UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Ciencias

Bitácora Seminario de Ciencias de la Computación

Grupo 7128

Seminario de Redes Neuronales y Neuronas

Análisis Predictivo de la Menopausia

Integrantes:

Hernández Norberto Dana Berenice 317163027
Lira González Rosa Linda 318074463

Profesores:
Humberto Andrés Carrillo Calvet
Jose Luis Jimenez Andrade
Luis Fermin Martinez

La menopausia es el momento en la vida de las mujeres qué marca el final de los ciclos menstruales de las mismas, representando no solo un momento fundamental a nivel biológico de una mujer, sino también a nivel social y afectivo, donde cambia la forma en que es percibida por la sociedad, nuestras formas de actuar se articulan por medio de las interacciones sociales que tenemos día con día, con base en las condiciones en las que vivimos y nos desarrollamos, en esta etapa, las interacciones sociales, biológicas y afectivas que tenemos con nuestro entorno, cambian y por lo tanto, nuestra calidad de vida y emocional.

La menopausia aún es considerada una enfermedad para algunos médicos, viviendo de esa forma por el resto de su vida las mujeres, teniendo una repercusión a nivel de percepción de su propio cuerpo y el autocuidado. En todo este discurso cabe la pregunta ¿la menopausia indica algún problema de salud?¿cuáles son los factores que influyen en su llegada y de qué manera?¿se encuentra relacionada con factores únicamente biológicos? Son preguntas que nos resultan interesantes de estudiar, no solo por tener un acercamiento a la medicina, sino también al entendimiento de un proceso natural tan estigmatizado.

El objetivo inicial del proyecto se enfoca en identificar los factores de incidencia de menopausia prematura, con el fin de mejorar no solo la calidad de vida de las mujeres, sino también, desestigmatizar el fenómeno vivido. Para ello planeamos utilizar algoritmos de aprendizaje supervisado y no supervisado, utilizando el aprendizaje supervisado para identificar los factores de incidencia, y el aprendizaje no supervisado para identificar si una persona incide o no en la menopausia, así como diferenciarla de algún desajuste hormonal como el climaterio.

Comenzando con la obtención de datos, encontramos información y artículos científicos relacionados a este tema, relacionados principalmente con el estigma social, la visión de este fenómeno, bases de datos relacionados con la terapia hormonal y los trastornos que este puede tener como consecuencias. Existen resultados previos de los que queremos lograr, debido a que es un tema médico, existen investigaciones biológicas al respecto y diversas teorías, como su correlación con la dieta, ejercicio, profesión e incluso país o clima en el que se desarrolla la persona.

Se tomó la decisión de elegir una base de datos referente a un estudio relacionado, enfocado principalmente al cáncer de ovario, de esa forma, podemos comenzar con el análisis de datos inicial.

Preparación y Análisis Inicial de los Datos

Para nuestro análisis inicial de datos, utilizamos 3 bases de datos de un mismo estudio, con las mismas pacientes pero con variables diferentes. Todos con datos referentes a la química sanguínea, recuentos hormonales y sustancias hepáticas.

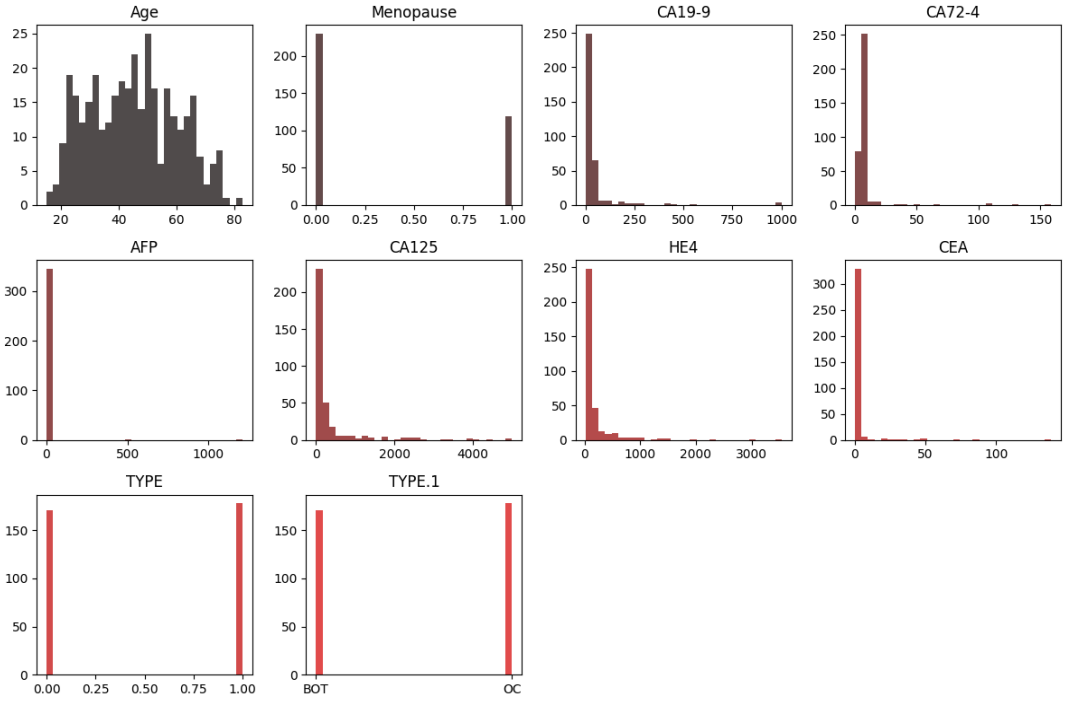
Obtenemos un total de 349 observaciones y 50 variables sin repetición para la base de datos completa. De la misma forma, verificamos la integridad de datos, es decir, identificamos los valores faltantes de la base, asegurando la integridad completa de nuestros datos para análisis posteriores. De esta manera comprobamos que no tenemos ningún dato faltante en ningúna variable, por lo tanto, no es necesario realizar imputación de los mismos.

