

Perinatal

DOCUMENTACIÓN

Autores: Iker Valdelvira, Aitziber Atutxa, Koldo Gojenola

Grupo de trabajo: grupo de investigación IXA en la Escuela de Ingeniería
de Bilbao de la Universidad del País Vasco (UPV/EHU)

Índice general

1. Caracterización de los datos (<i>datasets</i> Perinatal y ENSE):	5
1.1. Descripción del corpus	5
1.1.1. Perinatal dataset	5
1.1.2. ENSE 2017 dataset	10
1.2. Preproceso	11
1.2.1. Escalado	11
1.2.2. Imputación de los <i>missing values</i>	12
1.2.2.1. Perinatal dataset	12
1.2.2.2. ENSE 2017 dataset	15
1.2.3. Selección de atributos y atributos comunes entre ambos datasets	15
1.2.4. División del conjunto de datos y técnicas de remuestreo	18
1.3. Problemática con los conjuntos de datos	21
1.3.1. Datos desbalanceados	21
1.3.2. Mezcla de dos contextos en el <i>bajo peso</i>	22
2. Construcción de modelos de predicción	24
2.1. EXPERIMENTO 1: clasificación de <i>pesorec</i> + técnicas de remuestreo (oversampling, undersampling, múltiple undersampling personalizado...)	24
2.1.1. Algoritmos empleados	24
2.1.2. Planteamiento de la predicción	27
2.1.2.1. Tipo de clasificación	27
2.1.2.2. Resultados y evaluación de los modelos	27
2.1.2.3. Interpretabilidad (<i>feature significance</i>)	35
2.2. EXPERIMENTO 2: clasificación de <i>pesorec</i> con <i>fake features</i> de estimación del peso del feto	40
2.2.1. Algoritmos empleados	40
2.2.2. Planteamiento de la predicción	40
2.2.2.1. Tipo de clasificación	40
2.2.2.2. Resultados y evaluación de los modelos	40
2.2.2.3. Interpretabilidad (<i>feature significance</i>)	45
2.3. EXPERIMENTO 3: clasificación de <i>pesorec</i> agregando variables de ENSE 2017	47
2.3.1. Algoritmos empleados	54
2.3.2. Planteamiento de la predicción	54
2.3.2.1. Tipo de clasificación	55
2.3.2.2. Resultados y evaluación de los modelos	55
2.3.2.3. Interpretabilidad (<i>feature significance</i>)	56
2.4. EXPERIMENTO 4: clasificación de variables de ENSE 2017	57
2.4.1. Algoritmos empleados	58
2.4.2. Planteamiento de la predicción	58
2.4.2.1. Tipo de clasificación	58
2.4.2.2. Resultados y evaluación de los modelos	58
2.4.2.3. Interpretabilidad (<i>feature significance</i>)	61
Bibliografía	63

Índice de figuras

1.1. Distribución de <i>pesorec</i> en el <i>dataset</i> Perinatal	8
1.2. Distribución de <i>peson</i> en el <i>dataset</i> Perinatal	9
1.3. Porcentajes de <i>missing values</i> en las caraterísticas del <i>dataset</i> Perinatal útiles para la predicción del peso	9
1.4. Fórmulas para la estandarización mediante <i>StandardScaler</i>	12
1.5. Curva ROC y AUC de los modelos predictivos de <i>mimmi</i> sobre el conjunto Dev	13
1.6. Importancia de las variables del modelo predictivo Random Forest de <i>mimmi</i>	14
1.7. Distribuciones de <i>pesorec</i> y <i>peson</i> en el conjunto de entrenamiento	19
1.8. Distribuciones tras el primer undersampling	19
1.9. Distribuciones tras el segundo undersampling	20
1.10. Distribuciones tras el tercer undersampling	20
1.11. Distribuciones tras el cuarto undersampling	20
1.12. Distribuciones tras el quinto undersampling	21
1.13. Distribuciones de las clases <i>fuma</i> y <i>alcohol</i> sobre los datos de las mujeres en la ENSE	21
2.1. Ejemplo de un Random Forest	25
2.2. Ejemplo de una DNN	25
2.3. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (distribución original)	28
2.4. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (distribución original)	28
2.5. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling)	29
2.6. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling)	30
2.7. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (undersampling)	31
2.8. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (undersampling)	31
2.9. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))	32
2.10. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))	33
2.11. Curvas ROC y AUCs de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (múltiple undersampling personalizado)	34
2.12. Matrices de confusión del modelo predictivo DNN de <i>pesorec</i> sobre el conjunto Dev (múltiple undersampling personalizado)	34
2.13. Métricas de evaluación del modelo predictivo DNN de <i>pesorec</i> sobre el conjunto Dev (múltiple undersampling personalizado)	35
2.14. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (distribución original)	36
2.15. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (<i>feature ablation</i> eliminando <i>singleton</i>)	37
2.16. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (oversampling)	37
2.17. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (undersampling)	38
2.18. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (oversampling (10 %) / undersampling (50 %))	38

2.19. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (múltiple undersampling personalizado)	39
2.20. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando <i>peson_semanas</i>)	41
2.21. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando <i>peson_semanas</i>)	42
2.22. Curvas ROC y AUCs de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando <i>peso_semana_32</i> con la ecuación testeada sobre la población china [1])	43
2.23. Matrices de confusión del modelo predictivo DNN de <i>pesorec</i> sobre el conjunto Dev (agregando <i>peso_semana_32</i> con la ecuación testeada sobre la población china [1])	44
2.24. Métricas de evaluación del modelo predictivo DNN de <i>pesorec</i> sobre el conjunto Dev (agregando <i>peso_semana_32</i> con la ecuación testeada sobre la población china [1])	44
2.25. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (agregando <i>peson_semanas</i>)	46
2.26. <i>Feature significance</i> de los modelos predictivos Random Forest de <i>pesorec</i> (agregando <i>peso_semana_32</i> con la ecuación testeada sobre la población china [1])	46
2.27. Curva ROC y AUC de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev, y distribución de la clase <i>fumam</i> en mujeres de la ENSE 2017	48
2.28. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev	48
2.29. <i>Feature significance</i> del modelo predictivo Random Forest de <i>fumam</i>	49
2.30. Curva ROC y AUC de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev, y distribución de la clase <i>alcoholm</i> en mujeres de la ENSE 2017	50
2.31. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev	50
2.32. <i>Feature significance</i> del modelo predictivo Random Forest de <i>alcoholm</i>	51
2.33. Curva ROC y AUC de los modelos predictivos de <i>alcoholp</i> sobre el conjunto Dev, y distribución de la clase <i>alcoholp</i> en mujeres de la ENSE 2017	52
2.34. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>alcoholp</i> sobre el conjunto Dev	52
2.35. <i>Feature significance</i> del modelo predictivo Random Forest de <i>alcoholp</i>	53
2.36. Distribución de las variables predichas de la ENSE 2017 sobre el conjunto Perinatal	54
2.37. Curva ROC y AUC de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando variables ENSE 2017)	55
2.38. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando variables ENSE 2017)	56
2.39. <i>Feature significance</i> del modelo predictivo Random Forest de <i>pesorec</i> (agregando variables ENSE 2017)	57
2.40. Curva ROC y AUC de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	59
2.41. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	59
2.42. Curva ROC y AUC de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	60
2.43. Matriz de confusión y métricas de evaluación de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	61
2.44. <i>Feature significance</i> del modelo predictivo Random Forest de <i>fumam</i> (entrenado con todas las variables de ENSE 2017)	62
2.45. <i>Feature significance</i> del modelo predictivo Random Forest de <i>alcoholm</i> (entrenado con todas las variables de ENSE 2017)	63

Índice de cuadros

1.1.	Clasificación de las variables <i>estudiom</i> / <i>estudiop</i>	7
1.2.	Clasificación de las variables <i>profm</i> / <i>profp</i>	7
1.3.	Clasificación de las clases <i>fumam</i> / <i>fumap</i>	10
1.4.	Clasificación de la clase <i>alcoholm</i> / <i>alcoholp</i>	10
1.5.	Resultados de los modelos predictivos de <i>mimmi</i> sobre el conjunto Dev	13
1.6.	Compatibilidad de las variables <i>profm</i> / <i>profm</i> con <i>F19a_2</i> y <i>ACTIVA</i>	16
1.7.	Compatibilidad de las variables <i>estudiom</i> / <i>estudiop</i> con NVEST	17
2.1.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (distribución original)	27
2.2.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling)	29
2.3.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (undersampling)	30
2.4.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))	32
2.5.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (múltiple undersampling personalizado)	33
2.6.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando <i>person_semanas</i>)	41
2.7.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando <i>per-so_semana_32</i> con la ecuación testeada sobre la población china [1])	43
2.8.	Resultados de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev	48
2.9.	Resultados de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev	49
2.10.	Resultados de los modelos predictivos de <i>alcoholp</i> sobre el conjunto Dev	51
2.11.	Resultados de los modelos predictivos de <i>pesorec</i> sobre el conjunto Dev (agregando variables ENSE 2017)	55
2.12.	Resultados de los modelos predictivos de <i>fumam</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	58
2.13.	Resultados de los modelos predictivos de <i>alcoholm</i> sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)	60

1. Caracterización de los datos (*datasets* Perinatal y ENSE):

En el proyecto se han utilizado dos conjuntos de datos: el dataset **Perinatal** y el dataset **ENSE 2017**. El primero es el *dataset* principal del proyecto, el cual reúne todos los nacimientos registrados desde 1996 hasta 2019 en España. El objetivo inicial era construir un modelo predictivo del peso de un recién nacido (clasificación entre peso bajo, normal o alto) utilizando las características socioeconómicas de la madre y el padre incluidas en este conjunto de datos. El segundo *dataset* pertenece a la Encuesta Nacional de Salud de España (ENSE) realizada en 2017, el cual contiene información sobre hábitos y salud de los participantes españoles encuestados en ese año. Este último *dataset* se ha utilizado como complemento al conjunto de datos Perinatal, con el objetivo de mejorar el rendimiento del modelo de clasificación de peso creando añadiendo características de la ENSE.

A continuación se va a proceder a realizar la caracterización de los dos conjuntos de datos anteriormente mencionados.

1.1. Descripción del corpus

En esta sección se va a realizar una descripción de los dos conjuntos de datos utilizados. Se va a especificar el número de ítems, las características o atributos, el número de *missing values* y las clases y su distribución.

1.1.1. Perinatal dataset

Localización del fichero:

Perinatal_DatosResultados\Datos\Perinatal\Preprocess\dataPerinatal_converted.csv

El conjunto de datos Perinatal está formado por un total **10.311.494 ítems**, es decir, se ha registrado esa cantidad de nacimientos en España entre 1996 y 2019. Sin embargo, no todos los ítems del conjunto van a ser útiles para la creación del modelo predictivo de peso, ya que debido a los *missing values* se ha tenido que reducir (se explica en el apartado relacionado con la imputación de *missing values*).

A continuación se muestran los atributos que caracterizan las instancias del conjunto de datos. Los hemos dividido en tres grupos: los rasgos útiles para la predicción de peso del recién nacido, los descartados para la predicción del peso, y las clases.

Estos son los rasgos útiles para la creación del modelo predictivo de peso:

- **numhv**: Número de hijos vivos en partos anteriores de la madre. Es una variable numérica.
- **firstborn**: Indicador de primer parto de la madre. Es una variable binaria.
- **singleton**: Indicador de embarazo de un solo feto frente a un embarazo múltiple. Es una variable binaria.
- **propar**: Provincia del parto. Es una variable categórica donde cada categoría es el código de la provincia.

- **mespar:** Mes del parto. Es una variable categórica donde cada categoría es el número del mes del año.
- **anopar:** Año del parto. Es una variable numérica.
- **sexo:** Sexo del nacido. Es una variable binaria.
- **edadm:** Edad de la madre. Es una variable numérica.
- **edadm6:** Edad de la madre en 6 categorías: < 20, 20-24, 25-29, 30-34, 35-39, > 39. Es una variable categórica.
- **edadm35:** Indicador de edad mayor o igual a 35 años de la madre. Es una variable binaria.
- **edadp:** Edad del padre. Es una variable numérica.
- **edadp35:** Indicador de edad mayor o igual a 35 años del padre. Es una variable binaria.
- **mforeign:** Indicador de nacionalidad extranjera de la madre. Es una variable binaria.
- **fforeign:** Indicador de nacionalidad extranjera del padre. Es una variable binaria.
- **paisnacm:** País de nacionalidad de la madre. Es una variable categórica compatible con el diccionario geográfico en el que los países están clasificados a través de códigos.
- **paisnacp:** País de nacionalidad del padre. Es una variable categórica compatible con el diccionario geográfico en el que los países están clasificados a través de códigos.
- **mimmi:** Indicador de origen extranjero de la madre (teniendo en cuenta el país de nacimiento, no la nacionalidad). Es una variable binaria.
- **fimmi:** Indicador de origen extranjero del padre (teniendo en cuenta el país de nacimiento, no la nacionalidad). Es una variable binaria.
- **paisnxm:** País de origen de la madre. Es una variable categórica compatible con el diccionario geográfico en el que los países están clasificados a través de códigos.
- **paisnxp:** País de origen del padre. Es una variable categórica compatible con el diccionario geográfico en el que los países están clasificados a través de códigos.
- **estudiom:** Nivel de estudios de la madre. Es una variable categórica (ver Cuadro 1.1).
- **estudioip:** Nivel de estudios del padre. Es una variable categórica (ver Cuadro 1.1).
- **educm:** Nivel de estudios de la madre a través de la siguiente clasificación: estudios primarios o inferiores, estudios secundarios, o estudios universitarios. Es una variable categórica.
- **educp:** Nivel de estudios del padre a través de la siguiente clasificación: estudios primarios o inferiores, estudios secundarios, o estudios universitarios. Es una variable categórica.
- **profm:** Profesión de la madre. Es una variable categórica (ver Cuadro 1.2).
- **profip:** Profesión del padre. Es una variable categórica (ver Cuadro 1.2).
- **occupm:** Estado de ocupación laboral de la madre a través de la siguiente clasificación: inactiva, baja, o media/alta. Es una variable categórica.
- **occupp:** Estado de ocupación laboral del padre a través de la siguiente clasificación: inactiva/baja, o media/alta. Es una variable categórica.
- **casada:** Indicador de si está casada la madre. Es una variable binaria.
- **ecivm:** Estado civil de la madre a través de la siguiente clasificación: casada, soltera, separada/- divorciada, o viuda. Es una variable categórica.
- **pareja:** Indicador de si la madre tiene pareja. Es una variable binaria.
- **conviven:** Indicador de convivencia de la madre y el padre. Es una variable binaria.

Categoría	estudiom / estudiop
01	Analfabetos
02	Estudios primarios incompletos
03	Educación primaria
04	Primera etapa de educación secundaria y similar
05	Segunda etapa de educación secundaria con orientación general
06	Segunda etapa de educación secundaria con orientación profesional
07	Educación postsecundaria no superior
08	Enseñanzas de formación profesional, artes plásticas y diseño y deportivas de grado superior equivalentes; títulos propios universitarios que precisan del título de bachiller, de duración igual o superior a dos años
09	Grados universitarios de 240 créditos ECTS, diplomados universitarios, títulos propios universitarios de experto o especialista, y similares
10	Grados universitarios de más de 240 créditos ECTS, licenciados, másteres y especialidades en Ciencias de la Salud por el sistema de residencia, y similares
11	Másteres, especialidades en Ciencias de la Salud por el sistema de residencia y similares
12	Doctorado universitario

1.1. Cuadro: Clasificación de las variables *estudiom / estudiop*.

Categoría	profm / profp
01	Directores y gerentes
02	Técnicos y profesionales científicos e intelectuales
03	Técnicos profesionales de apoyo
04	Empleados, contables, administrativos y otros empleados de oficina
05	Trabajadores de los servicios de restauración, personales, protección y vendedores
06	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero
07	Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria
08	Operadores de instalaciones y maquinaria, y montadores
09	Ocupaciones elementales
10	Ocupaciones militares
11	Parados, personas que realizan o comparten las tareas del hogar
12	Estudiantes
13	Pensionistas, rentistas, jubilados y prejubilados
14	Invalidez permanente
15	Otra situación

1.2. Cuadro: Clasificación de las variables *profm / profp*.

Algunas características del conjunto de datos *Perinatal* han sido descartadas para la creación del modelo predictivo de peso, ya sea por razones como un alto porcentaje de *missing values* como por ser características posteriores al parto, por lo que no servían a la hora de realizar una predicción pre-parto del peso. Las variables descartadas son las siguientes:

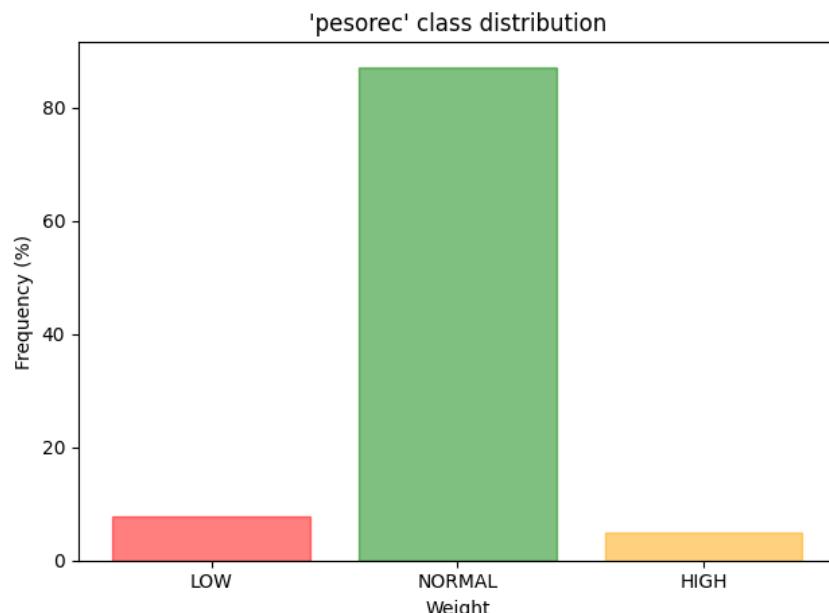
- **multipli:** Número de nacidos con o sin vida. Es una variable numérica.
- **nacvn:** Indicador de si el nacido nació vivo o muerto. Es una variable binaria.
- **v24hn:** Indicador de si el nacido vivió más de 24 horas. Es una variable binaria.
- **cesarea:** Parto con o sin cesárea. Es una variable binaria.
- **semanas:** Número de semanas del embarazo. Es una variable numérica.
- **gestage3:** Clasificación del parto por semanas de duración del embarazo: prematuro (< 37), a término (37-41), post-término (> 41). Es una variable categórica.

- **gestage4**: Clasificación del parto por semanas de duración del embarazo: muy prematuro (< 32), prematuro (32-36), a término (37-41), post-término (> 41). Es una variable categórica.
- **premature**: Indicador de parto prematuro (< 37 semanas de embarazo). Es una variable binaria.
- **normterm**: Indicador de parto a término (37-41 semanas de embarazo). Es una variable binaria.
- **postterm**: Indicador de parto post-término (> 41 semanas de embarazo). Es una variable binaria.
- **vpreterm**: Indicador de parto muy prematuro (< 32 semanas de embarazo). Es una variable binaria.
- **phecho**: Indicador de si la madre tiene pareja de hecho. Es una variable binaria.
- **mft**: Indicador de muerte fetal tardía. Es una variable binaria.
- **brank**: *No se sabe su descripción*.

Finalmente, estas son las clases que se pueden utilizar para realizar la predicción del peso del recién nacido:

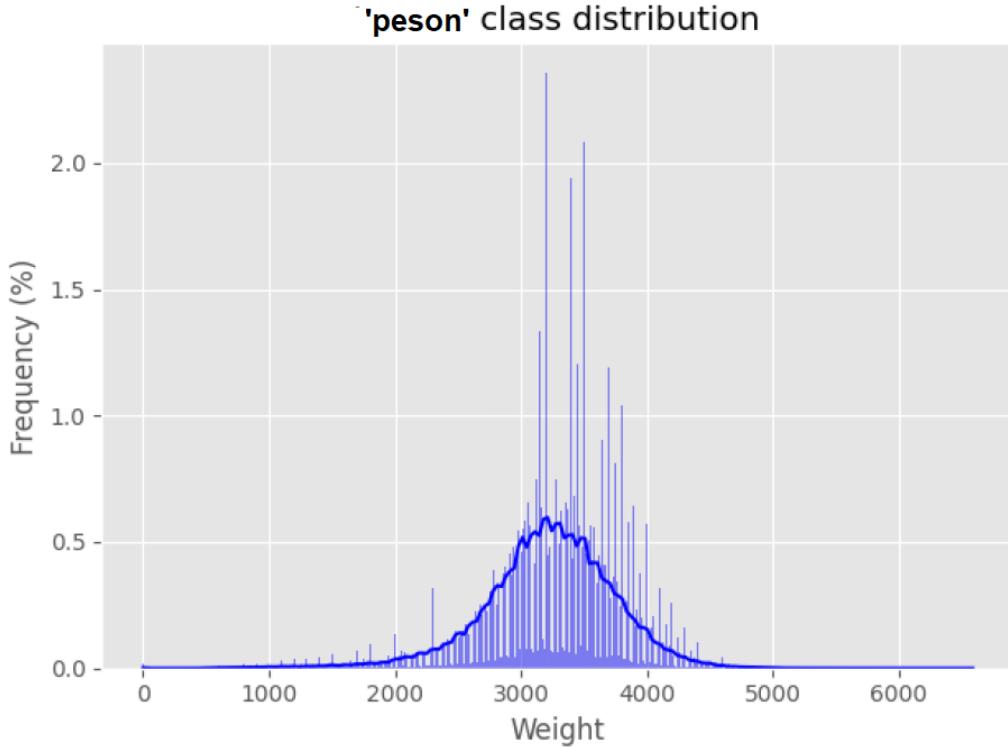
- **pesorec**: Clasificación del peso del nacido: bajo ($< 2500g$), normal (2500-4500g), alto ($> 4000g$).
- **lbw**: Indicador de peso bajo del nacido ($< 2500g$). Es una variable binaria.
- **nbw**: Indicador de peso normal del nacido (2500-4500g). Es una variable binaria.
- **hbw**: Indicador de peso alto del nacido ($> 4000g$). Es una variable binaria.
- **peson**: Peso del nacido en gramos. Es una variable numérica.

La distribución de la clase **pesorec** se muestra en la Figura 1.1. Como se puede observar más del 80% de los ítems del conjunto de datos está etiquetado con peso normal, entre 2500 y 4000 gramos. La frecuencia relativa de recién nacidos con bajo o alto peso se sitúa alrededor del 10%.



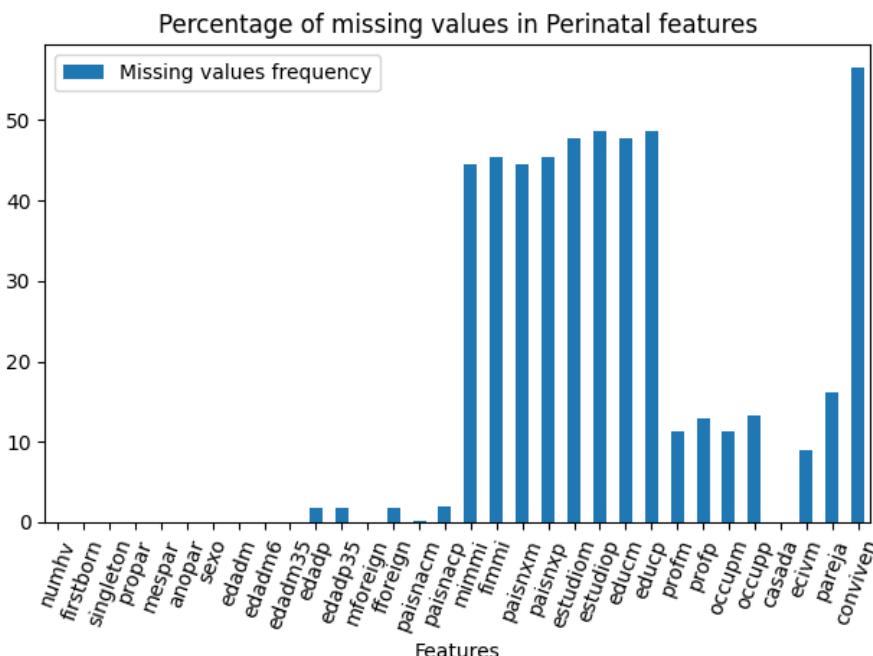
1.1. Figura: Distribución de **pesorec** en el *dataset* Perinatal

Por otro lado, la distribución de la clase ***peson*** se muestra en la Figura 1.2. Como se puede observar la gran mayoría de ejemplos tienen peso entre los 3000 y 4000 gramos (peso normal).



1.2. Figura: Distribución de ***peson*** en el *dataset* Perinatal

Respecto a los ***missing values***, se ha observado que hay un alto número de datos faltantes en algunas de las características. En la Figura 1.3 se muestran los porcentajes de *missing values* en cada variable útil para la predicción del peso.



1.3. Figura: Porcentajes de *missing values* en las características del *dataset* Perinatal útiles para la predicción del peso

Hay algunas características que en alrededor del 50 % de las instancias no tiene valor. Esto se debe a que entre los años 1996-2006 no se recogía esta información. Las variables con un alto número de *missing values* son: *mimmi*, *fimmi*, *paisnrm*, *paisnxp*, *estudiom*, *estudiop*, *educm*, *educp* y *conviven*.

1.1.2. ENSE 2017 dataset

Localización del fichero:

Perinatal_DatosResultados\Datos\ENSE2017\Preprocess\dataENSE2017Converted.csv

El conjunto de datos ENSE 2017 está formado por un total de **23.089 ítems**. Este *dataset* se ha utilizado para agregar características adicionales al conjunto Perinatal. En el estudio se han incluido los hábitos de **consumo de tabaco y alcohol** como experimento para observar las posibles mejoras a la hora de realizar las predicciones de peso del recién nacido.

Ya que en el conjunto Perinatal están recogidas las características socioeconómicas de la madre y el padre del recién nacido, los datos del conjunto ENSE 2017 se han tenido que dividir por sexo. De esta manera, el *dataset* ENSE 2017 formado por **mujeres** contiene **12.494 ítems (54 %)**, mientras que el conjunto formado por **hombres** tiene **10.595 ítems (46 %)**.

