ANÁLISIS SOBRE LA EVOLUCIÓN DE LA ENFERMEDAD DEL PÁRKINSON

TRABAJO DE FIN DE GRADO

GRADO EN ESTADÍSTICA EMPRESARIAL

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

Curso académico 2023 - 2024

Autora: Inmaculada Ramírez López

Tutora: María Asunción Martínez Mayoral

Índice de contenidos

1. Resumen

Este Trabajo de Fin de Grado aborda el análisis de una base de datos que recoge información sobre los resultados de las analíticas de 248 pacientes enfermos de Párkinson durante años de examen. El objetivo principal es identificar, mediante un análisis descriptivo exhaustivo y distintos algoritmos de Machine Learning desarrollados en Python, biomarcadores diagnósticos, pronósticos y de progresión válidos, así como correlaciones entre la evolución de la enfermedad y las variables predictoras. Además, compararemos todas las técnicas desarrolladas en este estudio con el fin de identificar aquella que proporcione mejores predicciones.

2. Palabras clave

Párkinson, análisis multivariado, predicción, evolución de la enfermedad, Machine Learning.

3. Introducción

La enfermedad del Párkinson es un trastorno neurodegenerativo del sistema nervioso central cuya principal característica es la muerte progresiva de neuronas en una parte del cerebro. Los síntomas más comunes están relacionados con temblores, bradicinesia, rigidez muscular y cambios en el habla, entre muchos otros.

La pérdida neuronal está marcada por una disminución de la dopamina en el cerebro, una sustancia sintetizada por las neuronas, originando así una disfunción en la regulación de las estructuras principales del cerebro implicadas en el control de los movimientos. La pérdida neuronal está asociada también a la formación de cuerpos de Lewy en el cerebro, aglomeraciones anormales de proteínas que se forman en el cerebro causando la muerte de las neuronas, y degeneración del cerebro.

Con esta investigación pretendemos identificar y validar biomarcadores diagnósticos y de progresión en proteínas y péptidos, con el fin de contribuir a la identificación de nuevas vías para la investigación y el desarrollo terapéutico.

Cabe destacar las investigaciones sobre el tema de Holden et al (2017), Shi et al (2015b) y Goetz, Tilley, Shaftman, Stebbins, Fahn, Martinez-Martin, et al (2008).

Actualmente, pese a las grandes inversiones en investigación y desarrollo, son pocos los tratamientos específicos aplicados a la enfermedad del Párkinson, de hecho no se ha alcanzado ningún hallazgo significativo que permita conocer las causas exactas de la muerte neuronal y formación de cuerpos de Lewy que provoca la enfermedad. Por ello, es crucial recurrir a asociaciones público-privadas sólidas así como voluntarios transformadores que puedan abordar de manera efectiva la problemática con el propósito de investigar sobre el tema en busca de una solución efectiva.

Las investigaciones previas indican que las anormalidades en proteínas o péptidos desempeñan un papel clave en el inicio y empeoramiento de la enfermedad. Esto se debe a que la proteína alfa-sinucleína, presente en condiciones normales en el cerebro, se pliega de manera anómala formando acumulaciones insolubles en las neuronas. Todo ello contribuye a la degradación de proteínas y la respuesta inmunitaria.

Nuestra base de datos procede de una competición de Kaggle titulada “AMP®-Parkinson's Disease Progression Prediction”. Los datos proporcionados serán el objeto de investigación, donde figuran las anormalidades en péptidos y proteínas en los pacientes analizados durante años de estudio. Este reto ha sido puesto a disposición de la comunidad internacional de analistas con el fin de obtener información útil que pueda generar algún cambio en el futuro de la enfermedad.

Sobre esta base de datos se plantea el siguiente Trabajo de Fin de Grado.

4. Objetivos

Utilizando como referente la base de datos "AMP®-Parkinson's Disease Progression Prediction" en Kaggle con información sobre la evolución diagnóstica y analítica de pacientes de párkinson, el objetivo general de este estudio consiste en identificar biomarcadores de progresión válidos, y construir modelos de predicción que permitan inferir sobre el pronóstico de un paciente en función de sus analíticas y características físicas.

4.1 Objetivos específicos

Puesto que el avance de la enfermedad se diagnostica en función de cuatro indicadores relacionados con capacidades motoras y no-motoras, planteamos como objetivos específicos:

- Estudiar la evolución de estos indicadores diagnósticos a lo largo del tiempo.
- Investigar su relación con el resto de variables predictoras.
- Identificar qué proteínas prevalecen en las analíticas a lo largo del tiempo, y por lo tanto se podrían asociar a la evolución de la enfermedad.
- Identificar qué péptidos prevalecen en las analíticas a lo largo del tiempo, y por lo tanto se podrían asociar a la evolución de la enfermedad.
- Plantear diversos modelos de aprendizaje automático para predecir cada uno de los indicadores diagnósticos en función de las variables analíticas y las características de los pacientes.
- Comparar los modelos ajustados en términos de varias métricas de calidad/evaluación, e identificar el que proporciona mejores predicciones.
- Por último, extraer las conclusiones pertinentes sobre el estudio realizado.

5. Información disponible

La información proporcionada como objeto de estudio, procede del programa conocido como *Accelerating Medicines Partnership® (AMP®) program*, impulsado por una

asociación público-privada formada por los Institutos Nacionales de Salud (NIH), múltiples empresas biofarmacéuticas y de ciencias de la vida, así como organizaciones sin ánimo de lucro.

Este programa tiene como objetivo entre otros, mejorar el diseño de los ensayos clínicos y contribuir en la identificación de nuevas vías para el desarrollo terapéutico.

AMP® propuso una competición de Kaggle titulada "AMP®-Parkinson's Disease Progression Prediction" para el análisis de la base de datos vinculada a este proyecto. La competición contó con la participación de 1805 equipos de voluntarios formados por entre una y cinco personas, que contribuyeron con sus hallazgos a través de la plataforma colaborativa Kaggle.

Esta competición comenzó en febrero de 2023 con una fecha límite de finalización en mayo de 2023, cuyo ganador sería el equipo que obtuviese la mejor predicción sobre la evolución de la enfermedad a partir de la información disponible. El equipo ganador del reto fue Connecting Dotts. Este equipo propuso como solución al reto dos modelos, LGB (Light Gradient-Boosting Machine) y NN (Neural Networks), ambos entrenados con las mismas características. Cabe destacar que el equipo menciona que ignoraron por completo los resultados de las analíticas de updrs en su investigación y han basado toda su investigación en los cambios que se produce en la enfermedad durante el tiempo, ya que ninguno de sus enfoques puede beneficiarse de las características de las variables de updrs lo suficientemente como para distinguirlos de las variaciones aleatorias.

La base de datos contiene una caracterización molecular profunda y un perfil clínico longitudinal de pacientes con la enfermedad del Párkinson. Está construída a partir de un total de 248 pacientes con Párkinson que fueron seguidos a lo largo de varios años mientras eran sometidos a evaluaciones sobre la evolución de la enfermedad. Las visitas de seguimiento con registro de datos se programaron cada 3 meses desde la entrada en el ensayo, e incluían evaluaciones diagnósticas y analíticas de proteínas y

péptidos, proporcionadas con la abundancia cuantificada para cada uno. Los valores de la abundancia de proteínas y péptidos se derivaron de la espectrometría de masas en muestras de líquido cefalorraquídeo (LCR).

Toda la información está organizada en diferentes ficheros, relacionados todos ellos por el registro del paciente y la visita en cuestión. Los ficheros de datos disponibles son:

- **train_peptides.csv**, que contiene datos sobre la abundancia de péptidos en el líquido cefalorraquídeo de los pacientes analizados a lo largo del estudio.
- **train_proteins.csv**, que contiene frecuencias de expresión de proteínas asociadas a partir de la cantidad de péptidos que posee cada paciente a lo largo del estudio.
- **train_clinical_data.csv**, que contiene datos sobre la evaluación del estado del paciente a lo largo del tiempo, así como si tomaba medicación o no durante el estudio.

La extensión de cada conjunto de datos varía, así como el número de variables, si bien todos contienen información sobre todos los pacientes y todas las visitas de exploración y seguimiento a las que asistieron. Cabe mencionar que todos los ficheros comparten 3 variables comunes asociadas al identificador del paciente (*patient_id*), el identificador de visita (*visit_id*) y el mes de visita (*visit_month*), que facilitan la combinación de información.

Las variables están presentadas con su nomenclatura original y junto a estas, una descripción que explica el valor que aporta cada una. A continuación, se describen todas las variables disponibles en los diferentes ficheros.

Fichero train_clinical_data.csv

Contiene información sobre cada uno de los pacientes, con un registro para cada una de las visitas realizadas desde reclutamiento, y en cada visita, información sobre la evaluación del paciente y si el paciente toma medicación.

La dimensión de la base de datos es de 2615 filas y 8 columnas, con un total de 248 pacientes distintos.

Variables en la base de datos:

- visit_id: Código identificativo de visita del paciente. Este oscila entre el mes de reclutamiento que se identifica con un valor de 0 hasta un máximo de 108 meses en el caso de algunos pacientes.
- visit_month: Mes de visita (desde reclutamiento).
- patient_id: Código identificativo del paciente.
- updrs_1, updrs_2, updrs_3 y updrs_4: La puntuación del paciente para la parte N de la Escala Unificada de Evaluación del Párkinson (UPDRS). Cada puntuación se obtiene de la suma de las puntuaciones asignadas en los ítems correspondientes, y la puntuación de cada ítem se registra en una escala decimal de 0.0 a 4.0, donde un 0.0 se considera en estado normal y un 4.0 en estado grave. La puntuación máxima total de UPDRS (en los 4 indicadores conjuntamente) es de 159.0 puntos. Cada subsección abarca unas puntuaciones y una categoría de síntomas distintas, estas categorías son:
    - updrs_1: aspectos no-motores de las experiencias de la vida diaria. Esta sección abarca una puntuación de entre 0.0 y 16.0.
    - updrs_2: aspectos motores de las experiencias de la vida diaria. Esta sección abarca una puntuación de entre 0.0 y 52.0.

  - updrs_3: exploración motora. Esta sección abarca una puntuación de entre 0.0 y 68.0.
  - updrs_4: complicaciones motoras. Esta sección abarca una puntuación de entre 0.0 y 23.0.

  En los cuatro indicadores las puntuaciones más altas van asociadas a un mayor deterioro.

- upd23b_clinical_state_on_medication: Indica si el paciente estaba tomando medicamentos como Levodopa durante la evaluación UPDRS. Respuestas posibles On/Off.

Fichero train_proteins.csv

Contiene información sobre las frecuencias de expresión de proteínas agregadas a partir de los datos a nivel de péptidos, con un registro para cada una de las visitas realizadas desde el reclutamiento en cada paciente. Cada registro posee un identificador de proteína asociada y la frecuencia comentada recopilada en una variable denominada NPX (Normalized Protein eXpression), que es un factor que indica si la expresión de proteínas en una muestra biológica es alta o baja.

La dimensión de la base de datos es de 232.742 filas y 5 columnas, con un total de 248 pacientes distintos.

Variables en la base de datos:

- visit_id: Código identificativo de visita del paciente. Este oscila entre el mes de reclutamiento que se identifica con un valor de 0 hasta un máximo de 108 meses en el caso de algunos pacientes.
- visit_month: Mes de visita (relativo a la primera visita del paciente).
- patient_id: Código identificativo del paciente.

