En esta primera parte se lleva a cabo el estudio sobre la predicción de cáncer haciendo uso de múltiples modelos de aprendizaje supervisado y no supervisado.

Nuestra investigación parte del estudio recogido en el artículo x donde establecen la modelo cancerA1De como modelo a raíz de su trabajo.

Nuestro objetivo es retomar los pasos donde dejaron el estudio e intentar profundizar utilizando nuevos modelos para mejorar los resultados que obtuvieron. Para ello compararemos los resultados obtenidos. También replicaremos alguno de los pasos seguidos para entender como responden sus algoritmos y a su vez partir de una misma base para poder llevar a cabo una comparación equitativa entre los resultados.

Esta primera parte de detección de cáncer se divide en 3 partes: el análisis y procesamiento de los datos y el entrenamiento de los modelos de aprendizaje supervisado y los modelos de aprendizaje no supervisado.

**Datos**

Y los datos utilizados para este estudio conforman un conjunto de 1817 registros con 32 variables.

Limpieza de los datos.

El conjunto de datos inicial se Encuentran bastante pulidos. Esto quiere decir que apenas hay valores nulos y únicamente encontramos hago un carácter indeseado fácilmente detectable y suprimible. Probamos varias técnicas de relleno de nulos pero vemos que finalmente lo más conveniente es la de la media. Llevamos a cabo una transformación del tipo de los datos para facilitarnos su posterior análisis.

Análisis de los datos

Para hacer el modelo más simple, manejable y por lo tanto más escalable aplicamos un algoritmo information gain para detectar qué variables tienen mayor peso a la hora de predecir el cáncer y reducir el total de los datos a un conjunto menor. Como resultados obtenemos Que los marcadores tumorales como CA-19-9, CA-125 y PSA, Se encuentran entre las variables que mayor peso tienen a la hora de predecir el cáncer. A su vez, para confirmar el resultado obtenido hallamos la correlación que existe entre las variables independientes y la variable objetivo aplicando un modelo de random forest. El resultado es sorprendente ya que, En este segundo enfoque, los marcadores tumorales dejan de ser los principales componentes predictivos del cáncer y pasan a serlo otros biomarcadores A priori secundarios. Esto refleja que se dan relaciones subyacentes Entre los distintos biomarcadores mostrando que detrás se encuentra una compleja relación entre ellas.

Uno de los objetivos que establecemos es verificar cuál de los dos enfoques ofrece un mejor resultado a la hora de predecir el cáncer. De esta manera, aplicaremos para cada 1 de los modelos ambos enfoques comparando al final cuál rinde mejor.

Para entender mejor las relaciones que existen entre las variables mostramos sus correlaciones. De acuerdo con nuestra creencia general, los marcadores de antígenos del cáncer están correlacionados positivamente con significancia estadística (p < 0,001). Curiosamente, se puede observar que TIMP-1, mieloperoxidasa (o MPO), OPN (o SPP1) y HGF forman un grupo de características correlacionadas positivamente. En el estudio original hacen uso del análisis de red STRING Para evidenciar las relaciones existentes entre estos cuatro biomarcadores Observando que se producen interacciones entre ellas

Transformación de los datos

* Omitimos el uso de la discretización de características usado en CancerA1DE y CancerA2DE basada en el principio MDL, donde La idea central es que el mejor modelo para una serie de datos es el que minimiza la longitud total de la descripción del modelo y los datos dados el modelo. Este procesado hace sus modelos sobresalientes ya que se trata de un enfoque donde se da mayor relevancia al intervalo que a la propia magnitud. No nos ha sido posible implementar El algoritmo MDL como herramienta para encontrar los mejores puntos de corte que minimicen la longitud de la descripción del modelo y los datos. A su vez investigamos otro posible enfoque mediante la utilización del módulo MDLP que permite la discretización de datos continuos pero, debido a un problema de compatibilidad de versiones con Python no pudimos llevar a cabo su implementación. Creemos que la implementación de una discretización sea para cualquiera de los dos enfoques aporte una mejoría notable en los resultados de los modelos.

Por tanto en cuanto a procesamiento de los datos acabamos optando por un enfoque clásico aplicando una estandarización para las variables numéricas Y un algoritmo de One hot encoder para las variables categóricas.

Más adelante observaremos que no aplicar una discretización de las variables conlleva una fuerte penalización en el rendimiento de los modelos a la hora de predecir, haciendo que estos no generalicen correctamente. A su vez ocasiona que los modelos sean más complejos y de manera indirecta influya generando mayor overfitting de estos.