Como ya se ha dicho, en el estudio se ha realizado un experimento con la información acerca del consumo de tabaco y alcohol, por lo tanto estas van a ser las dos variables a utilizar como clases, para la posterior predicción y agregación en los ítems del conjunto Perinatal. La información acerca del consumo de tabaco y alcohol se recogen en las variables categórica V121 y W127 de la ENSE, respectivamente. Para nuestro experimento hemos convertido esas variables de categóricas a binarias como se muestra en los Cuadros 1.3 y 1.4, respectivamente. De esta manera se han creado cuatro conjuntos de datos teniendo en cuenta el sexo, con las siguientes variables clase: *fumam* (indicador de consumo de tabaco de las mujeres), *fumap* (indicador de consumo de tabaco de los hombres), *alcoholm* (indicador de consumo de alcohol de las mujeres) y *alcoholp* (indicador de consumo de alcohol de los hombres).

fumam / fumap	V121
0: No fumo	03: No fumo actualmente, pero he fumado 04: No fumo ni he fumado nunca de manera habitual
1: Sí fumo	01: Sí, fumo a diario 02: Sí fumo, pero no a diario

1.3. Cuadro: Clasificación de las clases *fumam / fumap*.

alcoholm / alcoholp	W127
0: No bebo alcohol (o menos de 1-2 días por semana)	05: 2-3 días en un mes 06: Una vez al mes 07: Menos de una vez al mes 08: No en los últimos 12 meses, he dejado de tomar alcohol 09: Nunca o solamente unos sorbos para probarlo a lo largo de toda la vida
1: Sí bebo alcohol (mínimo 1-2 días por semana)	01: A diario o casi a diario 02: 5-6 días por semana 03: 3-4 días por semana 04: 1-2 días por semana

1.4. Cuadro: Clasificación de la clase *alcoholm / alcoholp*.

Localización de los ficheros:

Perinatal_DatosResultados\Datos\ENSE2017\FumaDatasets\dataENSE201_m_fuma.csv

Perinatal_DatosResultados\Datos\ENSE2017\FumaDatasets\dataENSE2017_p_fuma.csv

Perinatal_DatosResultados\Datos\ENSE2017\AlcoholDatasets\dataENSE2017_m_alcohol.csv

Perinatal_DatosResultados\Datos\ENSE2017\AlcoholDatasets\dataENSE2017_p_alcohol.csv

La Encuesta Nacional de Salud de España está formada por una gran cantidad de variables referentes a los hábitos y cuestiones de salud de las personas encuestadas. Además del consumo de tabaco y alcohol ya mencionados, se recogen variables de las siguientes categorías:

- Variables demográficas como la **edad**, el **sexo**, **comunidad autónoma** de residencia o **nacionalidad y país de origen**.
- Variables relacionadas con el **nivel de estudios o profesión/actividad económica actual**.
- Variables relacionadas con el **estado civil** o la **convivencia** con la pareja.
- Variables relacionadas con los siguientes campos:
 - Estado de salud (diagnósticos de múltiples enfermedades).
 - Accidentalidad.
 - Restricción de la actividad.
 - Limitaciones físicas, sensoriales y cognitivas.
 - Limitaciones para la realización de las actividades de la vida cotidiana.
 - Salud mental.
 - Consultas médicas y otros servicios ambulatorios.
 - Hospitalizaciones, urgencias y seguro sanitario.
 - Consumo de medicamentos.
 - Prácticas preventivas.
 - Necesidades de atención médica no cubiertas.
 - Características físicas como el peso, la altura y el IMC (Índice de Masa Corporal).
 - Actividad física.
 - Alimentación.
 - Apoyo afectivo y personal.
 - Cuidado a otras personas con problemas de salud.

1.2. Preproceso

En esta sección se da una explicación de las técnicas de preprocesado aplicadas sobre los dos conjuntos de datos (Perinatal y ENSE 2017) previas al entrenamiento de modelos predictivos.

1.2.1. Escalado

Tanto en los ítems del conjunto Perinatal como en los de ENSE 2017 se ha aplicado un escalado **Z-score (Standard score)** en todas sus variables. En general, los algoritmos de aprendizaje se benefician de la estandarización del conjunto de datos. Si algunos valores atípicos (*outliers*) están presentes en el conjunto, los transformadores o escaladores robustos son apropiados.

Para la aplicación del escalado Z-score se ha utilizado el objeto StandardScaler implementado en el paquete de preprocesado de datos de la librería Scikit-learn de Python. El objetivo del escalado Z-score es centrar los datos numéricos restando la media aritmética de cada variable, y después, dividiéndola por su desviación estándar.

Standardization:

$$z = \frac{x - \mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N} \sum_{i=1}^N (x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

1.4. Figura: Fórmulas para la estandarización mediante *StandardScaler*

1.2.2. Imputación de los *missing values*

En este apartado se introducen las estrategias de gestión de *missing values* realizadas en ambos conjuntos de datos.

1.2.2.1. Perinatal dataset

Como se ha mostrado en la Figura 1.3 hay un alto porcentaje de *missing values* en algunas variables del conjunto Perinatal (*mimmi*, *fimmi*, *paisn xm*, *paisn xp*, *estudiom*, *estudiop*, *educm*, *edu cp* y *conviven*). Por lo tanto, se va a tener que seguir una estrategia para tratar los *missing values* en estas variables.

La estrategia más sencilla para solucionar el problema de los datos flatantes sería eliminar todos aquellos ítems que contengan *missing values*. Sin embargo, seguir esta estrategia supondría eliminar alrededor del 50% de los datos del conjunto Perinatal. Por ello, se han estudiado las siguientes tres estrategias alternativas:

- **Descartar los datos de los nacimientos entre los años 1996-2006.** Es en este subconjunto en el que algunas variables están totalmente vacías de información. Sin embargo, se dejaría de tener en cuenta toda la información de los nacimientos entre esos años para la predicción del peso de los recién nacidos. Se mantienen solamente 3.229.972 instancias (el 31% del total) tras descartar los datos de 1996-2006 y el resto de ítems con *missing values* del periodo 2007-2019.
- **Eliminar las variables en las que hay alrededor de un 50% de *missing values*.** De esta manera, podrían utilizarse más datos de todo el periodo 1996-2019, al no haber variables con tanta falta de información. Tras eliminar esas variables y los ítems con *missing values* en el resto de características, se mantienen 6.904.522 ítems (el 67% del total).
- **Predecir los *missing values* en los ítems que no disponen de información en las variables con alrededor de un 50% de *missing values*: *mimmi*, *fimmi*, *paisn xm*, *paisn xp*, *estudiom*, *estudiop*, *educm*, *edu cp* y *conviven*.** Aplicando esta estrategia se utilizaría el mismo número de ítems que al eliminar las variables con *missing values* (el 67% del total del dataset), pero manteniéndolas e imputando valores predichos.

Localización de los ficheros:

Perinatal_DatosResultados\Datos\Perinatal\Preprocess\dataPerinatal_remove_items.csv

Perinatal_DatosResultados\Datos\Perinatal\Preprocess\dataPerinatal_remove_features.csv

Perinatal_DatosResultados\Datos\Perinatal\Preprocess\dataPerinatal_predictions.csv

Se han realizado tres experimentos entrenando modelos predictivos para la clasificación del peso del recién nacido como peso bajo, normal o alto (*pesorec*) con tres conjuntos de datos diferentes siguiendo las estrategias explicadas. Los resultados de estos tres modelos entrenados son muy similares, y por ello, **predecir los *missing values*** se ha escogido como la estrategia más adecuada, ya que no se elimina ninguna variable y se mantienen datos de los nacimientos durante todo el periodo 1996-2019.

A continuación, se va a realizar un resumen de los modelos predictivos entrenados para la predicción de las variables *mimmi*, *fimmi*, *paisnxml*, *paisnxp*, *estudiom*, *estudiop*, *educm*, *educp* y *conviven* en los ítems que contenían *missing values*. Se han empleado dos algoritmos en todos los casos: un Random Forest con capacidad extraer la importancia de las variables predictoras, y una red neuronal profunda DNN de dos capas ocultas. Se ha hecho una división de Train(80 %) / Dev(20 %) sobre el conjunto de datos.

Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\Preprocess\ModelsFeaturePredictors

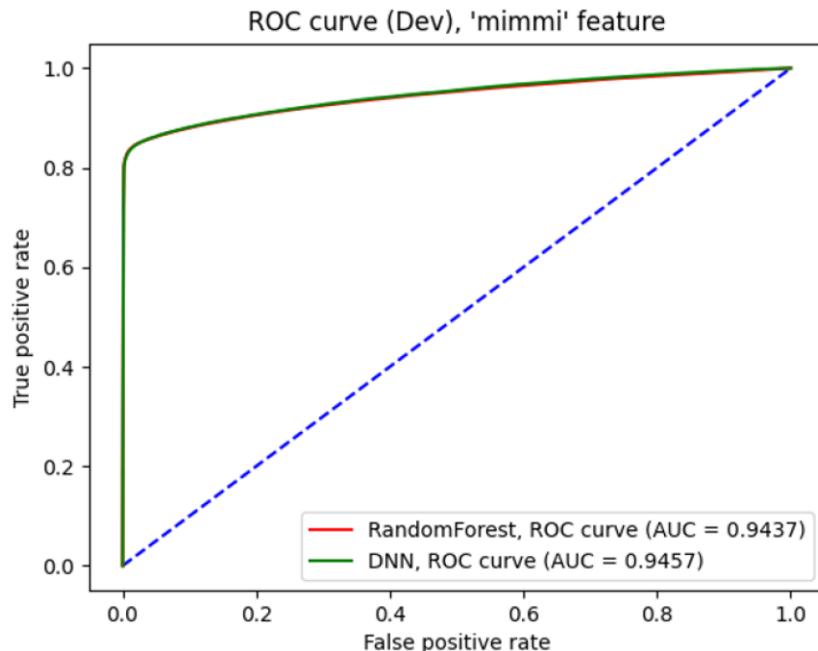
Predicción de *mimmi*

La variable *mimmi* (binaria) es el indicador de origen extranjero de la madre, teniendo en cuenta el país de nacimiento y no la nacionalidad.

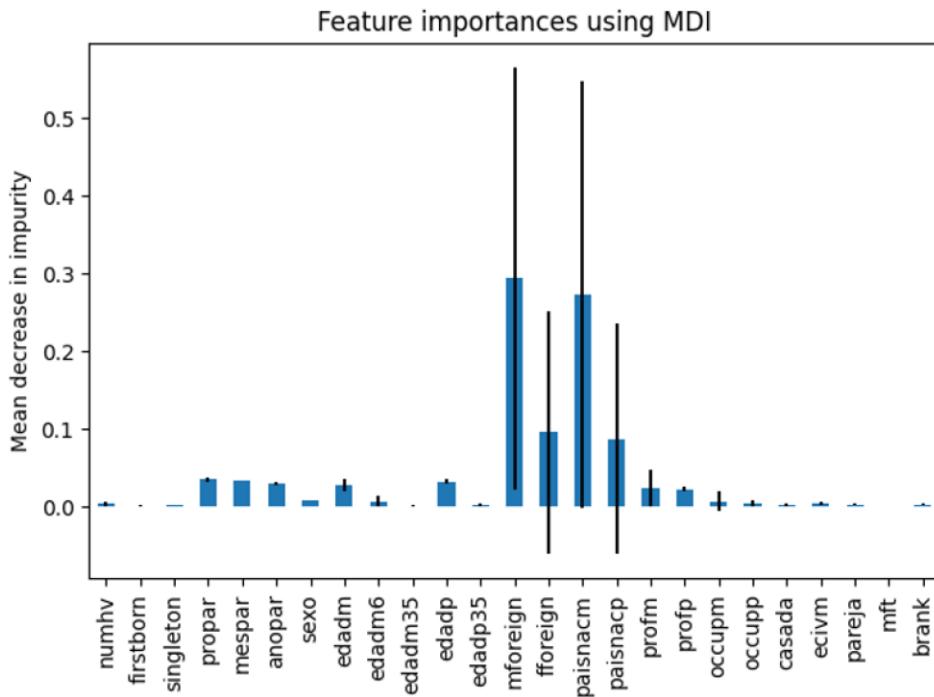
Como puede observarse en el Cuadro 1.5 y en las Figuras 1.5 y 1.6, los modelos predictivos aprenden a clasificar la variable *mimmi* de una manera muy precisa. Sin embargo, tiene una correlación muy fuerte con *mforeign* y *paisnacm* (variables relacionadas con la nacionalidad de la madre). Es más, si las coincidencias reales en el conjunto Dev entre *mimmi* (origen) y *mforeign* (nacionalidad) eran de un 96 %, las coincidencias tras las predicciones de *mimmi* son de un 99 %. En conclusión, podemos decir que las predicciones realizadas sobre *mimmi* van a ser los valores de *mforeign* en casi el 100 % de los casos.

<i>mimmi</i>	DEV		
	AUC	Accuracy	F1-score
Random Forest	0.9437	0.97	0.96
DNN	0.9457	0.97	0.96

1.5. Cuadro: Resultados de los modelos predictivos de *mimmi* sobre el conjunto Dev



1.5. Figura: Curva ROC y AUC de los modelos predictivos de *mimmi* sobre el conjunto Dev



1.6. Figura: Importancia de las variables del modelo predictivo Random Forest de *mimmi*

Predicción de *fimmi*

La variable *fimmi* (binaria) es el indicador de origen extranjero del padre, teniendo en cuenta el país de nacimiento y no la nacionalidad.

De la misma manera que ocurría con las predicciones de la variable *mimmi*, *fimmi* tiene una alta correlación con las variables *fforeign* y *paisnacp* (variables relacionadas con la nacionalidad del padre) y podemos decir que las predicciones realizadas sobre *fimmi* van a ser los valores de *fforeign* en casi el 100 % de los casos.

Predicción de *conviven*

La variable *conviven* (binaria) es el indicador de convivencia de la madre y el padre del recién nacido.

La importancia de las variables predictoras no es ninguna muy determinante para las predicciones; destacan entre otras: *pareja* (indicador de si la madre tiene pareja), *propar* (provincia del parto), y *edadp* (edad del padre). La clase positiva de *conviven* (la madre y el padre sí conviven) es muy mayoritaria, por lo que se predice en la gran mayoría de casos y la exactitud del modelo es muy alta. Sin embargo, falla la mayoría de predicciones de la clase negativa (la madre y el padre no conviven).

Predicciones de *educm* y *educp*

Las variables *educm* y *educp* expresan el nivel educativo de la madre y el padre, respectivamente, en tres categorías: primaria, secundaria y universidad.

La profesión (variables *profcp* y *profpm*) influye en las predicciones de *educm* y *educp*, aunque tampoco son muy determinantes. Hay bastantes errores en las clasificaciones de los ítems con clase *secundaria* y *universidad* al observar la matriz de confusión. Esto puede deberse a que generalmente la gente trabaja en profesiones que no requieren un nivel de estudios tan alto (*profesión < estudios*).

Predicciones de *estudiom* y *estudiop*

Las variables *estudiom* y *estudiop* expresan el nivel de estudios de la madre y el padre, respectivamente, con la clasificación que se muestra en el Cuadro 1.1.

La exactitud de los modelos es muy baja, está por debajo de 50 %, ya que hay 12 clases y la distribución es bastante heterogénea. No hay ninguna variable predictora que sea claramente determinante en las predicciones.

Predicciones de *paisn xm* y *paisn xp*

Las variables *paisn xm* y *paisn xp* contienen el código del país de origen de la madre y el padre, respectivamente.

Al igual que ocurría con las predicciones sobre las variables *mimmi* y *fimmi* (indicadores binarios de origen extranjero), *paisn xm* y *paisn xp* tienen una alta correlación con las variables *mforeign / paisn acm* y *fforeign / paisn acp*, respectivamente (variables relacionadas con la nacionalidad de la madre y el padre). Las predicciones realizadas sobre *paisn xm* y *paisn xp* van a ser los valores de *paisn acm* y *paisn acp* en el 98 % de los casos.

1.2.2.2. ENSE 2017 dataset

En el conjunto de datos ENSE 2017, todas las variables son binarias o categóricas, exceptuando la edad, el peso y la altura, las cuales son numéricas. Sin embargo, estas tres últimas variables no contienen *missing values*. Por ello, se ha utilizado la estrategia de imputación más simple para este tipo de atributos: imputación mediante la **moda (valor más frecuente)** de la variable.

Por otro lado, a la hora de crear de los *datasets* compatibles con el conjunto Perinatal (explicado en el apartado de atributos comunes), se han eliminado todos aquellos ítems que contenían *missing values*.

1.2.3. Selección de atributos y atributos comunes entre ambos datasets

En el apartado que describe el corpus del *dataset Perinatal* ya se han mencionado los atributos seleccionados para la predicción del peso de un recién nacido. Todos ellos son características socioeconómicas de la madre y el padre, o variables previas al parto. Las variables en el momento del parto o posteriores se han tenido que descartar para el entrenamiento de los modelos, ya que el objetivo es hacer una predicción del peso antes de que se dé el parto.

Por otro lado, a la hora de realizar el experimento agregando variables del *dataset ENSE 2017*, es necesario crear un *dataset* con los datos de este conjunto, pero solamente utilizando las variables que ambos *dataset* comparten. De esta manera lograríamos realizar predicciones sobre variables de la ENSE (como por ejemplo, indicadores binarios de consumo de tabaco y alcohol) utilizando los datos del conjunto Perinatal, y posteriormente, agregándolas a este último *dataset* como variables adicionales.

A continuación se va a realizar un listado de las variables comunes entre ambos *datasets*, y se explica la manera en la que se ha creado un conjunto de datos compatible con el de Perinatal partiendo de la información de la ENSE:

- Variables ***propar*** (Perinatal) y ***CCAA*** (ENSE 2017): La variable *propar* indica el código de provincia, mientras que la variable *CCAA* indica la comunidad autónoma. Por lo tanto, se han agrupado las provincias de *propar* en comunidades autónomas para hacerla compatible con la variable *CCAA*.
- Variables ***edadm*** (Perinatal), ***edadp*** (Perinatal) y ***EDADA*** (ENSE 2017): Las tres variables son compatibles ya que expresan la edad de forma numérica.
- Variables ***mforeign*** (Perinatal), ***fforeign*** (Perinatal) y ***E2_1b*** (ENSE 2017): La variable *E2_1b* de la ENSE, dependiendo del sexo de la persona, es compatible con las variables *mforeign* y *fforeign* del conjunto Perinatal. Las tres son indicadores binarios de nacionalidad extranjera o española.
- Variables ***mimmi*** (Perinatal), ***fimmi*** (Perinatal) y ***E1_1*** (ENSE 2017): La variable *E1_1* de la ENSE, dependiendo del sexo de la persona, es compatible con las variables *mimmi* y *fimmi* del conjunto Perinatal. Las tres son indicadores binarios de origen (país de nacimiento) extranjero o español.

- Variables *estudiom* (Perinatal), *estudiop* (Perinatal) y **NIVEST** (ENSE 2017): La variable *NIVEST* de la ENSE, dependiendo del sexo de la persona, es compatible con las variables *estudiom* y *estudiop* del conjunto Perinatal, realizando la clasificación que se muestra en el Cuadro 1.7. Las tres son variables categóricas que muestran el nivel de estudios.
- Variables *profm* (Perinatal), *profp* (Perinatal), **F19a_2** (ENSE 2017) y **ACTIVa** (ENSE 2017): Las variables *F19a_2* y *ACTIVa* de la ENSE, dependiendo del sexo de la persona, son compatibles con las variables *profm* y *profp* del conjunto Perinatal, realizando la clasificación que se muestra en el Cuadro 1.6. Las cuatro son variables categóricas que clasifican el tipo de profesión.
- Variables *ecivm* (Perinatal) y **E4b** (ENSE 2017): Las dos variables son compatibles ya que expresan el estado civil de la persona con las siguientes categorías: casada, soltera, separada/divorciada o viuda.
- Variables *casada* (Perinatal) y **E4b** (ENSE 2017): La variable *E4b* de la ENSE también puede utilizarse para crear un indicador binario compatible con la variable *casada* del conjunto Perinatal.
- Variables *conviven* (Perinatal) y **E4** (ENSE 2017): Las dos variables son compatibles ya que son indicadores de convivencia entre los dos miembros de la pareja.

Categoría	profm / profp	F19a_2	ACTIVa
01	Directores y gerentes	01: Directores y gerentes	-
02	Técnicos y profesionales científicos e intelectuales	02: Técnicos y profesionales científicos e intelectuales	-
03	Técnicos profesionales de apoyo	03: Técnicos profesionales de apoyo	-
04	Empleados contables, administrativos y otros empleados de oficina	04: Empleados contables, administrativos y otros empleados de oficina	-
05	Trabajadores de los servicios de restauración, personales, protección y vendedores	05: Trabajadores de los servicios de restauración, personales, protección y vendedores	-
06	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero	06: Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero	-
07	Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria	07: Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria	-
08	Operadores de instalaciones y maquinaria, y montadores	08: Operadores de instalaciones y maquinaria, y montadores	-
09	Ocupaciones elementales	09: Ocupaciones elementales	-
10	Ocupaciones militares	00: Ocupaciones militares	-
11	Parados, personas que realizan o comparten tareas del hogar	-	02: En desempleo 06: Las labores del hogar
12	Estudiantes	-	04: Estudiando
13	Pensionistas, rentistas, jubilados y prejubilados	-	03: Jubilado/a, prejubilado/a
14	Invalidez permanente	-	05: Incapacitado/a para trabajar
15	Otra situación	-	07: Otros

1.6. Cuadro: Compatibilidad de las variables *profm / profp* con *F19a_2* y *ACTIVa*.

Categoría	estudiom / estudiop	NIVEST
01	Analfabetos	02: No sabe leer o escribir
02	Estudios primarios incompletos	03: Educación primaria incompleta (ha asistido menos de 5 años a la escuela)
03	Educación primaria	04: Educación primaria completa
04	Primera etapa de educación secundaria y similar	05: Primera etapa de Enseñanza Secundaria, con o sin título (2º ESO aprobado, ECG, Bachillerato Elemental)
05	Segunda etapa de educación secundaria con orientación general	-
06	Segunda etapa de educación secundaria con orientación profesional	06: Estudios de Bachillerato
07	Educación postsecundaria no superior	07: Enseñanzas profesionales de grado medio o equivalentes
08	Enseñanzas de formación profesional, artes plásticas y diseño y deportivas de grado superior y equivalentes; títulos propios universitarios que precisan del título de bachiller, de duración igual o superior a dos años	08: Enseñanzas profesionales de grado superior o equivalentes
09	Grados universitarios de 240 ECTS, diplomados universitarios, títulos propios universitarios de experto o especialista, y similares	09: Estudios universitarios o equivalentes
10	Grados universitarios de más de 240 ECTS, licenciados, másteres y especialidades en Ciencias de la Salud por el sistema de residencia, y similares	-
11	Másteres, especialidades en Ciencias de la Salud por el sistema de residencia y similares	-
12	Doctorado universitario	-

1.7. Cuadro: Compatibilidad de las variables estudiom / estudiop con NIVEST.

Por lo tanto, los *datasets* con información de la ENSE compatibles con el conjunto Perinatal y que se utilizan para predecir variables de la ENSE como el consumo de tabaco y alcohol constan de las siguientes variables predictoras. Se diferencian dos conjuntos de datos dependiendo del sexo (madre o padre en Perinatal *dataset*).

- **ccaam / ccaap:** Comunidad autónoma (*CCAA*)
- **edadm / edadp:** Edad (*EDADA*)
- **mforeign / fforeign:** Indicador de nacionalidad extranjera (*E2_1b*)
- **mimmi / fimmi:** Indicador de origen extranjero (*E1_1*)
- **estudiom / estudiop:** Nivel de estudios (*NIVEST*)
- **profm / profp:** Tipo de profesión (*F19a_2* y *ACTIVa*)
- **casadam / casadop:** Indicador de casada/o (*E4b*)
- **evivm / evivp:** Estado civil (*E4b*)
- **convivenm / convivenp:** Indicador de convivencia en pareja (*E4*)

Localización de los ficheros:

Perinatal_DatosResultados\Datos\ENSE2017\Preprocess\dataENSE217_compatible_m.csv

Perinatal_DatosResultados\Datos\ENSE2017\Preprocess\dataENSE217_compatible_p.csv

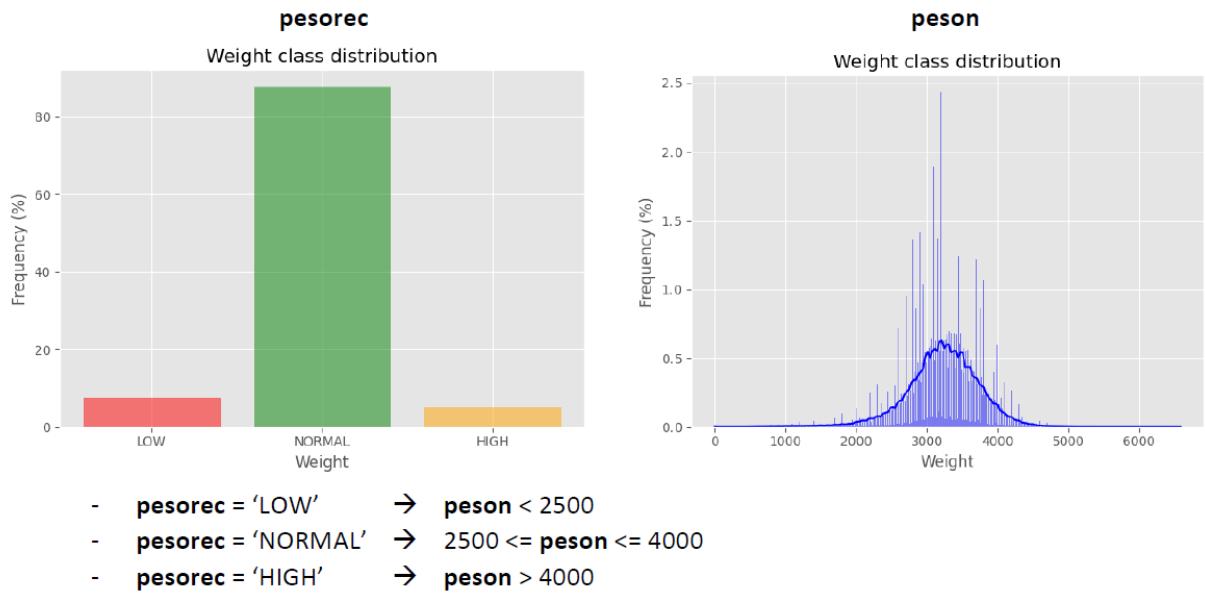
1.2.4. División del conjunto de datos y técnicas de remuestreo

En este apartado se explican las técnicas de división del conjunto de datos realizadas para el entrenamiento y evaluación de los modelos predictivos.

En todos los experimentos realizados, ya sea con el conjunto de datos Perinatal como con el conjunto ENSE 2017, se ha aplicado la misma división de datos: **Train (70 %), Dev (15 %) y Test (15 %)**. Los ítems pertenecientes a cada subconjunto han sido escogidos mediante muestreo aleatorio estratificado, es decir, se seleccionan de manera aleatoria pero manteniendo la distribución original de la clase. El 70 % de los datos se utiliza para el entrenamiento de los modelos predictivos, un 15 % para el conjunto de desarrollo sobre el que se evalúan los múltiples experimentos realizados para cada modelo predictivo entrenado, y el 15 % restante para validar el modelo final sobre otro conjunto con datos nunca antes vistos.

