tanto de origen genético como el adquirido por naturaleza
- Probar otro tipo de algoritmos de "Machine learning" y compararlos con los presentados en este estudio, siempre buscando el mejoramiento de la técnica de telemonitoreo; posiblemente una buena opción sería escalar en niveles de complejidad y potencia bioinformática probando con algoritmos basados en "redes neuronales" cuyos requerimientos son mayores, ya que se encuentran mejor estructurados y brindan resultados más completos.
- Examinar otro tipo de datos para el aprendizaje automático con fines predictivos, los síntomas que presenta el Parkinson son variados y no se remontan únicamente al movimiento involuntario, la actividad cerebral inusual puede resultar en un potente "vector" para la detección temprana de la enfermedad, muy probablemente con imágenes correspondientes a tomografías efectuadas a pacientes enfermos u otras fuentes de datos provenientes de la enfermedad.

REFERENCIAS BIBLIOGRÁFICAS

Abrahams, S., Miller, H. C., Lombard, C., Westhuizen, F. H. van der, & Bardien, S. (2021). Curcumin pre-treatment may protect against mitochondrial damage in LRRK2-mutant parkinson's disease and healthy control fibroblasts. *Biochemistry and Biophysics Reports*, *27*(1), 10–16. https://doi.org/10.1016/j.bbrep.2021.101035

Agapito, G., & Arbitrio, M. (2022). Microarray Data Analysis Protocol. *Methods in Molecular Biology (Clifton, N.J.)*, *2401*, 263–271. https://doi.org/10.1007/978-1-0716-1839-4_17

Ait Wahmane, S., Achbani, A., Elatiqi, M., Belmouden, A., & Nejmeddine, M. (2021). A meta-analysis of the prevalence of the mutation LRRK2 G2019S in patients with Parkinson's disease in Africa. *Gene Reports*, *24*(July). https://doi.org/10.1016/j.genrep.2021.101284

Álvarez, C. D. C. (2020). *Métodos para la selección de distribuciones a priori utilizando el estimador de James-Stein, planes de muestreo por atributo y modelos logísticos multinivel*. Universidad Nacional de Colombia.

Amarnath, B., Alias, A., & Balamurugan, S. (2016). Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. *Journal of Engineering Science and Technology*, *11*(11), 1639–1646.

Angelescu, R., & Dobrescu, R. (2021). MIDGET:Detecting differential gene expression on microarray data. *Computer Methods and Programs in Biomedicine*, *211*, 106418. https://doi.org/10.1016/J.CMPB.2021.106418

Arana, C. (2021). *Modelos de aprendizaje automático mediante árboles de decisión* (Vol. 45). Universidad del CEMA.

Arbez, N., He, X. F., Huang, Y., Ren, M., Liang, Y., Nucifora, F. C., Wang, X., Pei, Z., Tessarolo, L., Smith, W. W., & Ross, C. A. (2020). G2019S-LRRK2 mutation enhances MPTP-linked Parkinsonism in mice. *Human Molecular Genetics*, *29*(4), 580–590. https://doi.org/10.1093/hmg/ddz271

Arias-Molina, M. (2017). ¿Qué significa realmente el valor de p? *Rev Pediatría Atención Primaria*, *19*(76), 377–381.

Armstrong, R. (2020). What causes neurodegenerative disease? *Folia Neuropathologica*, *58*(2), 93–112. https://doi.org/10.5114/FN.2020.96707

Ayala, G. (2018). *Bioinformática estadística: Análisis estadístico de datos ómicos* (Issue 2). Universidad de Valencia.

Benamer, H. T. S., & De Silva, R. (2010). LRRK2 G2019S in the North African population: A review. *European Neurology*, *63*(6), 321–325. https://doi.org/10.1159/000279653

Betanzos, A. A., Bolón-Canedo, V., Morán-Fernández, L., & Maroño, N. S. (2019). A Review of Microarray Datasets: Where to Find Them and Specific

Characteristics. *Methods in Molecular Biology*, *1986*, 65–85. https://doi.org/10.1007/978-1-4939-9442-7_4

Beumer, J., & Clevers, H. (2021). Cell fate specification and differentiation in the adult mammalian intestine. *Nature Reviews Molecular Cell Biology*, *22*(1), 39–53. https://doi.org/10.1038/s41580-020-0278-0

Bhat, S., Acharya, U. R., Hagiwara, Y., Dadmehr, N., & Adeli, H. (2018). Parkinson's disease: Cause factors, measurable indicators, and early diagnosis. *Computers in Biology and Medicine*, *102*, 234–241. https://doi.org/10.1016/j.compbiomed.2018.09.008

Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, *397*(10291), 2284–2303. https://doi.org/10.1016/S0140-6736(21)00218-X

Bloomingdale, P., Karelina, T., Ramakrishnan, V., Bakshi, S., Véronneau-Veilleux, F., Moye, M., Sekiguchi, K., Meno-Tetang, G., Mohan, A., Maithreye, R., Thomas, V. A., Gibbons, F., Cabal, A., Bouteiller, J. M., & Geerts, H. (2022). Hallmarks of neurodegenerative disease: A systems pharmacology perspective. *CPT: Pharmacometrics and Systems Pharmacology*, *11*(11), 1399–1429. https://doi.org/10.1002/psp4.12852

Bonin, M., Poths, S., Osaka, H., Wang, Y. L., Wada, K., & Riess, O. (2017). Microarray expression analysis of gad mice implicates involvement of Parkinson's disease associated UCH-L1 in multiple metabolic pathways. *Molecular Brain Research*, *126*(1), 88–97. https://doi.org/10.1016/j.molbrainres.2004.03.025

Botta-Orfila, T., Tolosa, E., Gelpi, E., Sànchez-Pla, A., Martí, M. J., Valldeoriola, F., Fernández, M., Carmona, F., & Ezquerra, M. (2012). Microarray expression analysis in idiopathic and LRRK2-associated Parkinson's disease. *Neurobiology of Disease*, *45*(1), 462–468. https://doi.org/10.1016/j.nbd.2011.08.033