- UniProt: Código identificativo de UniProt para la proteína asociada. Este es un identificador único para cada proteína que identifica la composición de cada una. Tenemos tres tipos de código de proteína, estos empiezan únicamente por “O”, “P” y “Q”.
- NPX: Expresión de proteínas normalizada. La frecuencia de ocurrencia de la proteína en la muestra. Esta medida comprende un proceso de normalización que requiere una estandarización de los datos de expresión de proteínas utilizando un estándar de referencia. Se mide en una escala de decimales positivos que oscilan para la muestra considerada, entre un mínimo de 873.778 y un máximo de 122699.000. En este caso, un mayor NPX indica una respuesta inflamatoria y mayor degeneración neuronal en los enfermos.

Fichero train_peptides.csv

Contiene datos detallados sobre la presencia y la abundancia de péptidos en el líquido cefalorraquídeo de los pacientes enfermos de Parkinson, con un registro para cada paciente y cada una de las visitas realizadas desde el reclutamiento. Cada registro representa un péptido específico y su frecuencia en la muestra recogida. Cabe destacar que la combinación péptido-proteína es única para cada paciente, es decir, no se repiten.

La dimensión de esta base de datos es de 981.834 filas y 6 columnas, para un total de 248 pacientes distintos.

Variables en la base de datos:

- visit_id: Código identificativo de visita del paciente. Este oscila entre el mes de reclutamiento que se identifica con un valor de 0 hasta un máximo de 108 meses en el caso de algunos pacientes.
- visit_month: Mes de visita (relativo a la primera visita del paciente).

- patient_id: Código identificativo del paciente.
- UniProt: Código identificativo de UniProt para la proteína asociada. Este es un identificador único para cada proteína que identifica la composición de cada una. Tenemos tres tipos de código de proteína, que empiezan únicamente por “O”, “P” y “Q”.
- Peptide: Secuencia de aminoácidos incluida en el péptido. Cada secuencia está asociada a un Uniprot diferente. En total, tenemos 968 tipos de péptidos distintos.
- PeptideAbundance: Frecuencia del aminoácido en la muestra paciente. Esta es la medida que relaciona las cadenas cortas de aminoácidos conocidas como péptidos con la descomposición de proteínas en el sistema biológico. Esta abundancia se mide en una escala de decimales positivos que oscilan, para la muestra considerada, entre un mínimo de 10.9985 y un máximo de 178752000.0. Cuanto mayor sea la abundancia puede indicar la neurodegeneración característica de la enfermedad.

6. Metodología

A continuación se presentan de forma detallada los métodos escogidos para realizar el análisis completo de los datos, estructurado en análisis exploratorio y preprocesado, análisis descriptivo relacional, y modelización estadística para la predicción.

El análisis exploratorio y preprocesado de los datos es imprescindible para alcanzar mayor entendimiento sobre los datos, la información faltante y conseguir combinar toda la información disponible, de un modo estructurado, por paciente y visita.

El análisis descriptivo lo resolvemos básicamente a través de gráficos con los que descubrir la distribución de los datos en la muestra y las relaciones entre algunas de las variables predictoras y las respuestas. Cabe destacar que, no ha sido posible realizar

un análisis descriptivo de modo completo en la búsqueda de relaciones entre las distintas variables, dada la gran cantidad de predictores disponibles.

Tras el análisis descriptivo se plantean una serie de modelos de predicción para predecir las variables en las que se registra la evolución de la enfermedad en los pacientes (updrs[1-4]), en función de los resultados analíticos a nivel de proteína y péptido.

Detallamos a continuación cómo se desarrollan cada uno de estos procesos.

6.1 Análisis exploratorio y preprocesado de los datos

Para garantizar el uso eficaz de la base de datos, se exploran los diferentes ficheros en los que se identifican dimensiones, variables, respuestas distintas y datos faltantes. El procesado culmina en la combinación, en un único fichero, de toda la información disponible en los ficheros train_clinical_data.csv, train_proteins.csv y train_peptides.csv.

El análisis exploratorio se resuelve revisando, en cada fichero, cada una de las variables. Para abordar este proceso de forma exhaustiva, seguimos una serie de pasos que describimos a continuación.

Inicialmente, se examinan las variables categóricas con el fin de comprobar el número de respuestas distintas mediante la elaboración de tablas de frecuencias. Estas tablas se representan con gráficos que muestran y facilitan la visualización de la diversidad de respuestas. Estas tablas resultan útiles para identificar valores extraños que puedan requerir una atención distinta.

Seguidamente, se hace un análisis de las variables numéricas para comprobar su rango de variación y dimensiones. Este análisis sirve para identificar valores atípicos que puedan influir de alguna manera en los resultados de este estudio. Los resultados son clave para garantizar la validez e integridad del resto de la investigación, en caso de detectar valores inusuales.

Asimismo, cabe revisar la existencia de valores faltantes en todas las variables. Para ello, se elaboran unos mapas de calor (heatmap) con cada variable, que permiten observar cómo se distribuyen los valores faltantes en los datos. En caso de existir valores faltantes, será necesario comprender el porqué de esa falta de valores en las variables que contengan datos faltantes, y si es procedente y viable, se planteará la posible imputación de valores faltantes mediante la función *SimpleImputer()* de la librería Scikit-Learn. Esta función sustituye los valores faltantes por la media de los valores de su misma columna.

Una vez realizado el preprocesado, se procede con la combinación de toda la información en una única tabla, con los ficheros train_peptides.csv, train_proteins.csv y train_clinical_data.csv.

Para ello, tomamos como referencia el tratamiento de los datos dado en un estudio similar por Gusthema (2023). Este autor elabora una función en la que consigue un único registro por paciente y visita para los ficheros train_peptides.csv y train_proteins.csv. Seguidamente, pivota para añadir los valores de UniProt (en train_proteins.csv) y Peptide (en train_peptides.csv), relacionando ambas variables a través del indicador paciente-visita. Cabe destacar, que las variables UniProt y Peptide, originalmente eran categóricas, sin embargo, con esta combinación las transformamos en variables numéricas. Esto culmina en la combinación de los ficheros train_peptides.csv y train_proteins.csv en un único fichero con 1113 filas y 1202 columnas. Dado que las filas de train_clinical_data.csv poseen una distribución similar, se aplica la función *merge* y se almacena todo el contenido en un único fichero que incluye las 3 bases de datos y servirá para abordar los objetivos del análisis.

6.2 Análisis descriptivo de la evolución

Con el objetivo de investigar cómo evolucionan las distintas variables recopiladas conforme avanza el tiempo (plasmado en el ordinal de la visita realizada) se representan las distribuciones de las variables numéricas y de las variables categóricas

en diferentes instantes de tiempo a lo largo del periodo de observación de los pacientes, con uno u otro tipo de gráfico según sus características. Describimos a continuación los gráficos utilizados.

Para capturar gráficamente la tendencia y dispersión de los indicadores diagnósticos a lo largo del tiempo, usamos gráficos de puntos que representan la media de las puntuaciones por visita y la media más una desviación típica a ambos lados. Al unir las medias con líneas, y de forma similar los extremos dados por el rango de variación con más/menos una desviación típica, tenemos una visualización de la tendencia y dispersión a lo largo del tiempo. Diferenciamos además los puntos y líneas con colores, en función de si el paciente toma o no medicación. Del mismo modo, elaboramos este mismo gráfico de puntos y líneas pero representando la suma total de las puntuaciones, diferenciando de nuevo a los pacientes que tomaban medicación de los que no la tomaban.

Realizar todos los gráficos individualizados por paciente de los cuatro indicadores diagnósticos es poco ilustrativo, pero sí se ha escogido algún paciente al azar para mostrar cómo se interrelacionan sus valores de updrs a lo largo del tiempo. De ese modo, elaboramos gráficos de puntos y líneas para cuatro pacientes escogidos al azar, con el fin de representar los valores diagnósticos en cada uno de los meses en que se realizó visita.

Para estudiar la evolución de las variables analíticas PeptideAbundance y NPX, se representan sus valores en relación a la variable Peptide para PeptideAbundance, y UniProt para NPX. Las visualizaciones elaboradas para observar cómo se distribuyen a lo largo del tiempo consisten en gráficos de cajas y bigotes, que representan la abundancia de los péptidos o proteínas para cuatro péptidos y cuatro proteínas escogidas al azar. Representamos las abundancias de cada péptido y cada proteína a lo largo del tiempo, y así obtenemos unos gráficos de la distribución por mes.

Este tipo de gráfico de cajas y bigotes también ha sido utilizado para representar la distribución inicial de la suma de las puntuaciones de todos los pacientes en el mes de reclutamiento, en función de si el paciente tomaba o no medicación.

Para estudiar la relación entre las variables diagnósticas updrs y las variables analíticas Peptide y UniProt, representamos las correlaciones de Spearman con mapas de color (heatmap), donde los colores cálidos identifican correlaciones positivas y los colores fríos las correlaciones negativas, más fuertes a mayor intensidad de color.

En síntesis, el análisis exploratorio nos permite comprender cómo se interrelacionan entre sí las distintas variables y explorar la muestra disponible, proporcionando así un fundamento sólido para el análisis y la interpretación de los resultados posteriores.

6.3 Análisis estadístico. Modelización.

Con el fin de llevar a cabo un análisis detallado de los datos para predecir la evolución de la enfermedad en función de los parámetros analíticos y físicos medidos en los pacientes, hemos seleccionado una serie de algoritmos de Machine Learning de aprendizaje supervisado. Tras ajustarlos, comparamos los resultados según ciertas métricas que presentamos a continuación, y concluimos sobre el mejor modelo de predicción.

Las técnicas de modelización propuestas son:

1. Regresión lineal
2. Lasso
3. Ridge
4. K-Nearest Neighbors
5. Árboles de decisión

6. Bosques aleatorios
7. CatBoosting Regressor

Utilizamos como referente el estudio que realiza Lokeshparab (2023).

6.3.1 Regresión lineal

La regresión lineal (*linear regression*) es un técnica de aprendizaje supervisado que se utiliza para predecir y crear un modelo que describa, a través de un modelo lineal, la relación entre una variable dependiente numérica, denominada *y,* y una o más variables independientes, denominadas *x*. Dependiendo de cuántas variables independientes tengamos, hablamos de regresión lineal simple (una) o múltiple (más de una).

Cuanto mayor sea la relación lineal entre las variables independientes con la variable dependiente, mayor precisión habrá en la predicción (mayor será la proporción de varianza de la variable dependiente explicada por las variables independientes).

La regresión lineal se considera una técnica sencilla de implementar y con capacidad para proporcionar resultados confiables en presencia de asociaciones de tipo lineal.

6.3.2 Lasso y Ridge

Los métodos de regularización Lasso y Ridge son técnicas que nos permiten restringir el proceso de estimación en el modelo lineal, para controlar, principalmente, problemas de multicolinealidad o correlación entre las predictoras, y así evitar un posible sobreajuste del modelo (*overfitting*).

Ambos métodos utilizan un término de penalización en la función objetivo (error cuadrático medio), con el fin de reducir el tamaño de los coeficientes de variables que solapan información con otras y así reducir también la complejidad del modelo. De ese

modo, se reduce el número de predictores 'influyentes' y se pueden considerar como procedimientos de selección automática.

La regresión de Lasso aplica la penalización basada en la norma L1 reflejada en la siguiente función objetivo a minimizar:

$$\min_{\beta_0,\beta} \frac{1}{N}\|\mathbf{y} - \beta_0\mathbf{1} - \mathbf{X}\beta_j\|_2^2 + \lambda\|\beta\|_1$$

Este tipo de penalización tiene el efecto de forzar algunas de las estimaciones de coeficientes a ser iguales a cero cuando el parámetro **λ** de ajuste es suficientemente grande. Por lo tanto, la regresión de Lasso realiza selección de variables.

