TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Predicción de la enfermedad de Parkinson mediante análisis acústico

Autor: Carles Requena Sánchez

Directores: José Antonio González Alastrué (UPC), Mónica Giuliano (Universidad Nacional de Hurlingham/Universidad Nacional del Oeste, Buenos Aires)

Departamento: Statistics and Operations Research Dept

Convocatoria: Enero 2024

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen y palabras clave

La enfermedad de Parkinson es un trastorno neurodegenerativo progresivo que afecta principalmente al movimiento. Se caracteriza por síntomas como temblores, rigidez, bradicinesia (lentitud en los movimientos) y alteraciones posturales. Además de los síntomas motores, los pacientes pueden experimentar problemas en la voz y el habla, lo cual se puede utilizar como un marcador acústico para el diagnóstico de la enfermedad. El análisis de estas características acústicas puede proporcionar información valiosa para la detección temprana de la enfermedad.

En este estudio, se desarrolló un modelo predictivo para el diagnóstico de la enfermedad de Parkinson utilizando diversos marcadores acústicos. Los datos fueron obtenidos del proyecto mPower [4], que recopila grabaciones de voz mediante una aplicación móvil. Después del preprocesamiento de los datos, se aplicaron técnicas de selección de variables para reducir su dimensión, como el Análisis de Componentes Principales (PCA), el análisis de correlaciones y LASSO.

Finalmente, se utilizó el método de SVM para construir el modelo predictivo. Los resultados mostraron que, aunque el modelo con todas las variables originales obtuvo mejores resultados en términos de precisión y sensibilidad, la selección de variables redujo significativamente la complejidad del modelo, aunque sacrificando en gran medida su rendimiento. Este enfoque demuestra la poca viabilidad de utilizar grabaciones de voz obtenidas en entornos no profesionales para la detección de la enfermedad de Parkinson.

Palabras clave: PCA, SVM, LASSO, Parkinson, correlación, sensibilidad, especificidad.

Abstract and key words

Parkinson's disease is a progressive neurodegenerative disorder that primarily affects movement. It is characterized by symptoms such as tremors, rigidity, bradykinesia (slowness of movement), and postural instability. In addition to motor symptoms, patients may experience voice and speech problems, which can be used as acoustic markers for the diagnosis of the disease. Analyzing these acoustic characteristics can provide valuable information for early detection of the disease.

In this study, a predictive model for the diagnosis of Parkinson's disease was developed using various acoustic markers. The data was obtained from the mPower [4] project, which collects voice recordings through a mobile application. After preprocessing the data, variable selection techniques were applied to reduce its dimensionality, such as Principal Component Analysis (PCA), correlation analysis, and LASSO.

Finally, the SVM method was used to construct the predictive model. The results showed that, although the model with all the original variables achieved better results in terms of accuracy and sensitivity, the variable selection significantly reduced the complexity of the model, albeit greatly sacrificing its performance. This approach demonstrates the limited viability of using voice recordings obtained in non-professional environments for the detection of Parkinson's disease.

Key words: PCA, SVM, LASSO, Parkinson, correlation, specificity, sensitivity.

Clasificación AMS

68T10 Pattern recognition, speech recognition.

Índice

1. Introducción

La enfermedad de Parkinson es un trastorno neurodegenerativo progresivo que afecta principalmente al sistema motor. Se dice que es neurodegenerativo por qué afecta las células nerviosas del cerebro o de la médula espinal, y es progrsivo ya que se deterioran de forma gradual con el tiempo.

Se caracteriza por una pérdida progresiva de las neuronas dopaminérgicas en la sustancia negra del cerebro, lo que provoca una disminución en la producción de dopamina, un neurotransmisor crucial para la regulación del movimiento.

Los síntomas aparecen lentamente, y suelen ser pequeños temblores, rigidez o una disminución del movimiento. En fases más avanzadas los síntomas incluyen temblores en reposo, rigidez muscular, bradicinesia (lentitud en los movimientos) y alteraciones posturales. Además de estos síntomas motores, los pacientes también pueden experimentar problemas no motores, como alteraciones del sueño, depresión y disfunción autonómica.

Se desconoce la causa de la enfermedad de Parkinson, pero algunos factores como los genes o la exposición a ciertas toxinas como herbicidas o pesticidas parecen influir en su desarrollo. Además, los hombres tienen una mayor probabilidad de desarrollar la enfermedad.

Un aspecto menos conocido pero significativo de la enfermedad de Parkinson es su impacto en la voz y el habla. Las alteraciones vocales, como la reducción del volumen, la monotonía y dificultad para articular palabras, son comunes entre los pacientes. Estas alteraciones, por lo tanto, pueden servir como marcadores acústicos útiles para el diagnóstico temprano de la enfermedad.

Esta relación entre el habla y la enfermedad del Parkinson ha sido un tema de interés en la comunidad científica y de investigación, intentando ofrecer una solución útil al problema, intentando predecir la enfermedad en etapas tempranas y aplicando metodologías muy poco invasivas como es el caso del registro de audios.

Un factor común en las investigaciones es el de usar muestras de pacientes donde las grabaciones han sido realizadas en entornos muy controlados, por lo que suelen ser muestras no muy grandes.

Existen ejemplos como [1], en el que utilizando una base de datos de 108 personas, en entornos controlados y bajo supervisión, se obtuvieron resultados con capacidades predictivas muy notables, superiores al 80 %. O incluso existen otras líneas de trabajo como los informes de [2], que han sido desarrollados a partir de bases de datos mucho más grandes, con tamaños superiores a 1000 individuos y obtenidos mediante grabaciones con dispositivos móviles, en entornos no controlados (con ruido y posibles errores) y sin supervisión profesional, con capacidades predictivas que rondan el 60 %.

Parece que el origen de los datos y la forma en la que se obtienen puede jugar un papel importante en el desarrollo del modelo y en cómo de bien interpreta los datos. Además, siempre es interesante poder obtener modelos parsimoniosos, con las mínimas variables posibles sin afectar en la precisión del modelo, con la finalidad de hacer más fácil su interpretación.

2. Objetivos

El objetivo principal de este proyecto es desarrollar un modelo predictivo eficaz en el diagnóstico de la enfermedad, a través de diversos marcadores acústicos.

Los objetivos específicos son:

- Entender si a partir de una base de datos cruda, obtenida a partir de una aplicación sin ninguna supervisión de profesionales, se pueden obtener resultados que sean satisfactorios.
- Aprender acerca del proceso de selección de variables y realizar predicciones a partir de modelos de clasificación.
- Comparar el resultado de este estudio con el de otras investigaciones, en especial el de "*Selection of Dysphonia Measures for the Identification of Parkinson's Disease"* [1] en el proceso de selección de variables.
- Reducir la dimensionalidad de los datos para que sea un modelo parsimonioso facilitando su interpretación.

