UNIVERSIDAD
COMPLUTENSE
MADRID

ntic
master
revolucionamos la comunicación

MÁSTER BIG DATA Y DATA SCIENCE

TRABAJO FIN DE MÁSTER

Signos biológicos en sujetos fumadores

AUTOR: Hassan Chafi Xavier

Contenido

1. Análisis preliminar

El siguiente trabajo tiene como objetivo la identificación de personas fumadoras a partir de un conjunto de datos. Este conjunto de datos del que se dispone se trata de características biológicas para las cuales se han observado su valor para un conjunto amplio de individuos. En concreto se tienen 27 características las cuales adoptan un determinado valor para 55.692 individuos diferentes. Dichas características se tratan principalmente de marcadores biológicos tales como hemoglobina, glucosa en ayunas, colesterol… (**CONSULTAR ANEXO**$_1$). Gracias a la medición de estos marcadores de manera objetiva y que nos indican características diferenciales entre individuos, se puede trabajar en un modelo de predicción basado en un algoritmo de aprendizaje supervisado para un problema de clasificación binario (**CONSULTAR ANEXO**$_2$) capaz de realizar una selección de dichos individuos identificando qué sujetos son fumadores y cuales no lo son aprendiendo de la experiencia pasada para que, al proporcionarle nuevas observaciones de nuevos individuos, identifique correctamente si un determinado sujeto es o no es fumador.

2. Lectura y visualización de los datos

Nuestro problema comienza con una lectura de los datos para conocer qué características presentan los mismos, así como una amplia visualización sobre estos datos de manera que obtengamos una amplia visión de ellos y tener así el máximo conocimiento del problema que tenemos entre manos. En primer lugar, vemos que nuestro conjunto de datos dispone de 55.692 filas que no son más que observaciones, es decir, tenemos tal cantidad de individuos observados y para cada uno de esos individuos tenemos 27 columnas que no son más que características (variables). En resumen, diríamos que se han observado 27 características distintas de 55.692 individuos diferentes.

Existe cierto desbalanceo entre la cantidad de individuos que fuman y los que no fuman si observamos la proporción que representan cada uno de ellos, del total de individuos que tenemos, los no fumadores representan el 63,27% mientras que los fumadores son el 36,73%, esto hace que nuestro futuro modelo de predicción reconozca mejor a los individuos no fumadores que a los fumadores confundiéndose más a la hora de clasificar correctamente a los individuos fumadores. Esto se produce porque al tener mayor cantidad de individuos no

fumadores, el algoritmo tendrá mayor cantidad de datos sobre éstos para aprender. No obstante, este efecto se podrá minimizar posteriormente cuando se desarrolle el modelo predictivo de manera que podemos hacer que nuestro algoritmo tenga esta condición en cuenta indicando que debe tener en cuenta la clase minoritaria (los fumadores) penalizando de alguna manera la clase mayoritaria (los no fumadores).

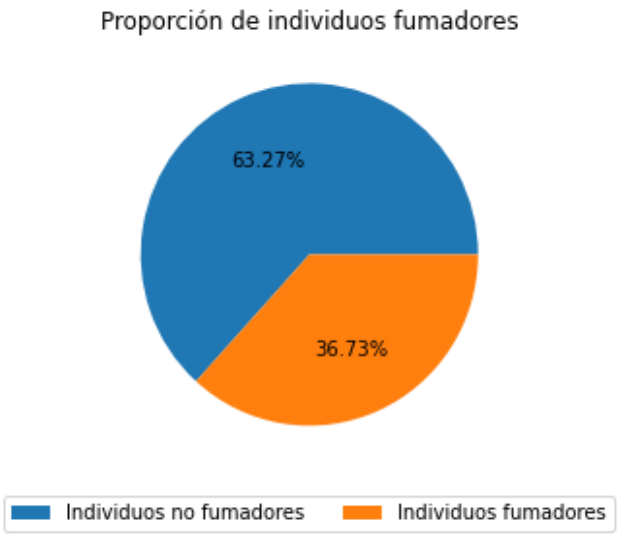

Imagen 1. Proporción de fumadores y no fumadores en nuestro conjunto de datos.

Siguiendo con la exploración de los datos que tenemos, debemos tener en cuenta dos aspectos relevantes desglosados para cada una de las variables (**CONSULTAR ANEXO**$_3$): la cantidad de valores nulos o faltantes y el tipo de dato que son. Estos dos puntos resultan de gran importancia, el primero para comprobar que ninguno de nuestros individuos observados, tengan una de sus características sin un valor; el segundo, para conocer si las características adoptan valores numéricos o si, por el contrario, se refieren a una categoría.

Vemos como todas nuestras variables tienen 55.692 valores no faltantes, coincidiendo con el número total de observaciones que tenemos en nuestro conjunto de datos, es decir, no existen valores nulos para ninguna de nuestras variables. Por otro lado, tenemos tres tipos de datos para nuestras variables, algunas son de tipo numérico entero, otras son de tipo numérico decimal y otras son de tipo categórico (designan una cualidad).

Podemos comenzar por la visualización de las variables numéricas que existen en nuestro conjunto de datos, para ello, podemos representar un histograma por

cada variable. Un histograma recoge la distribución de los valores de nuestras variables. En el eje de abscisas o eje horizontal se representan los posibles valores que adoptan nuestras variables mientras que en el eje de ordenadas o eje vertical tenemos la cantidad total de valores que tiene la variable para un valor determinado (**CONSULTAR ANEXO4**).