Por otro lado, la regresión de Ridge aplica el tipo de penalización basado en la norma L2 a la función objetivo a minimizar:

$$\min_{\beta_0,\beta} \frac{1}{N}\|\mathbf{y} - \beta_0\mathbf{1} - \mathbf{X}\beta_j\|_2^2 + \lambda|\beta|^2$$

Este tipo de penalización, usa el cuadrado del coeficiente en lugar del valor absoluto. No tiende a anular los coeficientes, sino reducirlos sin llegar a cero, a base de incrementar el valor de **λ**.

En resumen, ambos son métodos de regularización sirven para reducir la dimensión del modelo (el número o peso de las variables predictivas) y así el sobreajuste de los modelos de regresión con muchos predictores correlacionados, sin perder capacidad predictiva.

6.3.3 K-Nearest Neighbors

El algoritmo de los K-Vecinos más cercanos (K-Nearest neighbors) es una técnica de aprendizaje supervisado utilizada tanto para la clasificación como para el modelado predictivo.

Cuando la variable objetivo es de tipo numérico, y el objetivo es la predicción, esta se consigue con el promedio de las respuestas observadas en los K individuos más cercanos. En este caso, nos enfocaremos en KNN como algoritmo de predicción, pues también sirve como algoritmo de clasificación.

Los parámetros comunes en este algoritmo son K y la distancia. Las predicciones cambian en función de los parámetros escogidos. Para escoger el valor de K, se prueba con distintos valores y se elige el que mejores resultados proporciona en las muestras de entrenamiento/test. Para las distancias, se utilizan diversas distancias *Euclídea, Manhattan, Máxima, Mínima, Mahalanobis… ,* e igualmente se elige la que proporciona mejores errores de predicción.

Una vez identificados los K vecinos de cada punto, en lugar de considerar su clase cómo se produce en la clasificación, se considera el valor que toma la etiqueta para cada uno de ellos y se devolverá como predicción el valor medio de dichos valores.

KNN se caracteriza por su efectividad y simplicidad, y al ser un algoritmo no paramétrico, no hace suposiciones explícitas sobre la distribución de datos adyacentes.

6.3.4 Árboles de decisión

Un árbol de decisión *(decision tree)* es un algoritmo de aprendizaje supervisado utilizado para la clasificación y el modelado predictivo.

Los elementos principales de un árbol de clasificación son la regla de división, el criterio de parada y la asignación de clase.

Su estructura es semejante a un diagrama de flujo gráfico, con una estructura similar a la de un árbol. Se trata de dividir un nodo (nodo padre) en dos nodos (nodos hijos), pues el objetivo es disminuir el grado de impureza del nodo padre, esto es, la heterogeneidad de los sujetos que lo integran.

Cada nodo interno representa una característica respecto de la que se particionan los datos, y que a su vez se divide en dos nuevas ramas, si encuentra que los sujetos son distintos entre sí en los dos nodos resultantes, pero afines dentro de cada nodo. El proceso se itera hasta que los datos llegan a un nodo final o terminal, que no se ramifica más, bien porque es inviable reducir la impureza, bien porque cumple con el criterio de parada especificado.

Las medidas de impureza más comunes son:

- El índice de Gini: se distingue por evaluar qué tan homogéneos son los datos en un nodo determinado. Cuánto mayor sea el índice de Gini, mayor impureza del modelo.
- Entropía cruzada: sirve como medida de discrepancia entre los datos reales y los predichos por el modelo. Cuánta mayor entropía, mayor impureza.

Es común establecer un criterio de parada en función de la profundidad deseada para el árbol (número de niveles en que se distribuyen los nodos) o el número mínimo de sujetos con que ha de contar un nodo. Los árboles de decisión son modelos de aprendizaje automático muy populares porque pueden manejar tanto datos numéricos como categóricos, así como conjuntos de datos complejos, con relativa sencillez. Además son fáciles de interpretar. El único inconveniente que presentan es que no son muy robustos, pues una pequeña variación de los datos (como la selección de la muestra de entrenamiento) puede causar un gran cambio en el árbol final.

6.3.5 Bosques aleatorios

El algoritmo de bosques aleatorios *(random forest algorithm)* consiste en ajustar muchos árboles de decisión y combinarlos para mejorar la estabilidad y precisión de la clasificación y el modelado predictivo.

Los bosques aleatorios consisten en una colección de árboles que usan, cada vez, un conjunto de entrenamiento distinto. Cada conjunto de entrenamiento se selecciona a través del *bootstrapping* o remuestreo repetido sobre el conjunto original de entrenamiento.

Si utilizamos todos los predictores disponibles, el bosque aleatorio se conoce como *Bagging.*

Aunque pueden llegar a ser complejos y requerir mucho tiempo, los bosques aleatorios corrigen el problema de “sobreajuste” *(overfitting)* y aumentan la capacidad de generalización de los modelos al promediar las predicciones de múltiples árboles.

6.3.6 CatBoosting Regressor

CatBoosting Regressor es un algoritmo de aprendizaje automático basado en árboles de decisión. Este algoritmo utiliza una técnica de optimización de gradientes conocida como *boosting*, que guía la implementación de los sucesivos árboles de decisión, para combinarlos y conseguir predicciones más precisas y estables.

La idea principal del *boosting* es combinar secuencialmente muchos modelos que funcionan ligeramente mejor que la elección al azar y así, a través de una búsqueda guiada, crear un modelo predictivo competitivo y sólido. Debido a que el *boosting* es secuencial, los árboles ajustados conocen los errores de los árboles anteriores y, por lo tanto, dichos errores se van reduciendo. Además, destaca por su capacidad para lidiar con valores faltantes de manera robusta y eficiente para encontrar una solución óptima y evitar el sobreajuste.

6.3.7 Métricas de evaluación

En este estudio hemos establecido unas métricas de evaluación sobre cada modelo con el fin de evaluar la calidad de cada uno. Las métricas escogidas son el coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE). Consideramos ambas, respuestas numéricas que compararemos al ajustar las técnicas.

El coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE) sirven como medidas de calidad del ajuste. El coeficiente de determinación (R^2) indica cuánta variación en “y” puede ser explicada por “x” (valores altos del coeficiente de determinación indican una asociación lineal más fuerte entre “x” e “y”). La raíz del error cuadrático medio (RMSE) evalúa el error del modelo, pues refleja la magnitud del error en la predicción del modelo mediante el promedio de las diferencias al cuadrado entre los valores predichos y reales.

El coeficiente de determinación (R^2) oscila entre 0 y 1. Esto indicará cuánta variación en las variables diagnósticas updrs puede ser explicada por el resto de variables predictoras. Cuanto mayor sea el valor del coeficiente de determinación, más fuerte es la relación entre las variables.

La raíz del error cuadrático medio (RMSE) se calcula mediante la raíz cuadrada de la media de las diferencias al cuadrado entre las predicciones y los valores observados. Este cálculo mostrará un resultado numérico que cuanto menor sea, mejor será el modelo y sus predicciones.

6.4 Software y hardware

Este proyecto ha sido desarrollado en el lenguaje de programación Python, que es ampliamente utilizado en la actualidad en el desarrollo de software, aplicaciones web, Ciencia de Datos y Machine Learning.

Todo nuestro código ha sido programado en cuadernos Jupyter de Google Colab, que a diferencia de otros servidores, ofrece la posibilidad de compartir tu proyecto con otro

usuario de Gmail para poder trabajar de forma colaborativa, y no requiere de instalación de software.

Las librerías de Python necesarias para desarrollar este estudio son:

- Pandas, nos ofrece una amplia variedad de funcionalidades para cargar, seleccionar, transformar, manipular y analizar tablas de datos.
- Matplotlib, nos permite crear y personalizar todos los aspectos relacionados con la visualización de gráficos.
- Seaborn, nos proporciona gráficos que complementan a Matplotlib. Está especializada en la visualización de datos estadísticos.
- NumPy, incluye una amplia cantidad de funciones para realizar operaciones matemáticas.
- Scikit-learn (sklearn), nos proporciona una gran variedad de algoritmos que implementan modelos de Machine Learning.

Para almacenar todos los datos, debido al volumen de los ficheros de datos existentes en nuestro proyecto, ha sido necesario descargar e instalar la aplicación Github Desktop. Esta aplicación funciona como su versión online con la diferencia de que se pueden subir ficheros de más de 50 GB como es nuestro caso.

Además de Google Colab y Github, se ha hecho uso de Documentos de Google para desarrollar el informe del proyecto.

El hardware utilizado en este estudio es un portátil Lenovo IdeaPad con un procesador Intel Core i7.

7. Resultados

En esta sección, se detallan los resultados obtenidos de forma estructurada, con el objetivo de proporcionar unas conclusiones sólidas y rigurosas sobre el estudio.

- En primer lugar presentamos el análisis exploratorio inicial con el que abordamos el preprocesado de los datos, la identificación de valores faltantes, y la combinación de las tres bases de datos existentes en una única.
- A continuación presentamos el análisis descriptivo gráfico realizado, útil para comprender mejor la relación entre las variables de interés y a lo largo del tiempo.
- Después se presentan los resultados del ajuste de los modelos propuestos, junto con la evaluación de la calidad y una tabla comparativa de los resultados de predicción conseguidos. Dicha tabla, la presentamos para cada una de las medidas de updrs con el fin de observar qué modelo se ajusta mejor para cada una.
- Finalmente, exponemos un resumen de los resultados para poder concluir el Trabajo de Fin de Grado dando respuesta a las preguntas planteadas en los objetivos propuestos.

7.1 Análisis exploratorio y preprocesado

La eficiencia de la base de datos requería una modificación para facilitar el manejo de los datos en la posterior modelización. Una vez hemos obtenido un único fichero que contenga toda la información, realizamos el análisis exploratorio de los datos. Cabe destacar que la combinación de los datos en un fichero solo servirá para la modelización, pues el análisis exploratorio lo realizamos con la información de las bases de datos originales.

Posteriormente, se ha realizado un análisis exhaustivo de las bases de datos con el fin de encontrar valores faltantes. El único fichero que presenta valores faltantes es train_clinical_data.csv, para updrs_3, updrs_4 y upd23b_clinical_state_on_medication. En total, tenemos 2365 valores faltantes en esta base de datos y 1038 de estos valores pertenecen a las variables updrs_3 y updrs_4. La causa de la falta de valores en udprs_3 y updrs_4 es la ausencia de puntuación para algunos pacientes.

Sin embargo, previo a la imputación elaboramos un heatmap o mapa de intensidad de calor, que se muestra en la Figura 1, con el objetivo de observar los valores faltantes existentes en dichas variables. El eje X representa a la variable patient_id cuyo valor es faltante para las variables de updrs representadas en el eje Y. Las líneas negras resaltan la presencia de valores, y por el contrario, los espacios blancos señalan valores faltantes.

Figura 1: Gráfico de intensidad de calor (heatmap) para identificar los valores faltantes (en blanco) en los indicadores diagnósticos updrs_1, updrs_2, updrs_3 y updrs_4. Fuente: Elaboración propia.