3. Metodología

3.1 Diseño

En este TFG se ha diseñado como un estudio observacional transversal con el objetivo principal de desarrollar un modelo predictivo eficaz para el diagnóstico de la enfermedad de Parkinson utilizando diversos marcadores acústicos.

Además, se pretende comparar los resultados obtenidos con otras investigaciones, a diferencia que se han usando unos datos extraídos sin supervisión, quitándole validez a los mismos, ya que no ha habido ningún profesional supervisando el proceso.

3.2 Orígen de los datos

Los datos extraídos para poder realizar este TFG provienen del proyecto mPower [4], desarrollado por Sage Bionetworks [5], empresa sin ánimo de lucro situada en *Seattle* que tiene como lema promover la ciencia abierta y la participación del paciente en el proceso de investigación.

Este proyecto tiene como principal objetivo comprender la progresión de la enfermedad de Parkinson y entender esos patrones únicos que desarrolla cada paciente. Dicho estudio se basa en la utilización de una aplicación móvil, que recopila datos relevantes sobre la enfermedad de Parkinson mediante el uso de varios sensores y cuestionarios que deben ir completando los participantes de forma periódica.

El uso de este método de recopilación de datos no sólo permite una amplia participación, sino que también proporciona una gran cantidad de información esencial para el análisis y la investigación de patrones asociados con la enfermedad, además de poder realizar estudios más precisos y extensos.

El trabajo que se realizó, estuvo destinado a personas voluntarias, residentes de EEUU, mayores de edad, y que tuvieran poder de un dispositivo móvil compatible con la aplicación.
Tanto los pacientes de Parkinson como aquellos que están sanos tuvieron la necesidad de realizar ciertas tareas diariamente durante 14 días y repetir el proceso cada tres meses.
Todas las tareas de documentan a continuación gracias a la tabla:

Tabla 3.2.1: Descripción de las tareas a realizar por los voluntarios

Tarea	Tipo y frecuencia	Participantes únicos	Tareas únicas
Demografía	Encuesta - una vez	6.805	6.805
MDS-UPDRS	Encuesta - mensual	2.024	2,305
PDQ8	Encuesta - mensual	1.334	1.641
Memória	Actividad - t.i.d[1]	968	8.569
Táctil	Actividad - t.i.d	8.003	78.887
Voz	Actividad - t.i.d	5.826	65.022
Caminar	Actividad - t.i.d	3,101	35.410

[6]

- **Encuesta demográfica**: Incluye preguntas como la edad, el sexo, diagnóstico de la enfermedad por un profesional, la raza, situación laboral, estado civil, entre otras.
- **Encuesta MDS-UPDRS**: Cuestiones acerca de la salud motora, como por ejemplo número de actividades con ejercicio de la semana anterior, o puntuar el estado de salud en el que se encontraba el sujeto el día en cuestión.
- **Encuesta PDQ8**: Información sobre cómo está afectando la enfermedad, relevando información acerca de la imposibilidad de realizar tareas cotidianas como ir a trabajar o desvelando problemas motrices o de comunicación.
- **Actividad de memoria**: Se trata de una serie de juegos en la aplicación que miden la capacidad de memoria del usuario.
- **Actividad táctil**: Consiste en hacer una serie de ejercicios tocando la pantalla táctil del dispositivo, donde la aplicación registra, por ejemplo, el número de veces que el sujeto es capaz de tocar la pantalla en un intervalo de tiempo.
- **Voz**: Los individuos deben mantener una voz firme con la vocal 'a' durante 10 segundos.
- **Caminar**: Medidas del acelerómetro del dispositivo, cada vez que el individuo camina.

Cabe recalcar que a los usuarios con enfermedad, deben realizar las pruebas de memoria, táctil, de voz y caminar, en el momento previo a la medicación, el posterior, y en otro momento distinto a los anteriores, por lo que son realizadas tres veces al día para los enfermos, y únicamente una vez, para los sanos.
Además, de manera que los datos sean lo más útiles posible, en la aplicación se detalla meticulosamente cómo debe realizarse cada prueba, de manera que todas las medidas sean recogidas bajo las mismas condiciones.

Todos estos datos, siguiendo la filosofía de Sage Bionetworks [5], pueden ser utilizados por la comunidad científica y son depositados en la plataforma Synapse [6], usada para compartir

[1] t.i.d = tres veces al día

y analizar estos datos de manera colaborativa. Synapse es responsable del intercambio ético y seguro de datos con investigadores de todo el mundo, a través del proceso de almacenarlos, organizarlos y compartirlos.

En este presente estudio, enfocado únicamente a la información obtenida mediante a voz, se han obtenido los archivos de audio de todos los individuos, junto con las demás características demográficas interessantes como la edad y el sexo, sin tener en cuenta las demás características.

Es importante destacar que estos datos, a diferencia de otros estudios, han sido recogidos en entornos no profesionales, sin la supervisión de un experto, por lo que, de los 5.826 usuarios con información auditiva, muchos de ellos no eran válidos debido a interferencias con el sonido, possible ruido exterior o pruebas mal realizadas, por lo que, gracias a la colaboración de Luis Alberto Fernández, profesor de la universidad de Buenos Aires ha trabajado en el informe [1], podemos filtrar las grabaciones.

Luís nos proporcionó estos mismos audios peró filtrando aquellos en que el sonido estaba bien definido, y a su vez, descartando todos aquellos que no eran válidos para el estudio.

Debido a que es difícil calificar un audio si es correcto o no, la extracción de todos los audios seleccionados fué manual y a criterio de Luis Alberto Fernández.

Además, a raíz de que hay múltiples audios por persona, se ha escogido un audio único por participante, escogiendo en el caso de los enfermos, el momento previo a la toma de medicación, ya que es el momento en que la enfermedad está más latente. Por lo que, para aumentar la veracidad y precisión de los resultados, tuvimos que pasar de una base de datos con información de 5.826 voluntarios a 1514, 945 sin enfermedad, y 569 con enfermedad.

Tabla 3.2.2: Número total de participantes

	Participantes
Enfermos	569
Sanos	945
Total	1514

3.3 Lectura de la base de datos

Para poder leer los archivos de audio y procesar la información para su posterior análisis, se realizó un estudio comparativo previo, entre las dos principales herramientas usadas en el

sector, openSMILE y Matlab y así poder evaluar cual era más adecuada para nosotros.

Matlab es una herramienta con lenguaje de programación propio, que es ampliamente utilizada para el análisis de datos, simulaciones y desarrollo de algoritmos. De igual manera, está muy consolidada en el sector de investigación, mientras que openSMILE es una herramienta de código abierto diseñada específicamente para el procesamiento de señales de audio.