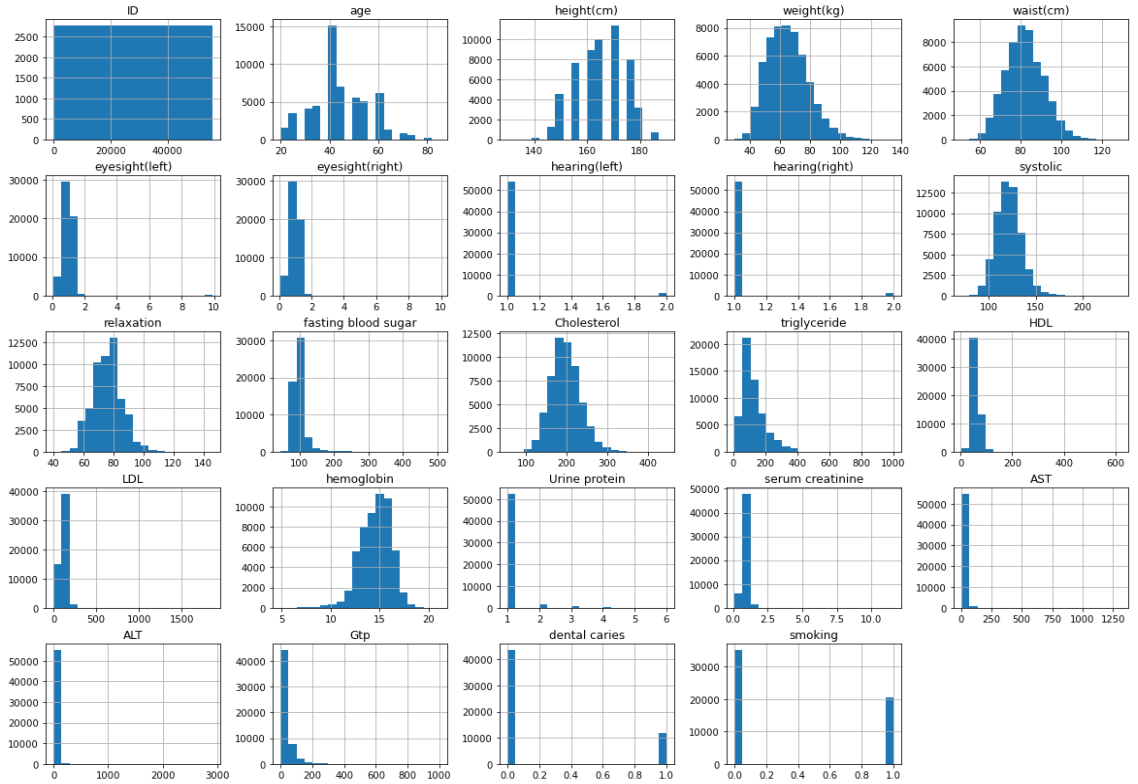

Imagen 2. Histograma de las variables numéricas.

Al observar el histograma para todas las variables numéricas, podemos observar que la variable *ID* tiene en total un valor para cada uno de los valores que dicha variable puede adoptar, es decir, existe un número identificador para cada individuo en el conjunto de datos, tenemos 55.692 individuos y existen 55.692 números identificadores diferentes entre sí, por tanto, se puede proceder a su eliminación puesto que no aporta ningún valor predictivo a nuestro futuro modelo.

Otro detalle que podemos observar y que podría llamar la atención es que algunas variables parecen tomar la mayoría de sus valores en torno a un rango específico pudiendo tomar algunos valores aislados. Esto nos lleva a representar diagramas de caja y bigotes para nuestras variables numéricas y conocer más sobre la distribución de sus valores atendiendo a una circunstancia importante, los valores atípicos o outliers.

Antes de la representación de los diagramas de caja y bigotes, resulta interesante destacar que la comparación de la media aritmética y la mediana (el valor central que toma la variable) de cada una de las características es importante a la hora de detectar posibles alteraciones en los datos. De este modo, cuando la diferencia entre ambos estadísticos es amplia, podría indicar la posible presencia de valores atípicos (valores que se salen del rango donde la variable toma la mayoría de sus valores). Cuando la media es llamativamente superior a la mediana, podría indicar la presencia de valores muy grandes en la variable, que arrastran del valor de dicha variable a la derecha, estaríamos ante valores atípicos grandes. En el caso contrario tenemos el caso en el que la media aritmética es bastante más pequeña que la mediana, en cuyo caso existe la posibilidad de que la variable adopte valores muy pequeños que arrastren la media de la variable a la izquierda, estaríamos ante valores atípicos pequeños.

En el caso de este estudio, ninguna de las características parece indicar una diferencia amplia entre la media aritmética y la mediana (**CONSULTAR ANEXO**[5]). Por ejemplo, a la vista del histograma de la variable *ALT*, veíamos como concentraba la mayor cantidad de sus valores entre 15 y 31, no obstante, la variable adoptaba valores que llegaban hasta 2.914 pero si observamos la media y la mediana de dicha variable, son 27,03 y 21 respectivamente. La diferencia entre ambos estadísticos no es demasiado amplia, pero vemos como la media es superior a la mediana, podríamos comprobar como el valor elevado que se comentaba anteriormente estaría arrastrando del valor de la media a la derecha, provocando que ésta sea superior a la mediana de dicha variable.

La representación de diagramas de caja y bigotes ayudará a identificar más detalladamente la presencia de dichos valores atípicos ya que se puede dar el caso en el que existan valores atípicos y que, al no ser suficientemente elevados en cuanto a cantidad, no tengan la capacidad de alterar significativamente la media aritmética de la variable. Vemos como casi todas las variables representadas tienen valores atípicos tanto por encima como por debajo de los bigotes de nuestros diagramas. El bigote superior nos indica el máximo, el bigote inferior nos indica el mínimo, la diferencia entre los bigotes es el rango y la diferencia entre el extremo superior y el extremo inferior de la caja es lo que se denomina rango intercuartílico, es decir, donde se encuentran el 50% de los

valores de la variable. Todo lo que sale tanto por debajo como por encima de los bigotes, son los denominados valores atípicos. En la mayoría de los casos se actuaría sobre ellos, posiblemente mediante su eliminación y manteniéndolos como valores nulos para una posterior imputación. No obstante, en el caso de este estudio, se puede decidir por dejar dichos valores como están actualmente ya que se podría argumentar de la siguiente manera: si, por ejemplo, observamos el diagrama de caja y bigotes de la variable *relaxation* indicativa de la presión arterial diastólica, vemos que todos los valores por encima de 100 y por debajo de 50 están considerados como valores atípicos, pero ¿no puede un individuo realmente presentar dichas presiones arteriales? La respuesta es afirmativa dado que pueden existir individuos que padecen hipertensión o hipotensión y tampoco podemos saber en la práctica real, en un panorama general de la población mundial, qué tan común es este padecimiento. En este sentido, se optaría por interpretar las demás variables del mismo modo además de que, una exhaustiva y rigurosa interpretación requeriría de conocimientos médicos avanzados.