Por tanto, decidimos aplicar ciertas técnicas para mitigar este sobreajuste. Para empezar, aumentamos el número de ‘folds’ de la validación cruzada, pasando a ser de 10. Por otra parte, añadimos un paso adicional al preprocesado de los datos, Una discretización haciendo uso de un árbol de decisión utilizando la librería DecissionTreeDiscretiser. Con ellos esperamos maximizar la ganancia de información y reducir la impureza en los nodos de forma que los puntos de corte generados por el árbol se utilicen para dividir la variable continua en intervalos discretos. Después de entrenar el árbol y generar los puntos de corte cada valor continuo se asignará a 1 de los intervalos generados. Con ello esperamos Reducir el abanico de valores de los datos, simplificar Y mejorar el rendimiento de los modelos y acelerar el proceso de entrenamiento. Utilizando este método esperamos capturar relaciones lineales entre la variable continua y la variable objetivo de forma que la discretización sea más informativa.

Por último, llevamos a cabo una división del conjunto de datos en entrenamiento validación y test en porcentajes de 80, 20 y 20 respectivamente. Y vamos a cabo esta división para poder ajustar los hiper parámetros en la validación y a posteriori comprobar si el modelo generaliza bien con el conjunto de datos de test.

**Modelos supervisados**

La siguiente parte del proceso trata del ajuste de los modelos supervisados. Para poder clasificar y comparar de una manera efectiva cada 1 de los modelos resultantes vamos guardando las métricas finales de cada 1 en una tabla dividida por tipo de modelo fase tipo de enfoque y las métricas resultantes. A su vez hemos creado una variable global score que es el resultado de la suma ponderada de cada una de las métricas que definen los modelos. Se trata de una variable orientativa no oficial que usamos para valorar el modelo de una forma rápida.

Para el ajuste de los hiper parámetros de los modelos utilizamos Algoritmos como GridSearchCV o RandomizedSearchCV. Hemos ido variando su uso dependiendo de la necesidad pero sobre todo en función de la cantidad de hiper parámetros presente para la búsqueda ya que RandomizedSearchCV a nivel computacional es menos demandante

Una vez llegado para cada modelo a un resultado satisfactorio estos son guardados en local mediante la librería Joblib, De manera que no tengan que volver a ser entrenados de nuevo.

Una vez terminados los modelos se observan que los resultados son positivos, pero al analizar con detenimiento algunas métricas, en particular el área bajo la curva AUC, se tiene que las gráficas de cada 1 de los modelos son muy picudas sin una transición paulatina. Esto en términos generales quiere decir que los modelos se encuentran sobre ajustados.

Como se comentó previamente en el tratamiento de los datos, se añade un paso más en el proceso de preprocesamiento de los datos, una discretización de las variables a través de un árbol de decisión.

**Resultados**

Los resultados de accuracy y área bajo de la curva AUC se muestras a continuación:

Al utilizar técnicas para mitigar el overfitting como, Rigde y Lasso (técnicas de regularización), cross validation de hasta 10 folds y discretización de los datos.

Hemos probado individualmente cada técnica y juntándolas todas.

* Rigde y Lasso no presentan una mejora significativa
* Cross validation mejora ligeramente el overfitiing
* Discretización de las variables presenta una mejora notoria

Hay una mejora significativa una vez se discretizan los datos.

|  |  |
| --- | --- |
| Sin discretizar | Discretizado |
| Tiempos de entramiento altos | Tiempos de entramiento bajos |
| Mayor complejidad de los modelos | Menor complejidad de los modelos |
| 4 modelos con muy buenas métricas | 7 modelos con muy buenas métricas |

Comparar enfoque 1 vs enfoque 2 : El enfoque 2 presenta una mejoría de en torno a 4 puntos de global score para todos los modelos. Por tanto, el conjunto de datos del enfoque 2 es mucho más informativo para la predicción del cáncer.

Comparar con el paper.

**Conclusiones**

Se han obtenido resultados alentadores, con modelos con métricas de evaluación bastante positivas y que a su vez generalizan medianamente bien en la fase de test.

Por regla general, el enfoque dos del estudio obtiene mejores resultados a la hora de predecir el cáncer.

Siendo los modelos bastante simples y poco costosos a nivel computacional, vamos a probar a combinarlos con modelos de aprendizaje no supervisado para ver si mejoran los resultados o encontramos alguna combinación que reduzca el overfitting presente en alguno de los modelos.