Matlab a pesar de que usa un lenguaje de programación de nivel alto, al estar más consagrado y tener un alto uso, posee una comunidad muy grande, con foros muy activos y un soporte completo, mientras que openSMILE posee una comunidad más reducida a los nichos de análisis de audio y procesamiento del habla, por lo que sus recursos no son muy abundantes.

En general, las dos herramientas son igual de útiles e igual de potentes, siendo openSMILE de código abierto, suele ser útil en muchos estudios en que no existe la necesidad de usar una herramienta de pago, peró en el caso de esta investigación, al poseer licencia educativa gratuita, y recibir el soporte del código de Matlab de la investigación de Tsanas [2], se decidió por esta aplicación.

El uso de este *script* permite una extracción precisa y consistente de los parámetros acústicos, proporcionando una base sólida para el análisis posterior en la detección de la enfermedad de Parkinson.

Estas rutinas de Matlab, una vez realizadas las modificaciones correspondientes en algunas funciones para poder adaptar el *script* y leer los archivos en nuestro ordenador con la versión más actualizada, R203b, correspondiente a diciembre de 2023, pudimos empezar a trabajar en el proceso de obtención de la información. Después de 10 días de trabajo computacional procesando la información de los 1514 audios, logramos obtener como resultado, de cada archivo, 339 parámetros. Posteriormente se añadió el género, la edad, y el diagnóstico, obteniendo una cifra final de 342 variables.

3.4 Variables

Las variables extraídas incluyen medidas como Jitter, Shimmer, la relación de ruido armónico (HNR), entre otras. Estas características son sensibles a los cambios en la voz que se asocian con la disfunción motora en la enfermedad de Parkinson.
A continuación, se detalla de forma breve algunos de los principales parámetros que se han extraído:

- **Jitter**: Medida de la inestabilidad de la frecuencia a corto plazo y de los cambios involuntarios en la frecuencia. Representa las pequeñas fluctuaciones o irregularidades en la periodicidad de la señal vocal. Un bajo nivel de Jitter indica que la voz tiene una frecuencia fundamental estable de un ciclo vocal al siguiente.

- **Shimmer**: Es la medida de la inestabilidad de la intensidad, es decir, mide las variaciones de amplitud o intensidad entre ciclos consecutivos de la onda de voz, por lo que representa la variabilidad de la amplitud a lo largo del tiempo. Por ejemplo, un valor bajo de Shimmer indica que la amplitud de la voz es bastante estable de un ciclo a otro. Por el contrario, un valor alto de Shimmer sugiere una mayor irregularidad en la amplitud de la voz, lo que puede indicar problemas vocales o de las cuerdas vocales, tales como tensión, fatiga.
- **HNR** (*Harmonics-to-Noise Ratio*): Mide la calidad sonora, es decir, compara la cantidad de sonido limpio en la voz con la cantidad de ruido o sonidos no armónicos presentes. Por ejemplo, un HNR alto indica una voz clara y nítida con pocos ruidos de fondo o distorsiones, lo que sugiere una buena calidad de la señal vocal. Por otro lado, un HNR bajo indica una voz ronca o áspera, con una mayor presencia de ruido en la señal de voz.
- **GNE** (*Glottal to Noise Excitation*): Mide la vibración de las cuerdas vocales y se compara con el ruido generado durante la producción de la voz. Un valor alto de GNE sugiere una voz clara con menos ruido. Por otro lado, un valor bajo de GNE sugiere una mayor proporción de ruido en la señal de voz.
- **NSR** (*Net Speech Region):* Mide los déficits en la temporización y ritmo del habla, es decir, la velocidad efectiva del habla, excluyendo micro pausas y silencios.
- **Entropía de shannon**: Mide la cantidad de información de la señal, es decir, la incertidumbre, por lo que es útil para analizar la riqueza de la voz y poder encontrar posibles patrones.
- **MFCC** (*Mel Frequency Cepstral Coefficients*): Coeficientes resultados del análisis del espectro de frecuencia de la señal de audio.
- **TKEO** (Teager-Kaiser Energy Operator): Evalúa la energía de las señales de voz, por lo que aporta información sobre los cambios de amplitud. Útil para detectar las posibles variaciones sutiles que otras métricas pasarían por alto.

Al tratarse de audios de 3 segundos, de las medidas principales, se obtienen subvariantes con relación a las mismas, por ejemplo las medias del periodo, la desviación o los percentiles.

3.5 Preprocesamiento de los datos

El preprocesamiento de datos es una etapa esencial en cualquier proyecto de análisis de datos. Este proceso implica convertir datos crudos y poco útiles en un formato más amigable para el análisis, sin perder el punto de vista de la calidad de los mismos.

En el caso de este estudio, como hemos comentado anteriormente, se ha realizado una etapa previa, filtrando aquellos audios que podrían estar mal grabados o podrían tener ruido excesivo. Además, ha sido necesario obtener una información más cuantitativa de cada audio, gracias a Matlab, para poder parametrizar toda la información, obteniendo ya un formato trabajable con la herramienta R Studio.

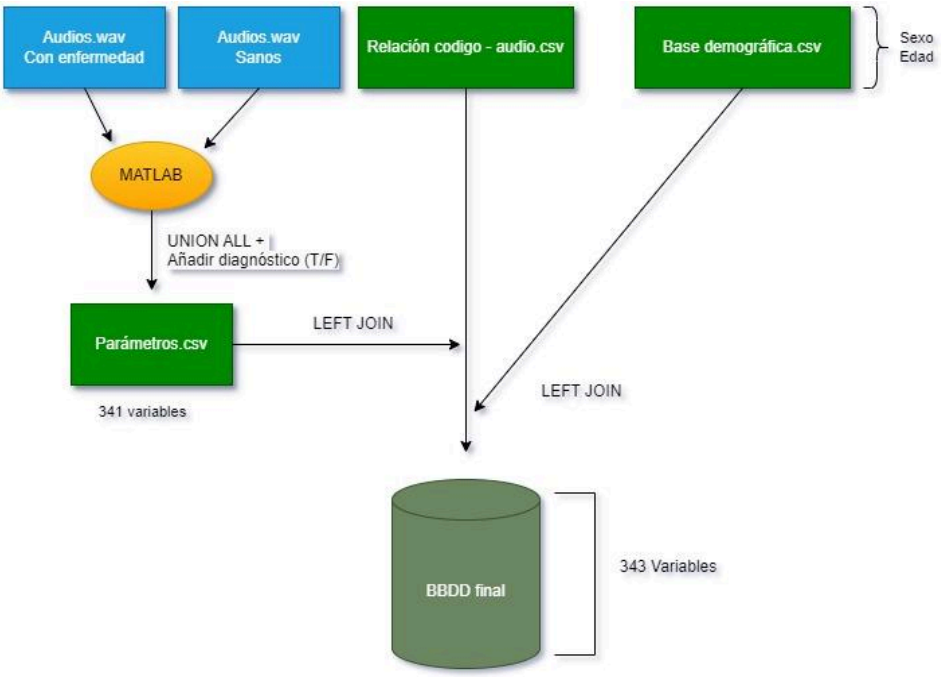
Una vez realizadas estas etapas y añadidas las variables de edad, sexo y diagnóstico, obtenemos una base de datos de 1514 voluntarios y 342 variables.