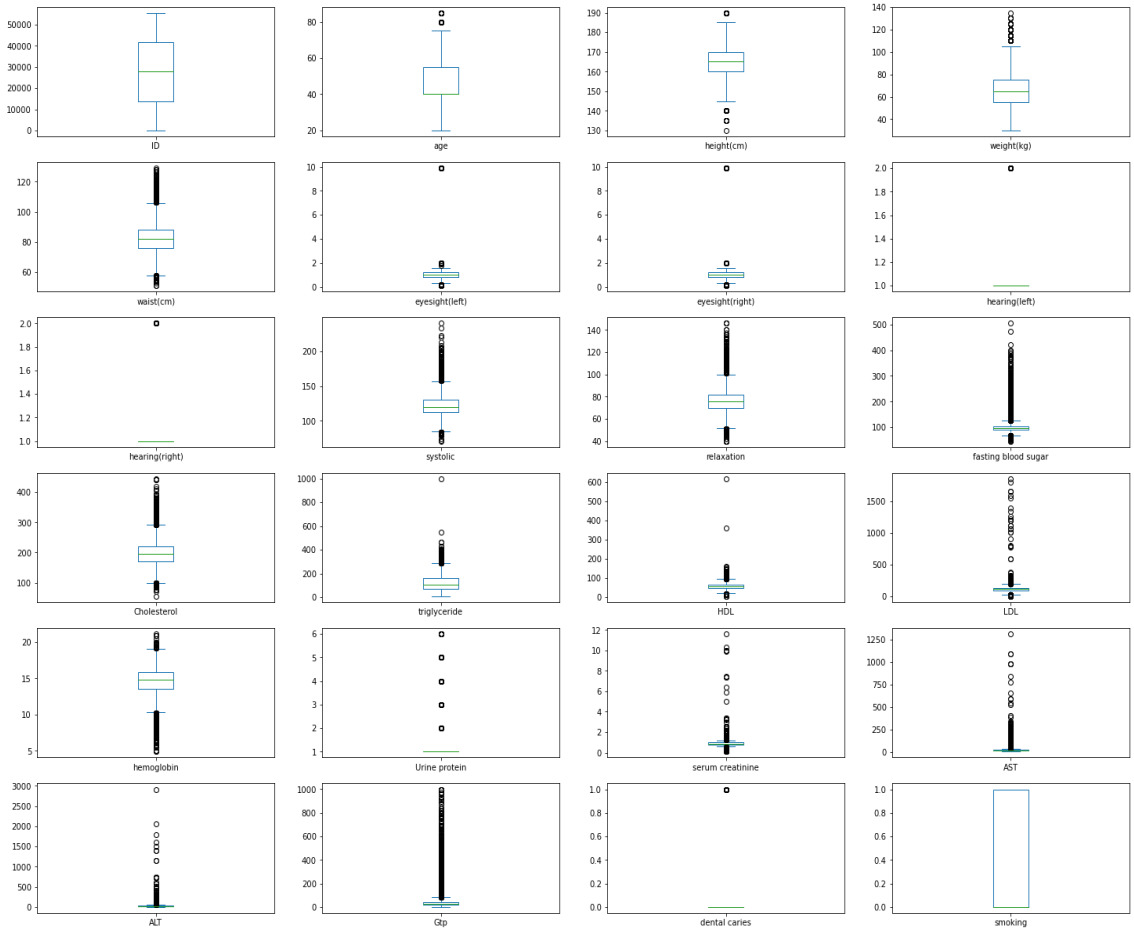

Imagen 3. Diagrama de caja y bigotes de variables numéricas.

Adicionalmente podemos representar el diagrama de caja y bigotes de una variable frente a la condición de fumador. Por ejemplo, en el siguiente diagrama observamos la hemoglobina en sangre diferenciando entre individuos fumadores y no fumadores, aprovechando que la hemoglobina está relacionada de alguna manera con el oxígeno en sangre y la disminución de oxígeno que provoca el tabaquismo. Vemos que el rango intercuartílico es superior para los fumadores y que los fumadores alcanzan picos superiores de hemoglobina que los no fumadores. Esto tiene su lógica ya que médicamente se conoce que el tabaquismo reduce el oxígeno en sangre provocando que el cuerpo produzca una cantidad mayor de esta célula.

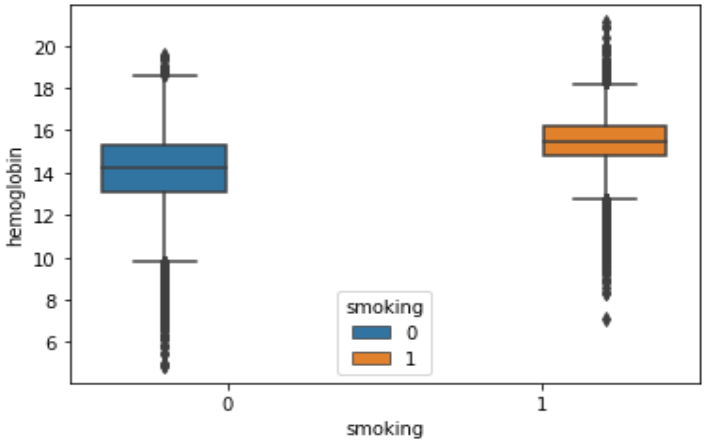

Imagen 4. Diagrama de caja y bigotes de hemoglobina frente a tabaquismo.

Cuando comenzamos a visualizar las variables categóricas de nuestro conjunto (**CONSULTAR ANEXO**[6]), observamos que existe una variable que tan solo tiene un único valor, la variable *oral*. Tal y como se comprueba en el gráfico siguiente donde comprobamos la distribución de las demás variables categóricas, vemos que la variable *oral* tan solo tiene el valor 'Y' y ese es el valor que tiene para todas las observaciones. Por otro lado, vemos como la variable *gender* consta de 35.401 hombres y 20.291 mujeres y la variable *tartar*, referente a la presencia de sarro, consta de 30.940 casos afirmativos y 24.752 casos negativos. Cabe señalar que, como ocurrió con la variable *ID,* la variable *oral* es candidata para eliminar, en esta ocasión porque una variable que solo tiene un único valor posible y es el mismo para todas las observaciones, no aporta valor predictivo a nuestro modelo ya que esta variable no aplica diferencia sobre las características de nuestros individuos.

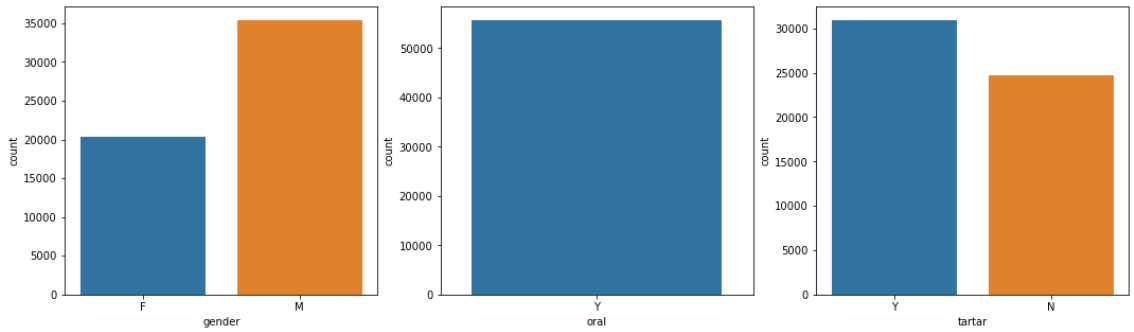

Imagen 5. Distribución de las variables categóricas.