En los experimentos de clasificación del peso del recién nacido (clase *pesorec* con categorías peso bajo, normal y alto) se ha observado un alto desbalance en la distribución de la clase (ver Figura 1.1. Por ello, además de los experimentos con la distribución original, se han realizado experimentos adicionales realizando los siguientes remuestreos en el conjunto de entrenamiento de los modelos predictivos:

- **Oversampling:** se duplican de manera aleatoria los ítems de las clases no mayoritarias (peso bajo y alto) hasta obtener el mismo número de ítems que la clase mayoritaria (peso normal). Se ha utilizado el objeto RandomOverSampler de la librería Imbalanced-learn de Python.
- **Undersampling:** se eliminan de manera aleatoria ítems de las clases no minoritarias (peso bajo y normal) hasta obtener el mismo número de ítems que en la clase minoritaria (peso alto). Se ha utilizado el objeto RandomUnderSampler de la librería Imbalanced-learn de Python.
- **Oversampling(10 %) / Undersampling(50 %):** se duplica de manera aleatoria el 10 % de los ítems de las clases no mayoritarias (peso bajo y alto), y se elimina de forma aleatoria el 50 % de los ítems de la clase mayoritaria (peso normal). Se han utilizado los objetos RandomOverSampler y RandomUnderSampler de la librería Imbalanced-learn de Python.
- **Oversampling con SMOTE:** se crean nuevos ítems de la clase minoritaria aplicando la técnica Synthetic Minority Oversampling TEchnique (SMOTE), la cual selecciona ejemplos que están cerca en el espacio de funciones, dibujando una línea entre los ejemplos en el espacio de funciones y dibujando una nueva muestra en un punto a lo largo de esa línea. De esta manera, se crean nuevos ítems sintéticos relativamente cercanos a los de clase minoritaria. Se ha utilizado el objeto SMOTE de la librería Imbalanced-learn de Python.
- **Undersampling con TomekLinks:** se eliminan ítems de la clase mayoritaria utilizando la técnica de Tomek's Links, la cual busca ítems de la clase mayoritaria que tienen la menor distancia Euclídea con las clases minoritarias (los datos de la clase mayoritaria más cercanos al resto de clases, difíciles de diferenciar), y los elimina. Se ha utilizado el objeto TomekLinks de la librería Imbalanced-learn de Python.
- **Múltiple undersampling personalizado:** se eliminan ítems de las clases no mayoritarias de manera iterativa, siguiendo el procedimiento que se explica a continuación:
 - Se van eliminando de forma iterativa los ítems que tienen el valor con mayor frecuencia en la clase que muestra el peso en gramos (variable *peson*). Por ejemplo, si el valor de *peson* que más apariciones tiene es 3500, se eliminan todos los ítems con ese valor, y a continuación, se eliminan los ítems con el siguiente valor más frecuente.
 - Los ítems que se van a eliminar en la gran mayoría de iteraciones son los pertenecientes a la clase mayoritaria de peso normal (variable *pesorec*). De esta manera, se consigue hacer un undersampling de la clase mayoritaria descartando los ítems con pesos más comunes, para que el modelo aprenda a clasificar mejor los ítems con valores menos frecuentes con peso bajo y alto.
 - El conjunto de entrenamiento original sobre el que se ha aplicado esta técnica de remuestreo está formado por 5.935.530 ítems con las siguientes distribuciones en las variables *pesorec* (peso en categorías) y *peson* (peso en gramos):

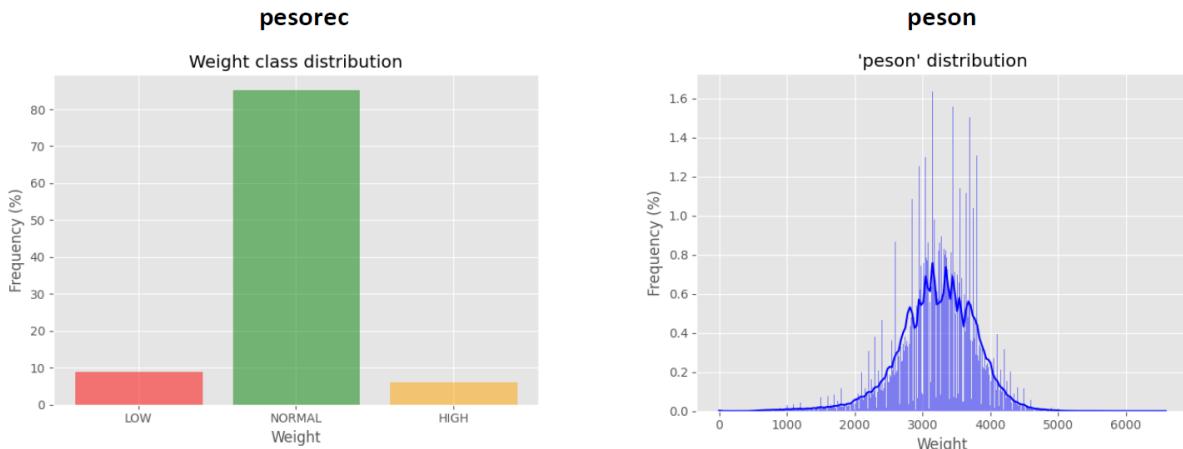


1.7. Figura: Distribuciones de *pesorec* y *peson* en el conjunto de entrenamiento

Por lo tanto, viendo las anteriores distribuciones, al eliminar los ítems con mayor frecuencia en la variable *peson*, se van a eliminar generalmente ítems con peso alrededor de los 3000g pertenecientes a la clase de peso normal, y se va a realizar un undersampling de esta clase.

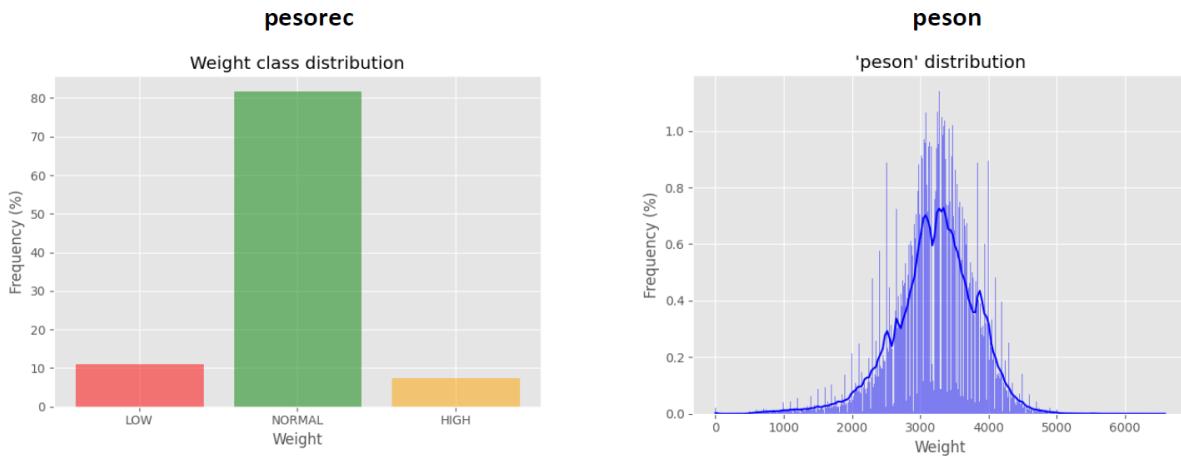
- Se han creado **5 conjuntos de entrenamiento** a medida que se iba realizando el proceso de undersampling:

Undersampling1: Obteniendo alrededor de 5 millones de ítems (4.930.438).



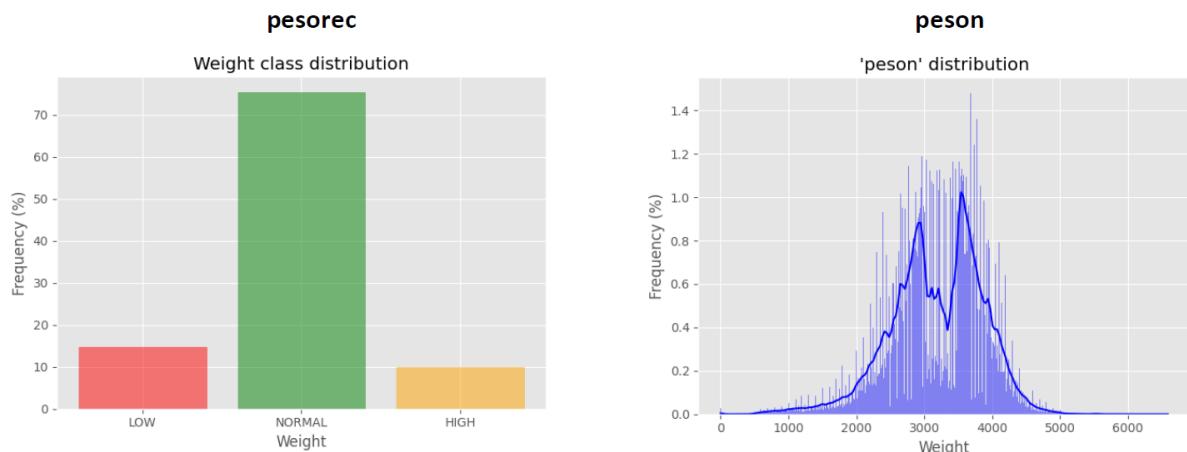
1.8. Figura: Distribuciones tras el primer undersampling

Undersampling2: Obteniendo alrededor de 4 millones de ítems (3.993.921).



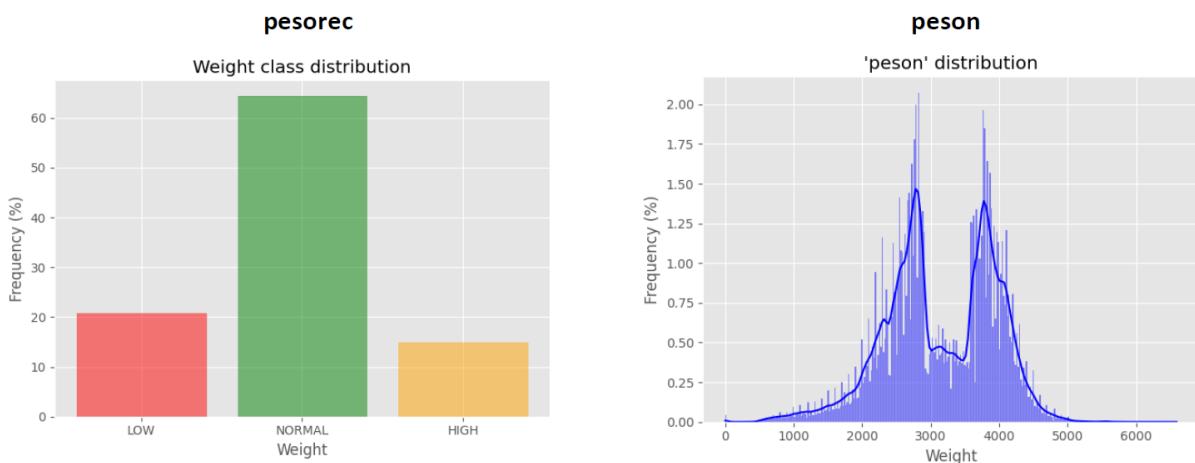
1.9. Figura: Distribuciones tras el segundo undersampling

Undersampling3: Obteniendo alrededor de 3 millones de ítems (2.976.958).



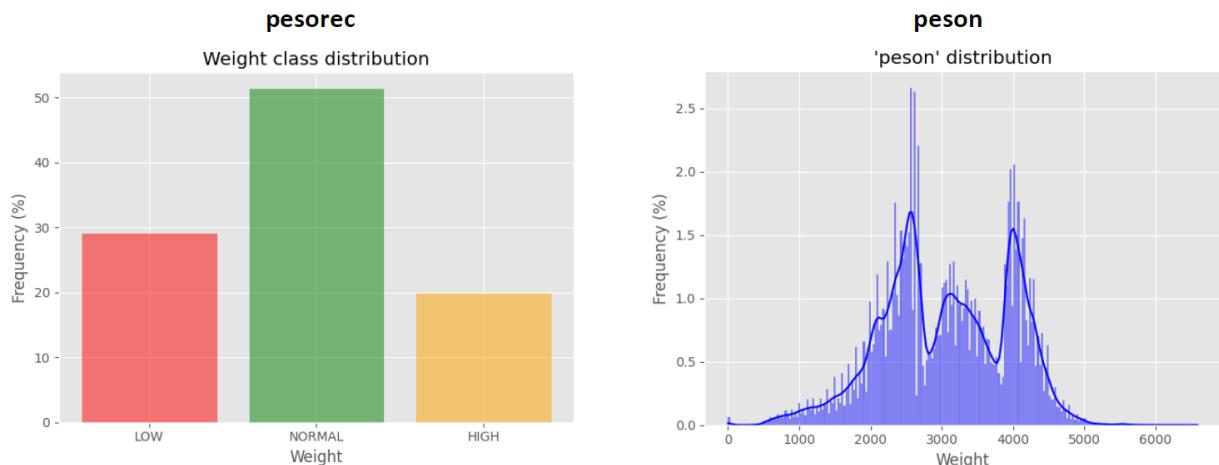
1.10. Figura: Distribuciones tras el tercer undersampling

Undersampling4: Obteniendo alrededor de 2 millones de ítems (1.997.886).



1.11. Figura: Distribuciones tras el cuarto undersampling

Undersampling5: Obteniendo alrededor de 1 millón de ítems (1.267.346), en el que se mantiene más la proporción entre las clases.

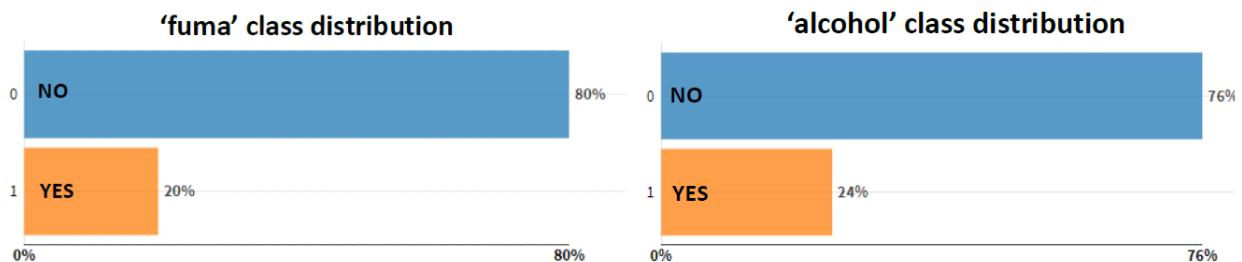


1.12. Figura: Distribuciones tras el quinto undersampling

Localización de los ficheros:

Perinatal_DatosResultados\Datos\Perinatal\MultipleUndersampling

Por otro lado, en los experimentos realizados para la clasificación de las variables que determinan el consumo de tabaco y alcohol en los datos de la ENSE del 2017, utilizando el resto de características del *dataset* ENSE 2017 como variables predictoras, se ha realizado un **undersampling aleatorio del 50 % sobre la clase mayoritaria** (no consume tabaco y no consume alcohol, respectivamente).



1.13. Figura: Distribuciones de las clases *fuma* y *alcohol* sobre los datos de las mujeres en la ENSE

1.3. Problemática con los conjuntos de datos

En este apartado se introducen dos de las grandes problemáticas surgidas a lo largo del proyecto con respecto a los datos de entrada. El primer problema ha sido el alto desbalance de los datos respecto a sus clases, y el segundo, la mezcla de contextos entre la prematuridad y el peso bajo en embarazos a término. A continuación se explican más detalladamente y se recogen las propuestas sugeridas para intentar reducir el impacto de estos problemas.

1.3.1. Datos desbalanceados

Como ya se ha comentado en apartados anteriores, los dos conjuntos de datos, Perinatal *dataset* y ENSE 2017 *dataset*, están altamente desbalanceados respecto a las variables clase sobre las que queremos realizar el ejercicio de clasificación, *pesorec* y *fuma/alcohol* respectivamente. Muestra de ello son las Figuras 1.1 y 1.13.

Por ello, se han aplicado las técnicas de remuestreo explicadas en el apartado acerca de división y remuestreo del conjunto de datos. De esta manera, dependiendo del caso, se realiza una reducción de

los ítems de la clase mayoritaria (undersampling) o un aumento de los ítems de la clase minoritaria (oversampling), para intentar que el algoritmo de clasificación aprenda de un conjunto de datos más equilibrado, y así, conseguir una mayor capacidad discriminatoria entre las diferentes clases.

1.3.2. Mezcla de dos contextos en el *bajo peso*

El segundo gran problema surgido durante el desarrollo del proyecto, a la hora de crear los modelos predictivos de peso de los recién nacidos, es la mezcla de contextos entre el **bajo peso** del recién nacido y la **prematuridad**.

El bajo peso al nacer y la semana del parto están muy correlacionados. Por ello, cuanto menos semanas dure el embarazo, es decir, cuanto más prematuro sea el nacimiento, menos peso va a tener el recién nacido. Es más, según los datos del conjunto Perinatal:

- El 59'90 % de los nacidos con bajo peso han sido prematuros.
- Y el 60'44 % de los nacimientos prematuros han sido con bajo peso.

La relación entre el peso y las semanas de embarazo es logarítmica, no es lineal, pero tratar esta relación como lineal y añadir a las variables predictoras una nueva variable artificial que exprese el peso que gana el feto por semana podría ser de gran ayudar en las predicciones del peso final del recién nacido.

De esta manera, se ha realizado un nuevo experimento creando la *false feature peson_semanas*. Esta nueva variable es el resultado del peso final del nacido en el parto (variable *peson*) dividido por el número de semanas que ha durado el embarazo (variable *semanas*).

Localización del fichero:

Perinatal_DatosResultados\Datos\Perinatal\peson_semanas

Sin embargo, como ya se ha mencionado, la ganancia de peso del feto no es lineal con el tiempo del embarazo. Por ello, se ha intentado mejorar la variable *peson_semanas*, y se han encontrado dos ecuaciones con las que se puede hacer una estimación del porcentaje del peso del feto en una semana concreta del embarazo (*proportionality growth function*).

Las fórmulas son las siguientes:

- Testeada sobre la población china y válida para las semanas 22-40 de gestación [1]:
 $Peso \% = 500,9 - (51,6 * semana) + 1,727 * (semana^2) - 0,01718 * (semana^3)$
- Testeada sobre la población inglesa y válida para las semanas 10-42 de gestación [2]:
 $Peso \% = \exp(0,578 + 0,332 * semana - 0,00354 * semana^2) / \exp(0,578 + 0,332 * mediana_semanas - 0,00354 * mediana_semanas^2) * 100$

La mediana de semanas de embarazo en el conjunto Perinatal es de 39.

El porcentaje de peso obtenido de estas fórmulas se aplicaría al peso del nacido en el momento del parto.

Por lo tanto, van a realizarse diferentes experimentos de predicción de peso del recién nacido introduciendo nuevas variables que expresen el peso estimado del feto en una semana concreta del embarazo (variable *peso_semana_XX*).

Se van a crear modelos predictivos para los siguientes experimentos:

- Estimando la feature *peso_semana_32* con la fórmula aplicada sobre la población china [1].
- Estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población china [1].
- Estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población inglesa [2].
- Estimando la feature *peso_semana_12* con la fórmula aplicada sobre la población inglesa [2].

Localización de los ficheros:

Perinatal_DatosResultados\Datos\Perinatal\peso_semana_XX

Se han escogido las semanas 12, 22 y 32 del embarazo ya que son las semanas en las que se le realizan la primera, segunda y tercera ecografía a la madre.

El objetivo de estos experimentos será crear *baselines* con dos únicas variables predictoras, la edad de la madre y el peso estimado del feto en una semana concreta del embarazo, y después, ir añadiendo características socioeconómicas (origen de los padres, estudios, profesión...) y ver como influyen estas en las predicciones del peso.

2. Construcción de modelos de predicción

En esta sección se introducen todos los experimentos realizados durante el proyecto. Por cada experimento, se va a dar una explicación de los algoritmos utilizados para el entrenamiento de los modelos predictivos, y se van a conocer las predicciones realizadas: tipo de clasificación, resultados y evaluación del modelo, e importancia de las variables predictoras.

2.1. EXPERIMENTO 1: clasificación de *pesorec* + técnicas de remuestreo (oversampling, undersampling, múltiple undersampling personalizado...)

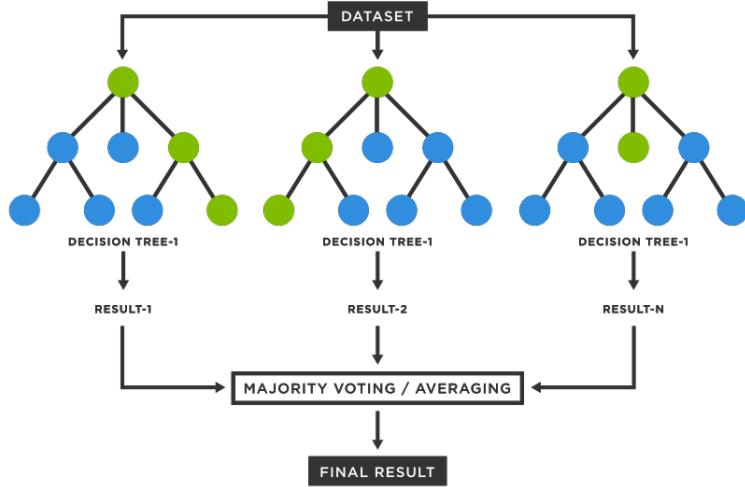
En este primer experimento se han entrenado modelos de predicción de peso de un recién nacido utilizando el conjunto de datos Perinatal y la clase *pesorec*, la cual clasifica el peso en bajo (< 2500g), normal (2500-4000g) o alto (> 4000g).

Los primeros modelos predictivos se han entrenado con el *dataset* Perinatal al completo y con su distribución original. Posteriormente, se han realizado experimentos adicionales aplicando las técnicas de remuestreo explicadas en el apartado *División del conjunto de datos y técnicas de remuestreo*. El *dataset* utilizado para la creación de todos los modelos ha sido el que contiene valores predichos en las variables con *missing values* en todas las ítems del periodo 1996-2006 de nacimientos.

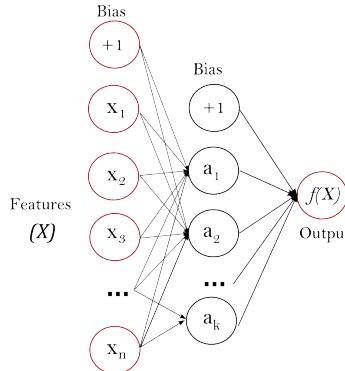
2.1.1. Algoritmos empleados

Para la creación de los modelos predictivos se han utilizado dos algoritmos de aprendizaje automático supervisado: **Random Forest** y **Deep Neural Network (DNN)**.

- **Random Forest:** Un Random Forest es un meta-estimador que se ajusta a varios clasificadores de árboles de decisión en varias submuestras del conjunto de datos y usa promedios para mejorar la precisión predictiva y controlar el sobreajuste (ver Figura 2.1). En Python se ha utilizado el objeto RandomForestClassifier de Scikit-learn.
- **Deep Neural Network (DNN):** Una red neuronal profunda es un algoritmo de aprendizaje supervisado que dado un conjunto de características $X = x_1, x_2, \dots, x_m$ y una variable de salida y , puede aprender un aproximador de función no lineal para clasificación o regresión. Es diferente de la regresión logística, ya que entre la capa de entrada y la de salida puede haber una o más capas no lineales, llamadas capas ocultas. La Figura 2.2 muestra una DNN de una capa oculta con salida escalar. En Python se ha utilizado el API Keras Deep Learning.



2.1. Figura: Ejemplo de un Random Forest



2.2. Figura: Ejemplo de una DNN

A la hora de crear la red neuronal profunda (DNN) se ha realizado un proceso de optimización de hiperparámetros, con el objetivo de construir la red neuronal que mejores resultados ofrece sobre el conjunto de datos. A continuación se hace un listado de los hiperparámetros a optimizar en la DNN y los valores probados:

- Número de capas ocultas: {1, 2, 3, 4, 5, 6}
- Número de neuronas en cada capa oculta: [32 – 512] con *step* de 32
- Función de activación de las neuronas de cada capa oculta: {ReLU, Tanh, Sigmoid}
- Learning Rate inicial: [0.01 – 0.0001]

Por otro lado, el método de optimización de hiperparámetros (*tuner*) utilizado ha sido **Bayesian Optimization**. Se ha hecho uso del objeto BayesianOptimization del API Keras para Deep Learning.

Al emplear otras técnicas de *tuning* como Random Search o Hyperband, todas las combinaciones de hiperparámetros se seleccionan al azar. Elegir los hiperparámetros de forma aleatoria ayuda a explorar el espacio de los hiperparámetros, pero no garantiza los hiperparámetros óptimos.

Bayesian Optimization, en lugar de escoger las combinaciones al azar, selecciona las primeras aleatoriamente, y después, basándose en el rendimiento de esos hiperparámetros, selecciona los siguientes mejores hiperparámetros posibles. Por lo tanto, tiene en cuenta la historia de los hiperparámetros que ya han sido probados. Las iteraciones escogiendo el siguiente conjunto de hiperparámetros basados en la historia y evaluando su rendimiento continúan hasta que el *tuner* alcanza los hiperparámetros óptimos o alcanza el número máximo de intentos establecido. El número máximo de intentos se configura con el parámetro *max_trials* del *tuner*.

Para la optimización de la DNN se han configurado los siguientes parámetros del *tuner* Bayesian Optimization:

- *objective=val_AUC*: el objetivo del *tuner* es optimizar el área bajo la curva ROC (AUC) obtenida sobre el conjunto de validación (20 % de los ítems del conjunto de entrenamiento).
- *max_trials=20*: 20 iteraciones como máximo en busca de los hiperparámetros óptimos.

Otros hiperparámetros configurados en la DNN son los siguientes:

- *batch_size=64*: la DNN se entrena en lotes de 64 ítems.
- *epochs=50*: la DNN se entrena con un máximo de 50 iteraciones.
- *ReduceLROnPlateau*: el Learning Rate se reduce con un factor de 0.1 en cada *epoch* que el AUC sobre el conjunto de validación no mejora respecto a la anterior *epoch*.
- *EarlyStopping*: el entrenamiento de la DNN se termina si el AUC sobre el conjunto de validación no mejora en dos *epochs* sucesivas.

Finalmente, tras el proceso de optimización de la red neuronal profunda (DNN), esta es su configuración óptima:

- **Número de capas ocultas: 6**
- **Número de neuronas en cada capa oculta:**
 - **Capa oculta 0:** 512 neuronas
 - **Capa oculta 1:** 32 neuronas
 - **Capa oculta 2:** 512 neuronas
 - **Capa oculta 3:** 160 neuronas
 - **Capa oculta 4:** 32 neuronas
 - **Capa oculta 5:** 320 neuronas
- **Función de activación de las neuronas de cada capa oculta:**
 - **Capa oculta 0:** ReLU
 - **Capa oculta 1:** ReLU
 - **Capa oculta 2:** ReLU
 - **Capa oculta 3:** ReLU
 - **Capa oculta 4:** ReLU
 - **Capa oculta 5:** Sigmoid
- **Learning Rate inicial:** 0.0001

Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\DNNTuning

Para concluir con la explicación de los algoritmos escogidos para la creación de los modelos predictivos, se debe razonar la elección de estos dos algoritmos y no otros. Por un lado, se ha querido emplear un algoritmo robusto de Machine Learning tradicional con capacidad de interpretación del modelo predictivo. Random Forest es capaz de mostrar las importancias de las variables predictoras del modelo basándose en las impurezas de las mismas. Cuanto más alta, más importante es la característica. La importancia de una característica se calcula como la reducción total (normalizada) del criterio aportado por esa característica. También se conoce como la importancia de Gini.