Para entender mejor cómo se han obtenido los datos finales, la figura 3.5.1 nos ayuda a comprender de donde parten los datos.

Primero de todo, partimos de dos archivos con los audios ya pre seleccionados, en formato wav, tanto de los voluntarios enfermos como de los sanos, y a su vez, tenemos una base de datos en formato csv con la base demográfica, es decir, con la totalidad de las respuestas referentes a la encuesta demográfica que se realiza al inicio del proyecto.

De esta base demográfica, a pesar de tener una gran cantidad de variables a cerca de cada usuario, se decidió solo incluir las variables edad y sexo, ya que hay variables fonatórias que dependen de la características físicas de cada sujeto y las consideramos relevantes. Además, existe una tercera tabla que será útil para poder unir las dos mencionadas anteriormente.

Figura 3.5.1: Diagrama de flujo de los datos



Los datos finales pueden tener valores atípicos, información faltante o incompleta que puede hacer disminuir la precisión y robustez del modelo, por lo que es fundamental realizar una serie de pasos de preprocesamiento para detectar estos errores.
Primero de todo, eliminamos de la base de datos la variable identificativa “ID”, que no nos aporta información necesaria. Por otra parte, debido al gran volumen de variables, para facilitar la identificación de las mismas, se añade un prefijo numérico delante de cada una de

ellas, “vX_” para poderlas identificar más claramente. Por ejemplo “v1_Jitter->F0_abs_dif”. Adicionalmente, se cambian a factor las variables género y diagnóstico.

Así que la base de datos, una vez realizados los pasos comentados anteriormente, y categorizada tal y como el equipo de Mónica establecieron, queda de la siguiente siguiente manera:

Tabla 3.5.1: Clasificación de las variables

Grupo	Medidas	Variables	Cantidad
G1	Variaciones de F0 (Jitter)	V1-V22, V49-V51, V155-336, V337, V339	209
G2	Variaciones de la amplitud	V23-V44	22
G3	Ruido	V45-V48, V52-V70, V338	24
G4	Problemas en la articulación	V71-V154	84

[1]

De igual manera, debido a la complejidad en la obtención de los parámetros mediante Matlab, era vital poder asegurarnos que los datos obtenidos eran lógicos, es decir, comprobar que cada variable estaba en la misma dimensión, o no existían valores que eran muy dispersos, cosa que indicaría que el *script* había podido fallar.

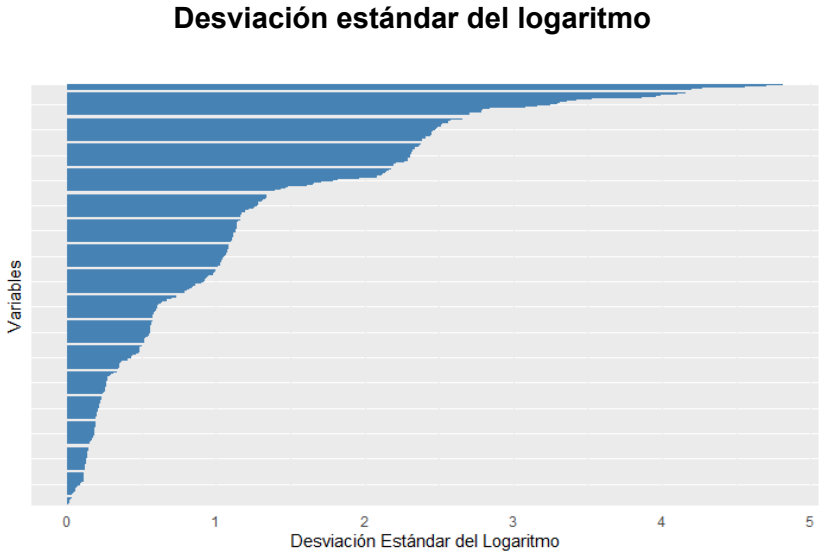
Para ello se generó una tabla con la media y la desviación del logaritmo de cada variable, pudiendo encontrar dos variables con un alto número de ceros, ya que al tener valores tan sumamente pequeños, se había reducido a 0.
Para evitar que este ruido pudiera afectar al análisis se procedió a eliminar las dos variables, “Jitter->F0_TKEO_prc75” y “v51_GQ->std_cycle_closed” y tener un conjunto de variables correctamente identificadas.

Tabla 3.5.2: Variables que han obtenido zeros, junto con el porcentaje

Variables	**Porcentaje de zeros**
v19_Jitter->F0_TKEO_prc75	93.79 %
v51_GQ->std_cycle_closed	12.81 %

Una vez eliminadas, y tal y como indica la 3.5.2, vemos que no existen valores muy grandes en la desviación del logaritmo de cada variable que pudieran indicar que los datos obtenidos no son correctos.

Figura 3.5.2: Desviación estándar del logaritmo de todas las variables



Además, se procedió a eliminar una fila de datos, correspondiente a las de un paciente sano, ya que la variable género era nula y podría ocasionar conflictos.

3.6 Estadística descriptiva

La etapa de elaboración del análisis descriptivo de los datos es fundamental en el proceso de investigación. El objetivo principal es entender qué características tiene la muestra, resumir los mismos, y por lo tanto es un proceso necesario para poder aprender de ellos y tener un resumen claro de cómo se distribuyen.

En esta etapa, nos centramos en las variables género y edad, ya que posteriormente, analizaremos más profundamente las variables numéricas mediante matrices de correlaciones.

3.7 Selección de variables

Este proceso consiste en identificar y elegir las variables que realmente importan y que mejoran el rendimiento del modelo, eliminando aquellas que no aportan información añadida o que son redundantes. Hacer una buena selección de variables no solo hace que el modelo sea más preciso y fácil de interpretar, sino que también reduce la complejidad del cálculo y el riesgo de sobreajuste.

En casos como el nuestro, donde tenemos un gran número de variables, tiene un plus interesante ya que el reducir considerablemente el número de variables hará que el modelo

sea computacionalmente más rápido y por lo tanto aplicable a un mayor número de situaciones.

Además, existe un gran número de parámetros que comparten orígen con la misma variable, por lo que muchas de ellas están correlacionadas y no aportan tanta información a los datos.