Podemos ver como se distribuyen los fumadores según el género y también según la presencia de sarro dental. Como vemos en la siguiente imagen, la proporción de hombres fumadores es mucho mayor que la de mujeres fumadoras, incluso existen más hombres fumadores en nuestros datos que hombres no fumadores. Por otro lado, se comprueba como la presencia de sarro dental es mayor en sujetos fumadores que no fumadores, tanto para hombres como para mujeres.

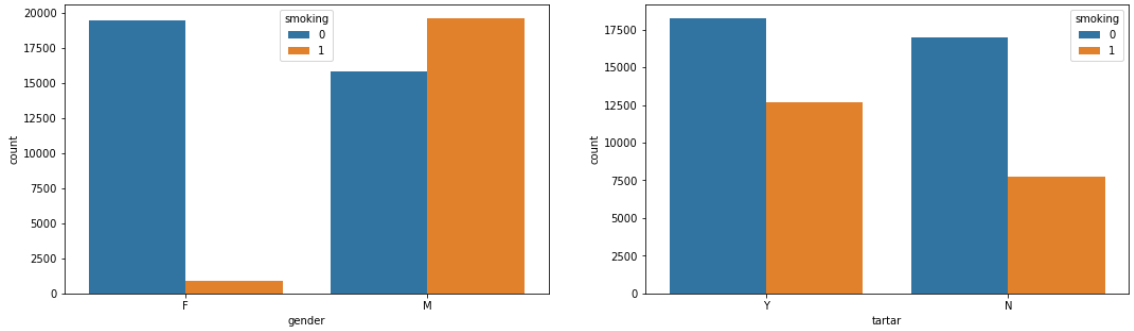

Imagen 6. Distribución de individuos fumadores por género y presencia de sarro dental.

Un aspecto relevante a la hora de analizar los datos que disponemos es la observación de correlaciones lineales. Una correlación lineal es un valor comprendido entre menos uno y uno de modo que cuanto más se acerque a menos uno indica correlaciones lineales negativas más altas, que en la práctica consiste en que cuando una variable aumenta su valor, la otra tiende a disminuir. En el caso contrario tenemos las correlaciones lineales positivas, cuando su valor se acerca a uno y cuyo caso significa que cuando una variable aumenta su valor, la otra variable también tiende a aumentar. En la siguiente imagen se representan las correlaciones lineales positivas más grandes respecto a la variable que contiene a los sujetos fumadores o no fumadores. Se puede sacar una clara conclusión y es que los individuos fumadores son los que tienden a

tener mayor cantidad de hemoglobina, son los individuos con mayor altura y también los de mayor peso corporal. Concretamente el valor de esta correlación lineal de las tres variables más correlacionadas que son la hemoglobina, altura y peso, entre otras (**CONSULTAR ANEXO**[7]), son del 0.40, 0.396 y 0.302

Asimismo, se observa en la siguiente imagen la correlación entre las variables más correlacionadas con la variable de fumadores donde podemos ver como existe correlación positiva entre la hemoglobina y la altura (cuanta más hemoglobina, más altura y viceversa) o la hemoglobina y el peso corporal (cuanta más hemoglobina, mayor peso corporal y viceversa), entre otras.

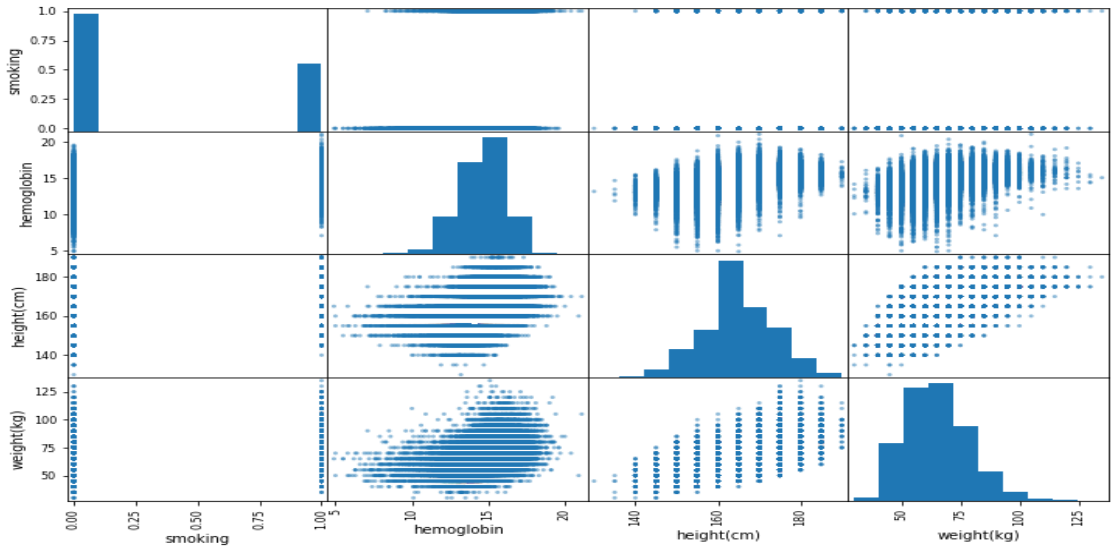

Imagen 7. Variables con las correlaciones lineales positivas más altas respecto a la variable objetivo.

Siguiendo con el análisis de las correlaciones lineales, podemos ver en el siguiente mapa de calor las correlaciones lineales entre par de variables. En él, se pueden observar cómo algunas correlaciones explican la lógica de los datos, como la relación existente entre el peso corporal y la altura, cuanto mayor es el peso, mayor es la altura y viceversa.

De esta imagen puede subrayarse, entre otras muchas:

- Una correlación lineal positiva elevada entre el peso corporal y la anchura de la cintura cuyo valor es de 0.82 o una correlación lineal positiva elevada entre la presión arterial diastólica y la presión sistólica cuyo valor es 0.76

- Una correlación lineal negativa entre la edad y la altura de -0.48 o una correlación lineal negativa entre la anchura de cintura y el parámetro sanguíneo *HDL*.

Variables que estuvieran muy correlacionadas podrían considerarse como variables que aportan poco valor predictivo al modelo ya que ambas variables seguirían una misma tendencia y sus valores serían intuitivos.

Cabe destacar que cuando el valor de la correlación lineal es igual a 1 como se observa en la diagonal principal, se refiere a par de variables que son la misma. Es decir, cuando se compara, por ejemplo, la edad con la edad, al tratarse de los mismos datos, el valor de la correlación lineal es uno.