Boulos, C., Yaghi, N., Hayeck, R. El, Heraoui, G. N. H. A., & Fakhoury-Sayegh, N. (2019). Nutritional risk factors, microbiota and parkinson's disease: What is the current evidence? *Nutrients*, *11*(8), 1–24. https://doi.org/10.3390/nu11081896

Brimicombe, A. (2000). Constructing and evaluating contextual indices using GIS: A case of primary school performance tables. *Environment and Planning A*, *32*(11), 1909–1933. https://doi.org/10.1068/a3316

Buś, S., Jędrzejewski, K., & Guzik, P. (2022). Using minimum redundancy maximum relevance algorithm to select minimal sets of heart rate variability parameters for Atrial fibrillation detection. *Journal of Clinical Medicine*, *11*(14). https://doi.org/10.3390/jcm11144004

Carbon, S., & Mungall, C. (2023). *Gene Ontology Data Archive*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.10162580

Cerri, S., Mus, L., & Blandini, F. (2019). Parkinson's Disease in Women and Men: What's the Difference? *Journal of Parkinson's Disease*, *9*(3), 501–515. https://doi.org/10.3233/JPD-191683

Chamorro-Poyo, C. (2019). *Análisis de datos de RNA-Seq empleando diferentes*

paquetes desarrollados dentro del proyecto Bioconductor para estudios de expresión génica diferencial. Universitat Oberta de Catalunya.

Chen, L., Zhang, Y. H., Wang, S. P., Zhang, Y. H., Huang, T., & Cai, Y. D. (2017). Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. *PLoS ONE*, *12*(9), 1–22. https://doi.org/10.1371/journal.pone.0184129

Chen, P. H., Lin, C. J., & Schölkopf, B. (2005). A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry*, *21*(2), 111–136. https://doi.org/10.1002/asmb.537

Chen, X., & Le, W. (2021). LRRK2 G2019S mutation amplifies protein aggregate propagation. *Brain*, *144*(5), 1289–1290. https://doi.org/10.1093/brain/awab146.

Chen, X., Li, H., Huang, Y., Han, W., Yu, X., Zhang, P., & Tao, R. (2023). Heart sound classification based on equal scale frequency cepstral coefficients and deep learning. *Biomed Tech (Berl)*, *15*(68), 285–295. https://doi.org/10.1515/bmt-2021-0254

Chen, Y., Sun, X., Lin, Y., Zhang, Z., Gao, Y., & Wu, I. X. Y. (2021). Non-genetic risk factors for parkinson's disease: An overview of 46 systematic reviews. *Journal of Parkinson's Disease*, *11*(3), 919–935. https://doi.org/10.3233/JPD-202521

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2), 1–12. https://doi.org/10.1167/tvst.9.2.14

Chung, M., Bruno, V. M., Rasko, D. A., Cuomo, C. A., Muñoz, J. F., Livny, J., Shetty, A. C., Mahurkar, A., & Dunning Hotopp, J. C. (2021). Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biology*, *22*(1), 1–23. https://doi.org/10.1186/s13059-021-02337-8

Clayton, B. L. L., & Tesar, P. J. (2021). Oligodendrocyte progenitor cell fate and function in development and disease. *Curr Opin Cell Biol*, *73*(2), 35–40. https://doi.org/10.1016/j.ceb.2021.05.003

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus database. *Methods in Molecular Biology*, *1418*(301), 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cookson, M. R. (2010). The role of leucine-rich repeat kinase 2 (LRRK2) in Parkinson's disease. *Nature Reviews Neuroscience*, *11*(12), 791–797. https://doi.org/10.1038/nrn2935

de Lima, D. A., Helito, C. P., de Lima, L. L., Clazzer, R., Gonçalves, R. K., & de Camargo, O. P. (2022). How To Perform a Meta-Analysis: a Practical Step-By-Step Guide Using R Software and Rstudio. *Acta Ortopedica Brasileira*, *30*(3),

1–9. https://doi.org/10.1590/1413-785220223003e248775

De Miranda, B. R., Goldman, S. M., Miller, G. W., Greenamyre, J. T., & Dorsey, E. R. (2022). Preventing Parkinson's Disease: An Environmental Agenda. *Journal of Parkinson's Disease*, *12*(1), 45–68. https://doi.org/10.3233/JPD-212922

Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, *365*(January 2020), 114227. https://doi.org/10.1016/j.geoderma.2020.114227

Deo, R. C. (2016). Machine learning in medicine. *Circulation*, *132*(20), 248–256. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.Machine

Diaz, L., & Rios, F. (2018). El valor p. Interpretación, orígenes y su utilización actual. *Revista Argentina de Terapia Intensiva*, *35*(3), 55–59.

Dimmer, E. C., Huntley, R. P., Barrell, D. G., Binns, D., Draghici, S., Camon, E. B., Hubank, M., Talmud, P. J., Apweiler, R., & Lovering, R. C. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, *25*(3045–3046), 3045–3046. https://doi.org/10.1093/bioinformatics/btp536

Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003*, *3*(2), 523–528. https://doi.org/10.1109/CSB.2003.1227396

Dinh, T. P., Pham-Quoc, C., Thinh, T. N., Nguyen, B. K. Do, & Kha, P. C. (2023). A flexible and efficient FPGA-based random forest architecture for IoT applications. *Internet de Las Cosas*, *22*(1). https://doi.org/https://doi.org/10.1016/j.iot.2023.100813

El Mestari, S. Z., Lenzini, G., & Demirci, H. (2024). Preserving data privacy in machine learning systems. *Computers and Security*, *137*(1). https://doi.org/10.1016/j.cose.2023.103605