Por otro lado, también se quería entrenar una red neuronal de Deep Learning, ya que estas son capaces de obtener mejores resultados que algoritmos de Machine Learning tradicionales cuando se trabaja con grandes cantidades de datos, como en este caso donde hay millones de instancias. El inconveniente de estas redes es su baja interpretabilidad.

2.1.2. Planteamiento de la predicción

En este apartado se van a mostrar los resultados y las evaluaciones de los modelos predictivos entrenados para este experimento.

2.1.2.1. Tipo de clasificación

Como ya se ha dicho, es un problema de **clasificación multiclas** donde pretende predecirse correctamente la clase *pesorec* del conjunto Perinatal. Las tres clases posibles son: **peso bajo**, **normal** y **alto**. El conjunto de datos está muy desbalanceado (ver Figura 1.1), por lo que además de un experimento con la distribución original, también se han realizado experimentos aplicando las técnicas **oversampling**, **undersampling**, **oversampling (10 %)** / **undersampling (50 %)**, y **múltiple undersampling personalizado**, todas ellas explicadas en el apartado *División del conjunto de datos y técnicas de remuestreo*.

2.1.2.2. Resultados y evaluación de los modelos

A continuación se muestran los resultados de los modelos predictivos entrenados y evaluados sobre el conjunto Dev.

Modelo entrenado con la distribución original

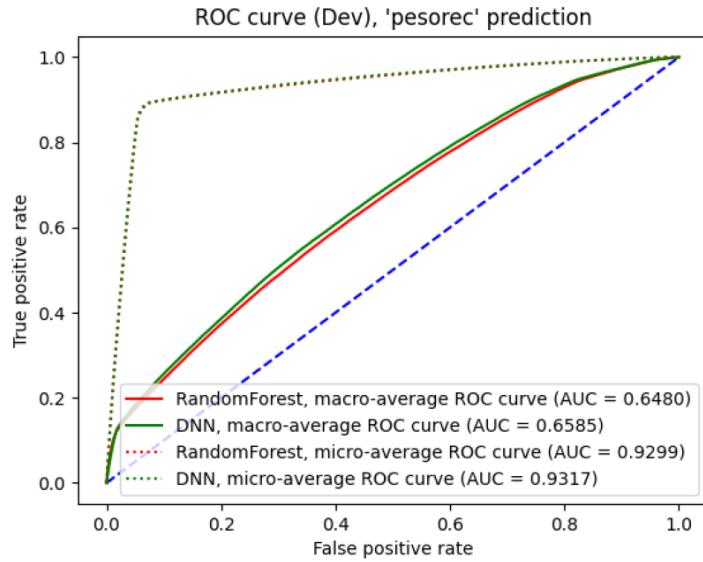
Localización de los resultados:

Perinatal.DatosResultados\Resultados\Perinatal\ExperimentsPesorec\FeatureAblationExperiments\AllFeatures

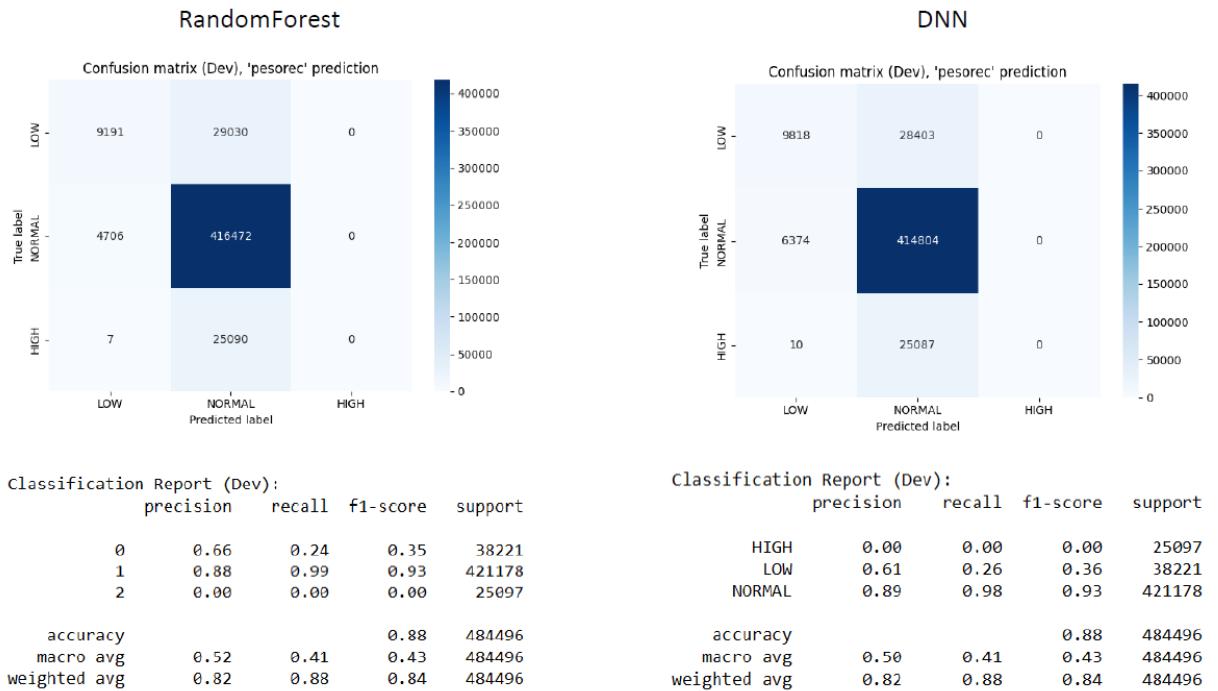
Los modelos se han entrenado con la distribución original de la clase *pesorec* (ver Figura 1.1).

pesorec	DEV		
	AUC (micro-average)	Accuracy	F1-score
Random Forest	0.9299	0.88	0.84
DNN	0.9317	0.88	0.84

2.1. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (distribución original)



2.3. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (distribución original)



2.4. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (distribución original)

Si nos fijamos en las métricas del Cuadro 2.1, da la impresión de que tanto el modelo Random Forest como la DNN aprenden a predecir bastante bien la variable *pesorec*, hay una tasa de acierto del 88% sobre las instancias del conjunto Dev. Sin embargo, los modelos no aprenden realmente a diferenciar los ítems entre las tres clases. Al clasificar la mayoría de las instancias con la clase mayoritaria (peso normal), la exactitud del modelo es alta, pero no significa que su capacidad discriminatoria sea buena. Como puede verse en las matrices de confusión de la Figura 2.4, los modelos cometen más errores que aciertos al clasificar los ítems de peso bajo, y ninguna instancia se clasifica con peso alto.

Modelo entrenado con remuestreo oversampling

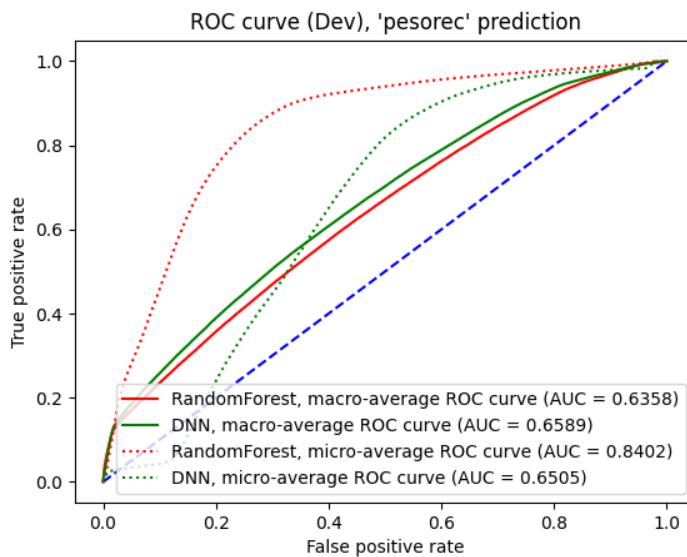
Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\3Experiments\Results\Oversampling\Experiment2

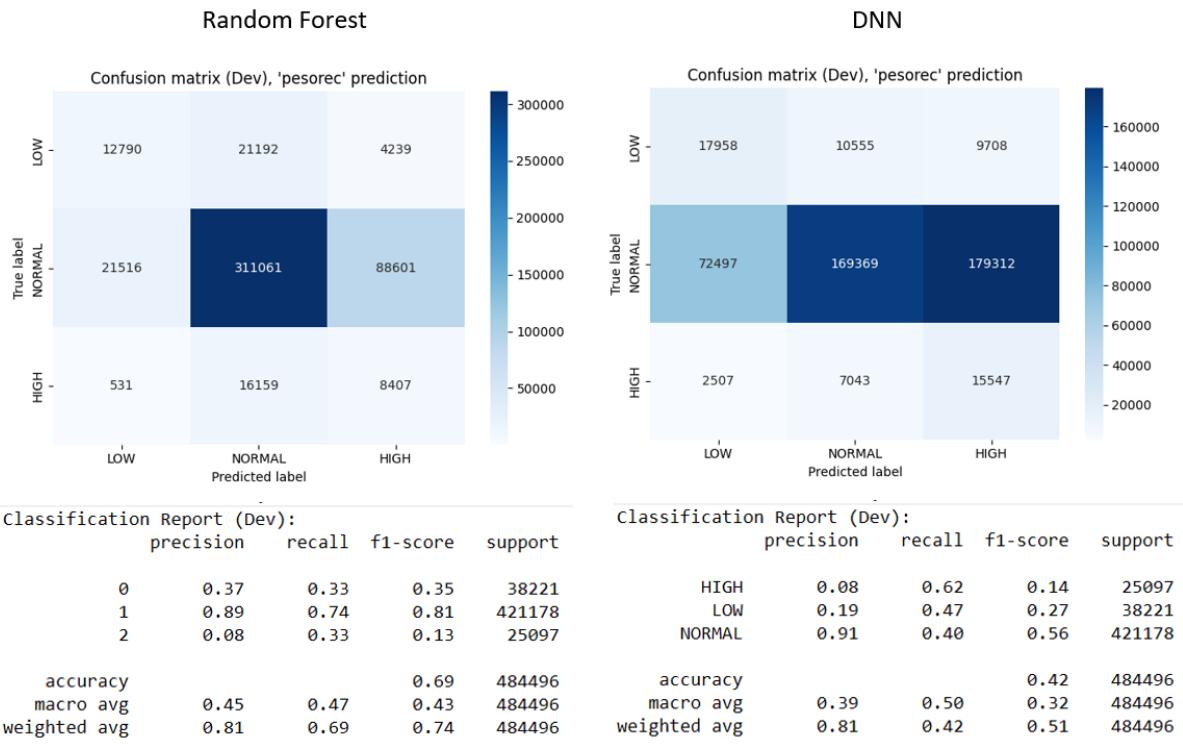
Los modelos se han entrenado balanceando la distribución de la clase *pesorec* al duplicar de forma aleatoria los ejemplos de las clases no mayoritarias (peso bajo y alto).

pesorec	DEV		
	AUC (micro-average)	Accuracy	F1-score
Random Forest	0.8402	0.69	0.74
DNN	0.6505	0.42	0.51

2.2. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling)



2.5. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling)



2.6. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling)

Al comparar estos resultados con los obtenidos al entrenar los modelos sin remuestreo en el conjunto de entrenamiento, se observa que las métricas del Cuadro 2.2 disminuyen al aplicar oversampling. Sin embargo, los modelos comienzan a predecir ítems de las tres clases de manera más equilibrada, ya no se predice en la gran mayoría de casos la clase mayoritaria (peso normal) solamente. Aun así, la gran cantidad de errores de clasificación (ver Figura 2.6) no hace viable este modelo predictivo.

Modelo entrenado con remuestreo undersampling

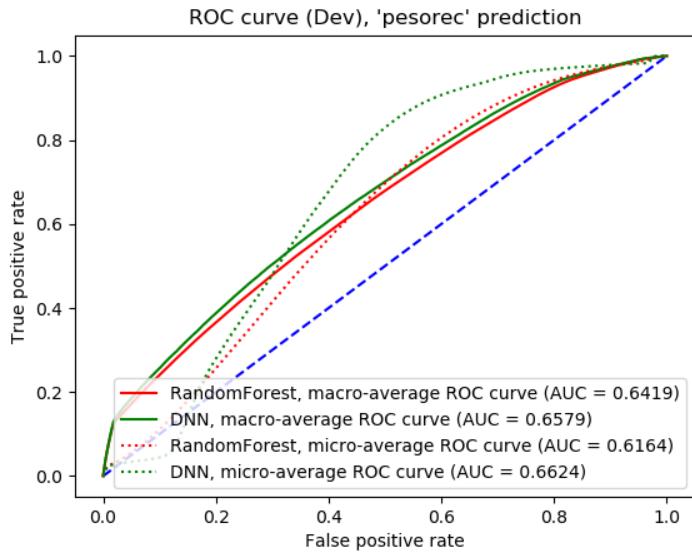
Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\3Experiments\Results\Undersampling\Experiment2

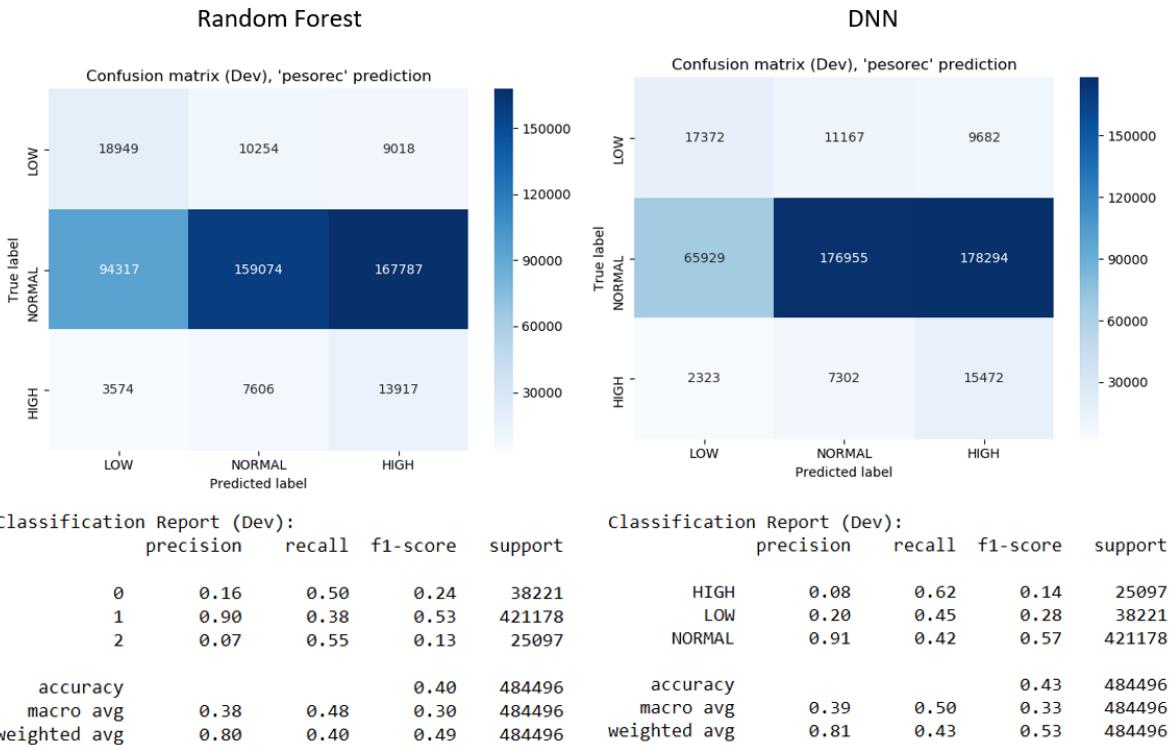
Los modelos se han entrenado balanceando la distribución de la clase *pesorec* al eliminar de forma aleatoria ejemplos de las clases no minoritarias (peso bajo y normal).

pesorec	DEV		
	AUC (micro-average)	Accuracy	F1-score
Random Forest	0.6164	0.40	0.49
DNN	0.6624	0.43	0.53

2.3. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (undersampling)



2.7. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (undersampling)



2.8. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (undersampling)

En estos experimento realizando undersampling en el conjunto de entrenamiento ocurre algo similar al experimento con aplicando oversampling. Las métricas del Cuadro 2.3 disminuyen aún más respecto a las obtenidas con la distribución original. Por otro lado, la capacidad discriminatoria de los modelos entre las tres clases sigue siendo baja, comete muchos errores de clasificación como puede observarse en la Figura 2.8.

Modelo entrenado con remuestreo oversampling (10 %) / undersampling (50 %)

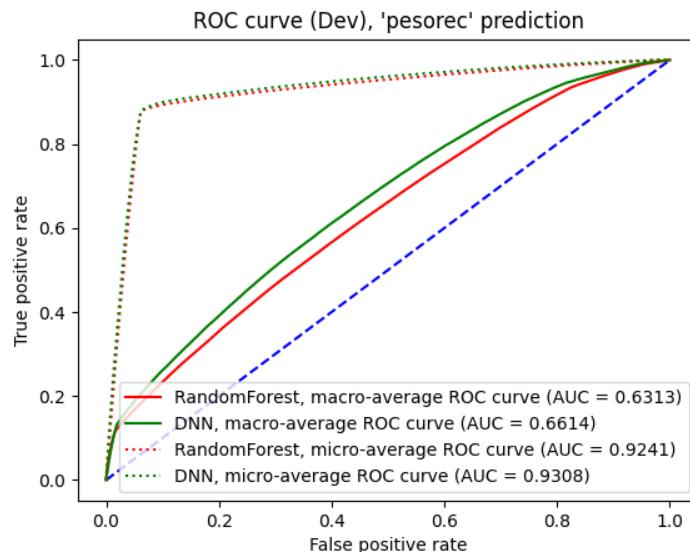
Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\3Experiments\Results\OverUndersampling\Experiment2

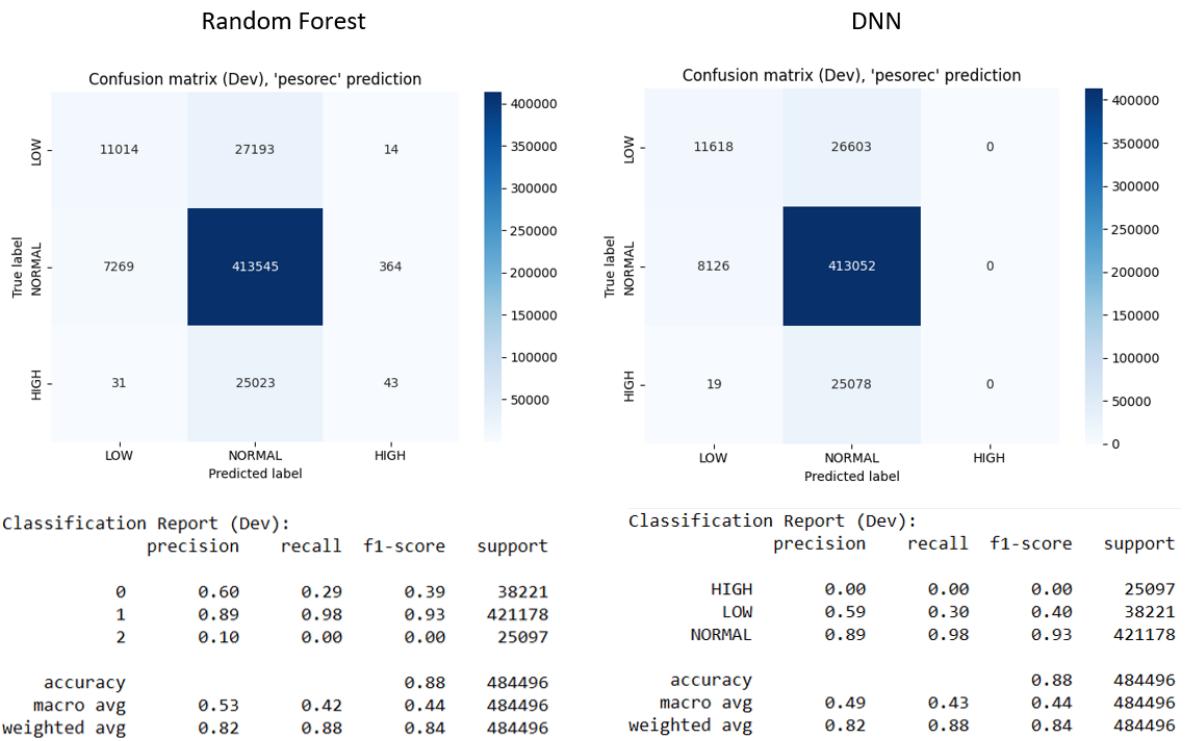
Los modelos se han entrenado balanceando la distribución de la clase *pesorec* al duplicar de forma aleatoria el 10 % de los ítems de las clases no mayoritarias (peso bajo y alto) en primer lugar, y al eliminar de forma aleatoria el 50 % de ejemplos de la clase mayoritaria (peso normal) en segundo lugar.

pesorec	DEV		
	AUC (micro-average)	Accuracy	F1-score
Random Forest	0.9241	0.88	0.84
DNN	0.9308	0.88	0.84

2.4. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))



2.9. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))



2.10. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (oversampling (10 %) / undersampling (50 %))

Los resultados que se muestran en el Cuadro 2.4 de este experimento, tras el remuestreo con oversampling (10 %) / undersampling (50 %), son muy similares a los obtenidos con la distribución original de la clase *pesorec* (ver Cuadro 2.1). Sin embargo, si nos fijamos en la matriz de confusión del modelo Random Forest en la Figura 2.10, vemos como en este caso el modelo comienza a clasificar algunos ítems con peso alto. Aun así, sigue cometiendo más errores que aciertos en las clasificaciones de peso alto y bajo.

Modelo entrenado con remuestreo múltiple undersampling personalizado

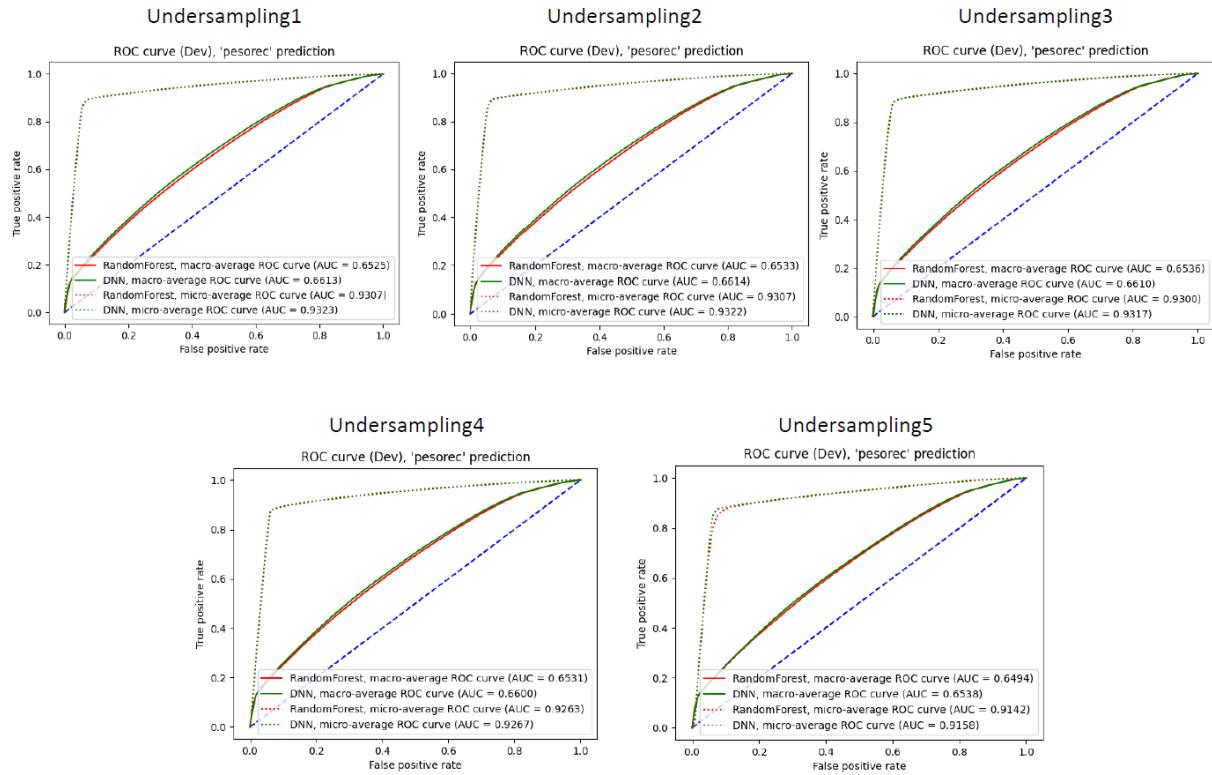
Localización de los resultados:

Perinatal.DatosResultados\Resultados\Perinatal\ExperimentsPesorec\MultipleUndersamplingExperiment

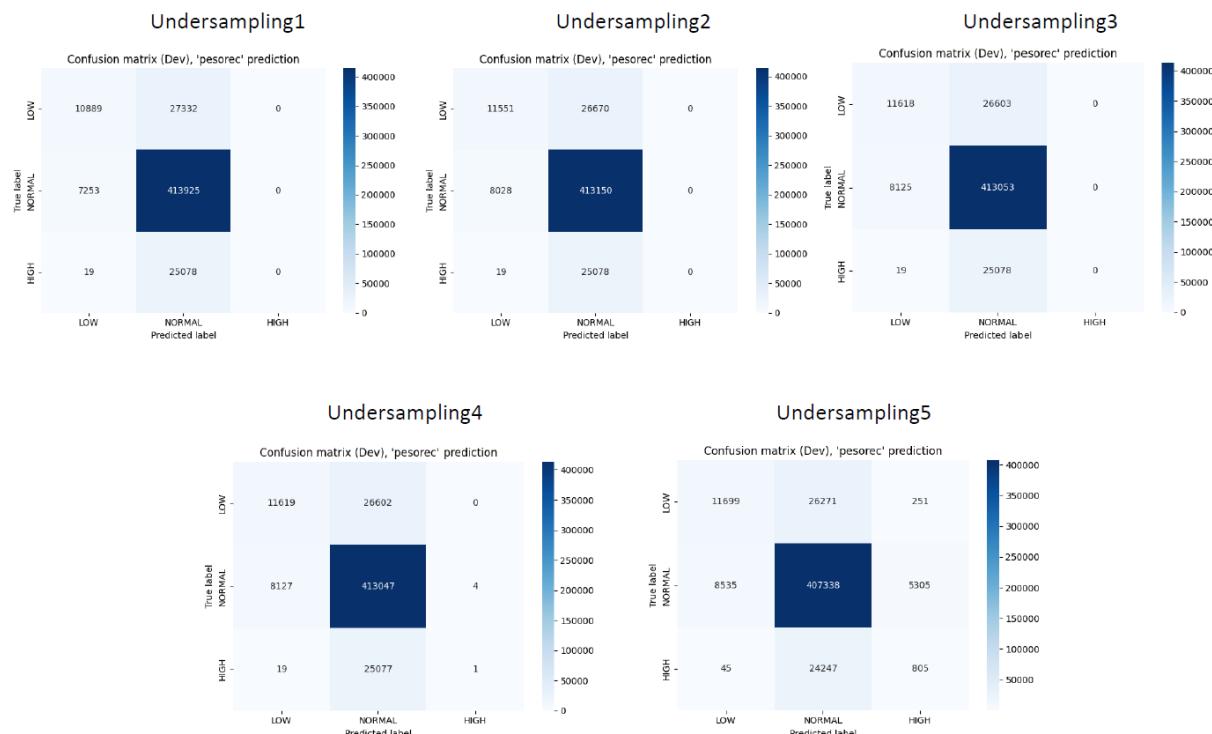
Los modelos se han entrenado balanceando la distribución de la clase *pesorec* utilizando el método de múltiple undersampling personalizado, explicado en la sección de remuestreo del conjunto de datos. A continuación se muestran los resultados de los 5 undersamplings realizados.