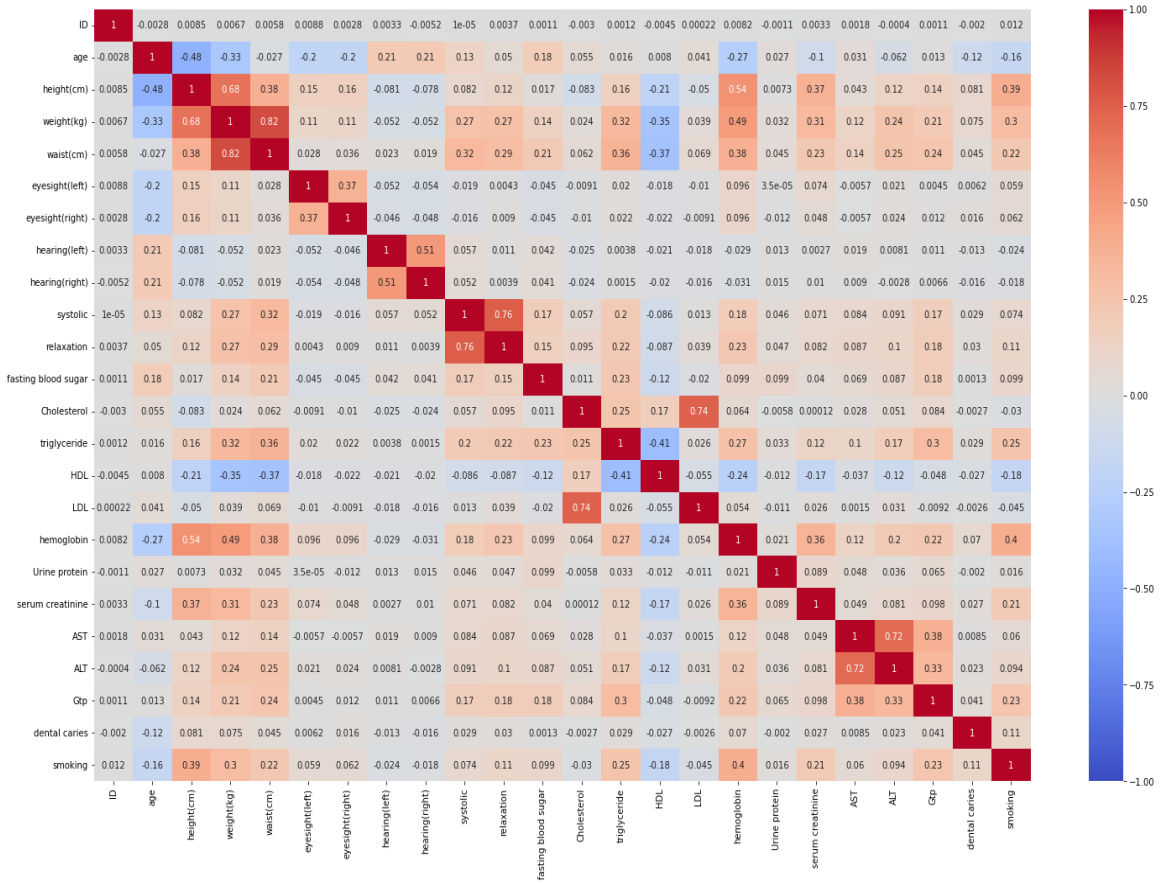

Imagen 8. Correlaciones lineales entre par de variables.

Una vez realizada una visualización general de nuestros datos, podemos profundizar más obteniendo gráficos más complejos que nos informe más acerca de nuestros datos. Tenemos en el siguiente gráfico una nube de puntos en tres

dimensiones donde se diferencia entre fumadores y no fumadores observado el valor que tienen para cada uno de ellos la hemoglobina en sangre, la presión arterial diastólica y la tensión arterial sistólica. Como se vio anteriormente, se vuelve a confirmar una mayor cantidad de hemoglobina en personas fumadores a diferencia de los que no fuman, para los fumadores también se diferencia una tendencia a presentar una mayor presión diastólica y también se aprecia ligeramente que las personas fumadoras presentan una mayor presión arterial sistólica. Puede resultar más difícil identificar a los sujetos fumadores ya que como se vio anteriormente, éstos representan una menor proporción sobre el total y son los no fumadores lo que sobresalen a simple vista en el diagrama de dispersión.

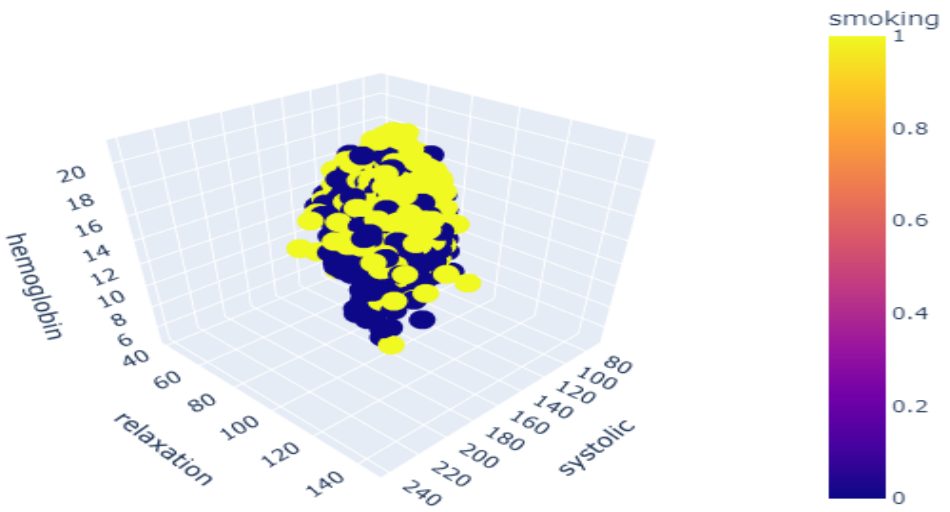

Imagen 9. Diagrama de dispersión de hemoglobina y presión arterial sistólica y diastólica por fumadores y no fumadores.

Podemos conocer también la relación que guardan los niveles de colesterol y triglicéridos con el consumo de tabaco. En el gráfico se aprecia como los fumadores presentan unos niveles de triglicéridos llamativamente superiores a los no fumadores. Con respecto a los niveles de colesterol, no se aprecian diferencias notables.