Elizondo-Cárdenas, G., Déctor-Carrillo, M. Á., Martínez-Rodríguez, H. R., Villarreal, L. M., & Esmer-Sánchez, M. del C. (2011). Genética y la enfermedad de Parkinson: Revisión de actualidades. *Medicina Universitaria*, *13*(51), 96–100. www.elsevier.es/en/node/2090153

Elsayad, A. M., Nassef, A. M., Al-Dhaifallah, M., & Elsayad, K. A. (2020). Classification of biodegradable substances using balanced random trees and boosted C5.0 decision trees. *International Journal of Environmental Research and Public Health*, *17*(24), 1–22. https://doi.org/10.3390/ijerph17249322

Espinosa Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, *21*(3), 1–16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*(2), 257–258. https://doi.org/10.1093/bioinformatics/btl567

Fan, L., Meng, H., Guo, X., Li, X., & Meng, F. (2018). Differential gene expression

profiles in peripheral blood in northeast Chinese han people with acute myocardial infarction. *Genetics and Molecular Biology*, *41*(1), 59–66. https://doi.org/10.1590/1678-4685-gmb-2017-0075

Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Arnan, C., Barnes, I., Banerjee, A., Bennett, R., Berry, A., Bignell, A., Boix, C., Calvet, F., Cerdán-Velez, D., Cunningham, F., Davidson, C., … Flicek, P. (2023). GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, *51*(D1), D942–D949. https://doi.org/10.1093/nar/gkac1071

Fu, Y., Zhao, D., Zhou, Y., Lu, J., Kang, L., Jiang, X., Xu, R., Ding, Z., & Zou, Y. (2022). Identification of differential expression genes between volume and pressure overloaded hearts based on bioinformatics analysis. *Genes*, *13*(7). https://doi.org/10.3390/genes13071276

Fujimoto, T., Kuwahara, T., Eguchi, T., Sakurai, M., Komori, T., & Iwatsubo, T. (2018). Parkinson's disease-associated mutant LRRK2 phosphorylates Rab7L1 and modifies trans-Golgi morphology. *Biochemical and Biophysical Research Communications*, *495*(2), 1708–1715. https://doi.org/10.1016/j.bbrc.2017.12.024

Ganz, N. B., Domínguez, F. A., Ares, A. E., & Kuna, H. D. (2019). Selección de características mediante la combinación de métodos para evaluar la precisión de clasificación en un conjunto de datos de implantes dentales. *Workshop de Investigadores En Ciencias de La Computación*, *1*(1).

Gao, L., Li, C., Yang, R. Y., Lian, W. W., Fang, J. S., Pang, X. C., Qin, X. M., Liu, A. L., & Du, G. H. (2015). Ameliorative effects of baicalein in MPTP-induced mouse model of Parkinson's disease: A microarray study. *Pharmacology Biochemistry and Behavior*, *133*(1), 155–163. https://doi.org/10.1016/j.pbb.2015.04.004

Genis-Mendoza, A. D., Martínez-Magaña, J. J., Bojórquez, C., Téllez-Martínez, J. A., Jiménez-Genchi, J., Roche, A., Bojorge, A., Chávez, M., Castañeda, C., Guzmán, R., Zapata, L., Aguilar-Méndez, D., Lanzagorta, N., Rebolledo, I., Castro-Chavira, S., Fernández, T., Orozco, L., Nicolini, H., & Martínez-Hernández, A. G. (2018). Programa de detección del alelo APOE-E4 en adultos mayores mexicanos con deterioro cognitivo. *Gaceta Medica de Mexico*, *154*(5), 555–560. https://doi.org/10.24875/GMM.18003784

Gentleman, R., Carey, V. J., Huber, W., & Hahne, F. (2023). *Genefilter: methods for filtering genes from high-throughput experiments* (1.80.3; p. 39). Bioconductor Package Maintainer.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 Oor 80/20 relation between training and testing sets : A pedagogical explanation. *Departmental Technical Reports (CS)*, *2*(1), 1–6.

Gilsbach, B. K., Eckert, M., & Gloeckner, C. J. (2018). Regulation of LRRK2: Insights from structural and biochemical analysis. *Biological Chemistry*, *399*(7), 637–642. https://doi.org/10.1515/hsz-2018-0132

Gitler, A. D., Dhillon, P., & Shorter, J. (2017). Neurodegenerative disease: Models, mechanisms, and a new hope. *DMM Disease Models and Mechanisms*, *10*(5), 499–502. https://doi.org/10.1242/dmm.030205

Gittes, G. K. (2009). Developmental biology of the pancreas: A comprehensive review. *Developmental Biology*, *326*(1), 4–35. https://doi.org/10.1016/j.ydbio.2008.10.024

González-Fernández, E., Álvarez-López, S., Garrido, A., Fernández-González, M., & Rodríguez-Rajo, F. J. (2022). Data mining assessment of Poaceae pollen influencing factors and its environmental implications. *Sci Total Environ*, *815*(1). https://doi.org/10.1016/j.scitotenv.2021.152874

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Rev Mol Cell Biol*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Guri, A. J., Mohapatra, S. K., Horne, W. T., Hontecillas, R., & Bassaganya-Riera, J. (2010). The Role of T cell PPAR γ in mice with experimental inflammatory bowel disease. *BMC Gastroenterology*, *10*. https://doi.org/10.1186/1471-230X-10-60

Harvey, J., Pishva, E., Chouliaras, L., & Lunnon, K. (2023). Elucidating distinct molecular signatures of Lewy body dementias. *Neurobiology of Disease*, *1*(23), 13–39. https://doi.org/10.1016/j.nbd.2023.106337

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*, *32*(6), 441–444. https://doi.org/10.1111/j.1532-5415.1984.tb02220.x

Hayes, M. T. (2019). Parkinson's Disease and Parkinsonism. *American Journal of Medicine*, *132*(7), 802–807. https://doi.org/10.1016/j.amjmed.2019.03.001