Hay muchas maneras de hacer esta selección, desde métodos estadísticos tradicionales hasta algoritmos de aprendizaje automático más avanzados. Peró en este trabajo, y con el objetivo de replicar también el método de [1] y que sea comparable, se realiza un PCA, con el posterior análisis de correlaciones para finalmente aplicar LASSO.

Se aplica un análisis de correlación previo a aplicar LASSO, ya que a pesar manejar la multicolinealidad en la base de datos, en los casos en que la multicolinealidad es muy grande, puede hacer que LASSO elimine de forma arbitraria alguna de las variables del modelo.

i. Análisis de componentes principales

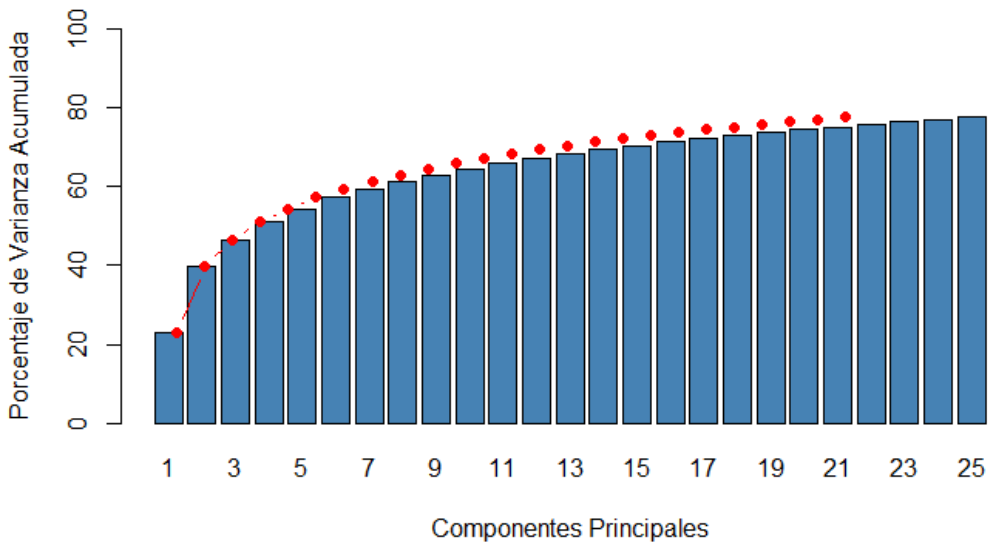
También conocido como PCA (*Principal Component Analysis*) se usa comúnmente para el preprocesamiento de datos antes de utilizar algoritmos de aprendizaje automático. Esto reduce la complejidad del modelo.

Al proyectar un conjunto de datos de gran dimensión en un espacio de características más reducidas, minimiza o elimina problemas comunes como la multicolinealidad y el sobreajuste. La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí, lo que puede ser problemático para el modelado. Los modelos sobre ajustados generalizan mal con datos nuevos, por lo que reduce mucho su precisión.

El PCA reúne la información de grandes conjuntos de datos en un conjunto más pequeño de variables no correlacionadas, conocidas como componentes principales. Estos componentes principales son combinaciones lineales de las variables originales que tienen la varianza máxima en comparación con las otras, capturando tanta información del conjunto de datos original como sea posible.

Para aplicar un PCA a nuestros datos, usaremos en R la función PCA, del paquete FactoMineR.
Observando los resultados obtenidos en la figura 3.7.1, nos quedamos con la información aportada por las 21 primeras componentes principales, que aportan un 77% de varianza explicada, valor que consideramos suficiente para nuestro modelo.

Figura 3.7.1 : Varianza acumulada por las primeras 25 componentes principales



De cara a interpretar los resultados obtenidos, las componentes principales, al ser combinaciones lineales de variables originales, dificultan la interpretación. De cara a la aplicabilidad del estudio, es esencial que los médicos u otros profesionales sepan qué variables específicas son las que están influyendo en el modelo, si se trata de variables que hacen referencia a las cuerdas vocales, al ritmo del habla, etc.

Por lo tanto, analizaremos las correlaciones entre las variables y las componentes principales y así conoceremos qué variables tienen más peso en la variabilidad de cada componente. El objetivo es obtener de cada componente qué variables tienen mayor impacto y así proporcionar una base de datos más sólida y robusta, seleccionando únicamente las variables necesarias.

Este proceso se realiza en R, y se ha desarrollado un pequeño algoritmo iterativo, en el que se seleccionan las tres variables con mayor correlación en valor absoluto de cada componente.

Debido a que la dimensionalidad de los datos continúa siendo elevada, aplicaremos LASSO para reducirla un poco. No sin antes extraer las variables correlacionadas, ya que un exceso de multicolinealidad podría afectar considerablemente al método LASSO.

ii. Análisis de correlaciones

En esta fase del proyecto, es clave detectar variables altamente correlacionadas, que pueden ser redundantes en la base de datos y afectar la efectividad del modelo. Además, ayudará a simplificar el modelo y reducir su dimensionalidad.

Para poder aplicar este análisis, mediante R, hemos seleccionado aquellas variables con una correlación en valor absoluto superior a 0.9, valor que consideramos alto y justo para este análisis.

Una vez tenemos las variables altamente correlacionadas localizadas, nos centraremos en incluir en el modelo únicamente una de cada grupo.

Esta selección se ha realizado de forma manual y no aleatoria, con el propósito de no incluir en el modelo variables de los mismos grupos. Además, se ha dictaminado el criterio de escoger, en aquellas variables iguales en las que solo varía el coeficiente, quedarnos con las del coeficiente más bajo. Por ejemplo, de las siguientes 3 variables, escogeremos la primera, ya que al ser un coeficiente menor, debería aportar más información:

- v309_app_LT_entropy_log_3_coef
- v311_app_LT_entropy_log_5_coef
- v310_app_LT_entropy_log_4_coef

.

Realizando este proceso, eliminamos 17 variables, por lo que reducimos aún más la dimensionalidad de los datos y tenemos los datos ya preparados para la siguiente fase, aplicar LASSO.

iii. LASSO

LASSO, en inglés Least Absolute Shrinkage and Selection Operator, es una poderosa técnica de selección automática de variables. Se trata de un método de regularización que aplica una penalización a las variables para prevenir el sobreajuste y mejorar la precisión de los modelos estadísticos.

Esto se logra añadiendo un término de penalización a la suma residual de los cuadrados (RSS), que luego se multiplica por el parámetro de regularización (lambda o λ). Este parámetro de regularización controla la cantidad de regularización aplicada. Valores altos de lambda aumentan la penalización, reduciendo más coeficientes a cero; esto hace que reduzca la importancia de algunas variables en el modelo o que las elimine por completo, generando así, un modelo de selección de variables automático. Por el contrario, valores menores de lambda reducen el efecto de la penalización, siendo más laxo en la selección de variables del modelo.