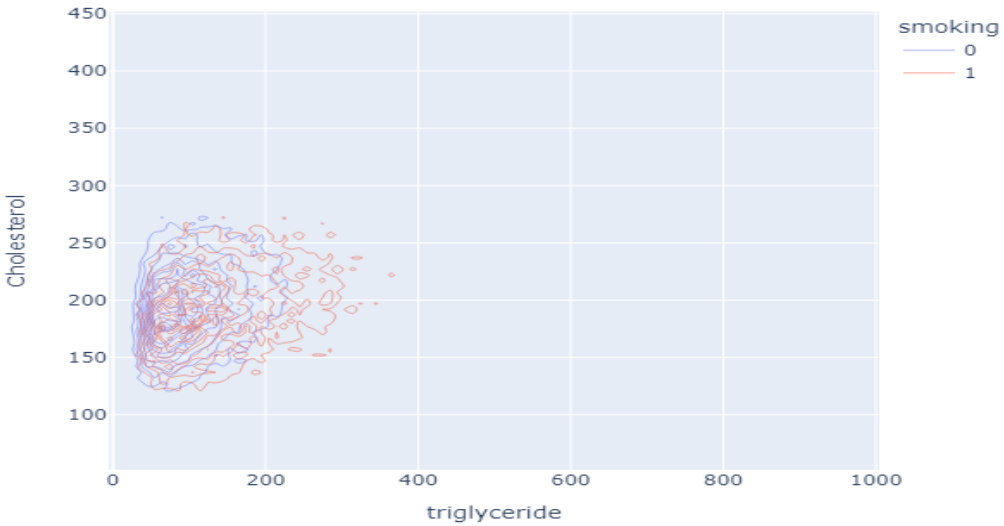

Imagen 10. Contorno de densidad del colesterol y triglicéridos en fumadores y no fumadores.

3. Feature Engineering

Para comenzar, podemos ampliar la información de la que disponemos trabajando sobre las variables que ya disponemos. Esto quiere decir que podemos crear nuevas características para los individuos observados, de esta manera podemos crear las siguientes variables (**CONSULTAR ANEXO**[8]):

- Variables calculadas a partir del valor de otras variables cuyos valores son valores numéricos

  - Una variable que recoja el Índice de Masa Corporal (IMC) que, si atendemos a su fórmula matemática, se trata de calcular la división del peso de cada individuo en kilogramos entre el cuadrado de su altura en metros.

- Variables creadas atendiendo al rango de los valores de otras variables de manera que, si una observación pertenece al rango específico que indica la nueva variable, el valor de la observación en esa variable será uno y en caso contrario, será cero.

  - Podemos trabajar sobre la variable de la edad creando tres nuevas variables que recoja cuando una persona es joven y pertenece a la primera edad (menor de 26 años), cuando es adulta y pertenece a la

segunda edad (entre 26 y 65 años) y cuando es anciana y pertenece a la tercera edad (mayor de 65 años).

- También podemos recoger en dos nuevas variables si una persona es diabética (más de 125 mg/dl) o presenta glucemia alterada en ayunas (entre 100 mg/dl y 125 mg/dl).
- A partir de la variable creada para el IMC, podemos recoger la condición de personas de peso bajo (IMC inferior a 18,5), persona de peso ideal (IMC entre 18,5 y 25), personas con sobrepeso (IMC entre 25 y 30) y personas obesas (IMC superior a 30) creando para ello cuatro nuevas variables.

Podemos visualizar las nuevas variables y ver la distribución que adquieren sus valores. En los histogramas de dichas variables podemos ver que la mayoría de los individuos tienen un peso que puede ser considerado como normal acorde al IMC, son adultos y que entre los individuos con problemas de glucemia en ayunas; existen más individuos con prediabetes que con diabetes.

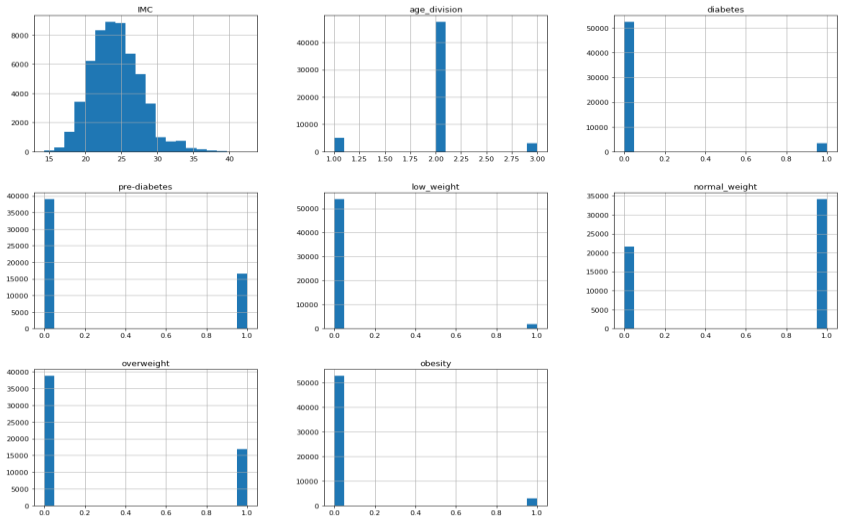

Imagen 11. Histograma de las nuevas variables.

4. División del conjunto de datos

De los 55.692 individuos que tenemos en nuestro conjunto de datos tenemos que realizar una división creando tres subconjuntos (**CONSULTAR ANEXO9**). El primer subconjunto se tratará del conjunto de entrenamiento y es a partir del cual nuestro algoritmo aprenderá para posteriormente realizar predicciones sobre datos que no ha visto anteriormente, constará de una muestra de 44.553

observaciones. De las 11.139 observaciones restantes saldrán los otros dos subconjuntos, uno de validación y otro de prueba. Con el de validación comprobaremos qué tan bueno es nuestro modelo realizando predicciones sobre observaciones que no ha visto anteriormente, este subconjunto tendrá 5.569 observaciones y validará nuestro modelo. El subconjunto de pruebas contendrá 5.570 observaciones y nos ayudará a comprobar que nuestro modelo tiene la capacidad de generalizar adecuadamente, esto es, que no provoca que aprende muy bien de los datos de entrenamiento, pero después es malo realizando predicciones de observaciones que no ha visto anteriormente, una circunstancia que se denomina sobreajuste del modelo. Tener este último subconjunto nos indicará que las predicciones correctas sobre el subconjunto de validación no es fruto del azar y que realmente nuestro modelo tiene una buena capacidad predictiva.

5. Preparación de los datos

Llegado esta fase, tras la exploración de los datos, se trata de realizar los cambios adecuados de manera que el algoritmo de aprendizaje ingiera los datos de la mejor manera posible. En primer lugar, no todos los algoritmos admiten tipos de datos considerados categóricos, de manera que tenemos que adecuar los datos que estuvieran considerados como categóricos y pasarlos a numéricos. De igual manera hay que observar que no existan valores nulos, es decir, observaciones que tuviera alguna de sus características faltante. Hay que adoptar una estrategia para actuar frente a los datos atípicos que vimos anteriormente. También hay que reflexionar acerca de si todas las características de las que disponemos nos servirán de cara al desarrollo de nuestro modelo predictivo.