Huang, Y., Liu, Y., Huang, Q., Sun, S., Ji, Z., Huang, L., Li, Z., Huang, X., Deng, W., & Li, T. (2022). TMT-Based quantitative proteomics analysis of synovial fluid-derived exosomes in inflammatory arthritis. *Frontiers in Immunology*, *13*(March), 1–14. https://doi.org/10.3389/fimmu.2022.800902

Hung, J. H., & Weng, Z. (2017). Analysis of microarray and RNA-seq expression profiling data. *Cold Spring Harbor Protocols*, *2017*(3), 191–196. https://doi.org/10.1101/pdb.top093104

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2012). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Selected Works of Terry Speed*, *4*(2), 601–616. https://doi.org/10.1007/978-1-4614-1347-9_15

Jahanshahi, H., & Baydogan, M. G. (2022). nTreeClus: un codificador de secuencia basado en árbol para agrupar series categóricas. *Neurocomputación*, *494*(1), 224–241. https://doi.org/https://doi.org/10.1016/j.neucom.2022.04.076

Jain, S., Shukla, S., & Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications*, *106*(1), 252–262. https://doi.org/10.1016/j.eswa.2018.04.008

Jankovic, J., & Tan, E. K. (2020). Parkinson's disease: Etiopathogenesis and treatment. *Journal of Neurology, Neurosurgery and Psychiatry*, *91*(8), 795–808. https://doi.org/10.1136/jnnp-2019-322338

Jiang, M., Wang, J., Hu, L., & He, Z. (2023). Random forest clustering for discrete sequences. *Pattern Recognition Letters*, *174*(1), 145–151. https://doi.org/https://doi.org/10.1016/j.patrec.2023.09.001

Jin, F., Ta, L., Liu, M., Sun, Y., Pan, Y., Li, Z., & Xu, D. (2022). Fluorescence microarrays for enzyme-free DNA detection based on web hybrid chain reaction. *Biosensors and Bioelectronics: X*, *11*(March), 0–5. https://doi.org/10.1016/j.biosx.2022.100151

Jo, S., Park, K. W., Hwang, Y. S., Lee, S. H., Ryu, H. S., & Chung, S. J. (2021). Microarray genotyping identifies new loci associated with dementia in parkinson's disease. *Genes*, *12*(12). https://doi.org/10.3390/genes12121975

Josso, P., Hall, A., Williams, C., Le Bas, T., Lusty, P., & Murton, B. (2023). Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean. *Ore Geology Reviews*, *162*(1). https://doi.org/10.1016/j.oregeorev.2023.105671

Ju, Z., & He, J. J. (2018). Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. *Analytical Biochemistry*, *550*(1), 1–7. https://doi.org/10.1016/j.ab.2018.04.005

Kadiri, S. R., & Alku, P. (2019). Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing. *The Journal of the Acoustical Society of America*, *146*(5), EL418–EL423. https://doi.org/10.1121/1.5131043

Kain, Z., & MacLaren, J. (2007). Valor de p inferior a 0.05: ¿qué significa en realidad? *Pediatrics (Ed. Española)*, *63*(3), 118–120.

Kessler, C., Atasu, B., Hanagasi, H., Simón-Sánchez, J., Hauser, A. K., Pak, M., Bilgic, B., Erginel-Unaltuna, N., Gurvit, H., Gasser, T., & Lohmann, E. (2018). Role of LRRK2 and SNCA in autosomal dominant Parkinson's disease in Turkey. *Parkinsonism and Related Disorders*, *48*, 34–39. https://doi.org/10.1016/j.parkreldis.2017.12.007

Khan, H. ur R., Khidmat, W. Bin, Hammouda, A., & Muhammad, T. (2023). Machine learning in the boardroom: Gender diversity prediction using boosting and undersampling methods. *Research in International Business and Finance*, *66*(1). https://doi.org/https://doi.org/10.1016/j.ribaf.2023.102053

Khatri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, *21*(18), 3587–3595. https://doi.org/10.1093/bioinformatics/bti565

Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y., & Shimizu, N. (1998). Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, *392*(6676), 605–608. https://doi.org/10.1038/33416

Kong, P., Lei, P., Zhang, S., Li, D., Zhao, J., & Zhang, B. (2018). Integrated microarray analysis provided a new insight of the pathogenesis of Parkinson's disease. *Neuroscience Letters*, *662*(1), 51–58. https://doi.org/10.1016/j.neulet.2017.09.051

Kruschke, J. K. (2015). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition. In *Elseiver*. https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Springer Science* (5th ed.). Springer Nature. https://doi.org/10.1007/978-1-4614-6849-3

Lacombe, D., & Rooryck-Thambo, C. (2018). Apport des puces à ADN et nouveaux syndromes microdélétionnels. *Bulletin de l'Académie Nationale de Médecine*, *202*(3–4), 693–705. https://doi.org/10.1016/s0001-4079(19)30310-3

Lahmiri, S. (2017). Parkinson's disease detection based on dysphonia measurements. *Physica A: Statistical Mechanics and Its Applications*, *471*(1), 98–105. https://doi.org/10.1016/j.physa.2016.12.009

Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine*, *129*, 104151. https://doi.org/10.1016/j.compbiomed.2020.104151

Lewis, P. A., & Spillane, J. E. (2019). Alzheimer's disease and dementia. In *The Molecular and Clinical Pathology of Neurodegenerative Disease*. https://doi.org/10.1007/978-3-030-23277-1_10

Li, X., Yi, S., Cundy, A. B., & Chen, W. (2022). Sustainable decision-making for contaminated site risk management: A decision tree model using machine learning algorithms. *Journal of Cleaner Production*, *371*(1). https://doi.org/https://doi.org/10.1016/j.jclepro.2022.133612

Lin, D., Liang, Y., Jing, X., Chen, Y., Lei, M., Zeng, Z., Zhou, T., Wu, X., Peng, S., Zheng, D., Huang, K., Yang, L., Xiao, S., Liu, J., & Tao, E. (2018). Microarray analysis of an synthetic α-synuclein induced cellular model reveals the expression profile of long non-coding RNA in Parkinson's disease. *Brain Research*, *1678*(1), 384–396. https://doi.org/10.1016/j.brainres.2017.11.007