El valor óptimo de λ se puede determinar con técnicas de validación cruzada. Este enfoque encuentra el valor de λ que minimiza el error cuadrático medio (MSE). A medida que λ aumenta, el sesgo del modelo aumenta mientras que la varianza disminuye, ya que a medida que λ se hace más grande, más coeficientes β se reducen a cero.

Para realizar la técnica de LASSO, en R, hemos usado el paquete glmnet, utilizando la función *cv.glmnet*, con validación cruzada. Además, para optimizar un poco más el proceso, generamos un pequeño proceso iterativo, programando una función que realice este método un total de 50 veces, y que en cada iteración, se guarden aquellas variables en que los coeficientes son diferentes de 0, que en nuestro caso, el valor 0 lo asumimos en aquellos coeficientes cuyo valor absoluto son superiores a 1×10^{-7}.

Después de finalizar la secuencia, y tenemos un listado con todas las variables que se han ido seleccionando, nos quedamos con todas aquellas que han aparecido un mínimo de 40 veces, es decir en el 80% de los casos.

3.8 Modelo

Una vez seleccionadas las variables, y obtenida una base de datos con una dimensión adecuada, procedemos a obtener un modelo capaz de predecir la presencia de la enfermedad de Parkinson a partir de señales de voz.

El modelo que se ha escogido es SVM (*Support vector Machine*), un método de aprendizaje supervisado para problemas de clasificación, muy útil en bases de datos con una dimensionalidad elevada como la nuestra.

El modelo de SVM se fundamenta en el *Maximal Margin Classifier*. Este algoritmo agrupa los puntos de cada lado según sus relaciones homogéneas, utilizando una línea llamada hiperplano. Se dice que estos puntos son linealmente separables si una línea recta puede dividirlos.

El SVM no solo busca cualquier línea que separe las clases, sino que encuentra la línea que maximiza la distancia (*Margin*) entre las dos clases más cercanas. Este margen maximizado ayuda a que el modelo sea más robusto y menos propenso a errores cuando se clasifiquen nuevos datos.

Los puntos de los datos que están más cerca de la línea de separación se llaman vectores de soporte. Estos puntos son cruciales porque definen la posición de la línea divisoria. Si movemos estos puntos, la línea también se moverá.

Además un problema con el que podría encontrarse el método SVM al modelar datos reales es la extensa no linealidad de un dato. La técnica del kernel, una característica del SVM, nos permite manipular esos datos fácilmente en datos linealmente separables, transformándolos a una dimensión superior donde se puedan separar linealmente.

Para poder generar el modelo y realizar las predicciones, continuamos usando la herramienta estadística R, y en este caso, se ha usado el paquete “caret” de R, uno de los más populares para tratar este tipo de algoritmos y que permite usar varios tipos de ellos,

como el SVM lineal, el polinomial o el radial, que son en las variantes en las que vamos a trabajar.
En la siguiente tabla mostramos las principales diferencias entre estos tipos de modelos SVM.

Tabla 3.8.1 : Tabla resumen de las variantes de SVM

Modelo	**Se basa en**	**Aplicación**	**Ventajas**	**Desventajas**
Lineal	Hiperplano lineal	Datos linealmente separables	Coste bajo computacional	No es muy útil con datos muy complejos
Polinomial	Hiperplano polinomial	Relación no lineal capturable por un polinomio	Muy flexible	Costoso computacional mente y riesgo de sobreajuste
Radial	Hiperplano de dimensión infinita	Separación no lineal compleja	Muy flexible	Costoso computacional mente

Se ha usado un 25 % de los datos para testear el modelo y un 75 % de ellos para el entrenamiento: Al tener una muestra de datos lo suficientemente grande, podemos dar un poco más de peso a la parte del testeo, creyendo que un 25 % de los datos es suficiente.

Además se realiza validación cruzada (10 × 10), proporcionando una estimación más robusta del rendimiento del modelo. Este procedimiento consiste en dividir los datos en varios subconjuntos, en este caso en 10 partes, entrenando el modelo en varios subconjuntos y testeando en los demás.

.

4. Resultados

En este apartado se pretende mostrar en más detalle y de forma clara y concisa, los hallazgos obtenidos a lo largo del trabajo.

4.1 Características de la muestra

Los datos de la tabla 4.1.1 indican cómo del total de voluntarios, hay más hombres que mujeres, sin embargo, la proporción de mujeres con la enfermedad es mayor que en el caso de los hombres. Por norma general, la enfermedad del Parkinson tiene por costumbre aparecer con mas frecuencia a personas de sexo masculino así que más adelante veremos si esta variable es lo suficientemente significativa como para introducirla en el modelo.

Tabla 4.1.1: Número de hombres y mujeres según si tienen diagnóstico positivo o negativo

	Hombres	Mujeres
Enfermos	377	192
Control	779	165
Totales	1156	357

Además, observamos que en la figura 4.1.1 no parecen existir diferencias claras entre la edad y el género, peró sí que existen, tal y como intuíamos, entre la entre la edad y el diagnóstico. La enfermedad del Parkinson se produce en edades avanzadas por lo que tenemos a voluntarios con enfermedad en franjas de mayor edad que en los grupos de control.

Figura 4.1.1: Boxplot de la distribución de la edad según el género

Boxplot Género - Edad

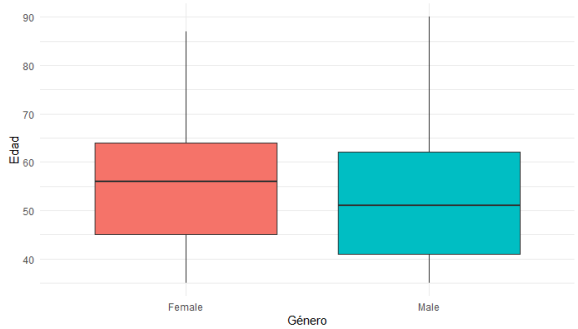
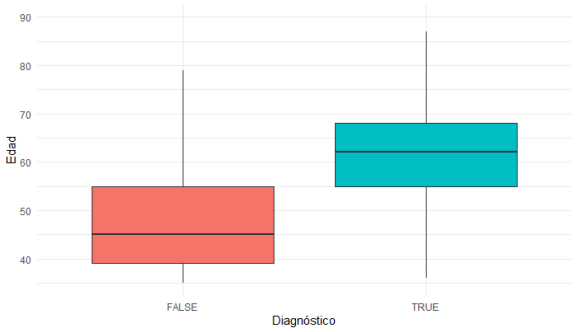