Como vimos anteriormente, ninguna de las variables de nuestro conjunto de datos dispone de valores nulos, en caso de que los hubiera, el procedimiento a seguir hubiera sido eliminar las observaciones asociadas a esas características con valor faltante, aunque hubiera supuesto una pérdida de información, solo si dichos valores nulos fuesen pocos en cuanto a cantidad, conllevando que las observaciones eliminadas hubiesen sido pocas. Otro procedimiento habría sido imputar dichos valores con algún estadístico, por ejemplo, la media o la mediana

de dicha variable para no provocar una pérdida de información y ganar la mayor capacidad predictiva posible.

Por otro lado, considerando las distintas variables categóricas que existen en nuestro conjunto de datos, tenemos dos características consideradas como datos categóricos, estos son ‘gender’ y ‘tartar’. Lo pasaremos a dato numérico creando una nueva característica para cada uno de los valores que adoptan estas variables y dicha característica tendrá el valor uno si la observación pertenece a tal categoría y en caso contrario valdrá cero; la variable categórica original se eliminará (**CONSULTAR ANEXO**[10]). Por ejemplo, para ‘gender’ habría dos nuevas características pertenecientes a hombre y a mujer. Si una determinada observación se trata de una persona de género femenino, el valor para la variable perteneciente a mujer valdrá uno y para hombre valdrá cero.

El tratamiento de los valores atípicos que vimos anteriormente puede dar lugar a interpretaciones más técnicas que determinen el modo de actuación frente a ellas. Como se explicó anteriormente, los diagramas de caja y bigotes nos indicaban la presencia de valores atípicos en muchas variables de nuestro conjunto de datos, de este modo, veíamos por ejemplo valores atípicos en la glucosa en ayuno, en parámetros como ALT o los triglicéridos, entre otros. Pero dichos valores pueden corresponderse totalmente con la realidad y una interpretación más exhaustiva de estos valores requeriría de criterios médicos avanzados. Por este motivo, se opta por no actuar sobre dichos valores.

Adicionalmente podemos llevar a cabo una estrategia para reducir el número de variables de nuestros datos y quedarnos únicamente con aquellas variables que mayor capacidad predictiva vayan a aportar a nuestro modelo, empezamos teniendo 26 características originalmente y, tras la fase de feature engineering y las modificaciones realizadas en esta fase, nos encontramos con 33 características. De este modo podemos quedarnos con las siguientes veintiuna variables que más aportan al modelo eliminando el resto (**CONSULTAR ANEXO**[11]**)** con el objetivo de simplificarlo también:

ALT, AST, Cholesterol, Gtp, HDL, IMC, LDL, age, dental caries, fasting blood sugar, gender_F, gender_M, height(cm), hemoglobin, overweight, relaxation, systolic, tartar_Y, triglyceride, waist(cm), weight(kg)

Hemos realizado todos los cambios descritos sobre nuestro conjunto de entrenamiento, pero ¿no se deberían realizar los mismos cambios sobre nuestros conjuntos de validación y de pruebas para que las estructuras de los tres conjuntos coincidan exactamente? La respuesta es afirmativa, y por ello se ha creado, para este caso, un transformador (**CONSULTAR ANEXO12**) que actúa como una especie de plantilla que contiene los cambios que deben realizar los conjuntos de datos que recibe y cuya función es realizar dichos cambios sobre el conjunto de dato que se le proporciona y nos lo devuelve con la estructura exacta que tiene nuestro conjunto de entrenamiento haciendo que no tengamos que volver a aplicar manualmente las transformaciones. Se pueden crear tantos transformadores como bloques de cambios hubiéramos realizado, como en este caso solo será necesario para pasar las variables categóricas a variables numéricas, sólo tendremos un transformador.

6. Obtención de un modelo de predicción

Tras experimentar con varios modelos utilizando distintos algoritmos (**CONSULTAR ANEXO13**), podemos seleccionar uno de ellos para realizar nuestras futuras predicciones atendiendo al valor de distintas métricas, unos valores que nos orientará acerca de lo adecuado que es el modelo.

En primer lugar, tenemos la matriz de confusión, una métrica que nos mostrará como se han clasificado nuestras observaciones. En ella observamos los individuos que se han clasificado como verdaderos negativos, es decir, individuos que no fuman y que el modelo de predicción ha detectado correctamente que no son fumadores. También podemos observar los verdaderos positivos, individuos que fuman y que el modelo ha clasificado correctamente como fumadores. Por otro lado, observamos los falsos positivos, individuos que no fuman pero que el modelo ha clasificado como fumadores. Por último, también los falsos negativos, individuos fumadores que el modelo detecta como no fumadores.

Por otro lado, tenemos la exactitud o accuracy, cuyo valor es deseable que se encuentre próximo a uno (con atención de que no se esté produciendo un sobreajuste). Esta métrica nos proporciona una visión acerca de las

clasificaciones correctas de nuestro modelo dado que su valor se calcula como la suma de verdaderos positivos y verdaderos negativos dividido entre el número total de observaciones que tenemos.

El gráfico de la representación de la curva ROC, es interesante comprobar que se encuentra por encima de la recta discontinua roja, ya que, si se encontrase por debajo, deberíamos considerar que las predicciones que realiza son fruto del azar.

El valor de la precisión nos proporciona una intuición sobre los falsos positivos que se producen en nuestro modelo mientras que el valor de la exhaustividad o recall nos informa sobre los falsos negativos, cuanto más alto sea el valor de ambas métricas mejor serán las clasificaciones que realiza nuestro modelo.

Otra métrica es el f1-score, que nos indica qué proporción de nuestras predicciones están correctamente clasificadas.

El valor de todas estas métricas junto con sus gráficos correspondientes los veremos a continuación cuando seleccionemos un modelo adecuado atendiendo a criterios de los valores de dichas métricas. Podemos ver a continuación una comparación de algunas de las métricas para cada uno de los modelos obtenidos. En línea con la explicación anterior sobre las métricas, parece evidente que el modelo seleccionado será el de Random Forest ya que el accuracy es del 0,833, la precision es del 0,766, tiene un recall del 0.796 y clasifica correctamente en el 78,06% de los casos puesto que su F1-Score es del 0,806, siendo el valor de todas las métricas superior al valor que se obtiene con los otros dos modelos.

	Regresión Logística	XGBoost	Random Forest
Accuracy	0.750943	0.786497	0.832825
Precision	0.660805	0.715596	0.766066
Recall	0.685879	0.711816	0.795869
F1-Score	0.673109	0.713701	0.780683

Tabla 1. Métricas principales de los modelos obtenidos.