pesorec		DEV		
		AUC (micro-average)	Accuracy	F1-score
Random Forest	Undersampling1	0.9307	0.88	0.84
	Undersampling2	0.9307	0.88	0.84
	Undersampling3	0.9300	0.88	0.84
	Undersampling4	0.9263	0.88	0.84
	Undersampling5	0.9142	0.85	0.83
DNN	Undersampling1	0.9323	0.88	0.84
	Undersampling2	0.9322	0.88	0.84
	Undersampling3	0.9317	0.88	0.84
	Undersampling4	0.9267	0.88	0.84
	Undersampling5	0.9158	0.87	0.84

2.5. Cuadro: Resultados de los modelos predictivos de pesorec sobre el conjunto Dev (múltiple undersampling personalizado)



2.11. Figura: Curvas ROC y AUCs de los modelos predictivos de *pesorec* sobre el conjunto Dev (múltiple undersampling personalizado)



2.12. Figura: Matrices de confusión del modelo predictivo DNN de *pesorec* sobre el conjunto Dev (múltiple undersampling personalizado)

Undersampling1					Undersampling2					Undersampling3				
Classification Report (Dev):					Classification Report (Dev):					Classification Report (Dev):				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
HIGH	0.00	0.00	0.00	25097	HIGH	0.00	0.00	0.00	25097	HIGH	0.00	0.00	0.00	25097
LOW	0.60	0.28	0.39	38221	LOW	0.59	0.30	0.40	38221	LOW	0.59	0.30	0.40	38221
NORMAL	0.89	0.98	0.93	421178	NORMAL	0.89	0.98	0.93	421178	NORMAL	0.89	0.98	0.93	421178
accuracy			0.88	484496	accuracy			0.88	484496	accuracy			0.88	484496
macro avg	0.50	0.42	0.44	484496	macro avg	0.49	0.43	0.44	484496	macro avg	0.49	0.43	0.44	484496
weighted avg	0.82	0.88	0.84	484496	weighted avg	0.82	0.88	0.84	484496	weighted avg	0.82	0.88	0.84	484496

Undersampling4					Undersampling5				
Classification Report (Dev):					Classification Report (Dev):				
	precision	recall	f1-score	support		precision	recall	f1-score	support
HIGH	0.20	0.00	0.00	25097	HIGH	0.13	0.03	0.05	25097
LOW	0.59	0.30	0.40	38221	LOW	0.58	0.31	0.40	38221
NORMAL	0.89	0.98	0.93	421178	NORMAL	0.89	0.97	0.93	421178
accuracy			0.88	484496	accuracy			0.87	484496
macro avg	0.56	0.43	0.44	484496	macro avg	0.53	0.44	0.46	484496
weighted avg	0.83	0.88	0.84	484496	weighted avg	0.83	0.87	0.84	484496

2.13. Figura: Métricas de evaluación del modelo predictivo DNN de *pesorec* sobre el conjunto Dev (múltiple undersampling personalizado)

Como puede observarse en el Cuadro 2.5, se obtienen prácticamente los mismos resultados en las métricas AUC, *accuracy* y *F1-score* en todos los experimentos con undersamplings diferentes, aunque es verdad que, según va reduciéndose el número de ítems de entrenamiento, las métricas van disminuyendo ligeramente.

Fijándonos en las matrices de confusión (ver Figura 2.12), en todos los experimentos, los modelos clasifican la gran mayoría de ítems con peso normal. A medida que se reduce el número de ítems en la clase mayoritaria, se clasifican ligeramente mejor los ítems de peso bajo y se cometan menos errores al clasificarlos con peso normal. Sin embargo, en ningún experimento se aprende a clasificar correctamente los pesos altos, exceptuando *Undersampling5*, en el resto de modelos no se hace ninguna clasificación de peso alto. Aunque es verdad que a medida que se reduce el número de ítems en la clase mayoritaria se comienzan a clasificar ítems con peso alto, la gran mayoría se clasifican erróneamente con peso normal.

Conclusiones

En conclusión, tras los múltiples modelos predictivos entrenado con la distribución original y los remuestreos realizados, no se ha obtenido ningún modelo capaz de discriminar entre peso bajo, normal y alto de manera suficientemente precisa. Da la impresión de que ni intentando solucionar el problema de desbalance de la clase *pesorec* se consiguen hacer buenas predicciones con la caracterización utilizada en los ítems del conjunto Perinatal. Los modelos hacen en la gran mayoría de casos predicciones de la clase mayoritaria (peso normal) y consiguen una pobre clasificación de los ítems con peso bajo. Los pesos altos no han sido capaces de ser clasificados.

2.1.2.3. Interpretabilidad (*feature significance*)

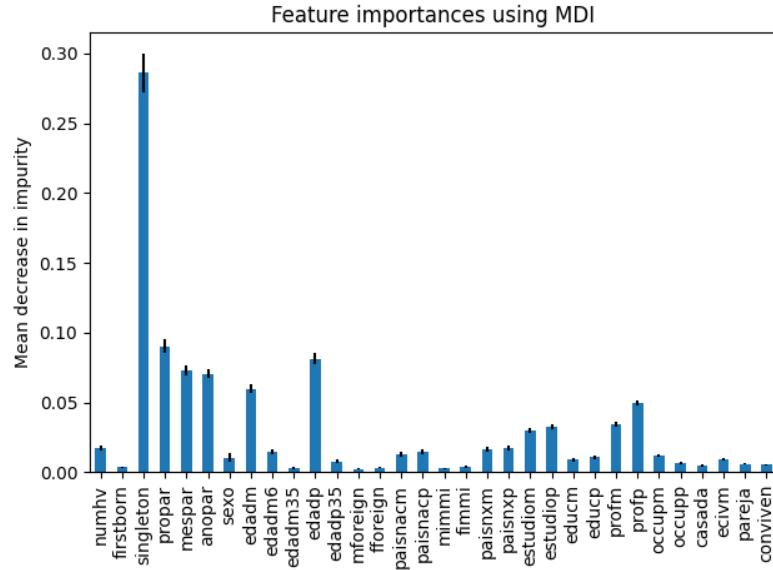
En esta última sección del análisis de los modelos predictivos entrenados se pretende interpretar lo aprendido por los algoritmos de clasificación a través de la importancia de las variables predictoras. Como ya se ha explicado, el algoritmo Random Forest permite conocer la importancia o determinación de las variables predictoras a la hora de realizar las predicciones a través de la importancia de Gini.

A continuación se muestra la importancia de las variables predictoras por cada modelo entrenado en los experimentos con la distribución original de la clase *pesorec* y las técnicas de remuestreo aplicadas.

Modelo entrenado con la distribución original

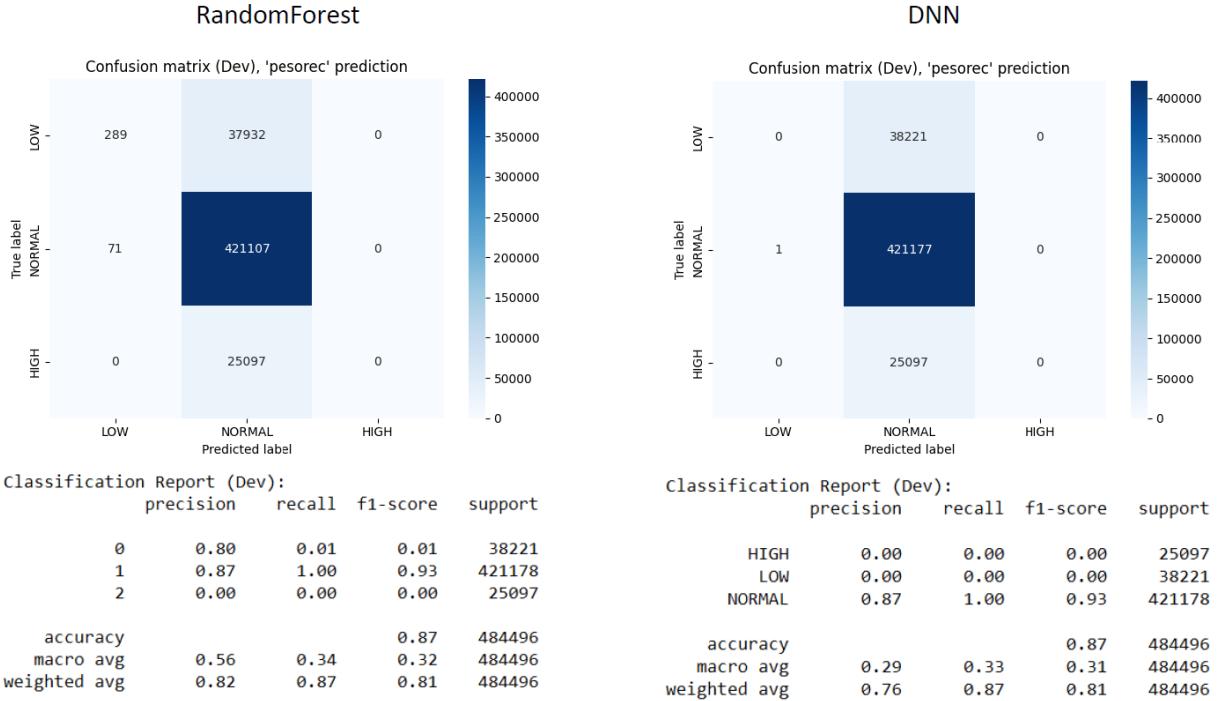
Al entrenar el modelo Random Forest para la clasificación de *pesorec* con su distribución original, la única variable predictora que parece relevante es *singleton* (indicador de hijo/a único/a o múltiple). Aunque sí es verdad que destacan entre otras variables como *propar* (provincia), *mespar* (mes), *anopar*

(año), *edadm* (edad de la madre) y *edadp* (edad del padre), ninguna de ellas parece determinante en las predicciones del peso del recién nacido (ver Figura 2.14).



2.14. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (distribución original)

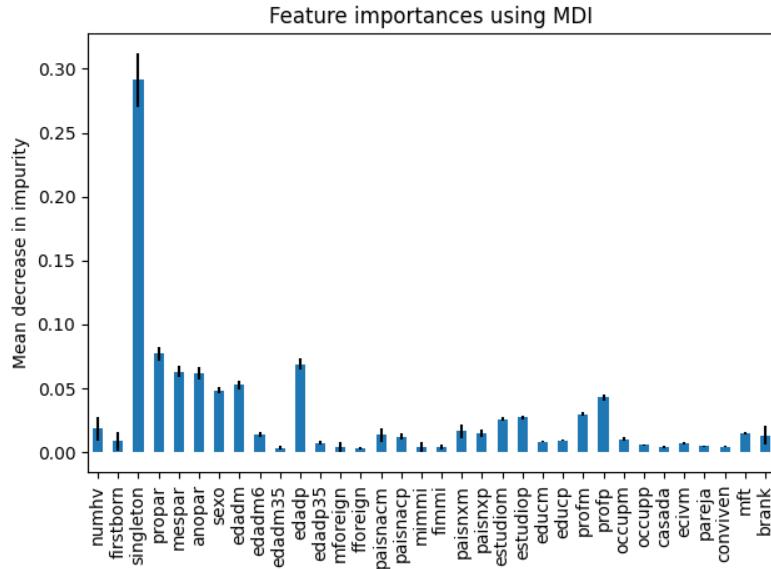
Tras haber observado que *singleton* es la única variable aparentemente determinante, se ha realizado un proceso de *feature ablation* eliminando esta variable y volviendo a entrenar el modelo. Como se puede observar en la Figura 2.15, al eliminar la variable *singleton* los modelos pierden totalmente la pobre capacidad de discriminación que tenían previamente entre las clases de peso bajo y normal, se clasifican todos los ítems con peso normal. Como consecuencia, podemos decir que la variable *singleton* era la única que aportaba en la distinción de las clases peso bajo y normal. También se ha realizado el proceso de *feature ablation* para el resto de variables predictoras, pero los cambios en los resultados son prácticamente irrelevantes.



2.15. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (*feature ablation* eliminando *singleton*)

Modelo entrenado con remuestreo oversampling

Tras entrenar el modelo Random Forest para la clasificación de *pesorec* aplicando la técnica oversampling de remuestreo, *singleton* (indicador de hijo/a único/a o múltiple) sigue siendo la única variable predictora que parece relevante (ver Figura 2.16).

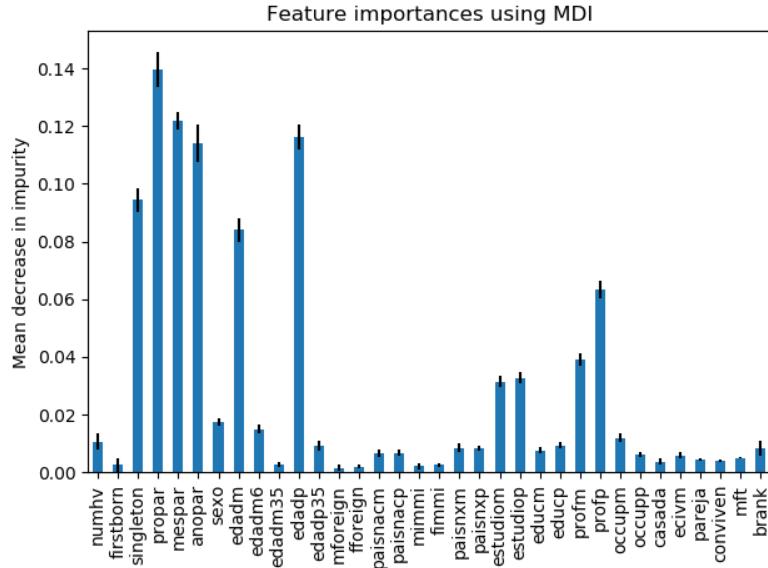


2.16. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (oversampling)

Modelo entrenado con remuestreo undersampling

Al entrenar el modelo Random Forest para la clasificación de *pesorec* aplicando la técnica undersampling de remuestreo, la variable *singleton* pierde determinación en las predicciones del modelo. Siguen

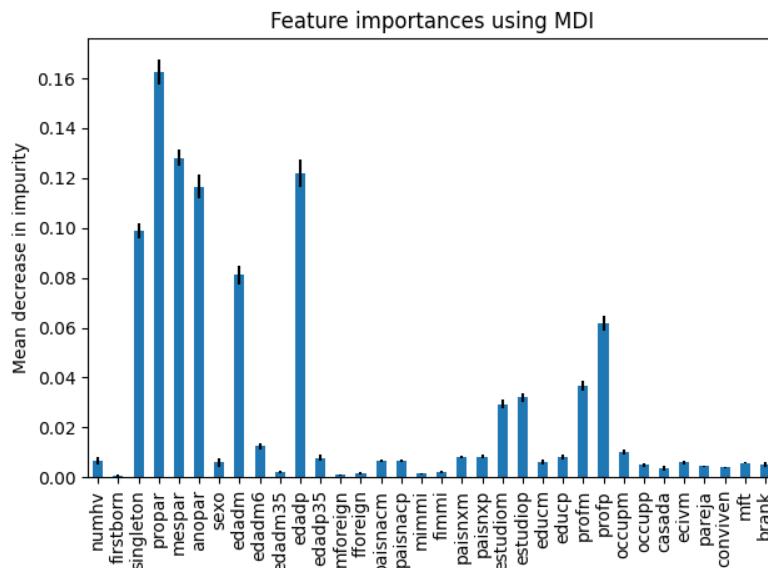
destacando por encima de otras variables como *propar* (provincia), *mespar* (mes), *anopar* (año), *edadm* (edad de la madre), *edadp* (edad del padre), *estudiom* (nivel de estudios de la madre), *estudiop* (nivel de estudios del padre), *profm* (tipo de profesión de la madre) y *profp* (tipo de profesión del padre) (ver Figura 2.17). Sin embargo, como se ha visto en el Cuadro 2.3, los resultados del modelo predictivo no son buenos, y a excepción de las edades de la madre y el padre, el resto de variables socioeconómicas no tienen una gran determinación sobre las predicciones.



2.17. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (undersampling)

Modelo entrenado con remuestreo oversampling (10 %) / undersampling (50 %)

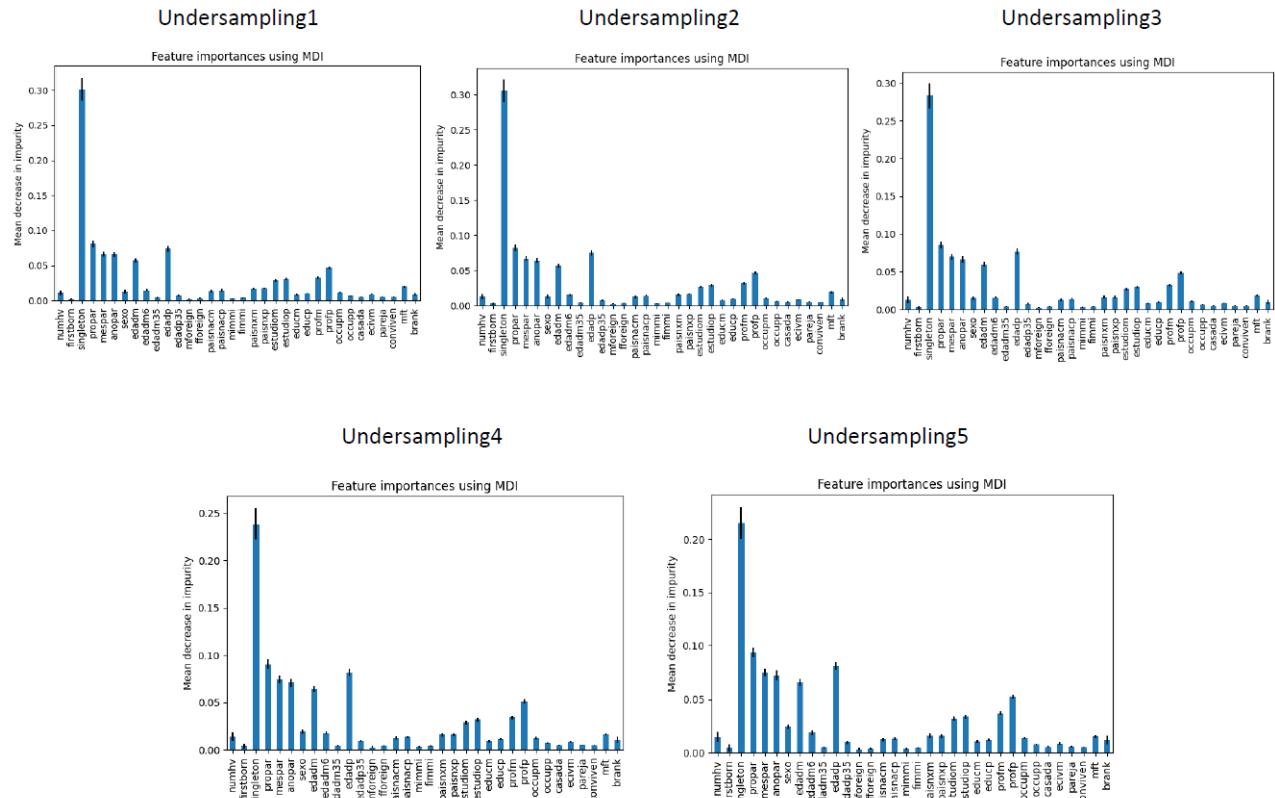
Al entrenar el modelo Random Forest para la clasificación de *pesorec* aplicando la técnica oversampling (10 %) / undersampling (50 %) de remuestreo, se da una situación muy similar a la que veíamos al aplicar undersampling. La provincia, el mes y el año del parto adquieren importancia, y entre las variables socioeconómicas, las edades de la madre y el padre, y sus estudios y profesiones parecen tener mayor relevancia comparadas con el resto.



2.18. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (oversampling (10 %) / undersampling (50 %))

Modelo entrenado con remuestreo múltiple undersampling personalizado

Por último, al entrenar modelos Random Forest para la clasificación de *pesorec* aplicando el proceso de múltiple undersampling personalizado, se observa que a medida que se reduce el número de ítems en la clase mayoritaria (se van eliminando los ítems con pesos más comunes), disminuye la importancia de *singleton* y aumenta ligeramente la de variables como *propar*, *mespar*, *anopar*, *edadm* o *edadp* (ver Figura 2.19). Aún así no se ve un aumento relevante en la importancia de las variables socioeconómicas, y ninguna de ellas llega a ser determinante en las predicciones.



2.19. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (múltiple undersampling personalizado)

Conclusiones

Tras el análisis de la importancia de variables predictoras en las predicciones de los modelos entrenados con la distribución original de la clase *pesorec* y los diferentes remuestreos aplicados, podemos llegar a la conclusión de que la variable *singleton* (indicador de hijo/a único/a o múltiple) es la única determinante en las clasificaciones de los modelos. Concretamente, tiene una gran importancia discriminatoria entre las clasificaciones de peso bajo y normal.

Las variables socioeconómicas referentes a la madre y el padre del recién nacido no parecen tener relevancia en las decisiones del modelo predictivo, aunque sí es verdad que variables como *edadm* (edad de la madre), *edadp* (edad del padre), *estudiom* (nivel de estudios de la madre), *estudiop* (nivel de estudios del padre), *profm* (tipo de profesión de la madre) y *profsp* (tipo de profesión del padre) destacan sobre el resto. También se ha observado que a medida que se balancea la distribución de la clase disminuyendo el número de ítems de entrenamiento (undersampling de la clase mayoritaria de peso normal), estas variables van aumentando ligeramente su relevancia. Sin embargo, no mejoran los resultados de las clasificaciones. Por último, se ha observado una mayor relevancia en las variables *propar*, *mespar* y *anopar* (provincia, mes y año del parto, respectivamente), las cuales no deberían suponer en principio una mayor importancia para las predicciones de peso del recién nacido.

2.2. EXPERIMENTO 2: clasificación de *pesorec* con *fake features* de estimación del peso del feto

Tras el *Experimento 1*, se ha observado que las variables predictoras socioeconómicas del conjunto de datos Perinatal no son capaces por sí solas de realizar buenas predicciones en los pesos de los recién nacidos. Por ello, este segundo experimento trata de incluir una variable predictora artificial (*fake feature*) que pretenda mejorar las predicciones de los modelos y observar cómo influye esta nueva variable en la importancia del resto de variables socioeconómicas.

Como se ha explicado en el apartado referente a la mezcla de contextos entre el bajo peso y la prematuridad, la idea propuesta ha sido agregar una nueva variable *peson-semanas* que exprese el peso que gana el feto por semana, tratando la relación entre la ganancia de peso y la duración del embarazo de manera lineal. Posteriormente, también se han realizado más experimentos añadiendo variables que estiman el peso del feto en una semana concreta del embarazo a través de las *porportionality growth functions* encontradas. Concretamente, se van a crear modelos predictivos para los siguientes experimentos:

- Estimando la feature *peso_semana_32* con la fórmula aplicada sobre la población china [1].
- Estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población china [1].
- Estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población inglesa [2].
- Estimando la feature *peso_semana_12* con la fórmula aplicada sobre la población inglesa [2].

En los cinco experimentos mencionados se van a entrenar, en primer lugar, *baselines* únicamente con la edad de la madre y el peso estimado del feto, y en segundo lugar, modelos predictivos añadiendo las características socioeconómicas del conjunto Perinatal.

2.2.1. Algoritmos empleados

Respecto a los algoritmos de clasificación supervisada para la creación de los modelos predictivos, se han empleado los mismos que en el *Experimento 1*: **Random Forest** y **Deep Neural Network (DNN)**. El algoritmo Random Forest nos será de gran ayuda para la interpretabilidad del modelo, mostrando la importancia de las variables predictoras en las predicciones realizadas por el modelo de clasificación; mientras que la red neuronal profunda (DNN) quizás nos sirva para obtener mejores resultados, ya que al trabajar con grandes conjuntos de datos las redes neuronales pueden ofrecer un mayor rendimiento.

2.2.2. Planteamiento de la predicción

En este apartado se van a mostrar los resultados y las evaluaciones de los modelos predictivos entrenados para este experimento.

2.2.2.1. Tipo de clasificación

El problema de **clasificación** sigue siendo **multiclasificación**, ya que pretende clasificarse correctamente la clase *pesorec* del conjunto Perinatal en tres categorías: **peso bajo, normal y alto**. En esta ocasión, para el entrenamiento de los modelos, va a utilizarse la distribución original del *dataset* Perinatal (ver Figura 1.1).

2.2.2.2. Resultados y evaluación de los modelos

A continuación se muestran los resultados de los modelos predictivos entrenados y evaluados sobre el conjunto Dev.