Al observar la Figura 1 cabe mencionar que tal y como muestran las lineas de color negro, para updrs_1 y updrs_2 no hay valores faltantes en ningún paciente. Sin embargo, para updrs_3 existen pocos valores faltantes en algunos pacientes, y en updrs_4 es donde encontramos la gran mayoría de valores faltantes. Una vez observados los valores faltantes, cabe realizar una imputación para ambas variables con el fin de realizar una correcta modelización posteriormente.

De ese modo, es necesario buscar un modo de sustituir dichos valores sin rellenarlos con ceros, ya que un cero para las variables updrs_3 y updrs_4, hace referencia a ausencia o mínima presencia de síntomas asociados a la enfermedad del Párkinson. Con el propósito de poder realizar la posterior modelización decidimos imputar dichos valores mediante la función *SimpleImputer()* de la librería Scikit-Learn y sustituimos los valores faltantes por la media de los valores de su misma columna.

Todas las variables numéricas están estandarizadas en una escala común de números decimales que no requiere ninguna modificación.

En el análisis descriptivo hemos estudiado las variables de los ficheros train_peptides.csv y train_proteins.csv. En la variable Peptide, descubrimos dos modificaciones postraduccionales que rompen la cadena de aminoácidos que representa esta variable. Estas modificaciones son “C(UniMod_4)” y “M(UniMod_35)”, ambas son modificaciones químicas que surgen durante la preparación de la muestra para estabilizar péptidos, es decir, forman parte de la composición de la cadena de aminoácidos. Por tanto, no cabe imputar ningún valor que sustituya esos caracteres, ya que su naturaleza no altera a la base de datos y no aporta una manipulación mejor. Posteriormente, analizamos UniProt y no consideramos necesaria ninguna modificación en la composición de la variable UniProt, ya que no aporta mayor eficiencia un cambio en su nomenclatura. Sin embargo, consideramos necesario para facilitar el manejo de los datos transformar UniProt y Peptide en variables numéricas como hemos mencionado anteriormente, con el fin de reducir nuestros datos en un único fichero fácil de manejar.

7.2 Evolución de la enfermedad: variables diagnósticas

En primer lugar, describimos cómo evolucionan las 4 variables numéricas de updrs a lo largo del tiempo para los distintos pacientes. Tal y como se expone en un artículo de Holden et al. (2017), cualquier hallazgo en cada medida de updrs depende de si el paciente estaba tomando medicación o no. De ese modo, era necesario subdividir la

muestra de nuestros pacientes en dos grupos diferentes, los que tomaban medicación y los que no la tomaban, considerando cualquier valor nulo que exista en nuestra base de datos como un estado desconocido. Para ello, se han elaborado los gráficos de la Figura 2 donde se muestra la distribución de las medias y las desviaciones típicas de cada medida en un gráfico distinto por medida, a lo largo de los meses para todos los registros de pacientes, diferenciados por los que que tomaban y no tomaban medicación.

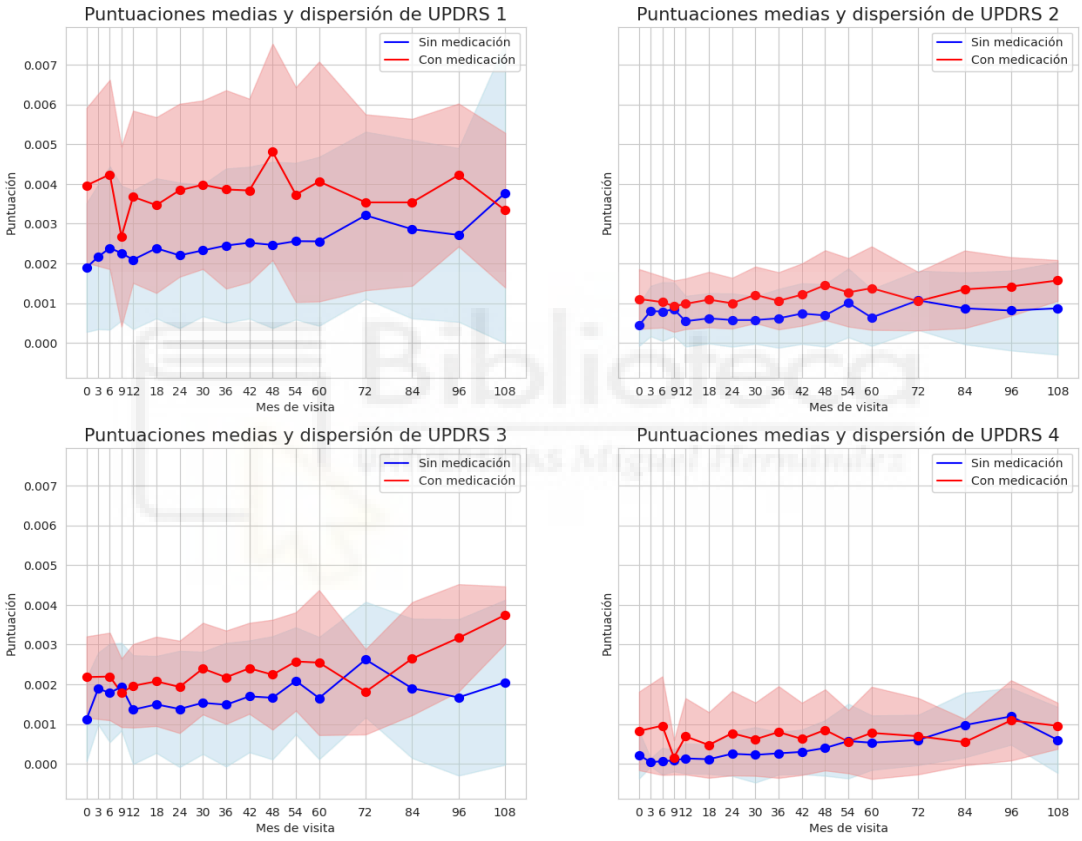


Figura 2: Gráfico lineal sobre la evolución de los indicadores updrs a lo largo de los meses, para los pacientes con y sin medicación. Los puntos unidos por líneas representan las medias en cada visita y las líneas punteadas que delimitan las sombras se calculan con más-menos una desviación típica. Fuente: Elaboración propia.

De la interpretación de la Figura 2 podemos destacar los siguientes aspectos:

- La tendencia evolutiva para cada updrs, tanto para pacientes que toman o no toman medicación, oscila y no se mantiene estable a lo largo del tiempo.
- En el instante inicial, las medias de los pacientes que reciben medicación manifiestan un estado más avanzado de la enfermedad en comparación a aquellos que no toman medicación.
- Los pacientes que toman medicación presentan por lo general puntuaciones más altas que los pacientes sin medicación, es decir, sus síntomas de la enfermedad son más graves.
- A largo plazo, en el mes 72 genera algo de mejora para los pacientes medicados, y equiparación con los no medicados, si bien en adelante evolucionan de modo diferente los distintos indicadores.
- Las líneas de tendencia de las medidas updrs_2 y updrs_4 se mantienen relativamente planas a lo largo del tiempo, y su dispersión es pequeña en comparación con los otros indicadores. En el mes 72, todos los pacientes parecen equiparar el estado de deterioro de la enfermedad, aunque posteriormente se traduzca de nuevo en un ligero aumento para los que toman medicación y una estabilización en los que no toman.
- Por el contrario, los puntos en las líneas de tendencia de las variables updrs_1 y updrs_3 presentan una alta dispersión a lo largo del tiempo. Del mismo modo que en el resto de indicadores, en el mes 72 las puntuaciones se equiparan en una ligera disminución que posteriormente supone de nuevo un aumento en los pacientes con medicación y una disminución que luego crece bruscamente en el mes 108 para los que no toman medicación.

El artículo de Holden et al. (2017), describe que en estudios sobre la escala unificada de la progresión del Párkinson, en el valor máximo de cada puntuación se observaba una progresión lineal notable a lo largo del tiempo . Por tanto, hemos decidido elaborar la Figura 3 que representa esta suma total, diferenciando de nuevo a los pacientes que

tomaban medicación de los que no la tomaban, con el fin de comprobar esta teoría sobre nuestros datos.

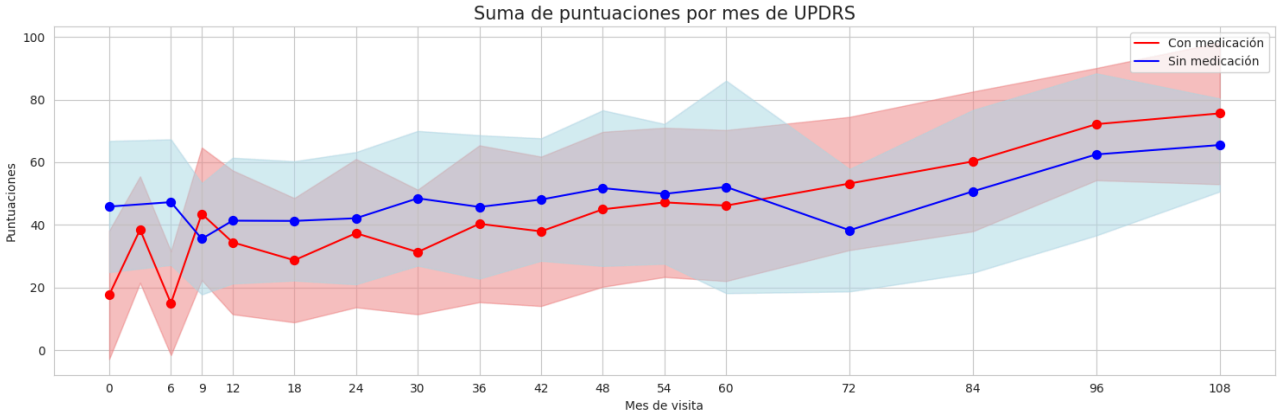


Figura 3: Gráfico lineal con la suma de las puntuaciones de los indicadores updrs a lo largo del tiempo, diferenciando si el paciente toma o no medicación. Los puntos unidos por líneas representan las medias en cada visita y las líneas punteadas que delimitan las sombras se calculan con más-menos una desviación típica. Fuente: Elaboración propia.

Como podemos observar, existe una ligera tendencia creciente de las puntuaciones de updrs en ambos casos a lo largo del tiempo. Para los pacientes que toman medicación ligeramente peor (con puntuación mayor en la suma de los updrs) que los que no la toman, hasta el mes 72. A partir de ese mes se aprecia una mejora sustancial de los pacientes que toman medicación frente a los que no toman, manteniendo una evolución de deterioro por debajo de los pacientes que no toman medicación. Esta puntualización confirma la teoría anterior y demuestra que la tendencia de los datos denota un avance de la enfermedad que a largo plazo se traduce en una ligera mejora para los pacientes que tomaban medicación frente a los que no tomaban.

Para aportar mayor información sobre las variables updrs_1, updrs_2, updrs_3 y updrs_4, se han elaborado un conjunto de cuatro gráficos lineales diferenciados por las puntuaciones obtenidas de un paciente escogido al azar en cada uno. A la vista del

gráfico, apreciamos la diversidad existente en la evolución de la enfermedad en distintos pacientes para cada uno de los índices de deterioro.