Liu, C.-X., Yu, G.-L., & Liu, Z. (2022). Machine learning models in phononic metamaterials. *Current Opinion in Solid State & Materials Science*, *26*(1). https://doi.org/10.1016/j.cossms.2023.101133

Liu, L., Chen, L., Zhang, Y. H., Wei, L., Cheng, S., Kong, X., Zheng, M., Huang, T., & Cai, Y. D. (2017). Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection. In *Journal of Biomolecular Structure and Dynamics* (Vol. 35, Issue 2). https://doi.org/10.1080/07391102.2016.1138142

Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, *249*(1). https://doi.org/10.1016/j.chemosphere.2020.126169

Lujan-Moreno, G. A., Howard, P. R., Rojas, O. G., & Montgomery, D. C. (2018).

Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. *Expert Systems with Applications*, *109*, 195–205. https://doi.org/10.1016/j.eswa.2018.05.024

Ma, A., Lau, K. K., & Thyagarajan, D. (2020). Voice changes in Parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, *72*(7), 1–7. https://doi.org/10.1016/j.jocn.2019.12.029

Ma, G., Liu, M., Du, K., Zhong, X., Gong, S., Jiao, L., & Wei, M. (2019). Differential expression of mRNAs in the brain tissues of patients with Alzheimer's disease based on GEO expression profile and its clinical significance. *BioMed Research International*, *2019*(45), 1–9. https://doi.org/10.1155/2019/8179145

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support Vector Machines (SVM). *Gesture*, *23*(6), 349–361. https://doi.org/10.1002/wics.049

Marjit, S., Bhattacharyya, T., Chatterjee, B., & Sarkar, R. (2023). Simulated annealing aided genetic algorithm for gene selection from microarray data. *Computers in Biology and Medicine*, *158*, 106854. https://doi.org/10.1016/J.COMPBIOMED.2023.106854

Martínez, G., & Torres, G. (2013). Reconocimiento de voz basado en MFCC , SBC y Espectrogramas. *Ingenius*, *1*, 12–20.

Mastracci, T. L., & Sussel, L. (2013). The Endocrine Pancreas: insights into development, differentiation and diabetes. *Wiley Interdiscip Rev Dev Biol*, *5*(1), 685–705. https://doi.org/10.1002/wdev.44.The

McHugh, M. L. (2012). Interrater reliability : The kappa statistic. *Biochemica Medica*, *22*(3), 276–282. https://hrcak.srce.hr/89395

McLachlan, G. J., Bean, R. W., & Ng, S. K. (2017). Clustering. In *Methods in Molecular Biology* (2nd ed., Vol. 1526, pp. 345–362). Methods in Molecular Biology,. https://doi.org/10.1007/978-1-4939-6613-4_19

Mehanna, R., & Jankovic, J. (2019). Young-onset Parkinson's disease: Its unique features and their impact on quality of life. *Parkinsonism and Related Disorders*, *65*, 39–48. https://doi.org/10.1016/j.parkreldis.2019.06.001

Miller, R. M., & Federoff, H. J. (2010). Microarrays in Parkinson 's disease : A systematic approach. *The American Society for Experimental NeuroTherapeutics*, *3*(1), 319–326.

Miranda, M. (2007). Mutación del gen LRRK2 se asocia a presentación autosómica dominante de la enfermedad de Parkinson en una familia chilena. *Revista Medica de Chile*, *135*(3), 406–407. https://doi.org/10.4067/s0034-98872007000300019

Mittapalle, K. R., Yagnavajjula, M. K., & Alku, P. (2023). Classification of functional dysphonia using the tunable Q wavelet transform. *Speech Communication*, *155*(August), 102989. https://doi.org/10.1016/j.specom.2023.102989

Miyashita, M., Bell, J. S. K., Wenric, S., Karaesmen, E., Rhead, B., Kase, M.,

Kaneva, K., De La Vega, F. M., Zheng, Y., Yoshimatsu, T. F., Khramtsova, G., Liu, F., Zhao, F., Howard, F. M., Nanda, R., Beaubier, N., White, K. P., Huo, D., & Olopade, O. I. (2023). Molecular profiling of a real-world breast cancer cohort with genetically inferred ancestries reveals actionable tumor biology differences between European ancestry and African ancestry patient populations. *Breast Cancer Research*, *25*(1), 1–13. https://doi.org/10.1186/s13058-023-01627-2

Mohapatra, S. K., & Krishnan, A. (2011). Microarray data analysis. In *Plant Reverse Genetics: Methods and Protocols* (1st ed., Vol. 678, Issue 1, pp. 211–227). Methods in Molecular Biology. https://doi.org/10.1007/978-1-60761-682-5

Montalvo, J., Montalvo, P., Albear, L., Intriago, E., & Moreira, D. (2017). Prevalencia de la enfermedad de Parkinson: Estudio puerta-puerta en la provincia de manabí-Ecuador. *Revista Ecuatoriana de Neurologia*, *26*(1), 23–26. http://scielo.senescyt.gob.ec/pdf/rneuro/v26n1/2631-2581-rneuro-26-01-00023.pdf

Morello-Da Cruz, A. N., Beber-Costa, B. C., Fagundes-Carvalho, V., Cielo-Aparecida, C., & Rieder, C. R. (2020). Dysphonia and Dysarthria in people with parkinson's disease after subthalamic nucleus deep brain stimulation: Effect of frequency modulation. *Journal of Voice*, *34*(3), 477–484. https://doi.org/10.1016/j.jvoice.2018.10.012

Moreno, V., & Solé, X. (2004). Uso de chips de ADN (microarrays) en medicina: Fundamentos técnicos y procedimientos básicos para el análisis estadístico de resultados. *Medicina Clinica*, *122*(SUPPL. 1), 73–79. https://doi.org/10.1157/13057538