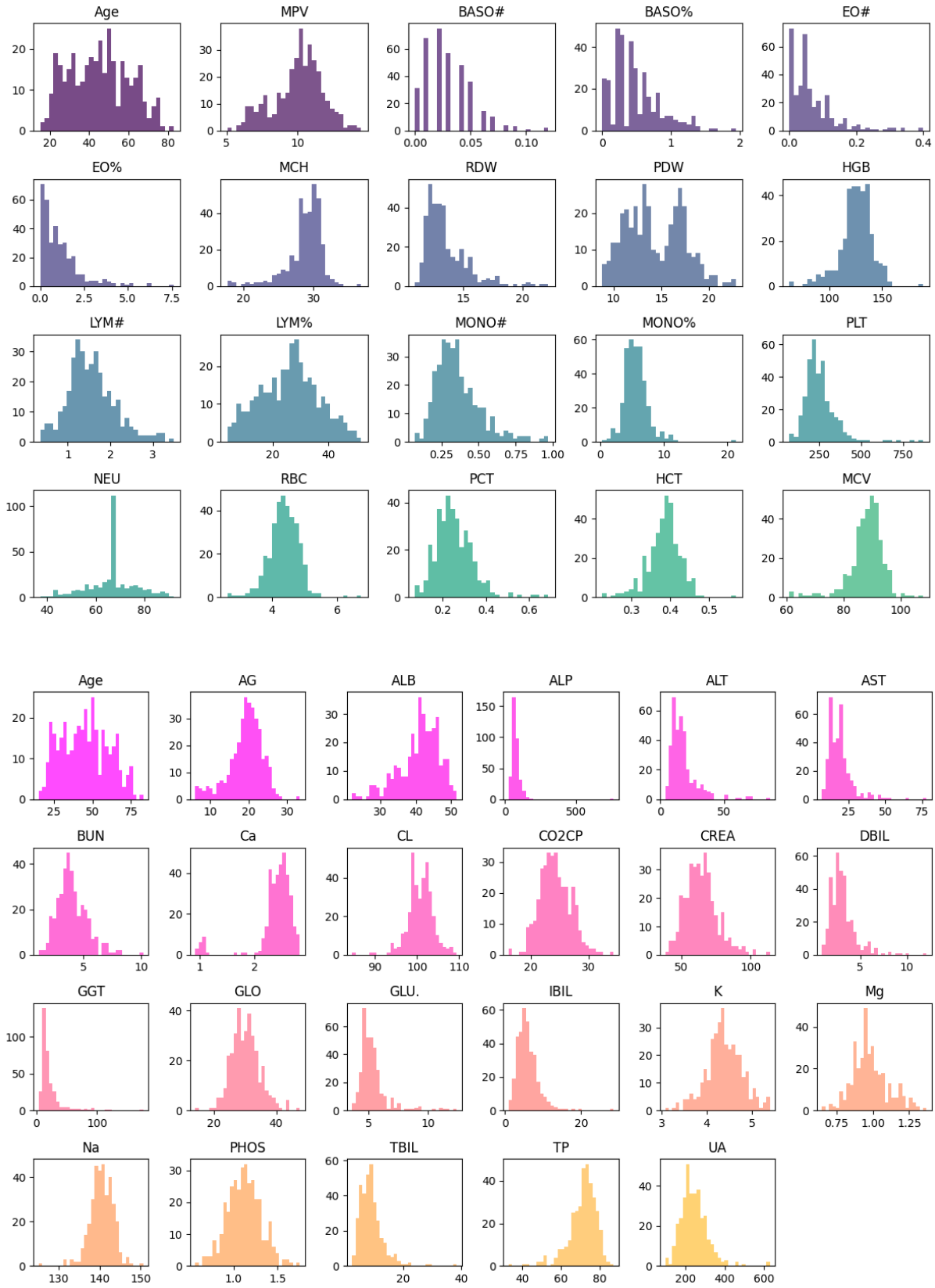
Análisis Exploratorio de Datos (EDA)

Inicialmente, identificamos el tipo de datos que posee cada columna. Obtuvimos que solo tenemos una variable categórica, la variable Type, la cual también se encuentra codificada en la base de datos de forma binaria, con dos tipos de respuesta:

- **BOT (Borderline ovarian tumor)**: Estos tumores también se conocen como tumores ováricos de bajo potencial maligno. Son tumores que se sitúan en un estado intermedio entre los tumores benignos y los malignos. A menudo, los tumores borderline tienen características que pueden ser preocupantes, como crecimiento en el epitelio ovárico, pero no invaden los tejidos circundantes ni se diseminan como lo haría un cáncer ovárico invasivo. Aunque tienen un potencial para volverse malignos, el pronóstico generalmente es mucho mejor que para los cánceres ováricos invasivos.
- **OC (Cáncer ovárico):** Estos son tumores malignos que se originan en el ovario. Pueden ser de diferentes tipos, como carcinoma epitelial (el tipo más común), carcinoma de células germinales, carcinoma de células claras, entre otros. Los cánceres ováricos invasivos tienen la capacidad de invadir los tejidos circundantes y diseminarse a otras partes del cuerpo, lo que los hace más agresivos y potencialmente mortales que los tumores borderline.

Veamos cómo en ambos casos tenemos un sesgo de información debido a que la población de estudio posee una anomalía de carácter médico más que una toma de población en general. De esta forma es posible que podamos redefinir el camino del proyecto hacia nuevos objetivos o encontrar resultados diferentes conforme lo vayamos trabajando. También elegimos dicha base de datos debido a que obtenemos de ella la variable base de nuestro estudio,es decir, si el paciente se encuentra en la menopausia o no es así.
Realizando histogramas de cada variable obtenemos los siguientes resultados:



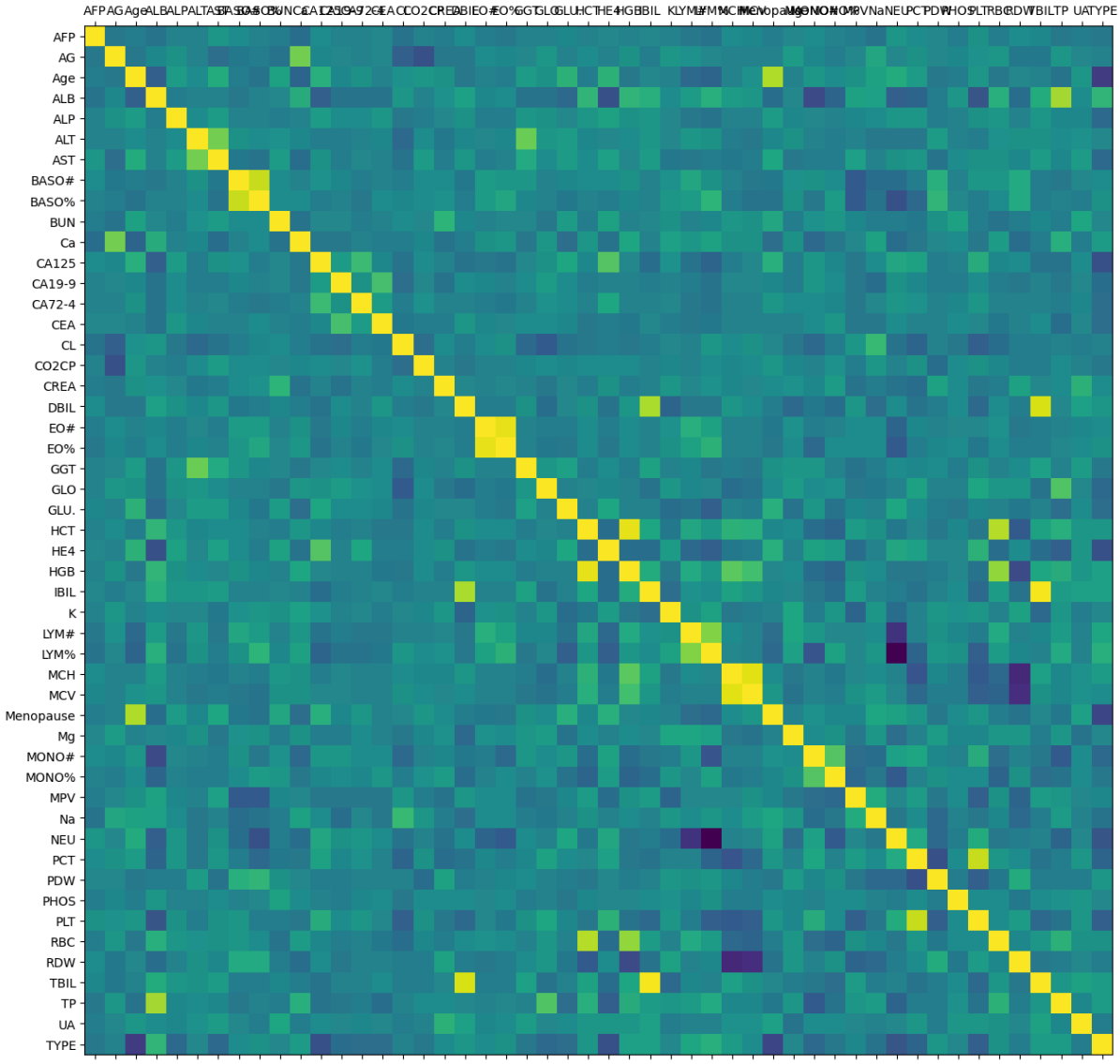


Primeramente, observemos como en la variable de edad obtenemos una distribución similar a una normal, sin embargo, no podemos asegurarlo, para comprobarlo es necesario realizar pruebas de hipótesis. Esto no quiere decir que la edad no está relacionada con la

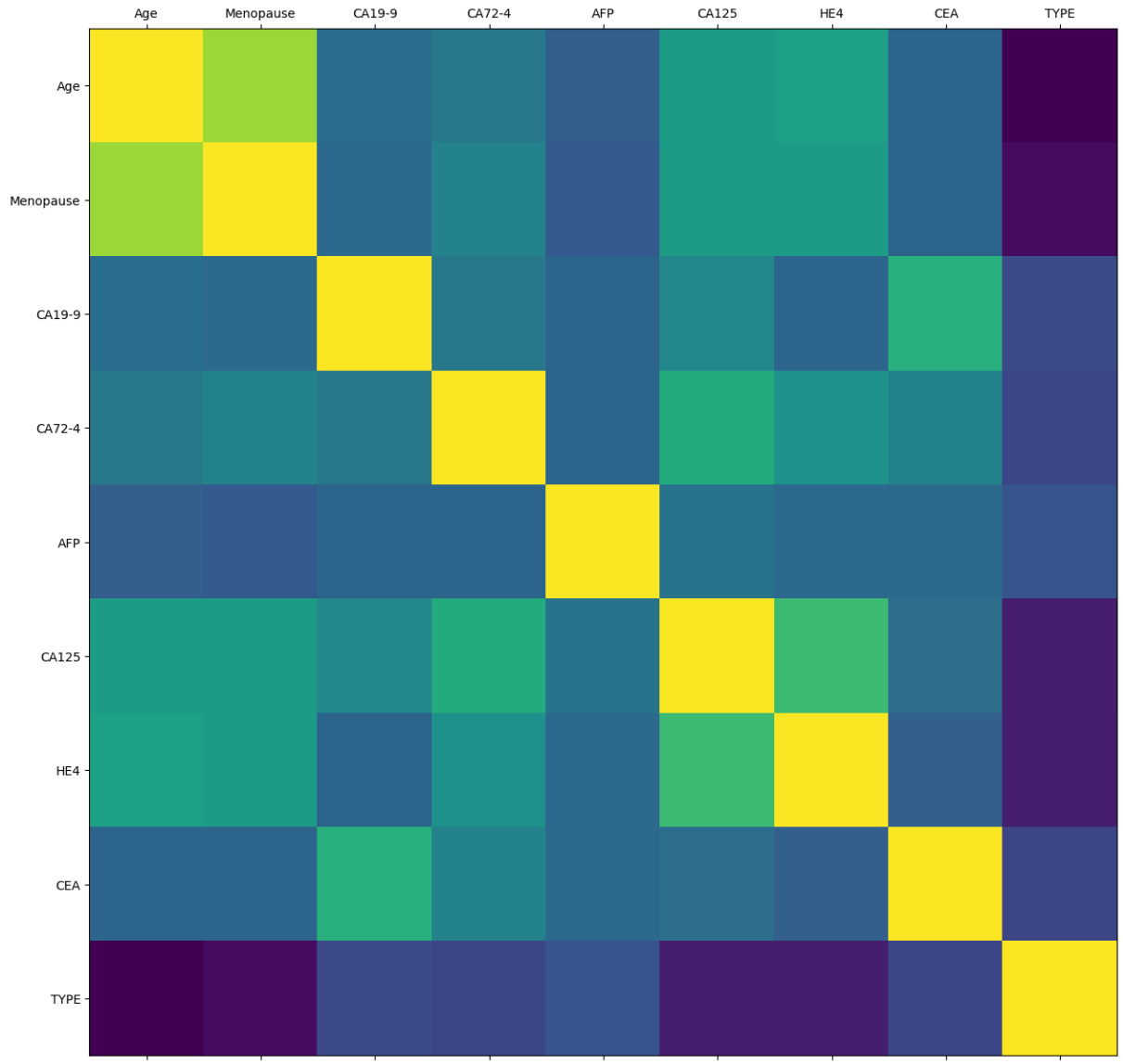
menopausia, sino que la incidencia de cáncer de ovario en las mujeres se encuentra presente en todas las edades, con una concentración particular en las mujeres de 50 años.