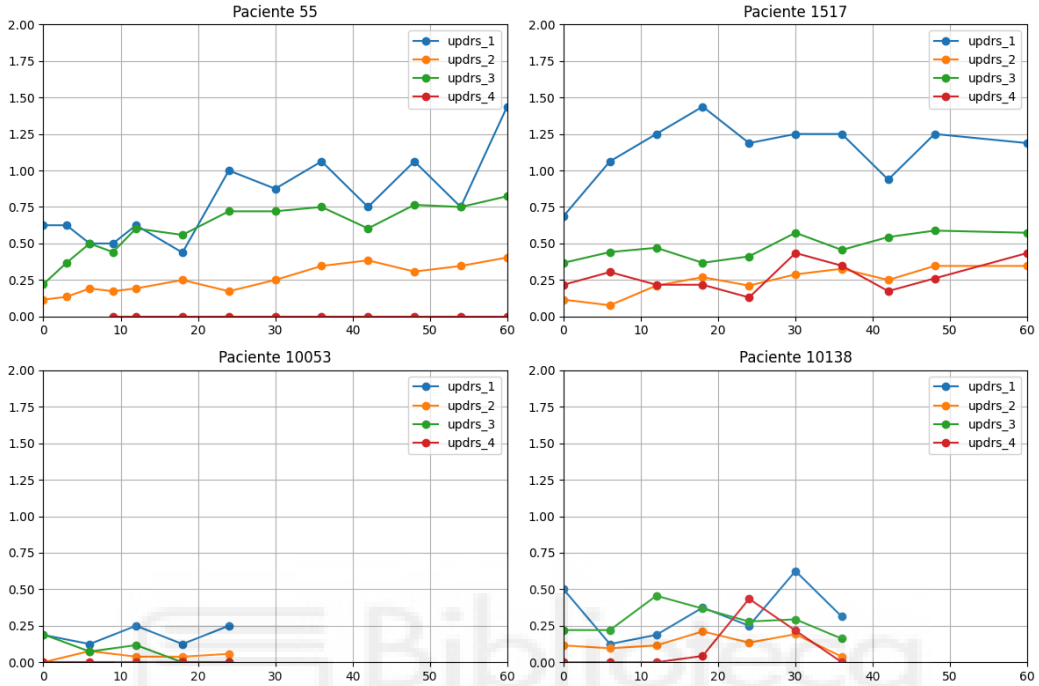


Figura 4: Gráfico lineal sobre la evolución de cada medida a lo largo de los meses para cuatro pacientes escogidos al azar. Los puntos unidos por líneas representan la distribución de cada medida diferenciadas por colores. Fuente: Elaboración propia.

Del conjunto de gráficos anteriores podemos destacar los siguientes aspectos:

- Si bien algunos evolucionan progresivamente empeorando, hay otros que muestran más picos y cambios de tendencia en los marcadores de deterioro.
- Las puntuaciones de updrs_1 (aspectos no-motores de las experiencias de la vida diaria) son las que tienen valores mayores que los otros indicadores (manifiestan mayor deterioro) y también oscilan de una forma más brusca a lo largo del tiempo. Esto justifica la alta variabilidad en las puntuaciones medias representadas en la Figura 2. Tanto updrs_2 como updrs_4, son las fases cuya línea de tendencia se mantiene más estable a lo largo del tiempo.

- Las puntuaciones de updrs_3, parecen tener una línea de tendencia estable en los pacientes observados con respecto a updrs_1.
- En definitiva, las puntuaciones de cada paciente se comportan de manera diferente a lo largo del tiempo pero con un patrón más o menos común.

Asimismo, en las visualizaciones anteriores apreciamos que en el mes de reclutamiento, los pacientes parten, en promedio, de un estado de la enfermedad peor en el grupo de los que toman medicación que en el de los que no toman. Por ello, elaboramos un gráfico de cajas y bigotes (Figura 5) que diferencie a los pacientes que tomaban medicación de aquellos que no tomaban medicación en el mes 0 y representamos la suma de las puntuaciones de updrs en ambos grupos.

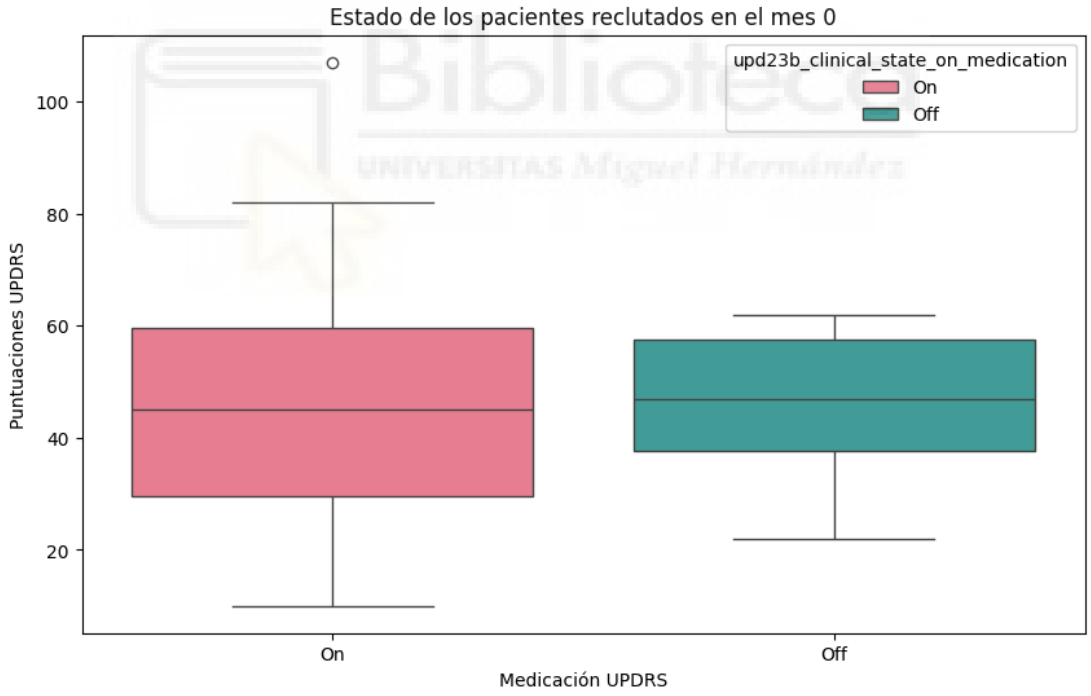


Figura 5: Gráfico de cajas y bigotes para la suma de las puntuaciones updrs en el reclutamiento, diferenciando a los pacientes que tomaban medicación (en rosa) de los que no (en azul). Fuente: Elaboración propia.

De la Figura 5 podemos destacar que prácticamente el 50% de los pacientes que toman y no toman medicación parten de un diagnóstico similar. Las diferencias se dan en la variabilidad, pues entre los pacientes que toman medicación hay más variabilidad, y encontramos un mayor porcentaje de pacientes con mejor situación (indicador muy bajo) que en el grupo sin medicación (el percentil 25 es inferior en el grupo con medicación que en él sin medicación). También en el 25% con peor estado, hay pacientes con medicación con un indicador global updrs mucho mayor que en el grupo de los que no tomaban medicación.

7.3 Descripción de las variables analíticas

En este apartado presentamos el análisis de la distribución de la abundancia de péptidos y proteínas. Estudiamos mediante el uso de gráficos de cajas y bigotes, la interacción entre las variables diagnósticas y las variables analíticas PeptideAbundance y NPX por separado, con el objetivo de identificar cómo se distribuyen a lo largo del tiempo y su efecto en la progresión de la enfermedad.

En primer lugar, analizamos la evolución de PeptideAbundance para cuatro péptidos diferentes a lo largo de los meses de estudio. Para ello, elaboramos cuatro gráficos de cajas y bigotes que muestran la distribución a lo largo del tiempo de la abundancia de los péptidos: “GAYPLSIEPIGVR”, “EPGLC(UniMod_4)TWQSLR”, “QPSSAFAAFVK” y “GLVSWGNIPC(UniMod_4)GSK”.

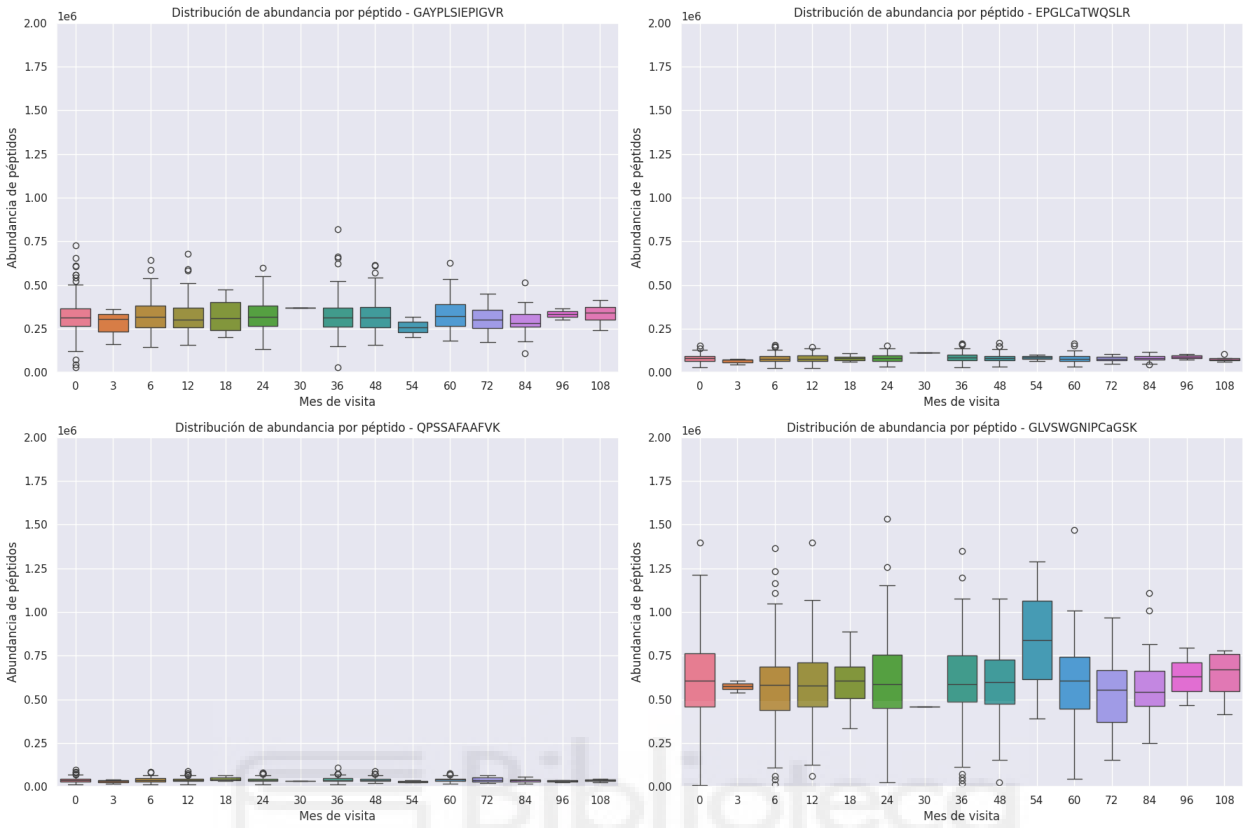


Figura 6: Gráfico de cajas y bigotes sobre la distribución de la abundancia de péptidos de cuatro péptidos diferentes a lo largo del tiempo. Cada caja representa la variabilidad de la abundancia del péptido en el mes correspondiente. Fuente: Elaboración propia.

De la representación anterior, podemos destacar los siguientes aspectos:

- El péptido “GAYPLSIEPIGVR” muestra variaciones significativas a lo largo del tiempo. En general, la abundancia del péptido fluctúa en los diferentes meses. Esto es, que hay meses en los que la abundancia es mucho mayor que en otros péptidos, por lo que podría indicar que hay péptidos cuya abundancia tiene mayor impacto en la progresión de la enfermedad.
- El péptido “EPGLC(UniMod_4)TWQSLR” muestra una mayor concentración de abundancia a lo largo de los meses que se distribuye de forma similar a lo largo del tiempo, pues no hay picos que denoten una fluctuación notoria entre meses.