Modelo agregando *peson_semanas*

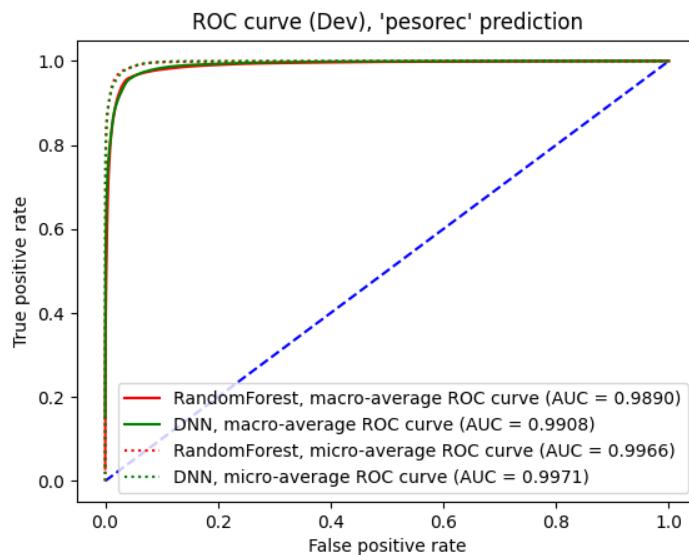
Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\PesonSemanasExperiments

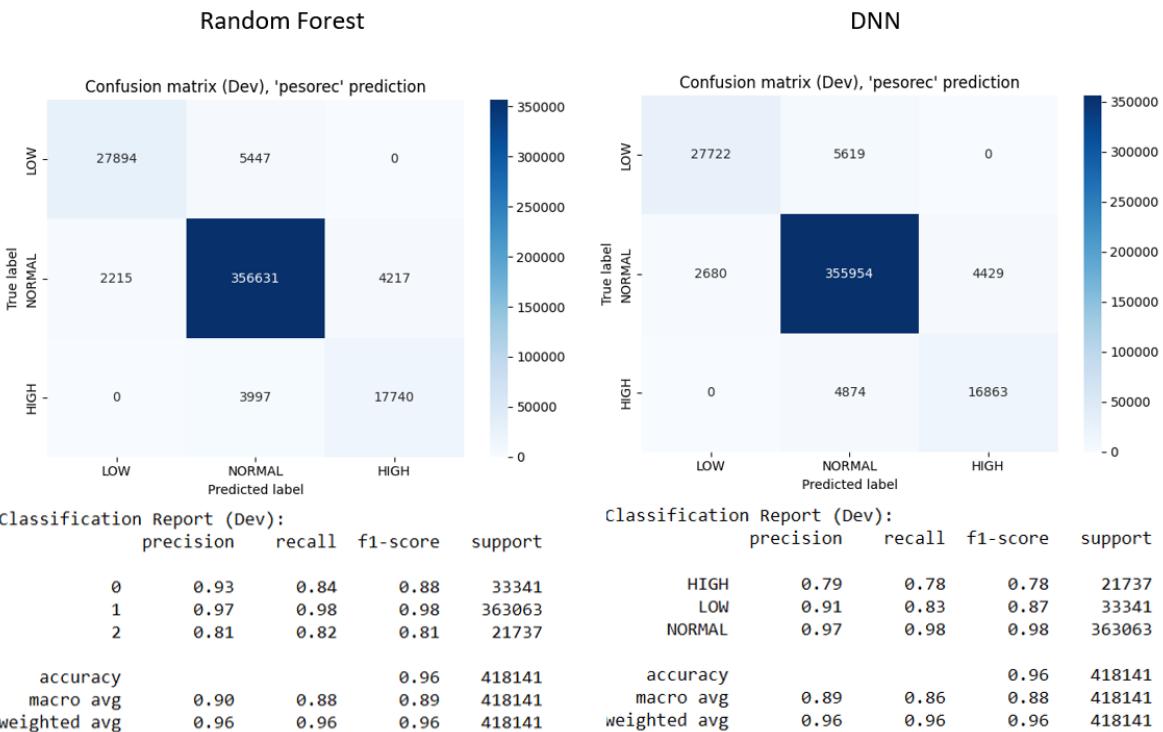
Los modelos se han entrenado tras agregar la variable *peson-semanas* que representa la ganancia de peso del feto por semana, suponiendo que la relación entre el peso y el tiempo fuera lineal, (explicado en la sección referente a la problemática entre el bajo peso y la prematuridad).

pesorec	DEV		
	AUC (micro-average)	Accuracy	F1-score
Random Forest	0.9966	0.96	0.96
DNN	0.9971	0.96	0.96

2.6. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando *peson_semanas*)



2.20. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando *peson_semanas*)



2.21. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando *peson_semanas*)

Como puede observarse en el Cuadro 2.6, se consiguen clasificaciones muy exactas, obteniendo una tasa de acierto del 96 % en las clasificaciones del peso. Si nos fijamos también en las matrices de confusión (ver Figura 2.21), los dos modelos aprenden a clasificar la gran mayoría de ítems correctamente en las tres clases: peso bajo, normal y alto. Estos buenos resultados nos llevan a la conclusión de que la nueva variable *peson_semanas* introducida es de gran ayuda para las predicciones (como era de esperar), pero quizás anula al resto de variables socioeconómicas, siendo esta la única determinante en el modelo.

Modelo agregando *peso_semana_XX*

Localización de los resultados:

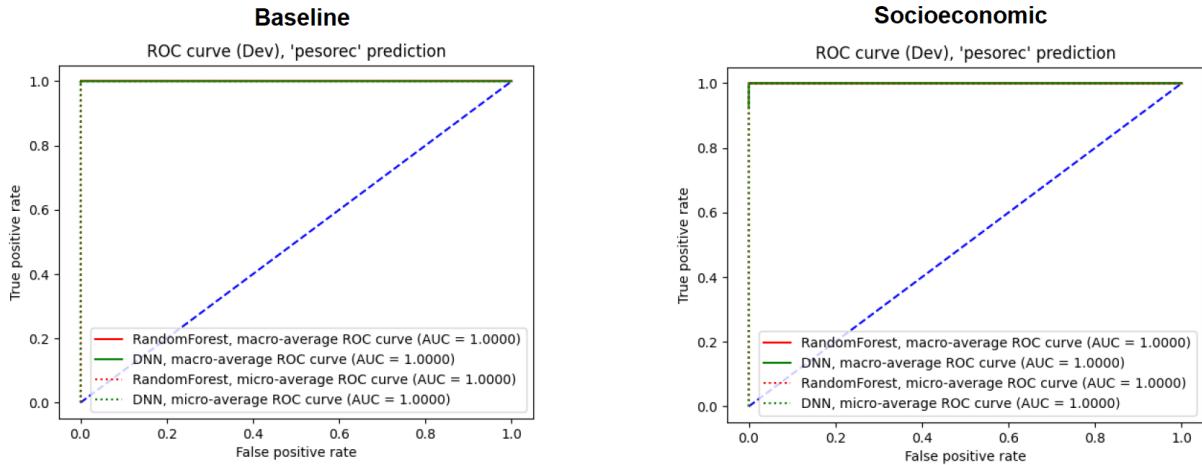
Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\PesoSemanaXXExperiments

En el apartado anterior, se ha tratado de manera lineal la relación entre la ganancia de peso del feto y la duración del embarazo. Sin embargo, sabemos que la relación es más bien logarítmica. Por ello, es más adecuado utilizar las *propotionality growth functions* mencionadas en el apartado referente a la problemática entre el bajo peso y la prematuridad, las cuales permiten estimar el peso del feto en una semana concreta del embarazo de manera más real.

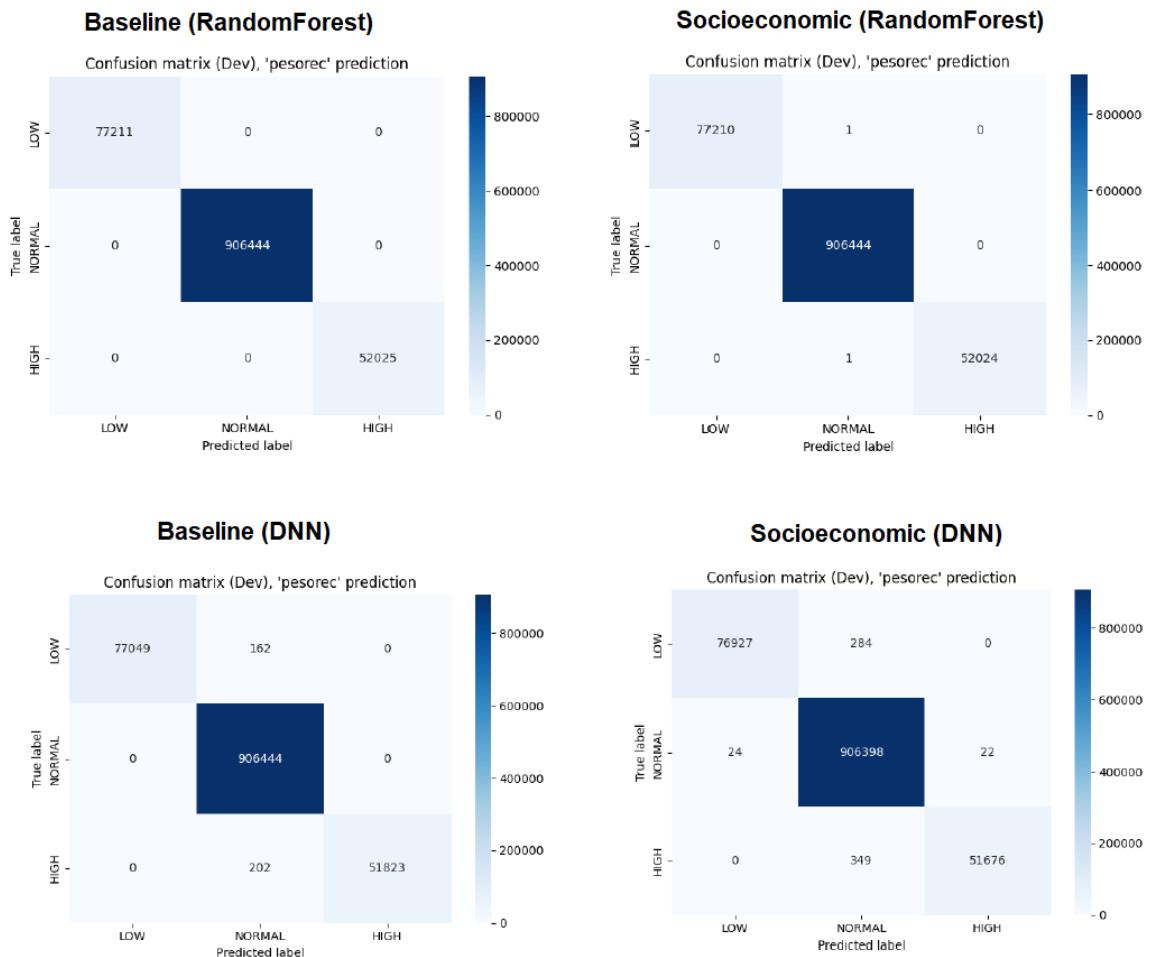
Los modelos evaluados en esta sección se han entrenado tras agregar la variable *peso_semana_XX* que representa la estimación del peso del feto en la semana XX del embarazo. Los resultados mostrados a continuación pertenecen a los modelos entrenados agregando la estimación del peso del feto en la semana 32 de embarazo y utilizando la *propotionality growth function* testeada sobre la población china [1].

pesorec		DEV		
		AUC (micro-average)	Accuracy	F1-score
Random Forest	Baseline	1	1	1
	Socioeconomic	1	1	1
DNN	Baseline	1	1	1
	Socioeconomic	1	1	1

2.7. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando *peso_semana_32* con la ecuación testeada sobre la población china [1])



2.22. Figura: Curvas ROC y AUCs de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando *peso_semana_32* con la ecuación testeada sobre la población china [1])



2.23. Figura: Matrices de confusión del modelo predictivo DNN de *pesorec* sobre el conjunto Dev (agregando *peso_semana_32* con la ecuación testeada sobre la población china [1])

Baseline (RandomForest)					Socioeconomic (RandomForest)				
Classification Report (Dev):					Classification Report (Dev):				
	precision	recall	f1-score	support		precision	recall	f1-score	support
LOW	1.00	1.00	1.00	77211	LOW	1.00	1.00	1.00	77211
NORMAL	1.00	1.00	1.00	906444	NORMAL	1.00	1.00	1.00	906444
HIGH	1.00	1.00	1.00	52025	HIGH	1.00	1.00	1.00	52025
accuracy			1.00	1035680	accuracy			1.00	1035680
macro avg	1.00	1.00	1.00	1035680	macro avg	1.00	1.00	1.00	1035680
weighted avg	1.00	1.00	1.00	1035680	weighted avg	1.00	1.00	1.00	1035680

Baseline (DNN)					Socioeconomic (DNN)				
Classification Report (Dev):					Classification Report (Dev):				
	precision	recall	f1-score	support		precision	recall	f1-score	support
HIGH	1.00	1.00	1.00	52025	HIGH	1.00	0.99	1.00	52025
LOW	1.00	1.00	1.00	77211	LOW	1.00	1.00	1.00	77211
NORMAL	1.00	1.00	1.00	906444	NORMAL	1.00	1.00	1.00	906444
accuracy			1.00	1035680	accuracy			1.00	1035680
macro avg	1.00	1.00	1.00	1035680	macro avg	1.00	1.00	1.00	1035680
weighted avg	1.00	1.00	1.00	1035680	weighted avg	1.00	1.00	1.00	1035680

2.24. Figura: Métricas de evaluación del modelo predictivo DNN de *pesorec* sobre el conjunto Dev (agregando *peso_semana_32* con la ecuación testeada sobre la población china [1])

Tal y como se muestra en el Cuadro 2.7, tanto con el modelo *baseline* únicamente entrenado con la edad de la madre y la variable *peso_semana_32* como con el modelo añadiendo las variables socioeconómicas de Perinatal, se obtienen unos resultados del 100 % de aciertos. Si nos fijamos en las matrices de confusión (ver Figura 2.23), prácticamente no se cometan errores de clasificación en ninguna clase, ni con el modelo *baseline* ni con el *socioeconomic*. Sin embargo, es verdad que los modelos entrenados con la DNN cometen una cantidad un poco mayor de errores, pero siguen siendo insignificantes.

Estos resultados tan buenos nos llevan a la conclusión de que la estimación del peso del feto en la semana 32 mediante la *propotionality growth function* testeada sobre la población china [1] es una gran predictora del peso final del recién nacido. Sin embargo, de la misma manera que ocurría al agregar la variable *peson_semanas*, podemos intuir que la variable introducida se lleva toda la determinación predictiva del modelo, anulando la influencia del resto de variables socioeconómicas.

El resto de experimentos realizados adelantando la semana de estimación del peso del feto,

- estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población china [1],
- estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población inglesa [2],
- y estimando la feature *peso_semana_12* con la fórmula aplicada sobre la población inglesa [2],

también han concluido con los mismos resultados.

Conclusiones

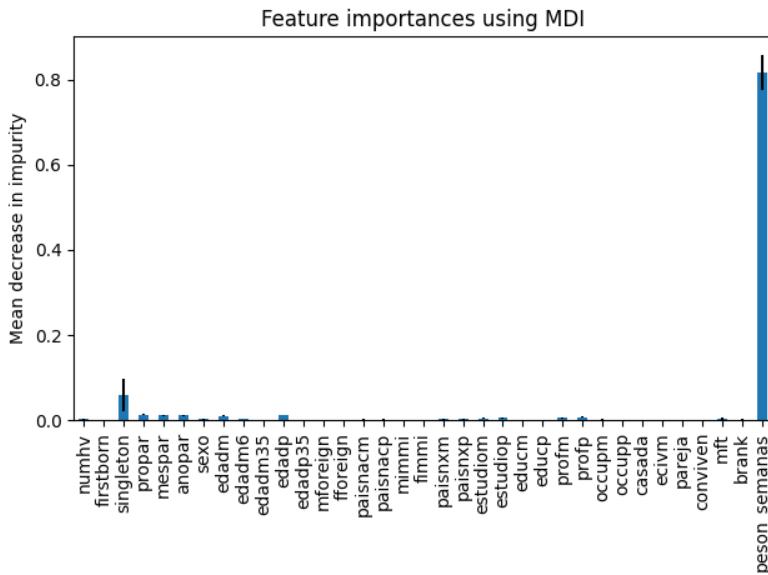
Agregar una *fake feature* que exprese la ganancia de peso del feto o el peso en una semana concreta del embarazo, convierte los modelos predictivos de peso final en óptimos, con una tasa de acierto cercana al 100 %. Sin embargo, esto era de esperar, ya que se está introduciendo una variable que tiene en cuenta, en gran parte, el peso final del recién nacido. Se han realizado experimentos adelantando la semana de estimación del peso del feto, intentando de esta manera ver si los resultados bajaban del 100 % de exactitud. Aun así, hasta con la estimación del peso en la semana 12 en la cual se debería de equilibrar más el peso de todos los fetos, los modelos consiguen hacer predicciones muy exactas gracias a la ayuda de esta variable.

2.2.2.3. Interpretabilidad (*feature significance*)

En este apartado del análisis de los modelos predictivos se va a interpretar la importancia de las variables predictoras a la hora de realizar las predicciones, a través de las importancias (Gini) otorgadas por el algoritmo Random Forest a cada una de ellas.

Modelo agregando *peson_semanas*

En el modelo agregando la variable *peson_semanas* (ganancia de peso del feto por semana de embarazo, relación lineal), se observa como esta variable introducida adquiere prácticamente toda la determinación a la hora de realizar las predicciones (ver Figura 2.25). Del resto de variables, solamente *singleton* (indicador de hijo/a único/a o múltiple) tiene una mínima influencia, pero es irrelevante. Como intuimos al conocer los resultados del modelo con una tasa de acierto del 96 % en las clasificaciones, el algoritmo aprende toda su capacidad discriminatoria de la variable *peson_semanas* agregada.



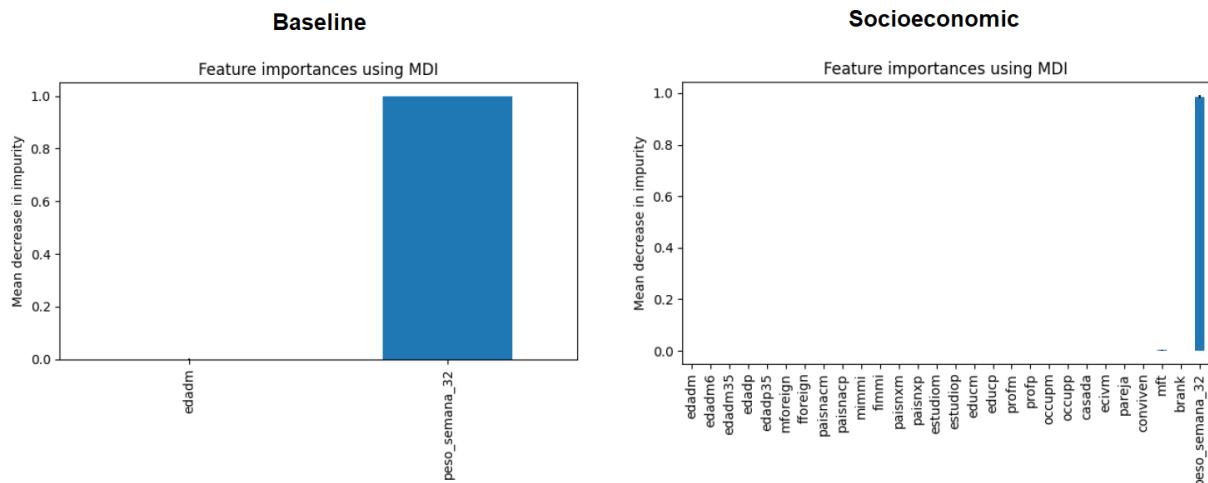
2.25. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (agregando *peso_semanas*)

Modelo agregando *peso_semana_XX*

La misma situación se da al agregar la variable *peso_semana_32*, la cual estima el peso del feto en la semana 32 del embarazo mediante la *propotionality growth function* testeada sobre la población china [1]. Tanto en el modelo *baseline* (solo añadiendo la edad de la madre) como en el modelo *socioeconomic* (añadiendo el resto de variables socioeconómicas de perinatal), la variable *peso_semana_32* toma toda la determinación de las predicciones del modelo (ver Figura 2.26).

Lo mismo ocurre en el resto de experimentos realizados adelantando la semana de estimación del peso del feto:

- estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población china [1],
- estimando la feature *peso_semana_22* con la fórmula aplicada sobre la población inglesa [2],
- y estimando la feature *peso_semana_12* con la fórmula aplicada sobre la población inglesa [2].



2.26. Figura: *Feature significance* de los modelos predictivos Random Forest de *pesorec* (agregando *peso_semana_32* con la ecuación testeada sobre la población china [1])

Conclusiones

En conclusión, tras ver las importancias de las variables predictoras en los modelos entrenados, al agregar una estimación del peso del feto durante el embarazo como variable, se anula prácticamente la determinación del resto de variables socioeconómicas en las predicciones del peso final de los recién nacidos. Por lo tanto, no nos ha servido este experimento para resaltar la influencia de las variables socioeconómicas al añadir estimaciones del peso fetal, ya que se convierten en irrelevantes para el modelo predictivo.

2.3. EXPERIMENTO 3: clasificación de *pesorec* agregando variables de ENSE 2017

En el *Experimento 1*, se han entrenado modelos predictivos de peso de un recién nacido utilizando las características originales del conjunto de datos Perinatal. En el *Experimento 2*, se ha agregado a esas características la estimación del peso del feto durante el embarazo, con el objetivo de mejorar las predicciones y analizar los cambios en la importancia del resto de variables. Sin embargo, en ninguno de los dos experimentos anteriores se han obtenido buenas predicciones solamente con el uso de las variables socioeconómicas del conjunto Perinatal, ninguna de ellas parecía ser realmente determinante.

Por ello, en este tercer experimento, van a agregarse variables de la Encuesta Nacional de Salud de España (ENSE) del 2017, con las cuales se esperan mejorar las predicciones sobre el peso de los recién nacidos. Como se ha explicado en el apartado referente a la descripción del corpus del conjunto ENSE 2017, las características de la ENSE que van a agregarse como variables del conjunto Perinatal en un principio van a ser los **indicadores de consumo de tabaco y alcohol de la madre y el padre: *fumam* y *fumap*** (consumo de tabaco de la madre y el padre, respectivamente), y ***alcoholm* y *alcoholp*** (consumo de alcohol de la madre y el padre, respectivamente).

Perinatal y ENSE 2017 son dos *datasets* diferentes, en los que sus ítems no pertenecen a los mismos sujetos. Por lo tanto, la única manera de agregar variables de la ENSE al conjunto Perinatal es a través de predicciones. Para ello, tal y como se ha explicado en el apartado *Selección de atributos y atributos comunes entre ambos datasets*, se ha creado un conjunto de datos a partir de los ítems de ENSE 2017 compatible con las características del conjunto Perinatal, y con este conjunto compatible se han creado modelos predictivos de las cuatro variables que queremos agregar al conjunto Perinatal. A continuación se hace un resumen de los modelos predictivos creados para las variables *fumam*, *fumap*, *alcoholm* y *alcoholp*.

Localización de los ficheros:

Perinatal_DatosResultados\Datos\ENSE2017\FumaDatasets\dataENSE2017_compatible_m_fuma.csv
Perinatal_DatosResultados\Datos\ENSE2017\FumaDatasets\dataENSE2017_compatible_p_fuma.csv
Perinatal_DatosResultados\Datos\ENSE2017\AlcoholDatasets\dataENSE2017_compatible_m_alcohol.csv
Perinatal_DatosResultados\Datos\ENSE2017\AlcoholDatasets\dataENSE2017_compatible_p_alcohol.csv

Localización de los resultados:

Perinatal_DatosResultados\Resultados\ENSE2017\PerinatalWithENSE\ModelsFeaturePredictors

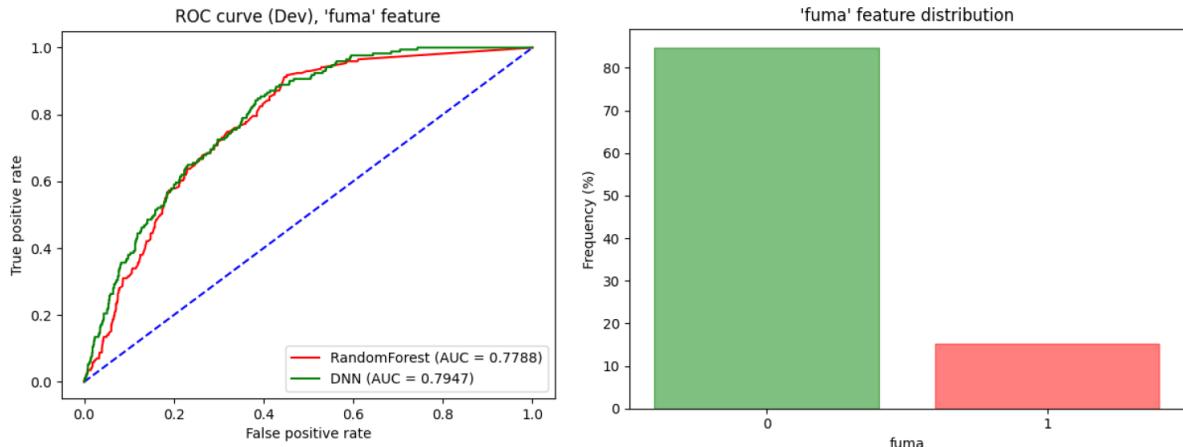
Predicción de *fumam*

La variable *fumam* (binaria) es el indicador de consumo de tabaco de las mujeres en el conjunto ENSE 2017. Sus predicciones sobre el conjunto de datos Perinatal se agregaán como los indicadores de consumo de tabaco de las madres.

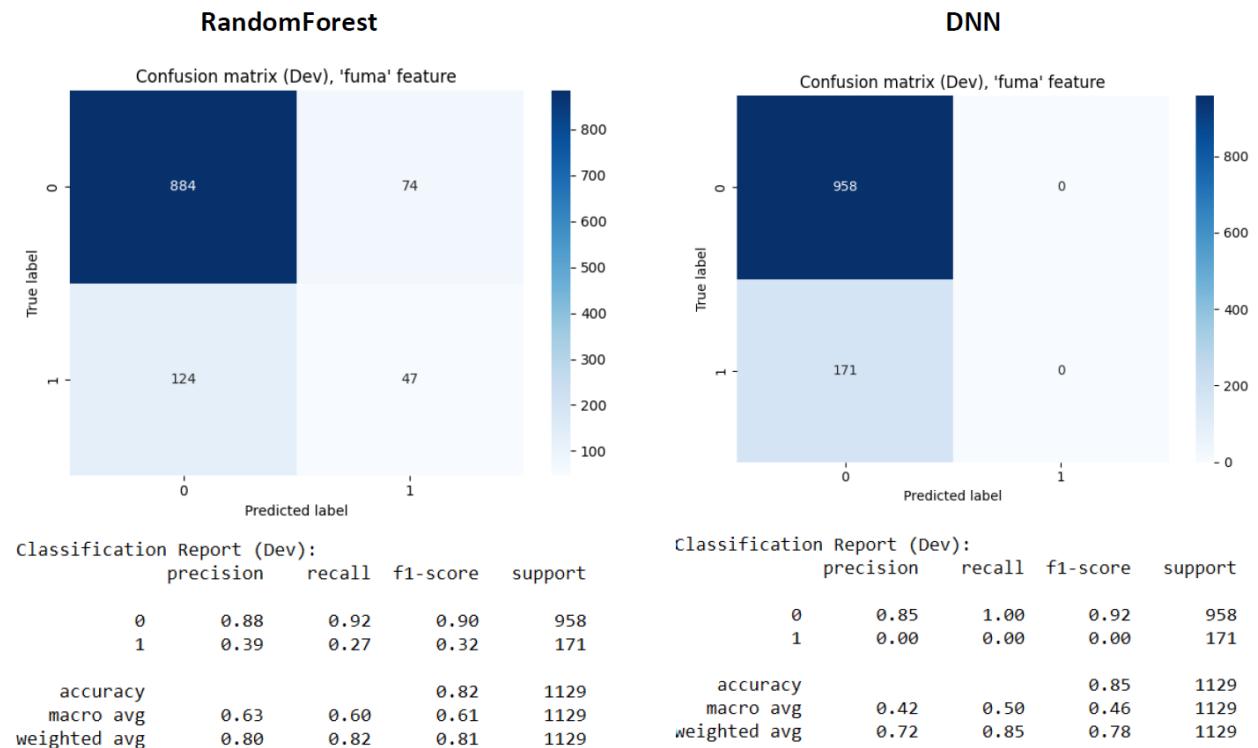
Como puede observarse en el Cuadro 2.8 y en las Figuras 2.27 y 2.28, los modelos predictivos no llegan a discriminar las dos clases (fumadora o no fumadora) de manera muy correcta, se clasifican la mayoría de ítems con la clase mayoritaria (no fumadora) que representa más del 80 % de los ejemplos. Las variables predictoras con más importacia son la edad (*edadm*), la comunidad autónoma (*ccaam*) y el nivel de estudios (*estudiom*) (ver Figura 2.29). Sin embargo, es evidente que solo con las variables utilizadas los modelos no son capaces de realizar buenas predicciones sobre el consumo de tabaco.