Figura 4.1.2: Boxplot de la distribución de la edad según el diagnóstico

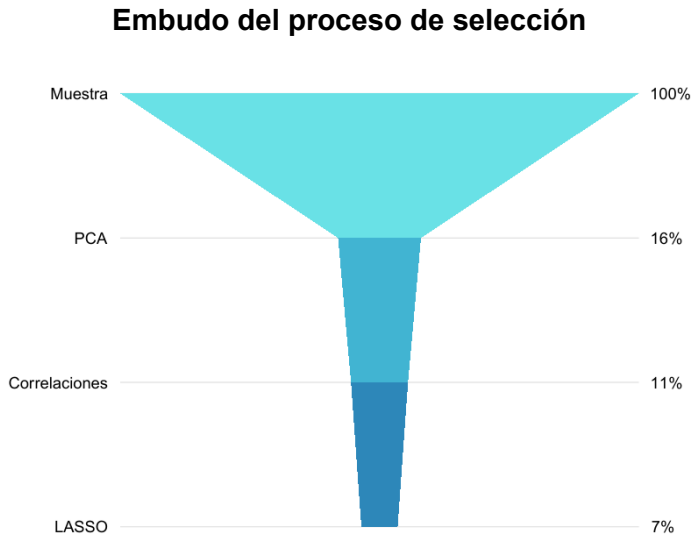
Boxplot Diagnóstico- Edad



4.2 Variables seleccionadas

Después de realizar el ACP, con posterior análisis de correlaciones y finalizando con la técnica LASSO, reducimos la dimensionalidad de los datos, de un total de 342 variables en la base de datos inicial, a 24, reduciendo su dimensionalidad un 93%.
En la primera selección, usando PCA, el resultado obtenido son 63 variables, que seleccionando las que no se repiten, acabamos obteniendo 54 variables, reduciendo considerablemente la dimensionalidad de los datos originales.
Al realizar la segunda fase, realizando el análisis de las correlaciones, la selección queda reducida a 37, y finalmente aplicando LASSO, la selección final cuenta con 24 variables.
En la siguiente figura, se muestra de forma esquemática el proceso de selección y como la dimensionalidad ha ido disminuyendo.

Figura 4.2.1: Gráfico de embudo del proceso de selección de variables



Por lo que finalmente, la muestra queda reducida a las siguientes variables, mucho más aplicables a un estudio de estas características, donde se busca que el modelo sea fácil de interpretar:

Tabla 4.2.1: Tabla resumen de las variables seleccionadas

Grupo	Medidas	Variables	Cantidad
G1	Variaciones de F0 (Jitter)	V5, V18, V205, V260, V261, V263, V295, V304, V309	9
G2	Variaciones de la amplitud	V34	1

G3	Ruido	V54, V63, V66, V67, V69,	5
G4	Problemas en la articulación	V71, V72, V73, V77, V99, V101, V133, V134, V135	9

4.3 SVM

Para el proceso de modelización, se han tenido en cuenta varias casuísticas.
Uno de los objetivos principales de este estudio era validar si los datos obtenidos de la aplicación de mPower [4] son lo suficientemente buenos para poder realizar predicciones de la enfermedad, es por eso, que este proceso se ha dividido en tres ramas de modelización usando:

- Las 24 variables seleccionadas anteriormente.
- Las variables del estudio *Selection of Dysphonia Measures for the Identification of Parkinson's Disease* [1].
- Todas las variables de la muestra.

Además, se ha realizado con las tres variantes de SVM comentadas anteriormente en el apartado de metodología, para entender qué variedad se adapta más a nuestros datos.

Tabla 4.3.1: Tabla resumen con las variables seleccionadas

Modelo	**Precisión**	**Sensibilidad**	**Especificidad**
SVM lineal	0.6720	0.3732	0.8517
SVM radial kernel	0.6905	0.4366	0.8432
SVM polynomial kernel	0.6508	0.3451	0.8347

Tabla 4.3.2: Tabla resumen con todas las variables

Modelo	**Precisión**	**Sensibilidad**	**Especificidad**
SVM lineal	0.7460	0.6338	0.8136
SVM radial kernel	0.7434	0.6690	0.7881
SVM polynomial kernel	0.7461	0.5704	0.8517

Tabla 4.3.3: Tabla resumen con las variables de *"Selection of Dysphonia Measures for the Identification of Parkinson's Disease"* [1]

Modelo	**Precisión**	**Sensibilidad**	**Especificidad**
SVM lineal	0.6323	0.0704	0.9703
SVM radial kernel	0.6640	0.2465	0.9153
SVM polynomial kernel	0.6534	0.1831	0.9364

Para entender las tablas anteriores, y a qué hacen referencia, es importante tener en cuenta que significa la precisión, la sensibilidad y la especificidad.

- La precisión es la métrica que mide la proporción de predicciones correctas sobre el total de realizadas.
- La sensibilidad mide la capacidad del modelo para identificar correctamente los casos positivos, en este caso los que presentan la enfermedad.
- La especificidad mide la capacidad del modelo para identificar a los casos negativos.

En este análisis, la sensibilidad juega un papel muy importante, ya que el objetivo principal es poder predecir correctamente a los enfermos, para poder intervenir en edades más tempranas y así promover un tratamiento lo antes posible.

No es de extrañar que el SVM Radial con Kernel, modelo que funciona bien con datos de gran dimensión donde las clases no son linealmente separables, indicativo de que los datos poseen una estructura más compleja.

Para entender mejor los resultados, en la tabla 4.3.4, tenemos la matriz de confusión con los valores obtenidos en el mejor modelo con nuestras variables. Se ve claramente como la mayoría de los datos apuntan a la predicción negativa, provocando una baja sensibilidad errando en los falsos negativos.

Tabla 4.3.3: Tabla de la matriz de confusión para el SVM Radial con nuestras variables

	Referencia	
Predicción	Negativo	Positivo
Negativo	199	80
Positivo	37	62

5. Conclusiones

En los últimos 10 años la comunidad científica ha realizado múltiples estudios relacionados con el Parkinson y la posible predicción de la enfermedad mediante los audios del paciente, por lo que la hipótesis sobre la predicción de la enfermedad está lo suficientemente validada.

En cuanto a la aplicabilidad de estos estudios en la vida real de las personas, no ha sido tan sencillo, ya que integrar estos análisis en la comunidad médica es complicado, es necesario tomar los audios en situaciones muy controladas, y no siempre es factible. Es por eso que Sage Bionetworks [5] desarrolló la aplicación para medir los audios y así dar más peso a los voluntarios y los futuros pacientes, tomando el riesgo de tener unos datos donde el método no ha sido supervisado por ningún profesional.

Cabe recalcar, además, la importancia de tener un modelo parsimonioso en este tipo de estudios, es decir, un modelo con el menor número de variables posible sin perder capacidad explicativa o sacrificando parte de ella para así ganar en interpretabilidad. De esa manera computacionalmente será menos complejo y será más fácil la implementación.