- El péptido “QPSSAFAAFVK” presenta valores pequeños en comparación con el resto, sin embargo, también presenta valores atípicos y ligeras diferencias en la abundancia de cada mes.
- El péptido “GLVSWGNIPC(UniMod_4)GSK” presenta una distribución de la abundancia que varía significativamente a lo largo del tiempo. La mayoría de los meses presentan grandes fluctuaciones y la variabilidad de la mayoría de meses es alta.
- Cabe destacar que en el mes 30, la variabilidad de los cuatro péptidos es mínima porque solo hay un paciente, el número 50611, estudiado durante ese mes.
- En general, cada péptido tiene un patrón de distribución diferente. Sin embargo, hay péptidos que destacan por indicar mayor neurodegeneración en los pacientes. Esto se observa en la variabilidad de la abundancia en cada mes.

Posteriormente, analizamos la distribución de NPX sobre diferentes UniProt. Para ello, hemos realizado el mismo tipo de representación que en el caso anterior. Representamos en la Figura 7, cuatro gráficos de cajas y bigotes que miden la frecuencia de ocurrencia a lo largo del tiempo para cuatro proteínas escogidas. Las cuatro proteínas seleccionadas son “P02768”, “P02787”, “Q13451” y “P02766”.

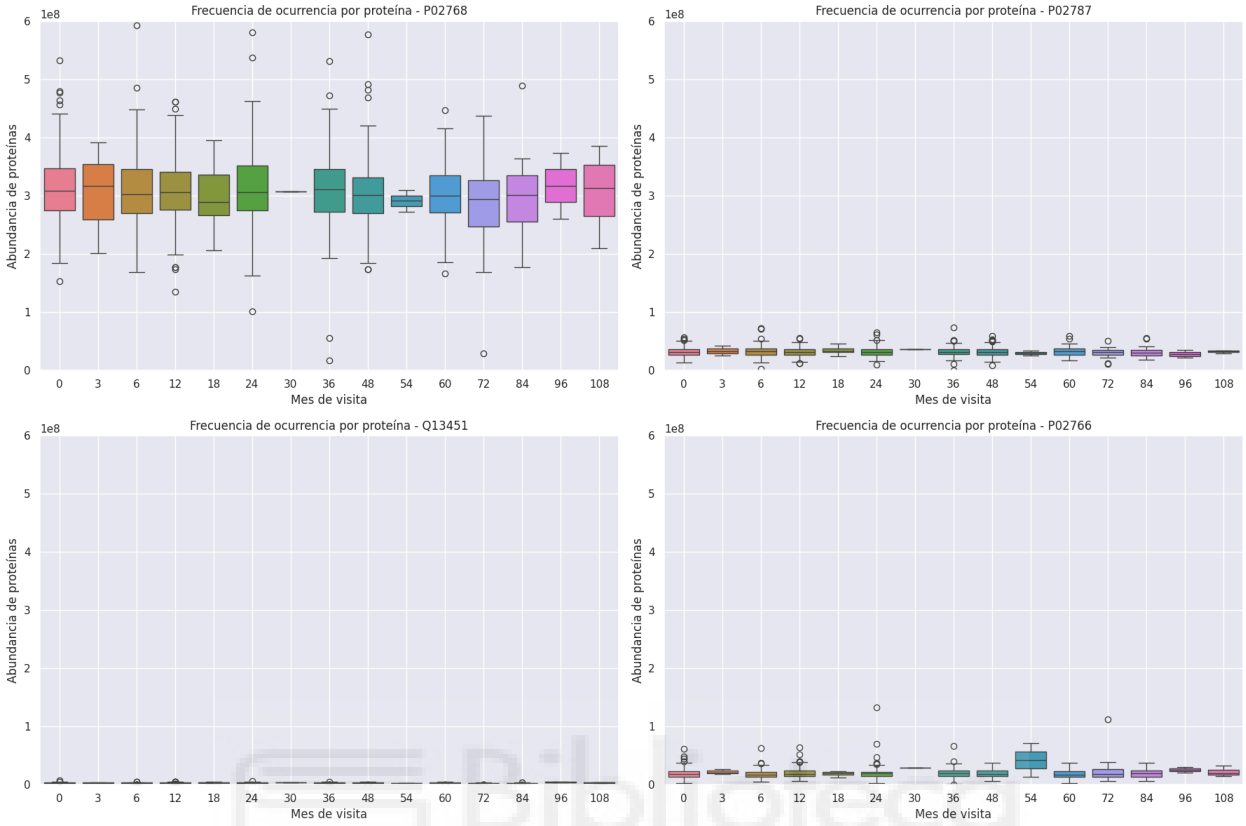


Figura 7: Gráfico de cajas y bigotes sobre la distribución de la abundancia de NPX para cuatro proteínas escogidas al azar. Cada caja representa la variabilidad de la abundancia de proteína en cada mes. Fuente: Elaboración propia.

De la representación anterior, podemos destacar los siguientes elementos:

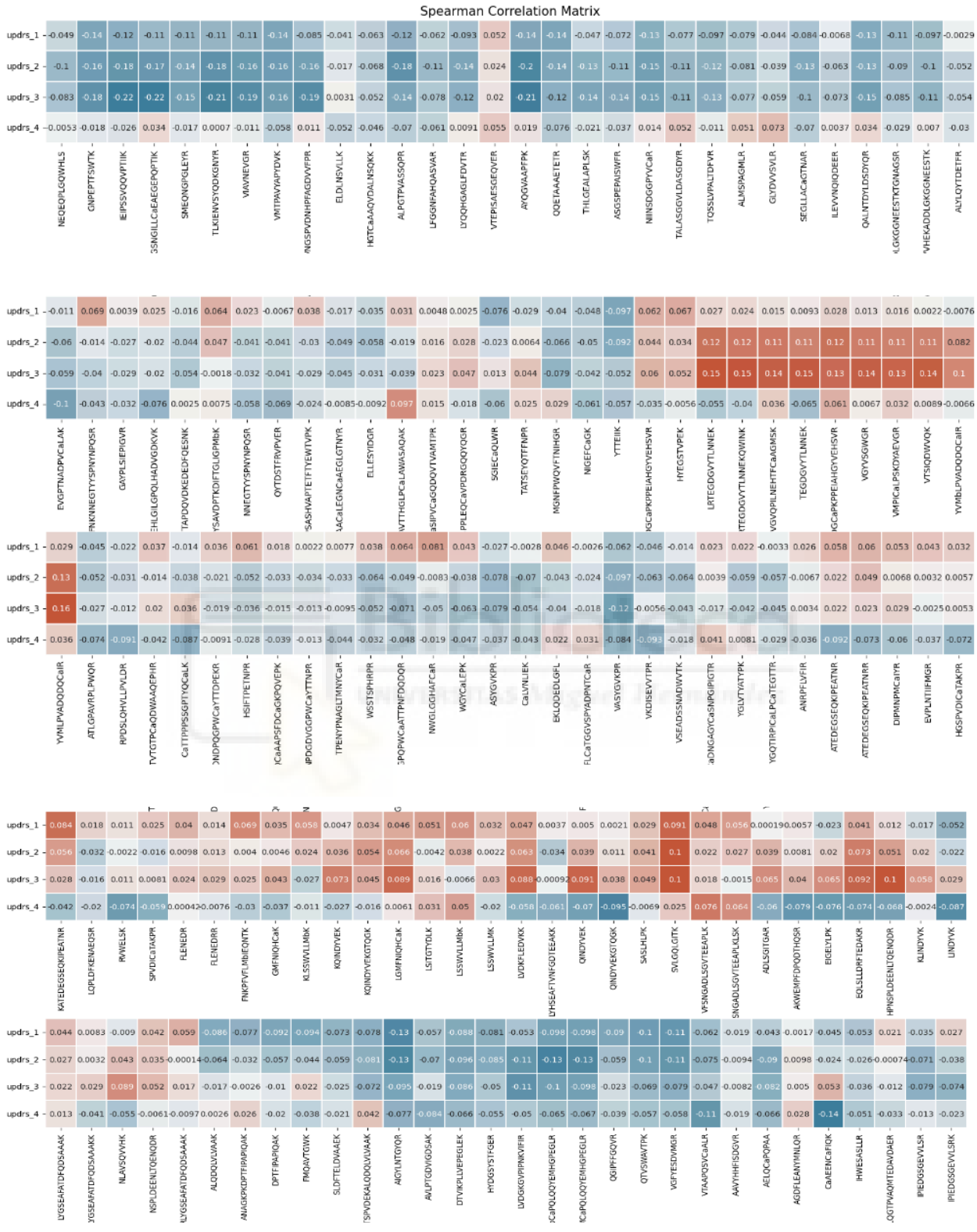
- La variabilidad de las proteínas refleja diferentes respuestas en la progresión de la enfermedad.
- La proteína “P02768” presenta valores atípicos y una distribución de la abundancia que varía significativamente a lo largo del tiempo. En general, la abundancia de la proteína fluctúa en los diferentes meses. En la mayoría de los meses la abundancia es mucho mayor que en otras proteínas, por lo que podría indicar que hay proteínas cuya progresión en la enfermedad tienen más impacto.
- La proteína “Q13451” muestra una mayor concentración de abundancia a lo largo de los meses que se distribuye de forma similar a lo largo del tiempo.

Aunque existan valores atípicos, no hay picos que denoten una fluctuación notoria entre meses.

- La proteína “P02787” presenta valores atípicos y ligeras diferencias en la abundancia de cada mes. Esta proteína sigue un patrón de tendencia similar a la proteína “P02766”. Ambas presentan valores atípicos y una concentración en la abundancia de NPX en cada mes que no varía significativamente a lo largo del tiempo.
- Cabe destacar que al igual que en el caso de los péptidos, en el mes 30, la variabilidad de las cuatro proteínas es mínima porque solo hay un paciente, el número 50611, estudiado durante ese mes.
- En general, cada proteína tiene un patrón de distribución diferente. Sin embargo, hay proteínas que destacan por indicar mayor neurodegeneración en los pacientes. Esto se observa en la variabilidad de la abundancia de NPX en cada mes.

7.4 Asociación entre variables analíticas y variables diagnósticas

La complejidad del estudio radica en la gran cantidad de datos disponibles, por lo que resulta difícil encontrar patrones y tendencias con claridad. Por ello, hemos indagado en distintos estudios sobre la enfermedad del Parkinson y hemos considerado de interés, la investigación realizada por Shi et al. (2015). De ese modo, decidimos elaborar un gráfico con una matriz de correlaciones de Spearman representado en la Figura 8 que compare cada variable numérica de updrs y una muestra de péptidos para ver si es posible observar una correlación entre las variables. Todo ello, despreciando el efecto tiempo, pues no tenemos en cuenta los meses de visita.