Nakajima, A., & Ohizumi, Y. (2019). Potential benefits of nobiletin, a citrus flavonoid, against Alzheimer's disease and Parkinson's disease. *International Journal of Molecular Sciences*, *20*(14), 1–14. https://doi.org/10.3390/ijms20143380

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Oerton, E., & Bender, A. (2017). Concordance analysis of microarray studies identifies representative gene expression changes in Parkinson's disease: A comparison of 33 human and animal studies. *BMC Neurology*, *17*(1), 1–14. https://doi.org/10.1186/s12883-017-0838-x

Olvera-Granados, C. P., Leo-Amador, G. E., & Hernández-Montiel, H. L. (2008). Páncreas y células beta: mecanismos de diferenciación, morfogénesis y especificación celular endocrina. ¿Regeneración? *Boletín Médico Del Hospital Infantil de México*, *65*(4), 306–324.

Origel-Rivas, G., Rendón-Lara, E., María Abundez-Barrera, I., & Alejo-Eleuterio, R. (2020). Redes neuronales artificiales y árboles de decisiónpara la clasificación con datos categóricos. *Research in Computing Science*, *149*(8), 541–554.

Pagès, H., Carlson, M., Falcon, S., & Li, N. (2023). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. In *R package version 1.64.1*.

https://doi.org/doi:10.18129/B9.bioc.AnnotationDb

Paredes, I., Vieira, J. R., Shah, B., Ramunno, C. F., Dyckow, J., Adler, H., Richter, M., Schermann, G., Giannakouri, E., Schirmer, L., Augustin, H. G., & Ruiz de Almodóvar, C. (2021). Oligodendrocyte precursor cell specification is regulated by bidirectional neural progenitor–endothelial cell crosstalk. *Nature Neuroscience*, *24*(4), 478–488. https://doi.org/10.1038/s41593-020-00788-z

Pischedda, F., Cirnaru, M. D., Ponzoni, L., Sandre, M., Biosa, A., Carrion, M. P., Marin, O., Morari, M., Pan, L., Greggio, E., Bandopadhyay, R., Sala, M., & Piccoli, G. (2021). LRRK2 G2019S kinase activity triggers neurotoxic NSF aggregation. *Brain*, *144*(5), 1509–1525. https://doi.org/10.1093/brain/awab073

Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. In *Machine Learning: Methods and Applications to Brain Disorders*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, *18*(1), 1–14. https://doi.org/10.1186/s12859-016-1423-9

Reich, S. G., & Savitt, J. M. (2019). Parkinson's Disease. *Medical Clinics of North America*, *103*(2), 337–350. https://doi.org/10.1016/j.mcna.2018.10.014

Ren, C., Ding, Y., Wei, S., Guan, L., Zhang, C., Ji, Y., Wang, F., Yin, S., & Yin, P. (2019). G2019S Variation in LRRK2: An ideal model for the study of parkinson's disease? *Frontiers in Human Neuroscience*, *13*(September), 1–6. https://doi.org/10.3389/fnhum.2019.00306

Rocha, E. M., Keeney, M. T., Di Maio, R., De Miranda, B. R., & Greenamyre, J. T. (2022). LRRK2 and idiopathic Parkinson's disease. *Trends in Neurosciences*, *45*(3), 224–236. https://doi.org/10.1016/j.tins.2021.12.002

Rodríguez-Martín, A. (2023). *Eficacia de la intervención logopédica en disfonía hiperfuncional en función del número de sesiones semanales* [Universidad deValladolid]. https://uvadoc.uva.es/handle/10324/61598

Ruiz-Villafranca, S., Roldán-Gómez, J., Carrillo-Mondéjar, J., Gómez, J. M. C., & Villalón, J. M. (2023). A MEC-IIoT intelligent threat detector based on machine learning boosted tree algorithms. *Computer Networks*, *233*(1). https://doi.org/10.1016/j.comnet.2023.109868

Ruiz Palmero, I. (2017). *Papel de la neurogenina 3 en las acciones neuritogénicas del estradiol en el hipocampo* [Universidad Autónoma de Madrid]. http://hdl.handle.net/10486/677470

Saffari, M., Khodayar, M., Saadabadi, M. S. E., Sequeira, A. F., & Cardoso, J. S. (2021). Maximum relevance minimum redundancy dropout with informative kernel determinantal point process. *Sensors*, *21*(5), 1–21. https://doi.org/10.3390/s21051846

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., & Apaydin, H. (2019). A comparative analysis of

speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing Journal*, *74*(1), 255–263. https://doi.org/10.1016/j.asoc.2018.10.022

Sakr, S., Elshawi, R., Ahmed, A., Qureshi, W. T., Brawner, C., Keteyian, S., Blaha, M. J., & Al-Mallah, M. H. (2018). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) project. *PLoS ONE*, *13*(4), 1–18. https://doi.org/10.1371/journal.pone.0195344

Salicrú, M., Ocaña, J., & Sánchez-Pla, A. (2011). Comparison of lists of genes based on functional profiles. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-401

Sánchez, S. del C. (2015). *Análisis de datos de RNA-Seq: comparación de métodos para el estudio de expresión génica diferencial* [Universidad de Sevilla]. https://idus.us.es/xmlui/handle/11441/40809

Sanders, L. H., Laganière, J., Cooper, O., Mak, S. K., Vu, B. J., Huang, Y. A., Paschon, D. E., Vangipuram, M., Sundararajan, R., Urnov, F. D., Langston, J. W., Gregory, P. D., Zhang, H. S., Greenamyre, J. T., Isacson, O., & Schüle, B. (2014). LRRK2 mutations cause mitochondrial DNA damage in iPSC-derived neural cells from Parkinson's disease patients: Reversal by gene correction. *Neurobiology of Disease*, *62*, 381–386. https://doi.org/10.1016/j.nbd.2013.10.013