Observando los resultados obtenidos en el modelo con nuestra selección de variables, los resultados no son demasiado buenos. Poniendo énfasis sobre todo en la sensibilidad, se ha observado un valor demasiado pequeño para otorgar la validez necesaria, es decir, el modelo predice de forma incorrecta a aquellos pacientes que realmente tienen la enfermedad, por lo que tiene dificultad para identificar a los casos positivos.

Asumiendo que a efectos prácticos, este tipo de modelos, no són interesantes ni aplicables en este tipo de estudios, se ha comprobado, que aumentando el número de variables a 50, los resultados se acercan mucho más a los obtenidos con la totalidad de la muestra (ver tabla 7.3 del anexo).

En cuanto a la comparativa con la selección de variables del estudio de Mónica Giuliano [1], no se han podido obtener tampoco buenos resultados en nuestra muestra.

Esto reafirma la idea de que los datos no son del todo representativos. Tomar las muestras en entornos muy controlados, con material profesional y con el soporte de personal especializado hacen que las muestras, a pesar de ser mucho más pequeñas, sean más representativas. La capacidad de captar el sonido de un teléfono móvil no es la misma que la de un micrófono profesional en una sala insonorizada. Por lo que podemos afirmar que la muestra obtenida en un entorno no controlado puede tener datos con excesivo ruido o errores, que dificultan mucho la representatividad de los mismos.

Para acabar de dar por finalizado este proyecto, es interesante marcar posibles pasos a seguir, y uno de ellos es intentar reducir la dimensionalidad de los datos con otras técnicas no utilizadas en este estudio y que sí puedan reducir el número de variables sin afectar a la precisión del modelo.

6. Bibliografía

1. Giuliano, M., Fernandez, L., & Pérez, S. (2020). Selection of Dysphonia Measures for the Identification of Parkinson's Disease. In 2020 IEEE ARGENCON
2. S.Aurora y A.Tsanas <Discrimination of Parkinson’s Disease aprticipants from healthy controls using telephone-quality voice recordings>. En: Mov Disord. 31, 20th Internacional Congress.2016
3. Mayo clinic [en línea]: Enfermedad del parkinson [consulta: 19 de enero de 2024]. <https://www.mayoclinic.org/es/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055 >.
4. mPower [en línea]: About the study [consulta: 19 de enero de 2024]. <https://parkinsonmpower.org/about>.
5. Sage Bionetworks [en línea]: Sage Bionetworks [consulta: 19 de enero de 2024]. <https://sagebionetworks.org/>.
6. Synapse [en línea]: mPower public research portal [consulta: 4 de febrero de 2024]. <https://www.synapse.org/Synapse:syn4993293/wiki/375988 >.
7. IBM [en línea]: Principal component analysis : Diciembre de 2023 [consulta: 4 de marzo de 2024]. <https://www.ibm.com/topics/principal-component-analysis >.
8. IBM [en línea]: Lasso Regression : Enero de 2024 [consulta: 7 de marzo de 2024]. <https://www.ibm.com/topics/lasso-regression >.
9. Joaquín Amat [en línea]: Máquinas de vector de soporte : Abril de 2017 [consulta: 8 de abril de 2024]. <https://cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines >.
10. GeeksforGeeks [en línea]: Implementación del clasificado de máquinas vectoriales en R : 8 de diciembre de 2022 [consulta: 8 de abril de 2024]. <https://www.geeksforgeeks.org/support-vector-machine-classifier-implementation-in-r-with-caret-package/ >.
11. Juan Barrios [en línea]: La matriz de confusión y sus métricas,: 26 de julio de 2019 [consulta: 20 de abril de 2024]. <https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/>.
12. José Berrendero [en línea]: Introducción a caret: [consulta: 15 de mayo de 2024]. <https://rpubs.com/joser/caret/>.
13. Rubén F Casal [en línea]: Métodos de regularización: [consulta: 21 de mayo de 2024]. <https://rubenfcasal.github.io/aprendizaje_estadistico/shrinkage.html#ejemplo-lasso>.
14. MATLAB [en línea]: Introducción a Matlab., [consulta: 21 de mayo de 2024]. <https://es.mathworks.com/help/matlab/getting-started-with-matlab.html>.
15. Andrés Rueda [en línea]: Código PCA: 08 de junio de 2022 [consulta: 21 de mayo de 2024]. <https://rpubs.com/andresss20/codigo_pca>.
16. ISCglobal [en línea]: Modelos de regularización: 08 de junio de 2022 [consulta: 21 de mayo de 2024].

<https://isglobal-brge.github.io/Aprendizaje_Automatico_1/modelos-de-regularizaci%C3%B3n.html>.
17. Máxima formación [en línea]: Análisis de correlación: Guía rápida en r, febrero de 2019 [consulta: 29 de mayo de 2024]. <https://www.maximaformacion.es/blog-dat/analisis-de-correlacion-guia-rapida-en-r/>.

7. ANEXO

Tabla 7.1: Tabla resumen de las variables seleccionadas

Grupo	Medidas	Variable
G1	Variaciones de F0	v5__Jitter->F0_PQ3_generalised_Schoentgen
		v18_Jitter->F0_TKEO_prc25
		v205_det_TKEO_std_10_coef
		v260_det_LT_entropy_shannon_4_coef
		v261_det_LT_entropy_shannon_5_coef
		v263_det_LT_entropy_shannon_7_coef
		v304_app_LT_entropy_shannon_8_coef
		v309_app_LT_entropy_log_3_coef
G2	Variaciones de la amplitud	v34_Shimmer->F0_abs0th_perturb
G3	Ruido	v54_GNE->SNR_TKEO
		v63_VFER->NSR_TKEO
		v66_IMF->SNR_TKEO
		v67_IMF->SNR_entropy
		v69_IMF->NSR_TKEO
G4	Problemas en la articulación	v71_mean_Log energy
		v72_mean_MFCC_0th coef
		v73_mean_MFCC_1st coef
		v77_mean_MFCC_5th coef
		v99_mean_delta delta log energy
		v101_mean_1st delta delta
		v133_std_5th delta
		v134_std_6th delta
		v135_std_7th delta

Tabla 7.2: Tabla resumen del proceso de selección de variables

Método	Selección	Porcentaje reducido contra la muestra original
PCA	54 variables	84%
Correlaciones	37 variables	89 %
LASSO	24 variables	93 %

Tabla 7.3: Tabla resumen de las pruebas realizadas con otro tipo de selecciones

Método	Variables	Modelo	Precisión	Especificidad	Sensibilidad
PCA (90 %)	180	SVM Rad	0.7196	0.8178	0.5563
PCA (90 %) + LASSO	51	SVM Rad	0.7407	0.7839	0.6690

Figura 7.1: Cómo varía la precisión del modelo SVM Radial con nuestras variables, en función del coste