<i>fumam</i>	DEV		
	AUC	Accuracy	F1-score
Random Forest	0.7788	0.82	0.81
DNN	0.7947	0.85	0.78

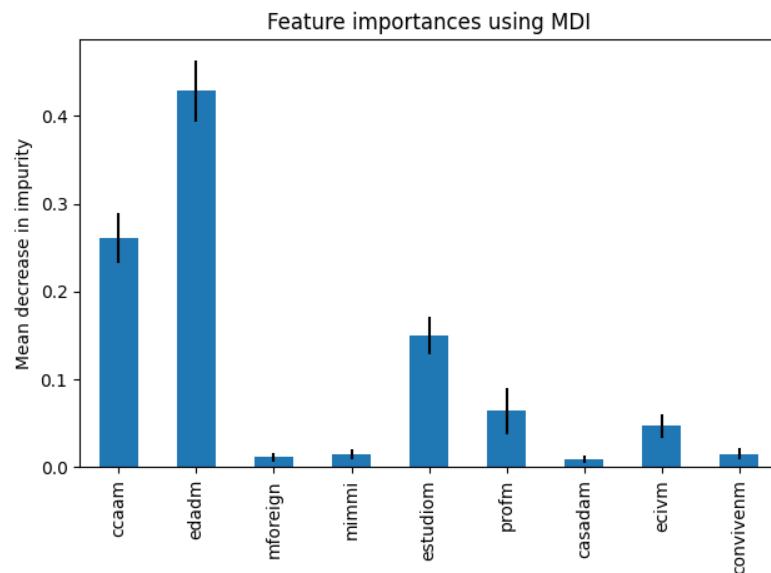
2.8. Cuadro: Resultados de los modelos predictivos de *fumam* sobre el conjunto Dev



2.27. Figura: Curva ROC y AUC de los modelos predictivos de *fumam* sobre el conjunto Dev, y distribución de la clase *fumam* en mujeres de la ENSE 2017



2.28. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *fumam* sobre el conjunto Dev



2.29. Figura: *Feature significance* del modelo predictivo Random Forest de *fumam*

Predicción de *fumap*

La variable *fumap* (binaria) es el indicador de consumo de tabaco de los hombres en el conjunto ENSE 2017. Sus predicciones sobre el conjunto de datos Perinatal se agregrán como los indicadores de consumo de tabaco de los padres.

Con el modelo predictivo de *fumap* ocurre lo mismo que con el anterior clasificador de consumo de tabaco para las mujeres. En este caso, más del 70% de los hombres encuestados no fuma, y el modelo no consigue discriminar de una manera muy correcta las dos clases (fumador y no fumador) a través de las variables predictoras.

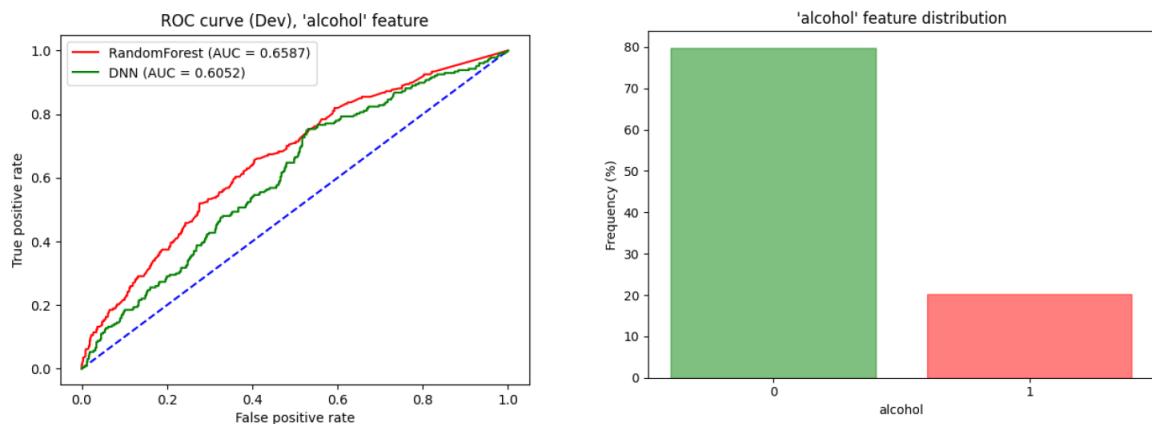
Predicción de *alcoholm*

La variable *alcoholm* (binaria) es el indicador de consumo de alcohol de las mujeres en el conjunto ENSE 2017. Sus predicciones sobre el conjunto de datos Perinatal se agregrán como los indicadores de consumo de alcohol de las madres.

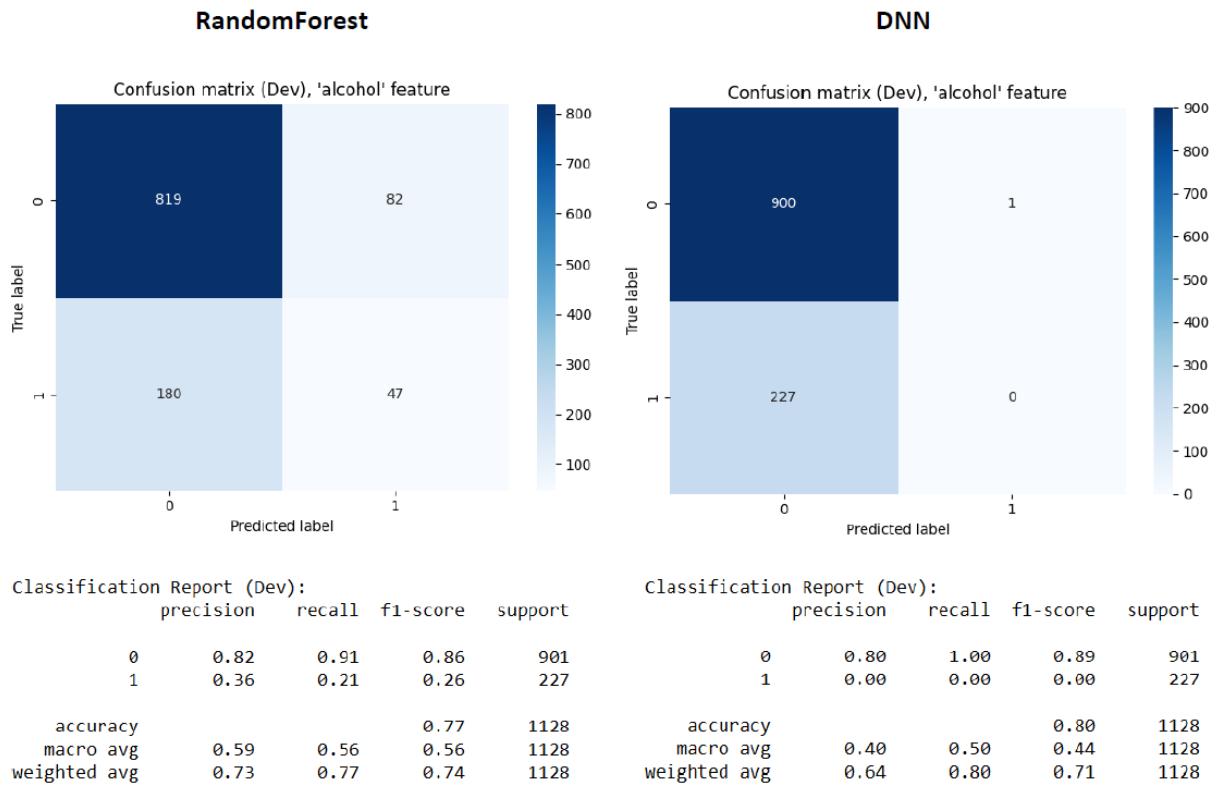
Como puede observarse en el Cuadro 2.9 y en las Figuras 2.30 y 2.31, los modelos predictivos tampoco consiguen discriminar las dos clases de manera muy correcta en el caso del consumo de alcohol, se clasifican la mayoría de ítems con la clase mayoritaria (no consumidora de alcohol) que representa el 80% de los ejemplos. Las variables predictoras con más importancia siguen siendo la edad (*edad*), la comunidad autónoma (*ccaa*) y el nivel de estudios (*estudios*) (ver Figura 2.32), pero parece que siguen sin ser suficientes para realizar buenas predicciones sobre el consumo de alcohol.

<i>alcoholm</i>	DEV		
	AUC	Accuracy	F1-score
Random Forest	0.6587	0.77	0.74
DNN	0.6052	0.80	0.71

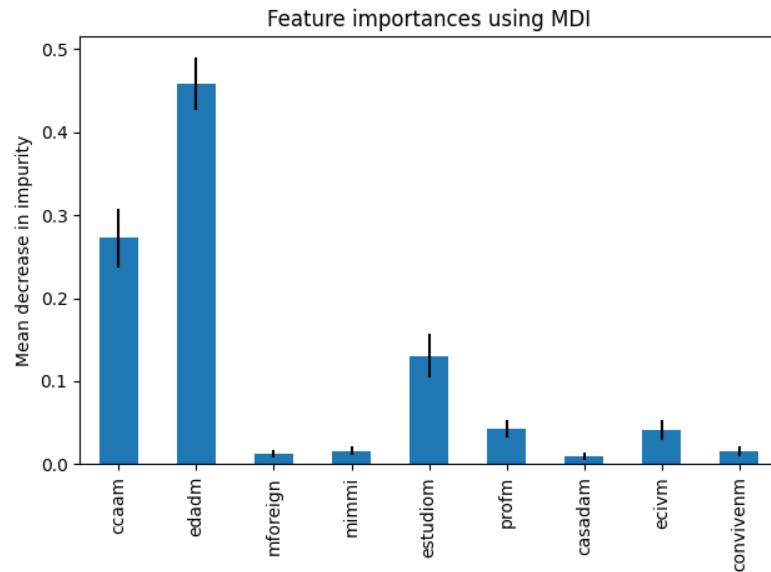
2.9. Cuadro: Resultados de los modelos predictivos de *alcoholm* sobre el conjunto Dev



2.30. Figura: Curva ROC y AUC de los modelos predictivos de *alcoholm* sobre el conjunto Dev, y distribución de la clase *alcoholm* en mujeres de la ENSE 2017



2.31. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *alcoholm* sobre el conjunto Dev



2.32. Figura: *Feature significance* del modelo predictivo Random Forest de *alcoholp*

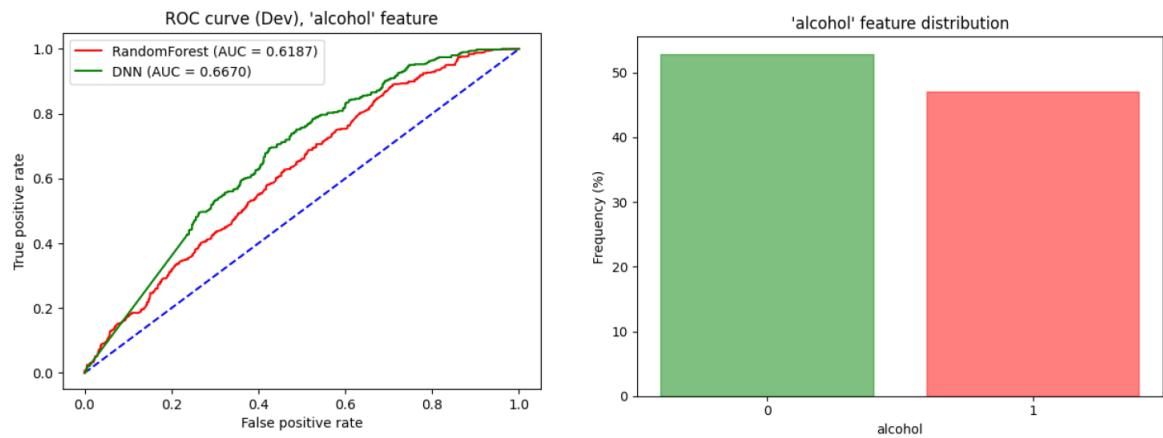
Predicción de *alcoholp*

La variable *alcoholp* (binaria) es el indicador de consumo de alcohol de los hombres en el conjunto ENSE 2017. Sus predicciones sobre el conjunto de datos Perinatal se agregrán como los indicadores de consumo de alcohol de los padres.

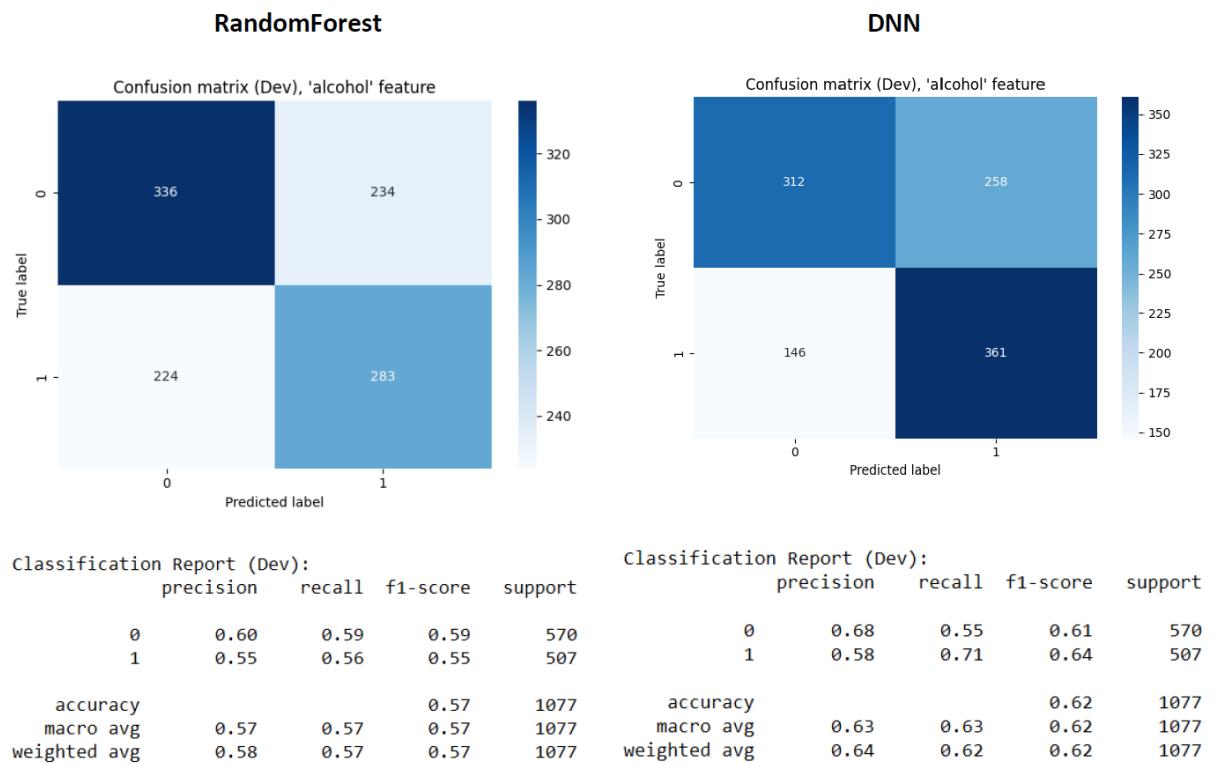
En este caso, la clase indicadora del consumo de alcohol en los hombres está más balanceada. Como puede verse en la Figura 2.33, alrededor del 50 % de los hombres consume alcohol frecuentemente. Sin embargo, tal y como se puede observar en el Cuadro 2.10 y las Figuras 2.33 y 2.34, siguen sin conseguirse buenas predicciones solamente utilizando las variables compatibles con el conjunto Perinatal. Se ha conseguido una tasa de acierto solamente del 60 %. Las variables predictoras más relevantes siguen siendo la comunidad autónoma (*ccaap*), la edad (*edadp*) y el nivel de estudios (*estudiop*) (ver Figura 2.35).

<i>alcoholp</i>	DEV		
	AUC	Accuracy	F1-score
Random Forest	0.6187	0.57	0.57
DNN	0.6670	0.62	0.62

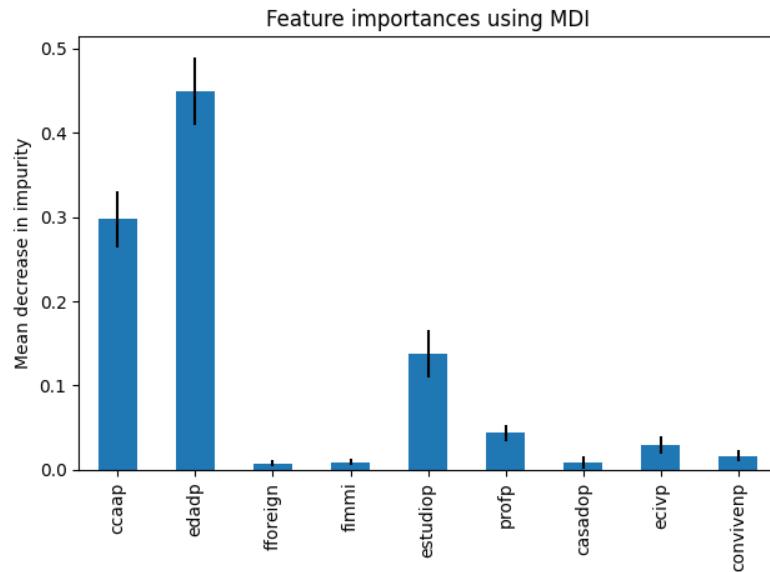
2.10. Cuadro: Resultados de los modelos predictivos de *alcoholp* sobre el conjunto Dev



2.33. Figura: Curva ROC y AUC de los modelos predictivos de *alcoholp* sobre el conjunto Dev, y distribución de la clase *alcoholp* en mujeres de la ENSE 2017



2.34. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *alcoholp* sobre el conjunto Dev

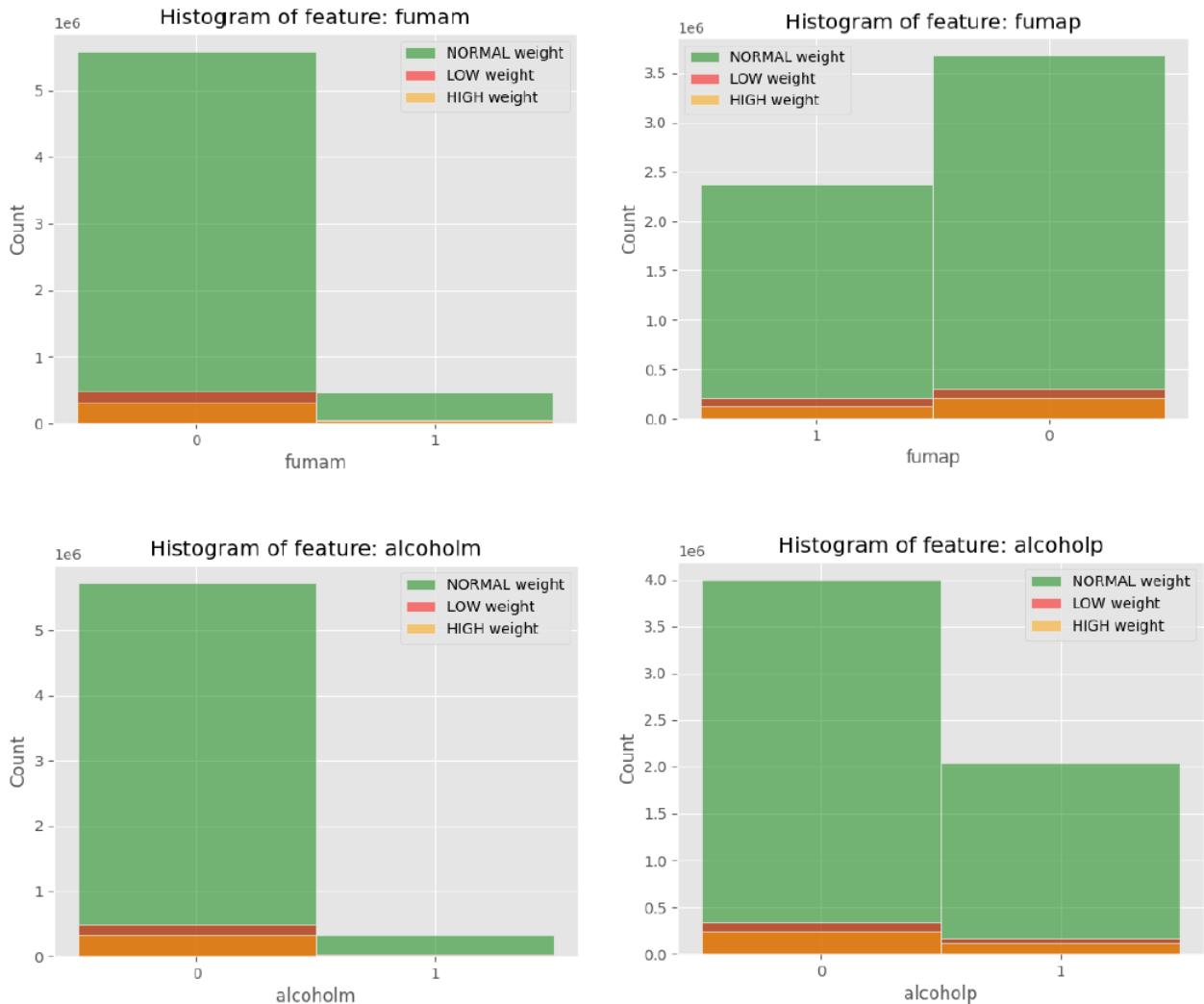


2.35. Figura: *Feature significance* del modelo predictivo Random Forest de *alcoholp*

Predicciones sobre las instancias de Perinatal

Los modelos entrenados para la predicción de las cuatro variables anteriores se han utilizado para predecir las características *fumam*, *fumap*, *alcoholm* y *alcoholp* en los ítems del conjunto de datos Perinatal, a pesar de no haber conseguido muy buenas predicciones. Las predicciones de estas variables se han realizado con los modelos Random Forest previamente entrenados, ya que eran mejores discriminadores de las clases que los modelos DNN.

A continuación, en la Figura 2.36, se muestran las distribuciones de las variables una vez añadidos los valores predichos al conjunto Perinatal, y respecto al peso del recién nacido (clase *pesorec*). Como se puede observar, en cuanto a las variables pertenecientes a las madres (*fumam* y *alcoholm*), hay muy pocas predicciones positivas (consumidora de tabaco y alcohol). En cuanto a las predicciones en los padres (*fumap* y *alcoholp*), la distribución de las predicciones del consumo de tabaco y alcohol están bastante más balanceadas.



2.36. Figura: Distribución de las variables predichas de la ENSE 2017 sobre el conjunto Perinatal

Una vez obtenido el conjunto de datos Perinatal con las predicciones en las nuevas variables *fumam*, *fumap*, *alcoholm* y *alcoholp* agregadas, se va a realizar el experimento de clasificación de peso del recién nacido respecto a la clase *pesorec*: peso bajo, normal o alto.

Localización del fichero:

`Perinatal_DatosResultados\Datos\ENSE2017\Preprocess\dataPerinatal_predictions_ENSE.csv`

2.3.1. Algoritmos empleados

Los algoritmos empleados para el entrenamiento de modelos predictivos de *pesorec* con el conjunto de datos Perinatal y las variables de la ENSE agregadas son los mismos que se han utilizado en los *Experimentos 1 y 2: Random Forest y Deep Neural Network (DNN)*. Como ya se ha dicho anteriormente, el algoritmo Random Forest ayuda a interpretar el modelo mostrando la importancia de las variables predictoras, mientras que la DNN puede ofrecer mejores resultados al tratar con grandes conjuntos de datos.

2.3.2. Planteamiento de la predicción

En este apartado se van a mostrar los resultados y las evaluaciones de los modelos predictivos entrenados para este experimento.

2.3.2.1. Tipo de clasificación

El problema de **clasificación** sigue siendo **multiclas**, ya que pretende clasificarse correctamente la clase *pesorec* del conjunto Perinatal en tres categorías: **peso bajo**, **normal** y **alto**. Se seguirá utilizando la distribución original del *dataset* Perinatal (ver Figura 1.1).