Spearman Correlation Matrix
updrs_1
updrs_2
updrs_3
updrs_4

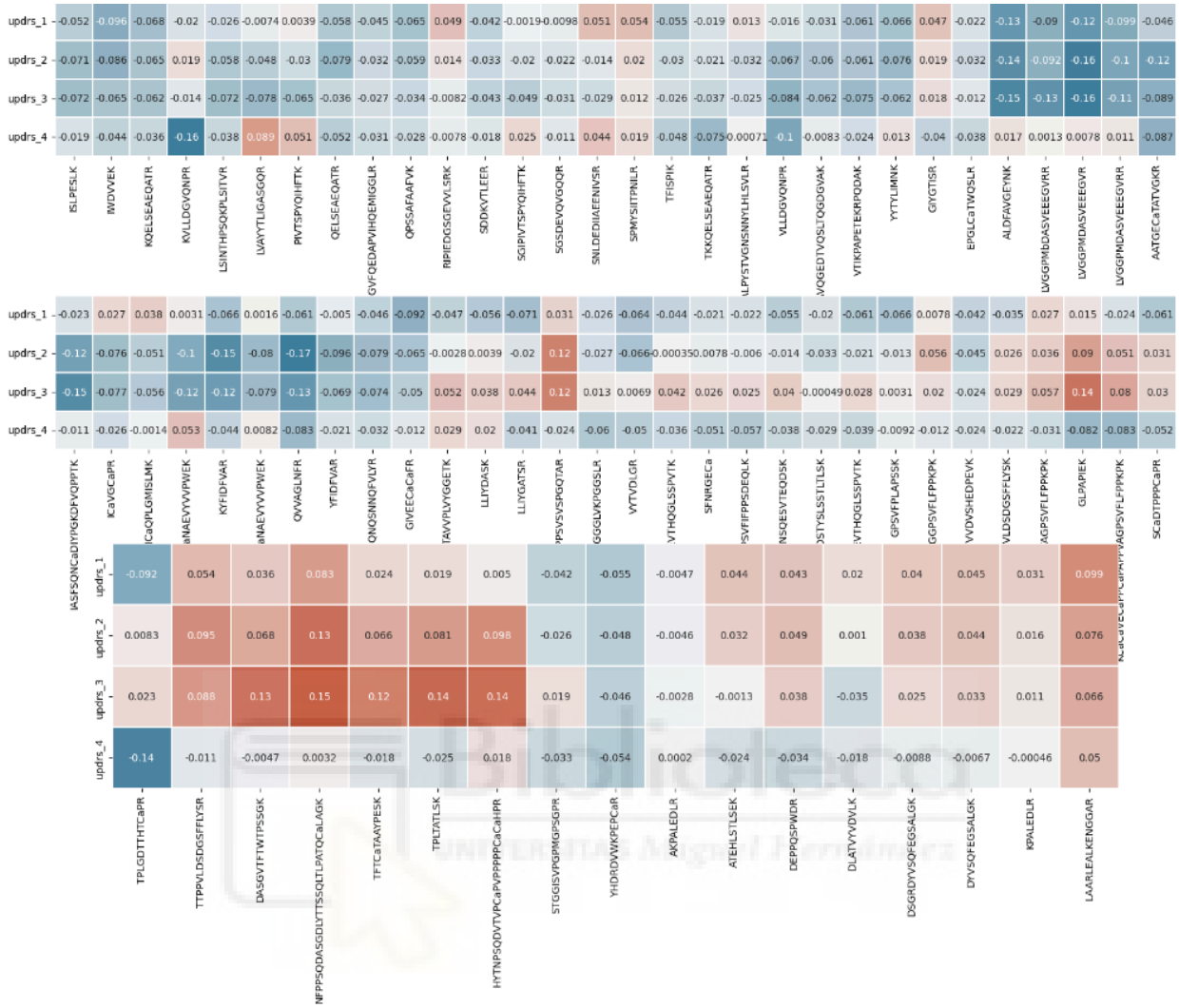


Figura 8: Matriz de correlaciones de Spearman para péptidos y UPDRS. Fuente: Elaboración propia.

Tabla 1: Máximas correlaciones positivas y negativas por puntuación para los péptidos. Fuente: Elaboración propia.

UPDRS_1	0.099 (LAARLEALKENGCAR)	-0.13 (ALDFAVGEYNK)
UPDRS_2	0.12 (NPPSQDASOOLYTTSSQLTLPATQC(UniMod_4)LACK)	-0.17 (QVVAGLNFR)
UPDRS_3	0.16 (YVMLPVADQDQC(UniMod_4)CR)	-0.22 (EIPSSVQQVPTIIK)
UPDRS_4	0.089 (LVAYYTLIGASGQR)	-0.16 (KVLLDGVQNPR)

Cabe destacar los siguientes aspectos sobre la figura y la tabla anterior:

- En general, updrs_3 es la variable que muestra correlaciones más fuertes, tanto positivas como negativas, pero en cualquier caso no son significativas y toman valores de 0.16 en positivo y -0.22 en negativo.
- No podemos decir que exista ninguna correlación fuerte, pues observando los resultados de la Tabla 1 los máximos de cada puntuación son correlaciones débiles.

Del mismo modo, representamos de nuevo en la Figura 9, una matriz de correlaciones de Spearman para medir posibles correlaciones entre las medidas de updrs y las proteínas.

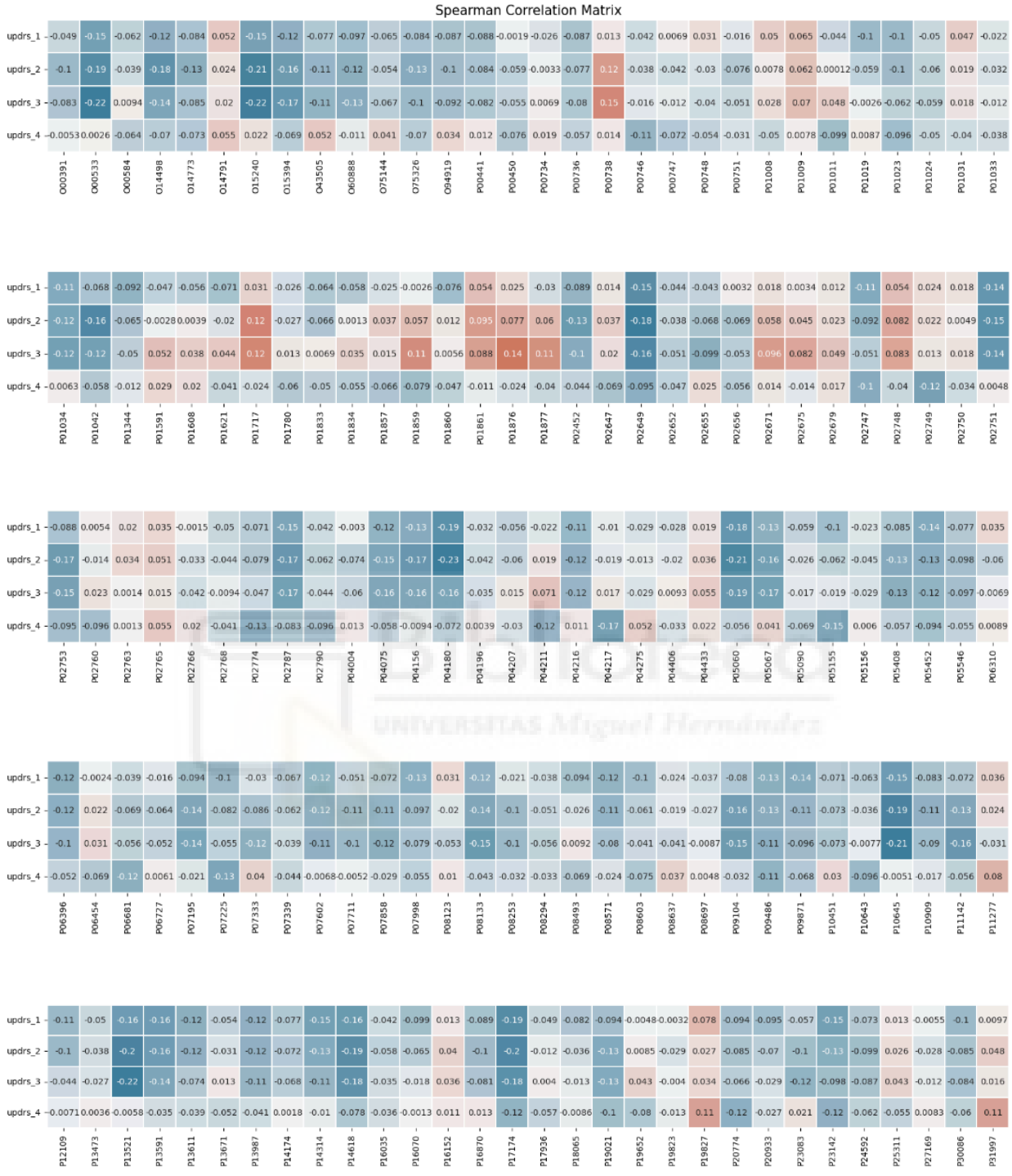

Spearman Correlation Matrix
updrs_1
updrs_2
updrs_3
updrs_4

Figura 9: Matriz de correlaciones de Spearman para proteínas y UPDRS. Fuente: Elaboración propia.

Tabla 2: Máximas correlaciones positivas y negativas por puntuación para las proteínas asociadas. Fuente: Elaboración propia.

UPDRS_1	0.12 (P01594)	-0.19 (P17174)
UPDRS_2	0.15 (P01594)	-0.23 (Q06481)
UPDRS_3	0.15 (P00738)	-0.22 (P13521)
UPDRS_4	0.11 (P19827)	-0.14 (P61626)

Destacamos los siguientes aspectos en base a la Figura 11 y la Tabla 2:

- La mayoría de las correlaciones entre las proteínas son negativas (el gráfico en general tiene tonalidades azules).

- No podemos decir que exista ninguna correlación fuerte, pues observando los resultados de la Tabla 2 los máximos de cada puntuación son correlaciones débiles, tanto para correlaciones positivas como negativas.

En resumen, las dos representaciones anteriores denotan una clara diferencia. Para los péptidos y las variables de updrs existen más correlaciones positivas que negativas por lo general. Por lo tanto, cuanto mayor sea el deterioro, mayor será la abundancia de proteínas como la alfa-sinucleína, cuya acumulación es anormal como hemos mencionado previamente. Por el contrario, para las proteínas y las medidas de updrs tenemos más correlaciones negativas, esto implica que a mayor deterioro de la enfermedad, menor abundancia de péptidos hallamos. La disminución de péptidos está asociado a un tipo de péptidos en específico, los péptidos neuroprotectores, que disminuyen a medida que la enfermedad progresa.

En definitiva, el análisis y la investigación desarrollada en este apartado ha extendido nuestros conocimientos sobre la base de datos estudiada, que nos serán de utilidad para la posterior interpretación de los resultados.

7.5 Análisis estadístico: modelización

A partir de la información disponible, ajustamos diversos modelos de aprendizaje automático y buscamos encontrar, entre ellos, el mejor modelo predictivo para cada una de las medidas de updrs que cuantifican el estado de deterioro del paciente debido a la enfermedad.

Para ello, hemos creado una función de entrenamiento que analiza cada una de las variables diagnósticas de updrs, cuya salida de código es un listado de las técnicas utilizadas, donde resaltamos la que mejor resultado proporciona. Como hemos mencionado previamente, el coeficiente de determinación y el RMSE son nuestras medidas de calidad en las que basamos el orden de los resultados. El modelo que

mejor se ajuste a los datos será el que mejor predicción tendrá en términos de las métricas utilizadas para las variables estudiadas.

Resolvemos cada técnica para las cuatro variables diagnósticas y obtenemos los resultados representados en la Tabla 3 que se muestran a continuación.

Tabla 3: Resultados de las métricas de calidad obtenidas para los diversos modelos de predicción ajustados sobre los cuatro indicadores. Fuente: Elaboración propia.