Para ellos vamos a preprocesar los datos usando los modelos no supervisados de forma que podamos explotar diferentes aspectos de los datos:

- clustering puede descubrir estructuras ocultas

- la reducción de dimensionalidad puede simplificar la información

- los autoencoders pueden encontrar representaciones más efectivas de los datos

De esta manera vamos a crear primero los modelos no supervisados y una vez ajustados, procederemos a combinarlos con los supervisados.

**Modelos no supervisados**

Tras analizar las métricas de los modelos no supervisados, no se cree conveniente profundizar el estudio por esta rama. Esto se debe a que los clústeres generados no generan agrupaciones de datos que respondan a patrones ocultos dentro de los datos. Por tanto empeora el hecho de preprocesar los datos mediante estos modelos, para luego ser explotados por los métodos supervisados desarrollados en la primera parte.

La reducción de componentes mediante estos algoritmos tampoco ofrece una ventaja considerable ya que, de por si, tenemos pocas variables al empezar el estudio de esta primera parte. Una reducción mayor resultaría en una pérdida de información valiosa que intercedería en una buena predicción del resultado.

No se han llegado a observar características relevantes que puedan contener los datos a través de los modelos creados, pero existe la posibilidad de que las haya. Mediante un ajuste más fino de los hiperparámetros y una investigación más profunda de los marcadores biométricos que conforman el estudio, es posible llegar a conclusiones que ni si quiera hemos llegado a atisbar durante el proceso.

Dejamos por tanto esta parte para ser explotada con mayor profundidad por futuros estudios sobre la materia.

**Resultados**

Mostrar tabla de resultados y comentar brevemente

**Aplicación Cáncer**

Explicación

Interfaz

Funcionamiento (modelos, generación de datos aleatorios bajo modelo CTGAN entrenado)

**Conclusiones generales**

- Partimos de la base de que los datos del estudio son de muy buena calidad y preparados "a mano" escogiendo las variables explícitamente para el estudio. El problema reside en que se tiene una muestra pequeña de datos de muy alta calidad. Sería conveniente tener acceso o poder recolectar más datos para poder entrenar los modelos y hacerlos más robustos y fiables.

- En caso de no tener acceso a nuevos datos reales, existe la posibilidad de generar nuevos datos sintéticos con el modelo CTGAN a expensas de acabar sesgando las predicciones del modelo ya que, a pesar de generar datos con distribuciones parejas a los datos originales, no se podrá replicar por completo la compleja aleatoriedad ni las relaciones que subyacen entre los biomarcadores.

- En cuanto al procesamiento de los datos:

- Se lleva a cabo un estudio de relleno de los valores nulos mediante distintas técnicas pero, al haber pocos registros faltantes, optamos por técnicas básicas para el relleno. Quizá exista la posibilidad de mejorar resultados abordando el sujeto con otro enfoque.

-

- Hubiera sido conveniente replicar el modelo CancerA1DE y CancerA2DE utilizados en el estudio original por comprender mejor el planteamiento, su funcionamiento y rendimiento. Tambén para poder compararlos de manera 'local' frente a nuestros modelos. Debido a dificultades técnicas y falta de tiempo para afrontarlas, no nos ha sido posible.

- Modelos y datos :

- Hay una tendencia general en los modelos a tender al sobreajuste debido a la restringida muestra de datos del estudio.

- Partimos de la base de que los datos del estudio son de muy buena calidad y preparados "a mano" escogiendo las variables explícitamente para el estudio. El problema reside en que se tiene una muestra pequeña de datos de muy alta calidad. Sería conveniente tener acceso o poder recolectar más datos para poder entrenar los modelos y hacerlos más robustos y fiables.

- En caso de no tener acceso a nuevos datos reales, existe la posibilidad de generar nuevos datos sintéticos con el modelo CTGAN a expensas de acabar sesgando las predicciones del modelo ya que, a pesar de generar datos con distribuciones parejas a los datos originales, no se podrá replicar por completo la compleja aleatoriedad ni las relaciones que subyacen entre los biomarcadores.

- Se puede continuar el estudio tanto ampliando la oferta de modelos como, ahondar en los estudiados para ajustar más finamente los hipérparámetros.

- De igual manera, pueden añadirse más modelos competentes al vosting classifier y así hacerlo más robusto.