En la distribución de los datos de ácidos biliares tenemos un sesgo sumamente notorio hacia la derecha, con datos irregulares mucho mayores a lo que se esperaría. Tomando en cuenta los valores normales de los mismos, estos se pueden considerar como datos atípicos y es necesario borrarlos de nuestra base, con el fin de obtener resultados más acertados.

Ahora, analicemos un poco la correlación de todas las variables de nuestra base de datos con un mapa de calor.



El mapa de calor completo no nos indica demasiado, sin embargo, al obtener un mapa de calor más pequeño podemos visualizar los siguientes resultados:

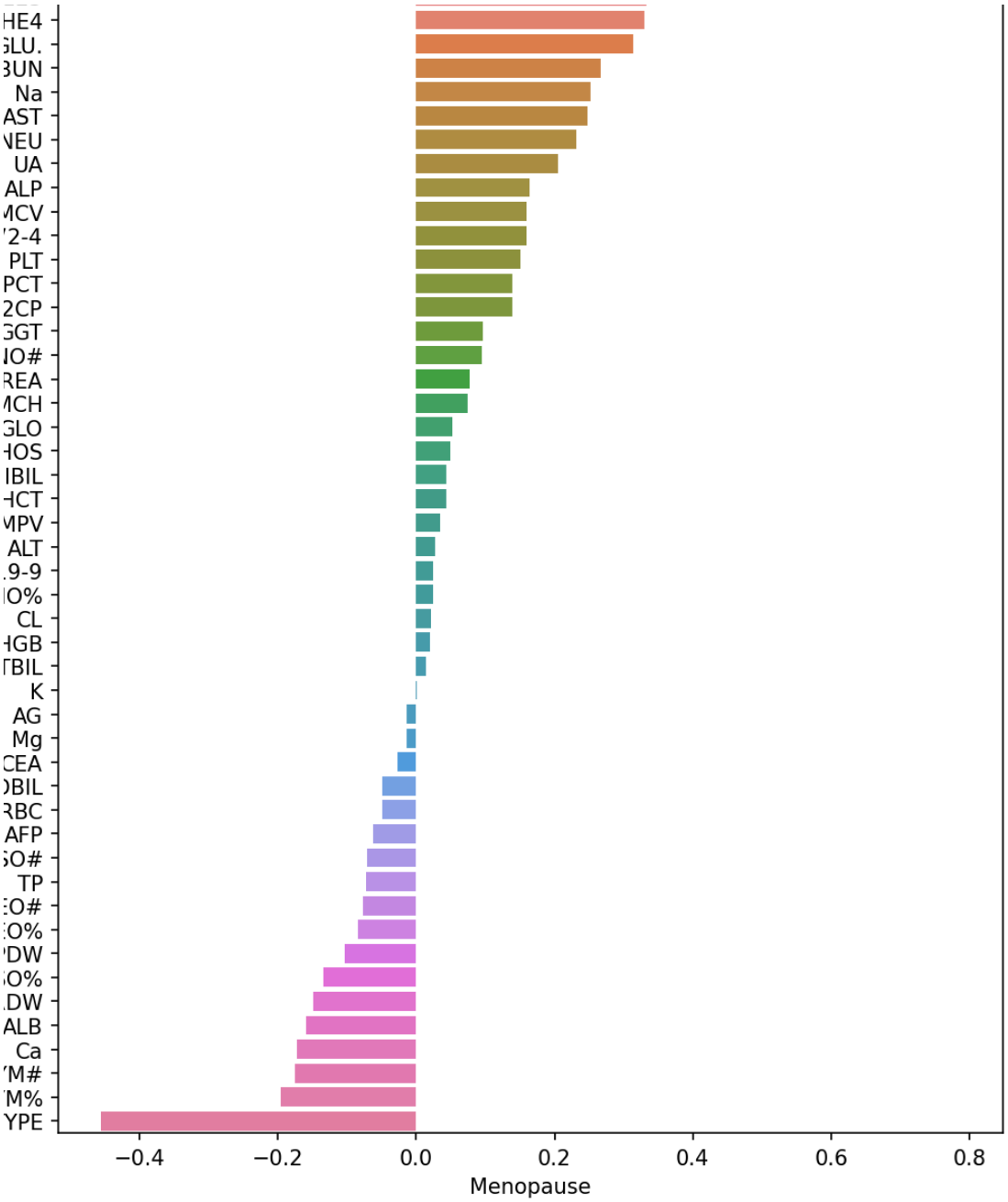


Ahora podemos ver de una forma más clara la correlación fuerte que existe entre la menopausia y la edad, con un coeficiente de correlación de 0.78. Parece ser que no tenemos alguna otra variable que sea relevante en el estudio, no obstante, vale la pena revisar la correlación que posee cada variable con respecto a la menopausia, aplicando el criterio de que el coeficiente de correlación sea mayor a 0.30 obtenemos las siguientes variables a considerar:

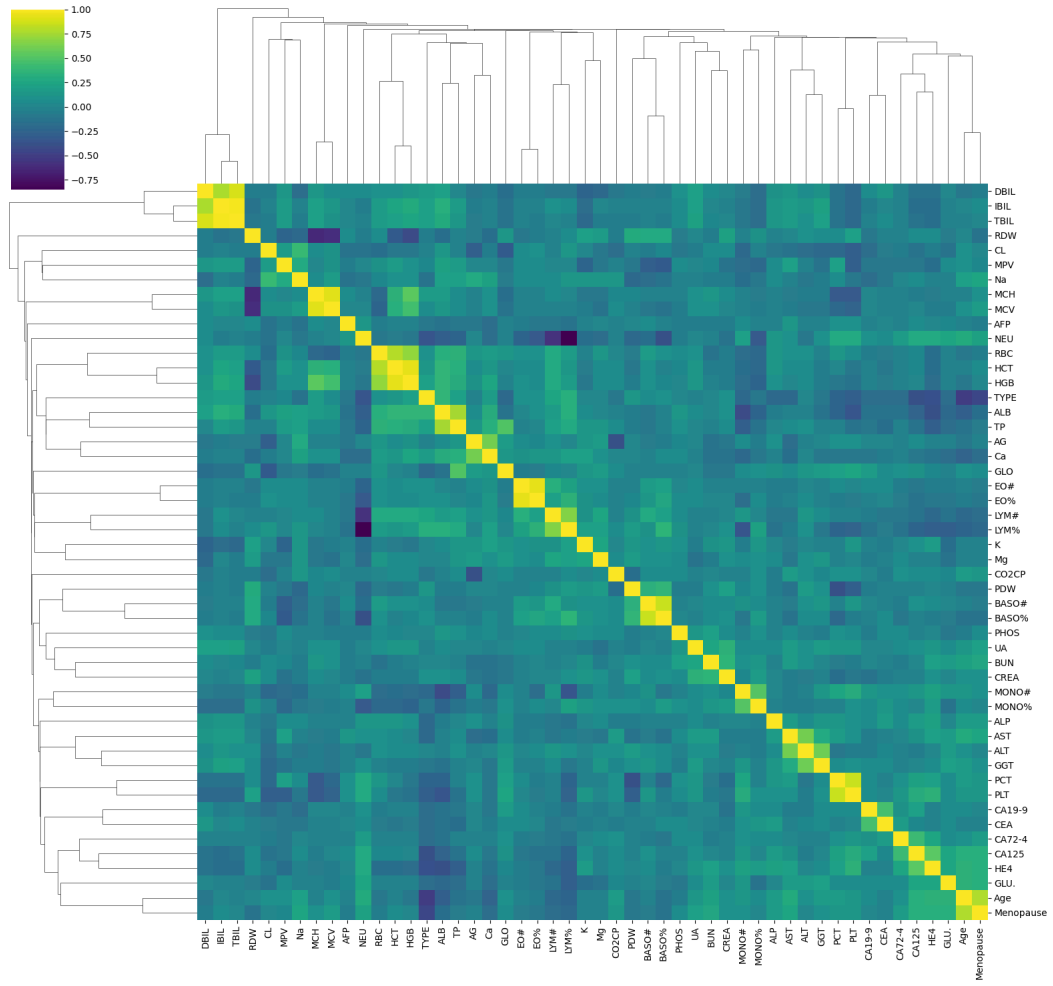
- **Edad:** La edad influye considerablemente en la menopausia, debido a que las mujeres salen de su periodo de edad reproductiva.
- **CA125:** El CA125 es un marcador tumoral que se utiliza en medicina para detectar y monitorear ciertos tipos de cáncer, especialmente el cáncer de ovario. El CA125 es una proteína que se produce normalmente en pequeñas cantidades por las células del revestimiento de los órganos internos, como los ovarios, el útero y las trompas de Falopio. Encontramos un coeficiente de correlación con la menopausia de 0.33.
- **Glucosa:** La glucosa es el azúcar principal que se encuentra en la sangre. Es la principal fuente de energía de su cuerpo, en este caso obtuvimos un coeficiente de

correlación de 0.31. Podemos tener la sospecha de que puede deberse de igual manera a la edad, dado que es más común encontrar a personas de una mayor edad con altos índices de glucosa en la sangre.

- **HE4:** HE4 es una proteína que se conoce como Human Epididymis Protein 4, que se produce en varias partes del cuerpo, incluyendo las células del epidídimo en los hombres y las células del epitelio de las trompas de Falopio en las mujeres. En medicina, HE4 se ha investigado como un marcador tumoral potencial para varios tipos de cáncer, especialmente para el cáncer de ovario. Los niveles de HE4 pueden estar elevados en mujeres con cáncer de ovario, particularmente en ciertos subtipos, como el carcinoma seroso de alto grado. Obteniendo una correlación de 0.32.
- **Type:** Como mencionamos anteriormente, esta variable indica el tipo de tumor que tiene el paciente, obteniendo una correlación negativa de -0.45.



De la misma manera, realizamos un análisis de clusters para identificar las variables que son similares, veamos que en general no tenemos variables altamente correlacionadas que signifiquen cosas similares o dependientes entre ellas. Un ejemplo son las variables de edad y menopausia, altamente relacionadas, sabemos que con el paso de los años la menopausia tiene un mayor índice incidencia, ya que es una culminación de la edad reproductiva de una mujer, por lo que vale la pena considerar la presencia de la variable en el análisis ya que resulta redundante.