MODELO	RMSE UPDRS_1	R2 UPDRS_1	RMSE UPDRS_2	R2 UPDRS_2	RMSE UPDRS_3	R2 UPDRS_3	R2 UPDRS_4	R2 UPDRS_4
CatBoosting	**3.5232**	**0.5806**	**3.0465**	**0.7557**	**9.3138**	**0.6451**	**2.3880**	**0.2088**
Random forest	3.6690	0.5452	3.1267	0.7427	9.4664	0.6334	2.4077	0.1957
Regresión lineal	3.7379	0.5279	3.1955	0.7312	10.0480	0.5869	2.4692	0.1541
KNN	3.7393	0.5276	3.2974	0.7138	10.1102	0.5818	2.4541	0.1644
Rigde	3.7434	0.5265	3.1955	0.7312	10.0482	0.5869	2.4718	0.1523
Lasso	3.8069	0.5103	3.2042	0.7297	10.0727	0.5849	2.4718	0.1523
Árbol de decisión	4.0168	0.4549	3.4745	0.6822	9.9898	0.5917	2.8041	-0.0908

Como podemos observar en la Tabla 3, la técnica que mejor se ajusta a los valores de updrs_1 es el CatBoosting, pues su RMSE indica mejor rendimiento en comparación con otros modelos al ser su valor 3.5232, menor que el resto. Además, como indica el valor de R^2 de dicho modelo, la variable updrs_1 queda explicada por el resto de variables independientes con un valor de 0.5806, el más cercano a 1 de todos los modelos.

Por el contrario, el peor método ajustado a nuestros datos según el análisis, es el árbol de decisión. Como podemos observar en los resultados de updrs_1, su RMSE es mayor que la del resto de modelos y su R^2 está más cerca de 0 que de 1.

En el caso de updrs_2, la técnica que mejor se ajusta a los valores es el CatBoosting. Su RMSE indica un valor de 3.0465, menor que en el resto de modelos, por lo que tiene mejor desempeño. Además, como indica el R^2 de dicho modelo, la variable updrs_2 queda explicada por el resto de variables predictoras, con un valor de 0.7557, el más cercano a 1.

Por el contrario, el peor método ajustado a nuestros datos según el análisis para updrs_2, es el árbol de decisión como en el caso de updrs_1. Su RMSE es mayor que la del resto de modelos y su valor de R^2 está más cerca de 0 que de 1.

Como podemos apreciar en los resultados de updrs_3, el algoritmo que mejor ajuste tiene sobre los valores es el CatBoosting, del mismo modo que para las variables anteriores. Su RMSE denota mayor eficiencia con respecto a otros modelos con un valor de 9.3138, menor que en el resto de modelos. Además, como indica el valor de R^2 de dicho modelo, la variable updrs_3 queda explicada por el resto de variables, con un valor de 0.6451, el más cercano a 1. Esto quiere decir que CatBoosting ha hecho unas predicciones más precisas que el resto de algoritmos para los datos observados.

Por el contrario, el peor método ajustado a nuestros datos según el análisis, es el K-Nearest Neighbors. Como podemos observar en los resultados de updrs_3, su RMSE es mayor que el del resto de modelos y su R^2 está más cerca de 0 que de 1.

Por último, para updrs_4 podemos observar de nuevo que la técnica que mejor se ajusta a los valores es el CatBoosting. Su RMSE denota ser más eficiente que el de otros modelos al ser su valor 2.3880, menor que el resto. Cabe destacar que el valor de R^2 de dicho modelo para la variable updrs_4 presenta un valor de 0.2088 más cercano

a 0 que a 1. En comparación con el resto de variables estudiadas, no es un R^2 significativamente alto.

Por el contrario, el peor método ajustado a nuestros datos según el análisis, es el árbol de decisión, como en el caso de updrs_1 y updrs_2. Tal y como se aprecia en los resultados de updrs_4, su RMSE es mayor que el del resto de modelos y su R^2 es negativo, esto último indica que el modelo es impreciso o inexacto para los datos observados.

En base a la modelización y el análisis expuesto, podemos destacar los siguientes aspectos:

- Para cada una de las medidas de updrs el algoritmo que mejor se ajusta a las variables es el CatBoosting. En este caso, nuestra base de datos presenta una gran cantidad de valores atípicos, y este algoritmo se caracteriza por su gran capacidad para manejar y tolerar los valores atípicos. Además, destaca por su buen rendimiento predictivo y la robustez al sobreajuste.
- Las técnicas que peor ajuste presentan son K-Nearest Neighbors para updrs_3 y los árboles de decisión para el resto de variables. En el caso de K-Nearest Neighbors, puede funcionar mal principalmente por su sensibilidad a los valores faltantes. Además, es una técnica que maneja mejor un conjunto de datos pequeño y estructurado con pocas características, todo lo contrario a nuestra base de datos. Por otro lado, los árboles de decisión pueden funcionar mal por su inestabilidad y su poca capacidad para manejar datos continuos de forma eficiente como en el caso de cada medida de updrs.
- Cabe mencionar que los resultados de updrs_3 son relativamente altos en comparación al resto de medidas. Esto se debe a que el RSME de updrs_3 es más alto que el resto de puntuaciones para todos los modelos. Todo ello indica que en la fase 3 de estudio de la enfermedad, muchos pacientes presentan gravedad en los síntomas motores de la enfermedad, como son la bradicinesia,

temblor o rigidez. Por el contrario, los resultados en updrs_4 son relativamente bajos con respecto al resto de medidas de updrs, incluso encontramos un coeficiente de determinación negativo.

- Mencionar que el valor del R2 no es especialmente cercano a 1, como resultaría para un modelo de predicción preciso, pero dada la multidimensionalidad y la debilidad de las relaciones detectadas entre las variables predictoras y la respuesta, era esperable.

7.6 Conclusiones

Para concluir con el estudio, es necesario resaltar los resultados más significativos acordes a los objetivos planteados al principio.

Los objetivos propuestos en el estudio trataban de identificar biomarcadores diagnósticos, pronósticos y de progresión válidos en los parámetros analíticos medidos en los pacientes. Todo ello, mediante el estudio de la evolución de las variables updrs_1, updrs_2, updrs_3 y updrs_4 a lo largo del tiempo, en función del resto de variables predictoras.

En primer lugar, para poder realizar un análisis descriptivo detallado y riguroso sobre la información disponible en nuestras bases de datos, era necesario realizar un preprocesado de datos junto a un análisis exploratorio. Esto ha servido para profundizar en nuestros datos y adquirir una total comprensión sobre las variables de cada base de datos, así como para organizar toda la información en una única base de datos fácil de manejar para la posterior modelización.

En segundo lugar, se ha realizado el análisis descriptivo de los datos del cual podemos extraer las siguientes conclusiones:

- Todos los pacientes parten de un estado avanzado de la enfermedad. El promedio de las puntuaciones denota que los que presentan mayor avance en la

enfermedad son los pacientes que toman medicación desde el mes de reclutamiento.

- Un gran porcentaje de péptidos y proteínas presentan una alta abundancia así como variabilidad a lo largo del tiempo. Esto sugiere que existe mayor degeneración neuronal a lo largo del tiempo para una gran cantidad de pacientes.
- El mes 72 representa un cambio respecto a los meses anteriores en la suma de las puntuaciones. Parece que la neurodegeneración de la enfermedad se estabiliza y las puntuaciones de los pacientes así como la abundancia de péptidos y proteínas disminuyen significativamente, equiparando ambas puntuaciones. A partir del mes 72, las puntuaciones se traducen de nuevo en un ligero aumento para los pacientes que tomaban medicación y una disminución temporal que luego vuelve a aumentar bruscamente en el mes 108 para los que no tomaban medicación.
- Las medidas de updrs y los distintos péptidos y proteínas presentan correlaciones débiles. No es esperable en consecuencia, un modelo de predicción preciso.
- Por último, de entre todas las técnicas utilizadas con el fin de encontrar el mejor modelo predictivo para cada una de las medidas de updrs, identificamos a una de estas como la que mejor predicciones aporta. Los resultados de los coeficientes de determinación y del RMSE nos permiten identificar a CatBoosting Regressor como el algoritmo que mejores predicciones da sobre estas cuatro variables.

Hemos observado que en general la progresión de la enfermedad es lenta a lo largo del tiempo, y en términos globales parece tener un efecto positivo la medicación, pero sólo a largo plazo. Hoy en día son pocos los tratamientos existentes que influyen en la mejoría de esta enfermedad. Por ello, es importante investigar en mayor profundidad la

enfermedad, pues cualquier tipo de hallazgo sobre este tema puede suponer un cambio en el futuro de las personas que padecen esta enfermedad.

Referencias

AMP®-Parkinson's Disease progression prediction. (s/f). Kaggle.com. https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/discussion/388322

EFSA (European Food Safety Authority) 2014. Guidance on Statistical Reporting. EFSA Journal 2014 12 (12): 3908. doi: 10.2903/j.efsa.2014.3908

http://www.efsa.europa.eu/en/efsajournal/pub/3908

Regresión lineal de los mínimos cuadrados (OLS). (s. f.-b). XLSTAT, Your Data Analysis Solution. https://www.xlstat.com/es/soluciones/funciones/regresion-lineal-de-los-minimos-cuadrados-ols

13 Bosques aleatorios (Random Forest) | Machine Learning: teoría y práctica. (s. f.). https://bookdown.org/victor_morales/TecnicasML/bosques-aleatorios-random-forest.html

Thiesen, S. (2021, 31 diciembre). CatBoost regression in 6 minutes - Towards Data Science. *Medium*.

https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

Coursera. (2023, 15 junio). *7 algoritmos de machine learning que hay que conocer:*

Guía para principiantes.

https://www.coursera.org/mx/articles/machine-learning-algorithms

Gusthema. (2023, 3 mayo). *Parkinson's Disease Progression Prediction w TFDF*. Kaggle. https://www.kaggle.com/code/gusthema/parkinson-s-disease-progression-prediction-w-tfdf

Lokeshparab. (2023, 8 mayo). *Parkinsons disease analysis*. Kaggle. https://www.kaggle.com/code/lokeshparab/parkinsons-disease-analysis

Home | AMP-PD. (s. f.). https://amp-pd.org/

Holden, S. K., Finseth, T., Sillau, S., & Berman, B. D. (2017). Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort. *Movement Disorders Clinical Practice*, *5*(1), 47-53. https://doi.org/10.1002/mdc3.12553

Shi, M., Movius, J., Dator, R. P., Aro, P., Zhao, Y., Pan, C., Lin, X., Bammler, T. K., Stewart, T., Zabetian, C. P., Peskind, E. R., Hu, S. C., Quinn, J. F., Galasko, D., & Zhang, J. (2015b). Cerebrospinal Fluid Peptides as Potential Parkinson Disease Biomarkers: A Staged Pipeline for Discovery and Validation*. *Molecular & Cellular Proteomics*, *14*(3), 544-555. https://doi.org/10.1074/mcp.m114.040576

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martínez-Martín, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R. G., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A. J., Leurgans, S. E., LeWitt, P. A., Nyenhuis, D. L., . . . LaPelle, N. R. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129-2170. https://doi.org/10.1002/mds.22340

Anexos

El código desarrollado para la realización de este estudio está disponible en mi repositorio de GitHub. Todo el código ha sido desarrollado en cuadernos de Jupyter de Google Colab. Asimismo, se encuentran en ese mismo repositorio las bases de datos originales que contienen toda la información relativa al estudio.