- Trás el estudio, llegamos a la conclusión de que los modelos no supervisados no presentan una mejora significativa para la predicción del cáncer. Un posible enfoque es desechar esta conclusión y profundizar en los modelos no supervisados y su posible influencia en la predicción.

- Vemos que en la aplicación final, para cada modelo, la probabilidad difiere considerablemente según qué registros de entrada. Mantenemos que puede deberse a un sobreajuste de los modelos qué, aún obteniendo una misma predicción, la probabilidad de esta difiere quizás en demasía.

Imágenes interesantes:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

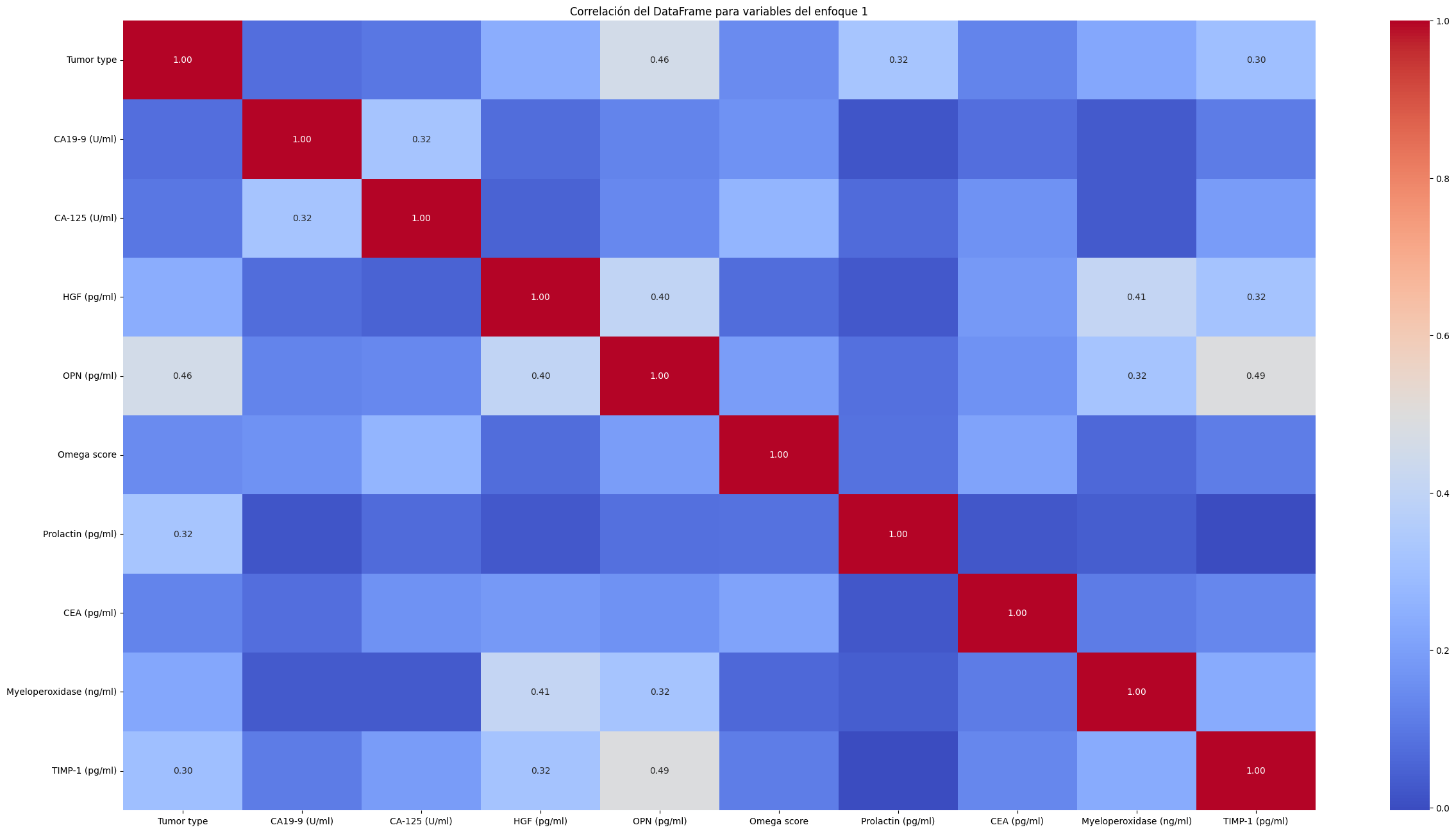


Imagen que contiene Escala de tiempo

Descripción generada automáticamente