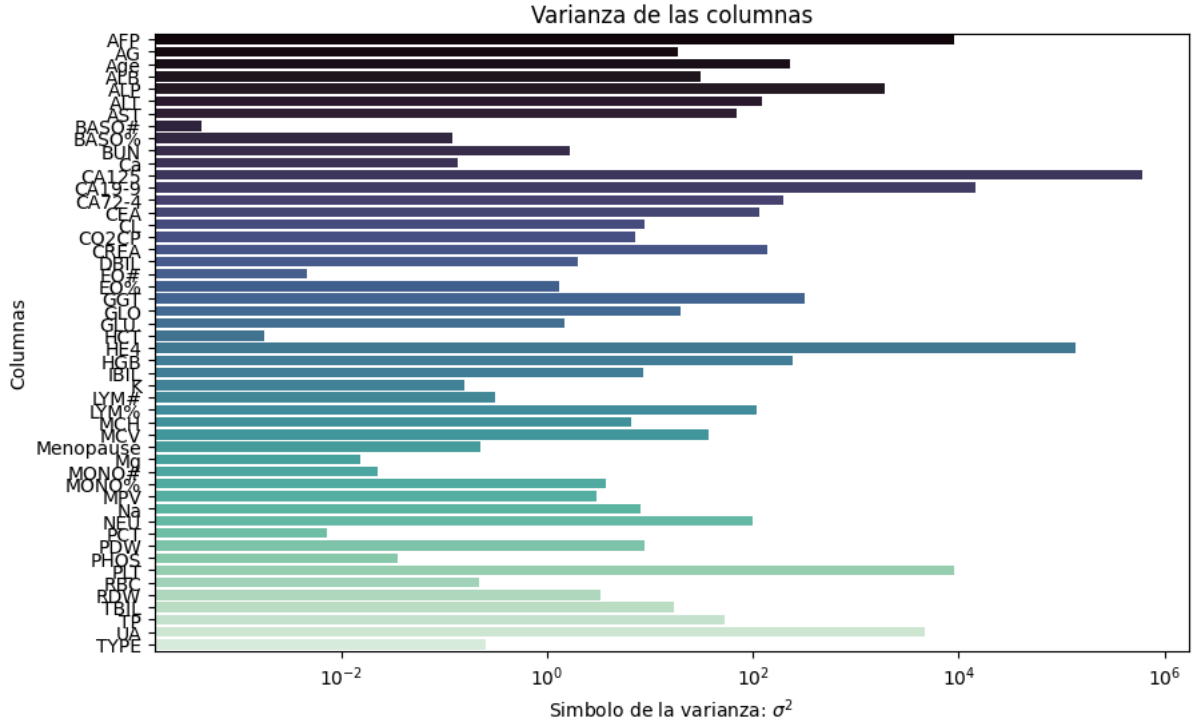


Como conclusión del análisis exploratorio de datos decidimos conservar los datos atípicos debido a que los datos normales se encuentran dentro de un rango normal, por lo que estos datos atípicos podrían otorgarnos información relevante en nuestros resultados finales.

Preparación de datos

En la base de datos utilizada no tenemos datos faltantes por lo cual no es necesario realizar imputación además de que todos los datos ya son numéricos, por lo tanto, no es necesario hacer una conversión de texto a un dato numérico, en el caso de la variable objetivo “Menopause” se encuentra representada de forma binaria con un 0 si no se encuentra en la menopausia y 1 si se encuentra en ella. Con un total de 349 observaciones, tenemos a 230 mujeres que no se encuentran en la menopausia y 119 que sí, información que resultará relevante al seleccionar el conjunto de entrenamiento y prueba.

Continuamos el análisis con la varianza de las variables, estos datos nos permiten conocer qué tan volátiles son y además si tenemos diversidad de datos, es decir, si no tenemos el mismo dato para todas las observaciones de una variable. Realizando una gráfica de varianza de las columnas podemos observar como existen algunas variables con poca diversidad, por ello, tenemos que identificarlas y establecer un umbral para retirarlas o no de nuestra base.



Veamos como una variable altamente correlacionada, el CA125 posee la mayor varianza de nuestro dataset, esto puede ser debido a los datos atípicos que posee. Ahora bien, como mencionamos anteriormente, determinamos cuales son las variables con poca varianza, con un umbral de 1.

La variable objetivo aparece dentro de ellas, sin embargo, no la borraremos por razones obvias, con respecto a las demás, analizaremos si estas variables están poco o nada relacionadas con la variable objetivo, con el fin de conocer si podemos y es conveniente retirarlas de nuestra base de datos.

BASO#	0.000436
BASO%	0.119875
Ca	0.133206
EO#	0.004584
HCT	0.001779
K	0.157448
LYM#	0.315225
Menopause	0.225357
Mg	0.015288
MONO#	0.022651
PCT	0.007236
PHOS	0.035212
RBC	0.217057
TYPE	0.250618

Determinamos que las variables con una correlación menor a 0.3, mayor a -0.3 y que además tuvieran una varianza muy baja, las retiraremos de nuestra base, debido a que no se les considera relevantes en el estudio, nos quedamos con 38 variables.

Ahora, con respecto a la transformación de datos, hemos elegido utilizar como modelo de aprendizaje Random Forest, para ello estandarizamos nuestros datos con StandarScaler(), ya que creemos que de esta forma obtendremos mejores resultados a pesar de que Random Forest no es tan sensible a los datos atípicos.

Ajuste del modelo

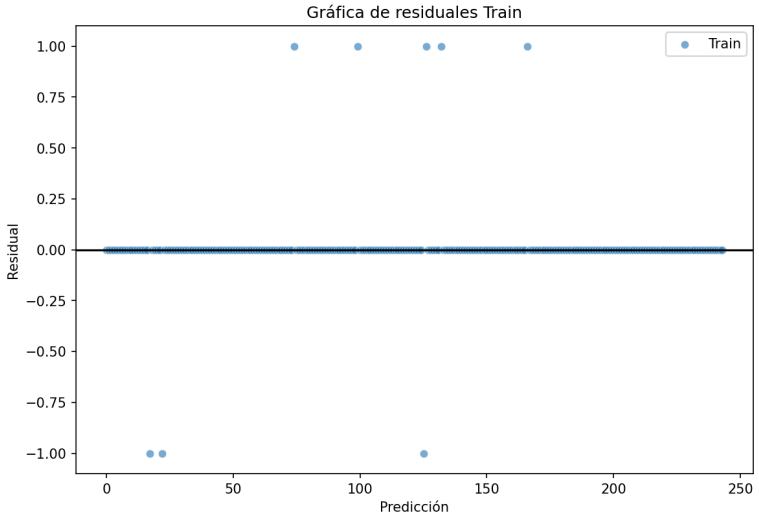
Como mencionamos anteriormente, elegimos como modelo de entrenamiento principal el algoritmo de Random Forest, no obstante, de igual manera lo compararemos con el algoritmo de regresión logística, con el fin de reconocer cuál es el modelo más adecuado para cumplir con nuestro objetivo.

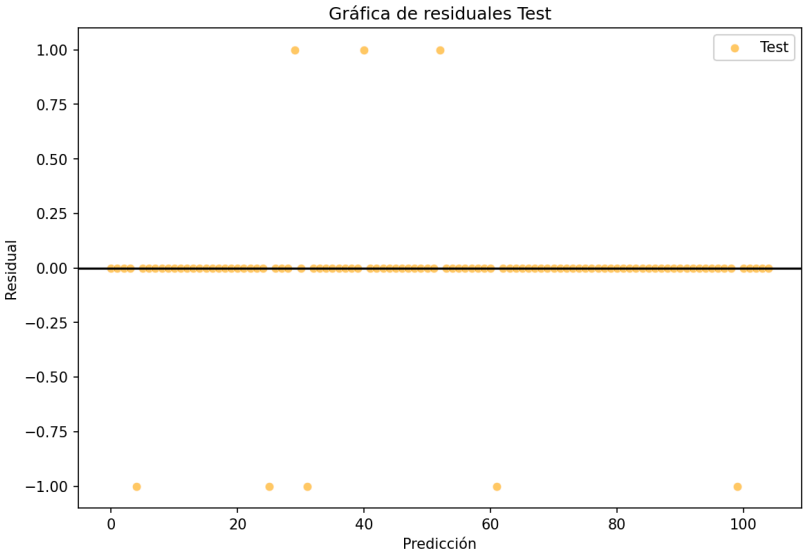
Anteriormente pudimos darnos cuenta de que el 65% de las pacientes no se encontraban en la menopausia, con ello podemos concluir que nuestros datos se encuentran sesgados hacia las pacientes pre-meno o no menopausia. Por ello decidimos utilizar el muestreo estratificado para seleccionar nuestros conjuntos de entrenamiento y prueba, seleccionando el 70% como el conjunto de entrenamiento.