Podemos ver a continuación como la curva ROC de nuestro modelo queda por encima del área que marca la línea roja discontinua. Concluimos que las predicciones realizadas por nuestro modelo no se deben a una cuestión de azar.

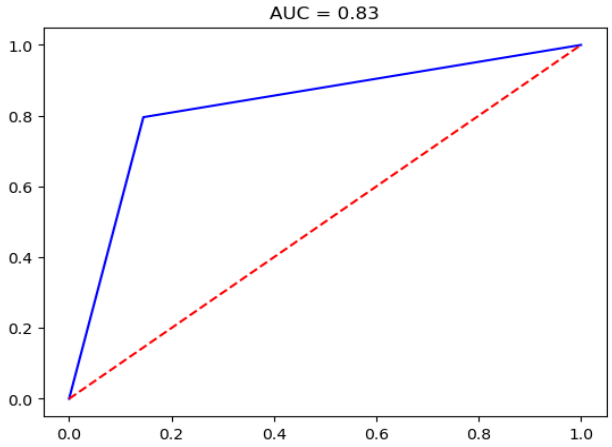

Imagen 12. Curva ROC del modelo.

En la matriz de confusión que clasifica a los individuos tenemos:

- 2.981 individuos no fumadores que nuestro modelo ha identificado correctamente.
- 506 individuos no fumadores que el modelo identifica como fumadores.
- 1.657 individuos fumadores que nuestro modelo identifica correctamente.
- 425 individuos fumadores que el modelo clasifica como no fumadores.

Vemos, según la escala de colores, como los verdaderos negativos han alcanzado un valor óptimo, no obstante, para el caso de los verdaderos positivos, la situación es mejorable.

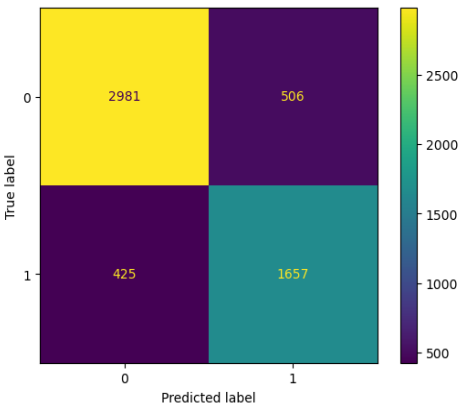

Imagen 13. Matriz de confusión del modelo.

Podemos visualizar también la curva PR referente a la precisión y exhaustividad del modelo. La situación perfecta sería donde la curva alcanza el extremo superior derecho, donde tanto la precisión como la exhaustividad alcanzan el máximo posible.

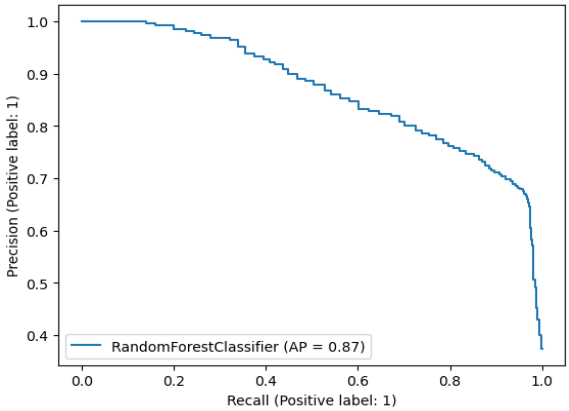

Imagen 14. Curva PR del modelo.

7. Mejora del modelo

Tras haber creado nuevas variables para aumentar la capacidad predictiva del modelo podemos seguir por modificar y probar ciertos hiperparámetros del modelo, una cuestión que entraña aspectos más técnicos (**CONSULTAR ANEXO$_{14}$**). Respecto al modelo obtenido en el apartado anterior y el valor de las métricas, obtenemos aquí un nuevo modelo donde se mejora levemente los valores para el accuracy (de 0.8328 a 0.8369), la precision (de 0.766 a 0.773), el recall (de 0.7958 a 0.7982), el F1-Score (0.7806 a 0.7854) y el valor de la curva ROC (de 0.8253 a 0.8291). Podemos ver en la matriz de confusión como también mejora levemente la clasificación tanto de verdaderos negativos como de verdaderos positivos.

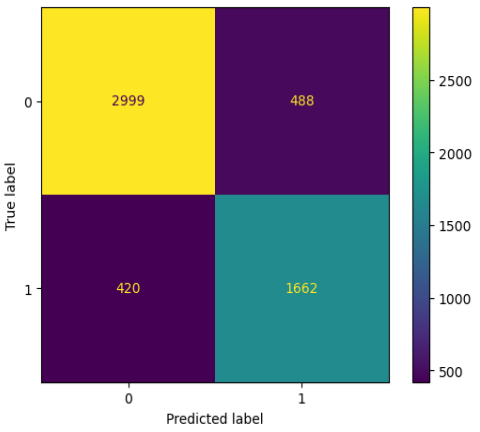

Imagen 15. Matriz de confusión del modelo mejorado.

8. Predicción sobre el conjunto de pruebas

Con el objetivo de confirmar que en nuestro modelo no se está produciendo el sobreajuste que se explicó anteriormente, a continuación, se aplica el modelo

sobre el último subconjunto que se obtuvo, el conjunto de pruebas. En esta ocasión, al ser el valor de las métricas similar a las que se obtuvieron para el conjunto de validación con un accuracy del 0.8414, una precision del 0.787, un recall del 0.7881, una F1-Score del 0.7875 y un valor de la curva ROC del 0.8306 (**CONSULTAR ANEXO15**), se confirma que el modelo no está sobreajustado y que es capaz de generalizar para nuevos datos. Del mismo modo, podemos ver en la siguiente imagen la matriz de confusión donde observamos que la distribución es similar a la matriz de confusión para las predicciones con el conjunto de validación.

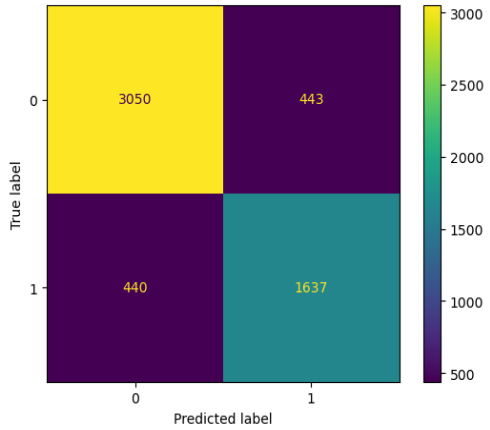

Imagen 16. Matriz de confusión de las predicciones con el conjunto de pruebas.