Santos-García, D., de Deus Fonticoba, T., Cores Bartolomé, C., Feal Painceiras, M. J., García Díaz, I., Íñiguez Alvarado, M. C., Paz, J. M., Jesús, S., Cosgaya, M., García Caldentey, J., Caballol, N., Legarda, I., González Aramburu, I., Ávila Rivera, M. A., Gómez Mayordomo, V., Vela, L., Escalante, S., Mendoza, Z., Martínez Castrillo, J. C., … Mir, P. (2023). Response to levodopa in Parkinson's disease over time. A 4-year follow-up study. *Parkinsonism & Related Disorders*, *116*, 105852. https://doi.org/10.1016/J.PARKRELDIS.2023.105852

Schootemeijer, S., Coker, D., Shelton, J. F., Chanoff, E., Rowbotham, H. M., Darweesh, S. K. L., Bloem, B. R., Cannon, P., & de Vries, N. M. (2023). Exercise knowledge, barriers and motivators among LRRK2 G2019S mutation carriers. *Parkinsonism and Related Disorders*, *113*(May), 0–3. https://doi.org/10.1016/j.parkreldis.2023.105497

Scionti, F., Arbitrio, M., Caracciolo, D., Pensabene, L., Tassone, P., Tagliaferri, P., & Di Martino, M. T. (2022). Integration of DNA Microarray with Clinical and Genomic Data. *Methods in Molecular Biology (Clifton, N.J.)*, *2401*, 239–248. https://doi.org/10.1007/978-1-0716-1839-4_15

Shin, J. H., Ko, H. S., Kang, H., Lee, Y., Lee, Y. Il, Pletinkova, O., Troconso, J. C., Dawson, V. L., & Dawson, T. M. (2011). PARIS (ZNF746) repression of PGC-1alpha contributes to neurodegeneration in parkinson's disease. *Cell*, *144*(5), 689–702. https://doi.org/10.1016/j.cell.2011.02.010

Siknun, G. P., & Sitanggang, I. S. (2016). Web-based classification application for forest fire data using the shiny framework and the C5.0 algorithm. *Procedia Environmental Sciences*, *33*(1), 332 – 339.

Simon, D. K., Tanner, C. M., & Brundin, P. (2020). Parkinson Disease Epidemiology, Pathology, Genetics and Pathophysiology. *Clin Geriatr Med*, *36*(1), 139–148. https://doi.org/10.1016/j.cger.2019.08.002

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1). https://doi.org/10.2202/1544-6115.1027

Snow, G., & Guardiani, E. (2019). Movement disorders and voice. *Otolaryngologic Clinics of North America*, *52*(4), 759–767. https://doi.org/10.1016/j.otc.2019.03.018

Sol, J., Aaen, M., Sadolin, C., & Bosch, L. ten. (2023). Towards automated vocal mode classification in healthy singing voice: An XGBoost decision tree-based machine learning classifier. *Journal of Voice*, *11*(1). https://doi.org/https://doi.org/10.1016/j.jvoice.2023.09.006

Suguiura-Rivero, F. O. (2022). Árbol de decisión en aprendizaje automático decision tree in machine learning. *Revista Varianza*, *19*(1), 39–46. https://ojs.umsa.bo/ojs/index.php/revistavarianza/article/view/433/365

Sun, Q., Li, X., Xu, M., Zhang, L., Zuo, H., Xin, Y., Zhang, L., & Gong, P. (2020). Differential expression and bioinformatics analysis of circRNA in non-small cell lung cancer. *Frontiers in Genetics*, *11*(1), 1–11. https://doi.org/10.3389/fgene.2020.586814

Sun, Y., Ye, L., Zheng, Y., & Yang, Z. (2018). Identification of crucial genes associated with Parkinson's disease using microarray data. *Molecular Medicine Reports*, *17*(3), 3775–3782. https://doi.org/10.3892/mmr.2017.8305

Tan, L. F., Ng, S. E., & Merchant, R. (2018). Atypical Cause of Functional Decline in Parkinson's Disease. *American Journal of Medicine*, *131*(6), e243–e244. https://doi.org/10.1016/j.amjmed.2018.01.036

Tell-Marti, G., Sarda, S. P., & Puig-Butille, J. A. (2021). Gene expression microarray: Technical fundamentals and data analysis. In *Comprehensive Foodomics* (pp. 291–312). Elsevier. https://doi.org/10.1016/B978-0-08-100596-5.22905-3

Tolosa, E., Vila, M., Klein, C., & Rascol, O. (2020). LRRK2 in Parkinson disease: challenges of clinical trials. *Nature Reviews Neurology*, *16*(2), 97–107. https://doi.org/10.1038/s41582-019-0301-2

Treviño-Cantú, J. A. (2022). Alternativas de estandarización para índices compuestos espacio-temporales. El caso del rezago educativo en los estados de México, 2000 a 2020. *Investigaciones Geográficas*, *109*(11). https://doi.org/10.14350/rig.60615

Trujillano, J., Sarria-Santamera, A., Esquerda, A., & Badía, M. (2010). Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. *Gaceta Sanitaria*, *22*(1), 65–70. https://doi.org/10.1157/13115113

Valente, E. M., Abou-Sleiman, P. M., Caputo, V., Muqit, M. M. K., Harvey, K.,

Gispert, S., Ali, Z., Del Turco, D., Bentivoglio, A. R., Healy, D. G., Albanese, A., Nussbaum, R., González-Maldonado, R., Deller, T., Salvi, S., Cortelli, P., Gilks, W. P., Latchman, D. S., Harvey, R. J., … Wood, N. W. (2004). Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science*, *304*(5674), 1158–1160. https://doi.org/10.1126/science.1096284