Regresión logística

El algoritmo de regresión logística es bueno para nuestro modelo debido a que nos permite modelar respuestas categóricas, en este caso, binaria. Nos ayudará a determinar a qué categoría pertenece una observación dada, en este caso, si el paciente se encuentra o no en la menopausia.

Procedimos con el ajuste de hiperparámetros usando *GridSearchCV* y *RandomizedSearchCV*. El Grid Search proporcionó un modelo con un score de validación cruzada de 95.49%, mientras que el Randomized Search mejoró ligeramente este score a 96.72%. Obtuvimos un error absoluto medio de 0.032 para el conjunto de entrenamiento y 0.076 para el conjunto de prueba.





Veamos que ya con los hiperparametros ajustados, tenemos muy buenos resultados, reflejados tanto en las gráficas de residuales, como en el error absoluto medio obtenido.

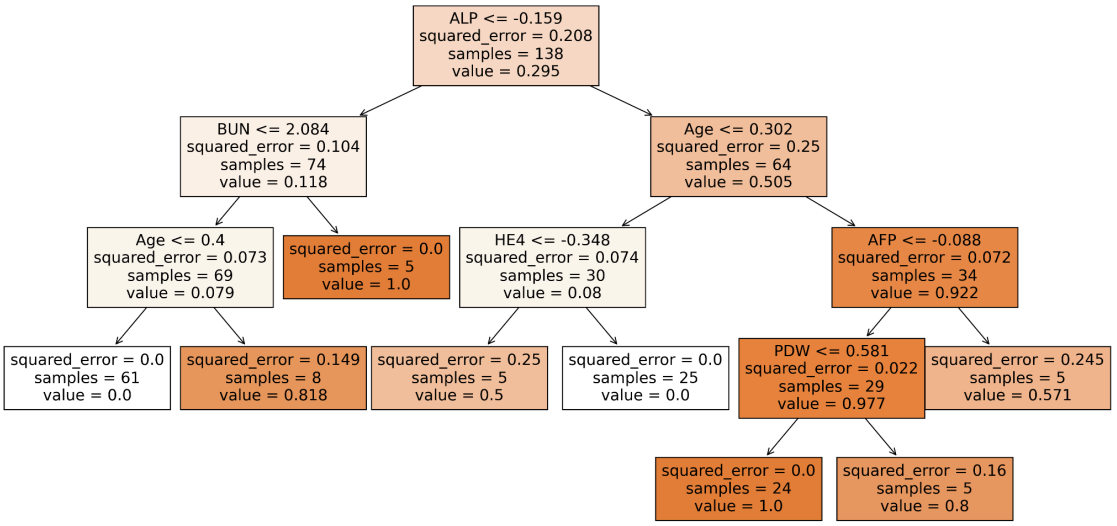
Random Forest

El algoritmo Random Forest es un algoritmo de generación de múltiples árboles de decisión, estos suelen ofrecer una alta precisión en la clasificación debido a la combinación de múltiples árboles de decisión. Cada árbol contribuye con su clasificación y la decisión final se toma por votación mayoritaria, lo que reduce el riesgo de sobreajuste, además es menos sensible a los datos ruidosos y a los valores atípicos.

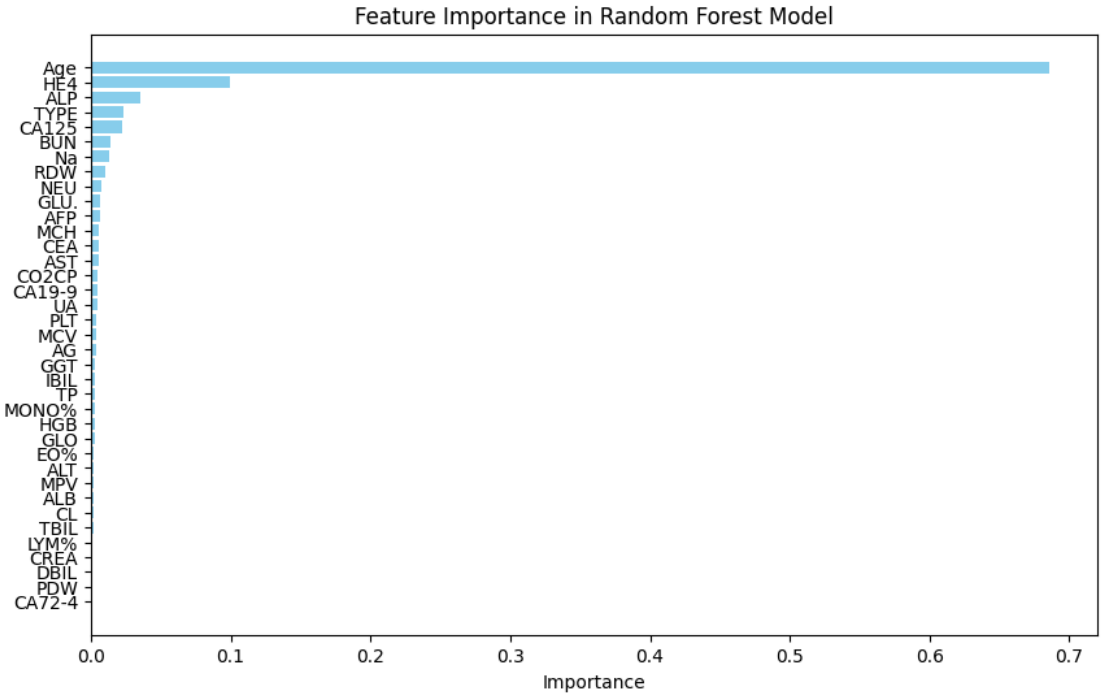
Una desventaja puede llegar a ser la gran cantidad de hiperparametros que tenemos que ajustar, los hiperparametros que utilizamos para el modelo son los siguientes:

- **n_estimators:** Número de árboles para crear en el algoritmo.
- **max_depth:** Profundidad máxima de los árboles.
- **min_samples_leaf:** Mínimo de muestras en una hojas.
- **max_features:** Máximo de variables a considerar en cada split.
- **max_samples:** Máximo de muestras a considerar.
- **random_state:** Semilla.
- **n_jobs:** Uso de todos los núcleos.

Establecimos hiperparámetros iniciales, basandonos en el criterio del libro “Hands on machine learning”, donde recomiendan inicializar 10 veces el número de variables que tenemos en nuestra base de datos árboles aleatorios, por lo tanto, ya que tenemos 37 variables, inicializamos 370 árboles. De esa forma, imprimimos el primer árbol del modelo entrenado.

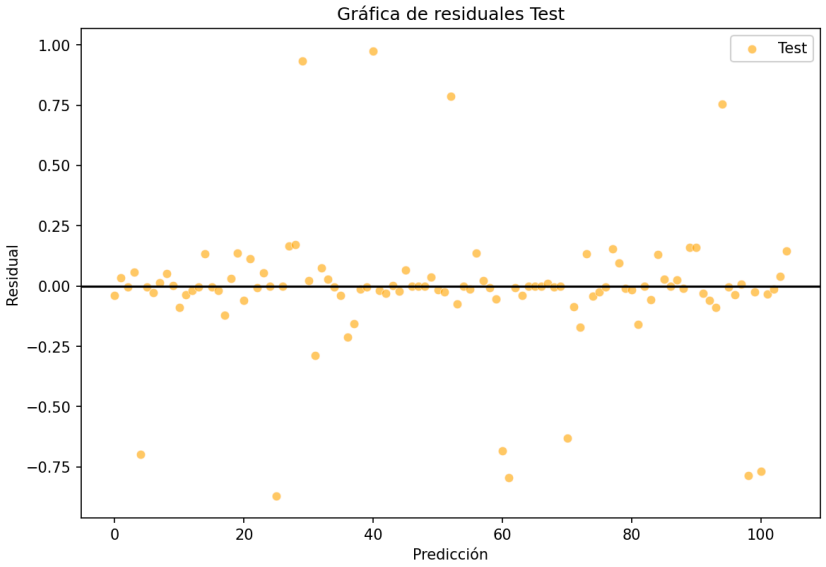


Obtenemos de esa forma la importancia de cada una de las variables de nuestra base de datos.



Observemos como la variable de edad es la que toma una mayor relevancia, por otro lado, también existen las variables HE4 y ALP. En este caso, una variable la cual habíamos determinado que poseíamos una correlación importante era la HE4, con ello tenemos pruebas para considerar relevante esta variable en el fenómeno de estudio. Con un error medio absoluto de 0.093 para el conjunto de entrenamiento y 0.14 en el conjunto de prueba

Continuamos con el ajuste de hiperparametros para mejorar los resultados obtenidos, usando como en el caso anterior *GridSearchCV,* usando como scoring el error absoluto medio, con el fin de comparar ambos modelos correctamente. Calculando nuevamente el error absoluto medio de nuestro modelo, ahora con los hiperparametros ajustados obtenemos un error de 0.075 para el conjunto de entrenamiento y 0.12 para el conjunto de prueba, mejorando de esa forma los resultados.



Resultados

Obtuvimos como resultado que el HE4 es una proteína importante en la incidencia temprana o general de la menopausia. El HE4 se encuentra siempre en la sangre y se encuentra en niveles altos cuando existe el riesgo de sufrir cáncer de ovario. Junto con la proteína CA125 son indicadores importantes, no existe información científica que nos permita determinar qué tipo de factores inducen el cáncer de ovario, fuera de una predisposición genética.
Sin embargo, se indica que el HE4 puede ser reducida o controlada con el consumo de alimentos antiinflamatorios como el pescado graso, nueces, y aceites saludables, además del consumo de alimentos ricos en fitoestrógenos como los garbanzos o las semillas de calabaza.

Conclusiones

A pesar de las mejoras en el ajuste de hiperparámetros, el modelo requiere ajustes adicionales para mejorar su capacidad de generalización. Sugerimos explorar más modelos y técnicas de feature engineering.

¿Qué modelo creemos mejor?:

- **Regresión Logística vs. Random Forest:**
  - La Regresión Logística ofrece una mayor interpretabilidad, lo cual es beneficioso cuando es importante entender la influencia directa de las variables.
  - Random Forest demostró una capacidad superior para manejar la complejidad de los datos sin un preprocesamiento intensivo, ofreciendo robustez y consistencia en el rendimiento.
  - La elección entre los dos modelos depende de las prioridades del proyecto, ya sea la necesidad de explicar los resultados del modelo o la prioridad de maximizar la precisión y la robustez.

Ambos modelos son viables, si la prioridad es la interpretabilidad y la simplicidad del modelo, la Regresión Logística podría ser más adecuada. Por otro lado, si la prioridad es la robustez, el manejo de no linealidades, y la capacidad de capturar complejidades en los datos sin un preprocesamiento extensivo, el Random Forest sería más eficiente. Por lo tanto, Random Forest podría ser nuestro preferido para situaciones que requieren un manejo robusto de complejidades en los datos.

Referencias:

- http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0036-36342001000500004
- http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0026-17422018000200051
- http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1029-30192014001000011#:~:text=La%20menopausia%20no%20es%20una,negativa%2C%20depende%20del%20contexto%20social.
- https://www.emergenresearch.com/industry-report/menopausal-therapy-market
- https://es.statista.com/estadisticas/581668/numero-de-mujeres-con-trastornos-de-la-menopausia-por-edad-espana/
- https://www.kaggle.com/code/swabbie8/dt-ovarian-cancer-bio-marker/input?select=OC_Marker.csv
- https://scikit-learn.org/stable/
- *"Python Machine Learning" por Sebastian Raschka y Vahid Mirjalili*
- *https://github.com/danahdzn/Analisis-predictivo-de-menopausia*