Volta, M., Beccano-Kelly, D. A., Paschall, S. A., Cataldi, S., Macisaac, S. E., Kuhlmann, N., Kadgien, C. A., Tatarnikov, I., Fox, J., Khinda, J., Mitchell, E., Bergeron, S., Melrose, H., Farrer, M. J., & Milnerwood, A. J. (2017). Initial elevations in glutamate and dopamine neurotransmission decline with age, as does exploratory behavior, in LRRK2 G2019S knock-in mice. *ELife*, *6*(2), 1–17. https://doi.org/10.7554/eLife.28377

Walter, J., Bolognin, S., Poovathingal, S. K., Magni, S., Gérard, D., Antony, P. M. A., Nickels, S. L., Salamanca, L., Berger, E., Smits, L. M., Grzyb, K., Perfeito, R., Hoel, F., Qing, X., Ohnmacht, J., Bertacchi, M., Jarazo, J., Ignac, T., Monzel, A. S., … Schwamborn, J. C. (2021). The Parkinson's-disease-associated mutation LRRK2-G2019S alters dopaminergic differentiation dynamics via NR2F1. *Cell Reports*, *37*(3). https://doi.org/10.1016/j.celrep.2021.109864

Wang, H., Yao, Y., & Salhi, S. (2020). Tension in big data using machine learning: Analysis and applications. *Technological Forecasting and Social Change*, *158*(May 2019), 120175. https://doi.org/10.1016/j.techfore.2020.120175

Watanabe, R., Buschauer, R., Böhning, J., Audagnotto, M., Lasker, K., Lu, T. W., Boassa, D., Taylor, S., & Villa, E. (2020). The In Situ Structure of Parkinson's Disease-Linked LRRK2. *Cell*, *182*(6), 1508-1518.e16. https://doi.org/10.1016/j.cell.2020.08.004

Williams-Gray, C. H., & Worth, P. F. (2023). Parkinson's disease and related conditions. *Medicine*, *51*(9), 645–651. https://doi.org/10.1016/J.MPMED.2023.06.004

Yu, Z., Sakai, M., Fukushima, H., Ono, C., Kikuchi, Y., Koyama, R., Matsui, K., Furuyashiki, T., Kida, S., & Tomita, H. (2023). Microarray dataset of gene transcription in mouse microglia and peripheral monocytes in contextual fear conditioning. *Data in Brief*, *46*, 108862. https://doi.org/10.1016/j.dib.2022.108862

Zeve, D., Stas, E., de Sousa Casal, J., Mannam, P., Qi, W., Yin, X., Dubois, S., Shah, M. S., Syverson, E. P., Hafner, S., Karp, J. M., Carlone, D. L., Ordovas-Montanes, J., & Breault, D. T. (2022). Robust differentiation of human enteroendocrine cells from intestinal stem cells. *Nature Communications*, *13*(1), 1–20. https://doi.org/10.1038/s41467-021-27901-5

Zhao, F., Hao, J., Zhang, H., Yu, X., Yan, Z., & Wu, F. (2024). Quality recognition method of oyster based on U-net and random forest. *Journal of Food Composition and Analysis*, *125*(1). https://doi.org/https://doi.org/10.1016/j.jfca.2023.105746

Zhao, J., He, K., Du, H., Wei, G., Wen, Y., Wang, J., Zhou, X., & Wang, J. (2022). Bioinformatics prediction and experimental verification of key biomarkers for diabetic kidney disease based on transcriptome sequencing in mice. *PeerJ*,

10(1). https://doi.org/10.7717/peerj.13932

Zheng, B., Liao, Z., Locascio, J. J., Lesniak, K. A., Roderick, S. S., Watt, M. L., Eklund, A. C., Zhang-James, Y., Kim, P. D., Hauser, M. A., Grünblatt, E., Moran, L. B., Mandel, S. A., Riederer, P., Miller, R. M., Cantuti-Castelvetri, I., Young, A. B., Vance, J. M., Davis, R. L., … Scherzer, C. R. (2011). PGC-1alpha, a potential therapeutic target for early intervention in parkinson's disease. *Science Translational Medicine*, *2*(52), 52–73. https://doi.org/10.1126/scitranslmed.3001059.PGC-1

ANEXOS

La información correspondiente a los Anexos 1, 2, 3 y 4 se encuentra adjunta en el siguiente enlace https://doi.org/10.5281/zenodo.10457024

Anexo 1. Código empleado para la ejecución del análisis de expresión diferencial.

Archivo titulado como: "Anexo 1. Análisis de Datos Ómicos"

Anexo 2. Código empleado para la ejecución del análisis de Mínima redundancia / Máxima relevancia (mRMR)

Archivo titulado como: "Anexo 2. Análisis mRMR"

Anexo 3. Características y observaciones seleccionadas para ser utilizadas dentro del análisis de Aprendizaje Autónomo

Archivo titulado como: "Anexo 3. Resultados análisis mRMR"

Anexo 4. Código empleado para la ejecución del análisis de Aprendizaje Autónomo o "Machine learning"

Archivo titulado como: "Anexo 4. Análisis Machine Learning"

Anexo 5. Visualización de resultados, genes de mayor significancia en análisis *up - down regulated*

Nivel de expresión	Symbol	ID	Descripción	Categoría
Down Regulated	PAX6	GC11M031784	Paired Box 6	Protein Coding
	TFAP2B	GC06P119195	Transcription Factor AP-2 Beta	Protein Coding
	FRZB	GC02M182833	Frizzled Related Protein	Protein Coding
	ADD2	GC02M070626	Adducin 2	Protein Coding
	CAS1	GC07P094509	CAS1 Domain Containing 1	Protein Coding
	NMU	GC04M055595	Neuromedin U	Protein Coding
Up Regulated	PLOD2	GC03M146035	Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2	Protein Coding
	NKX2-2	GC20M021511	NK2 Homeobox 2	Protein Coding
	FOXA2	GC20M022581	Forkhead Box A2	Protein Coding
	SELENOP	GC05M042800	Selenoprotein P	Protein Coding