2.3.2.2. Resultados y evaluación de los modelos

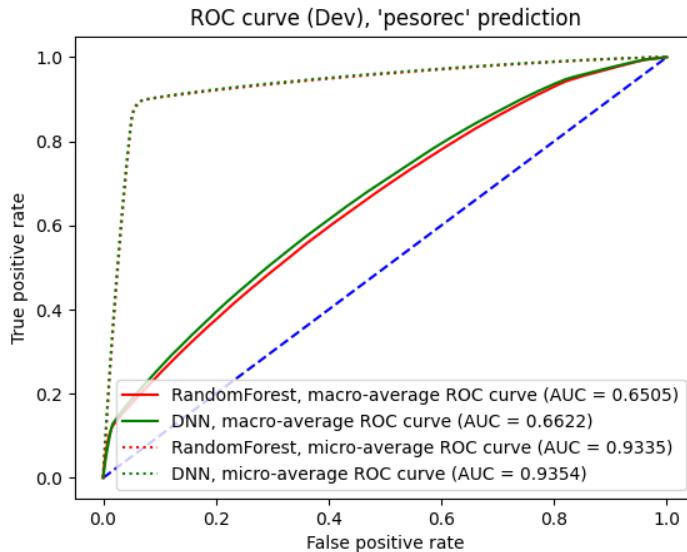
Localización de los resultados:

Perinatal_DatosResultados\Resultados\Perinatal\ExperimentsPesorec\FeatureAblationExperiments\ENSE\AllFeatures

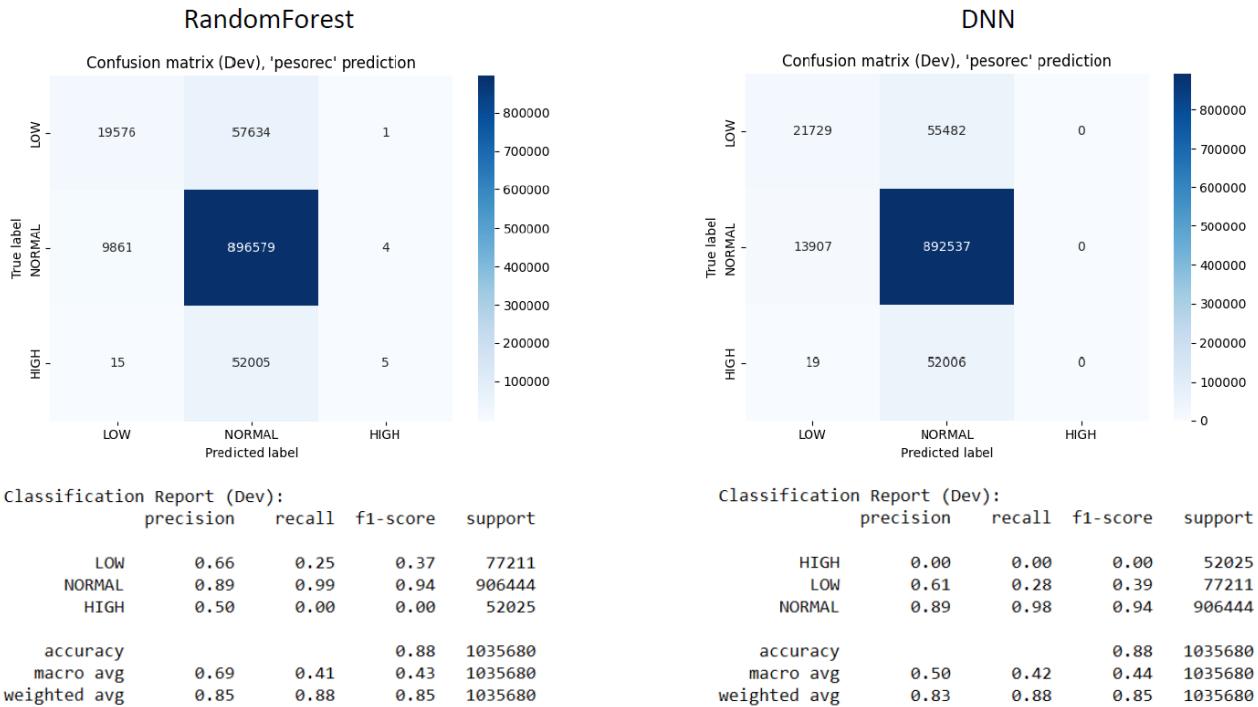
A continuación se muestran los resultados de los modelos predictivos entrenados y evaluados sobre el conjunto Dev. Los modelos se han entrenado tras agregar las variables *fumam* (indicador de consumo de tabaco de la madre), *fumap* (indicador de consumo de tabaco del padre), *alcoholm* (indicador de consumo de alcohol de la madre) y *alcoholp* (indicador de consumo de alcohol del padre), todas ellas predichas a partir de los datos de la ENSE 2017.

pesorec	DEV					
	AUC (micro-avg.)	variación sin ENSE 2017	Accuracy	variación sin ENSE 2017	F1-score	variación sin ENSE 2017
Random Forest	0.9335	+0.0036	0.88	0	0.85	+0.01
DNN	0.9354	+0.037	0.88	0	0.85	+0.01

2.11. Cuadro: Resultados de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando variables ENSE 2017)



2.37. Figura: Curva ROC y AUC de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando variables ENSE 2017)

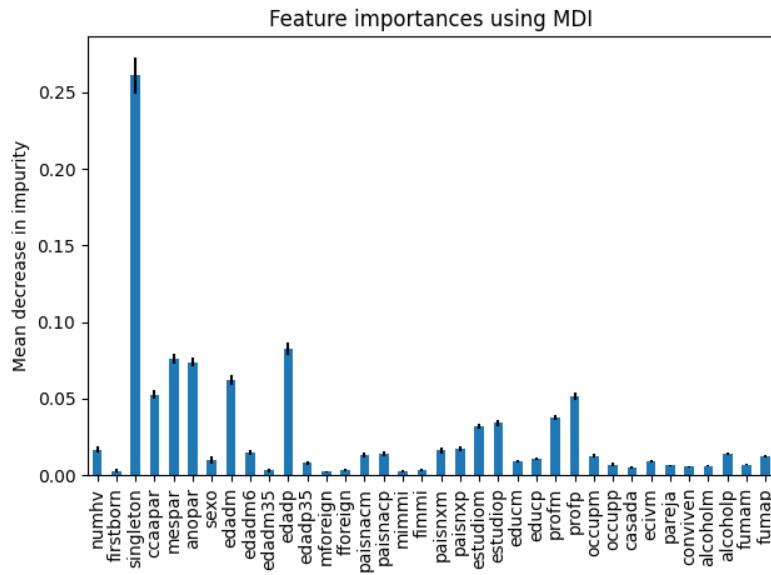


2.38. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *pesorec* sobre el conjunto Dev (agregando variables ENSE 2017)

Como puede observarse en el Cuadro 2.11 y en las Figuras 2.37 y 2.38, los cambios en los resultados de estos modelos respecto a los modelos entrenados solamente con las variables originales del conjunto Perinatal son prácticamente irrelevantes. Es verdad que mejora muy ligeramente el rendimiento (+1% en el F1-score), pero las mejoras son insignificantes. Los modelos siguen clasificando la mayoría de instancias con peso normal y se clasifican más instancias de peso bajo de manera incorrecta que correctamente. En la clasificación de peso alto, el algoritmo Random Forest consigue clasificar 5 ejemplos correctamente (previamente no se había clasificado ninguno), pero es totalmente insignificante cuando se trata de clasificar decenas de miles de ítems de peso alto.

2.3.2.3. Interpretabilidad (*feature significance*)

En este apartado del análisis de los modelos predictivos se va a interpretar la importancia de las variables predictoras a la hora de realizar las predicciones, a través de las importancias (Gini) otorgadas por el algoritmo Random Forest a cada una de ellas.



2.39. Figura: *Feature significance* del modelo predictivo Random Forest de *pesorec* (agregando variables ENSE 2017)

En la Figura 2.39 se puede observar que las variables agregadas y predichas a partir del conjunto de datos ENSE 2017 (*fumam*, *fumap*, *alcoholm* y *alcoholp*) tienen una importancia prácticamente nula sobre las predicciones de peso del recién nacido. La variable *singleton* (indicador de embarazo con hijo/a único/a o múltiple) sigue siendo la única altamente determinante en las predicciones de *pesorec*.

Era de esperar que estas nuevas variables agregadas de la ENSE 2017 no llegaran a ser determinantes, ya que los modelos predictivos de estas mismas no ofrecen un muy buen rendimiento al haber sido únicamente entrenados con las variables comunes entre los dos conjuntos de datos (Perinatal y ENSE 2017). Una opción para mejorar las predicciones de estas variables podría ser el entrenamiento de los modelos con todas las variables disponibles en la ENSE, pero no podrían realizarse posteriormente las predicciones sobre las instancias del conjunto Perinatal, ya que no se dispone de esas mismas características en los ítems de este último *dataset*.

2.4. EXPERIMENTO 4: clasificación de variables de ENSE 2017

En el *Experimento 3* se ha realizado una clasificación de peso de los recién nacidos del conjunto de datos Perinatal agregando variables con valores predichos a partir de los datos de la ENSE 2017. Concretamente, se han agregado las siguientes variables: *fumam* (indicador de consumo de tabaco de la madre), *fumap* (indicador de consumo de tabaco del padre), *alcoholm* (indicador de consumo de alcohol de la madre) y *alcoholp* (indicador de consumo de alcohol del padre). Sin embargo, no han sido variables predictoras determinantes en la clasificación del peso como bajo, normal o alto (clase *pesorec*), y la mejora en los resultados de los modelos predictivos ha sido irrelevante.

Una razón por la que la agregación de variables de la ENSE no ha resultado satisfactoria puede ser la escasez de características comunes que tiene los dos conjuntos de datos (Perinatal y ENSE 2017). Esto hace que los modelos predictivos de variables ENSE entrenados únicamente con las variables de Perinatal no sean lo suficientemente precisos. Por este motivo, en este cuarto experimento se van a volver a entrenar los **modelos predictivos de consumo de tabaco y alcohol** con el conjunto de datos ENSE 2017, pero esta vez **utilizando todas las variables predictoras** disponibles (ver sección *Descripción del corpus de ENSE 2017*). El objetivo será ver si mejoran los resultados y qué otras variables no incluidas en el *dataset* Perinatal adquieran importancia. Se van a entrenar modelos predictivos para la clasificación de las siguientes variables: *fumam* (indicador binario de consumo de tabaco en las mujeres de la ENSE 2017) y *alcoholm* (indicador binario de consumo de alcohol en las mujeres de la ENSE 2017).

2.4.1. Algoritmos empleados

Los algoritmos empleados para el entrenamiento de modelos predictivos de *fumam* y *alcoholm* con el conjunto de datos ENSE 2017 y todas sus variables son los mismos que se han utilizados en los *Experimentos 1, 2 y 3*: **Random Forest** y **Deep Neural Network (DNN)**. Como ya se ha dicho anteriormente, el algoritmo Random Forest ayuda a interpretar el modelo mostrando la importancia de las variables predictoras, mientras que la DNN puede ofrecer mejores resultados al tratar con grandes conjuntos de datos.

2.4.2. Planteamiento de la predicción

En este apartado se van a mostrar los resultados y las evaluaciones de los modelos predictivos entrenados para este experimento.

2.4.2.1. Tipo de clasificación

Se trata de una **clasificación binaria**, ya que pretenden clasificarse correctamente las clases *fumam* y *alcoholm*, las cuales son **indicadores binarios de consumo de tabaco y alcohol**, respectivamente. Los modelos predictivos van a entrenarse con la distribución original de estas variables. En la Figura 1.13 se ha mostrado la distribución desbalanceada de las clases *fumam* y *alcoholm*, y por este motivo, se ha realizado un remuestreo en los conjuntos de datos de entrenamiento, concretamente undersampling, reduciendo al 50 % el número de ítems de la clase mayoritaria (no fumadora y no consumidora de alcohol, respectivamente).

2.4.2.2. Resultados y evaluación de los modelos

A continuación se muestran los resultados de los modelos predictivos entrenados y evaluados sobre el conjunto Dev.

Modelo predictivo de *fumam*

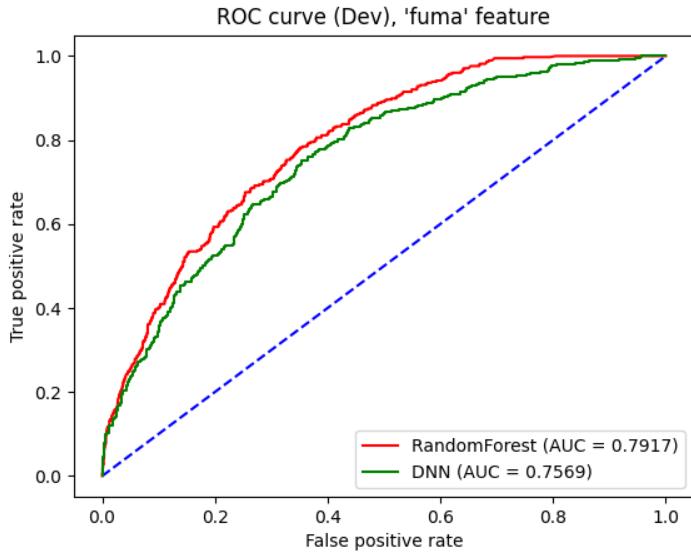
Localización de los resultados:

Perinatal_DatosResultados\Resultados\ENSE2017\ExperimentsFumam

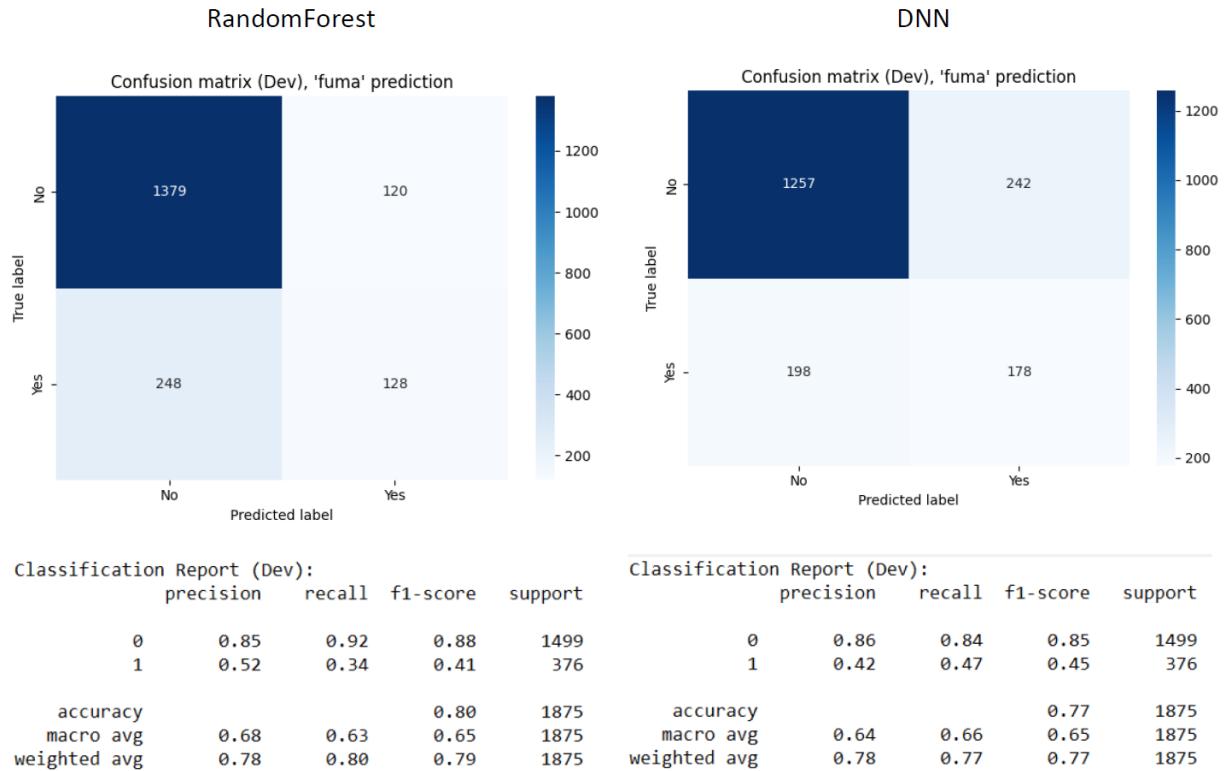
Los modelos se han entrenado con los ítems pertenecientes a las mujeres y todas las variables de la ENSE 2017, y tras haber aplicado el preprocesado de datos explicado en la sección *Preproceso*.

<i>fumam</i>	AUC	DEV				
		variación variables compatibles Perinatal	Accuracy	variación variables compatibles Perinatal	F1-score	variación variables compatibles Perinatal
Random Forest	0.7917	+0.0129	0.80	-0.02	0.79	-0.02
DNN	0.7569	-0.0379	0.77	-0.08	0.77	-0.01

2.12. Cuadro: Resultados de los modelos predictivos de *fumam* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)



2.40. Figura: Curva ROC y AUC de los modelos predictivos de *fumam* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)



2.41. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *fumam* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)

Como puede observarse en el Cuadro 2.12 y las Figuras 2.40 y 2.41, los resultados obtenidos son peores a los esperados. Se esperaba una mejora significativa en las predicciones realizadas al pasar de entrenar los clasificadores de *fumam* solamente utilizando las variables compatibles con el *dataset* Perinatal (ver sección *Predicción de fumam en el Experimento 3*) a entrenarlos utilizando todas las características disponibles en el *dataset* ENSE 2017. Sin embargo, los resultados de las predicciones realizadas son

prácticamente iguales o incluso empeoran ligeramente. Se consigue una alta precisión en las predicciones de la clase negativa (no fumadora), pero al clasificar la clase positiva (fumadora), se comete alrededor de un 50 % de errores.

Modelo predictivo de *alcoholm*

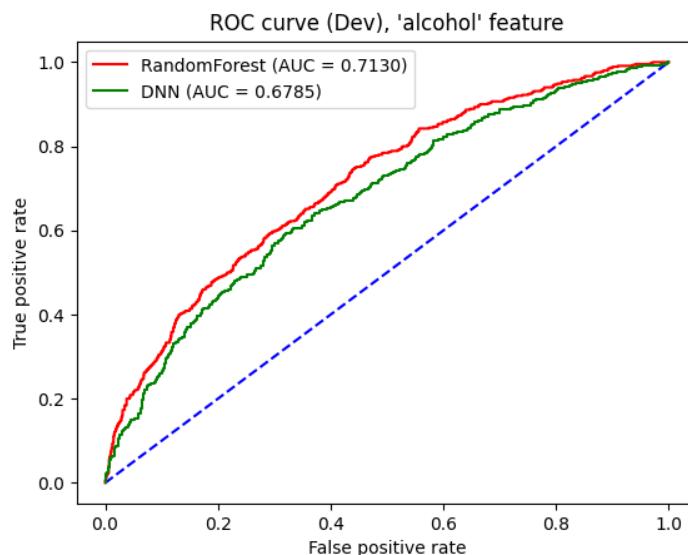
Localización de los resultados:

Perinatal_DatosResultados\Resultados\ENSE2017\ExperimentsAlcoholm

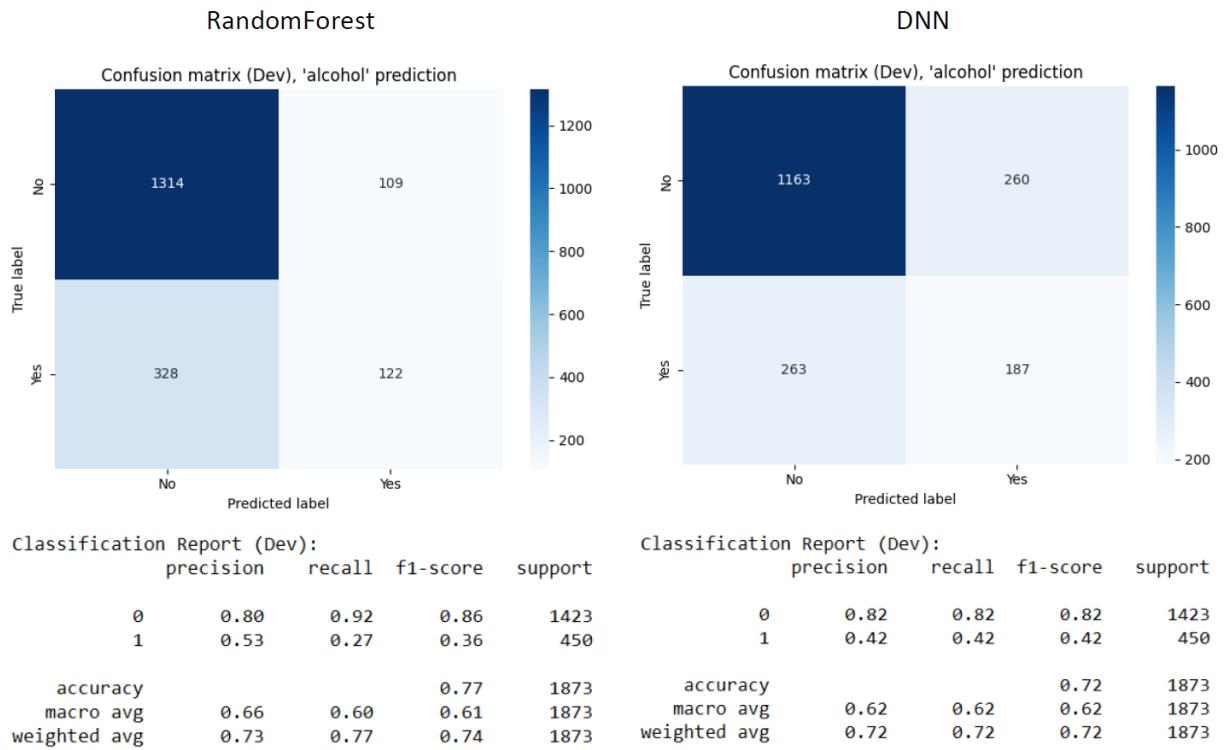
Los modelos se han entrenado con los ítems pertenecientes a las mujeres y todas las variables de la ENSE 2017, y tras haber aplicado el preprocesado de datos explicado en la sección *Preproceso*.

<i>alcoholm</i>	AUC	DEV				
		variación variables compatibles Perinatal	Accuracy	variación variables compatibles Perinatal	F1-score	variación variables compatibles Perinatal
Random Forest	0.7130	+0.0543	0.77	0	0.74	0
DNN	0.6785	+0.0733	0.72	-0.08	0.72	+0.01

2.13. Cuadro: Resultados de los modelos predictivos de *alcoholm* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)



2.42. Figura: Curva ROC y AUC de los modelos predictivos de *alcoholm* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)



2.43. Figura: Matriz de confusión y métricas de evaluación de los modelos predictivos de *alcoholm* sobre el conjunto Dev (entrenado con todas las variables de ENSE 2017)

En el caso de los clasificadores de *alcoholm* sucede algo muy similar a lo observado con los modelos predictivos de *fumam*. Como se muestra en el Cuadro 2.13 y las Figuras 2.42 y 2.43, en esta ocasión sí que mejoran ligeramente los resultados respecto a los modelos predictivos de *alcoholm* únicamente entrenados con las variables compatibles con el *dataset* Perinatal (ver sección *Predicción de alcoholm en el Experimento 3*). Sin embargo, los clasificadores siguen cometiendo errores de clasificación muy similares, clasificando más del 80 % de los ítems de clase negativa (no consumidora de alcohol) de manera correcta, pero solamente alrededor del 50 % de los ítems de clase positiva (consumidora de alcohol).

Conclusiones

Por lo tanto, no se han conseguido grandes mejoras en los modelos predictivos de consumo de tabaco y alcohol en las mujeres habiendo sido entrenados con la gran variedad de características que ofrece la Encuesta Nacional de Salud de España del 2017. Aun así, en la siguiente sección, se van a analizar las variables predictoras más determinantes en estos modelos, a la espera de encontrar nuevas variables de las que el *dataset* Perinatal no dispone y podrían ser de ayuda para predecir factores como el consumo de tabaco y alcohol.

2.4.2.3. Interpretabilidad (*feature significance*)

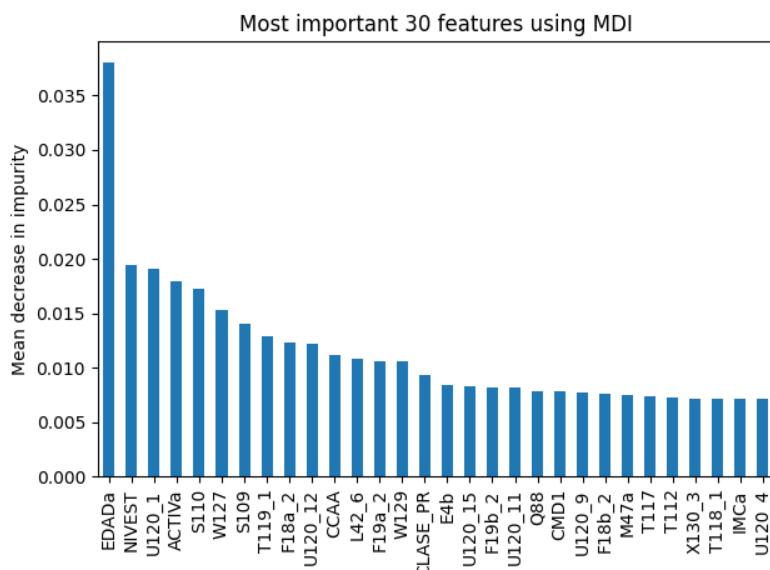
En este apartado del análisis de los modelos predictivos se va a interpretar la importancia de las variables predictoras a la hora de realizar las predicciones, a través de las importancias (Gini) otorgadas por el algoritmo Random Forest a cada una de ellas.

Modelo predictivo de *fumam*

En la Figura 2.44 se muestra que la variable predictoría más determinante en la clasificación del consumo de tabaco de las mujeres es la edad (**EDADA**), con el doble de influencia que el resto de variables. Las demás variables tienen una influencia muy similar sobre las predicciones del modelo. A continuación se hace un listado de las siguientes 10 variables más relevantes:

- **NIVEST**: Nivel de estudios (variable categórica)

- **U120_1**: Frecuencia de consumo de fruta fresca (excluyendo zumos) (variable categórica)
- **ACTIVa**: Actividad económica actual (variable categórica)
- **S110**: Peso en kg (variable numérica)
- **W127**: Frecuencia de consumo de alcohol en los últimos 12 meses (variable categórica)
- **S109**: Altura en cm (variable numérica)
- **T119_1**: Tiempo sentado en un día normal (Horas por día) (variable numérica)
- **F18a_2**: Actividad de la empresa en la que trabaja (código CNAE2009, 3 dígitos) (variable categórica)
- **U120_12**: Frecuencia de consumo de refrescos con azúcar (variable categórica)
- **CCAA**: Comunidad Autónoma de residencia (variable categórica)



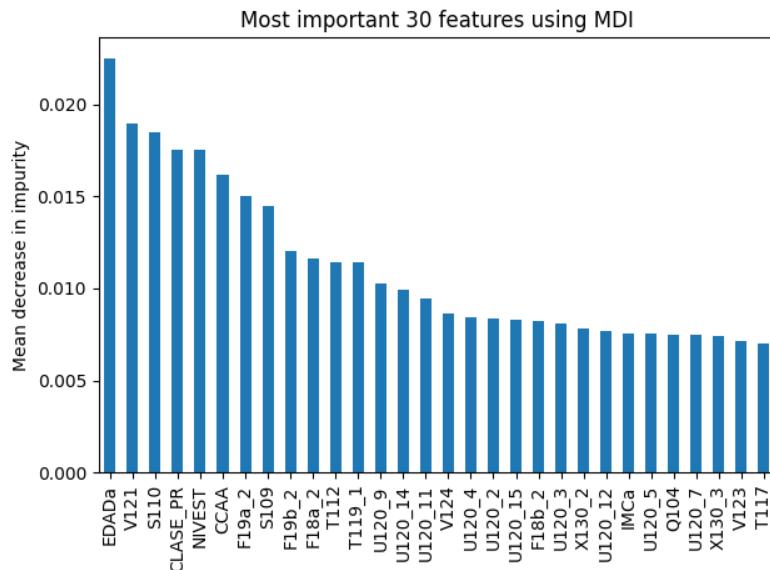
2.44. Figura: *Feature significance* del modelo predictivo Random Forest de *fumam* (entrenado con todas las variables de ENSE 2017)

Modelo predictivo de *alcoholm*

En la Figura 2.45 se observa que la edad (**EDADA**) también es la variable predictoría más determinante en la clasificación del consumo de alcohol de las mujeres. Sin embargo, la diferencia de relevancia que adquiere la edad con el resto de variables no es tan amplia como en el caso del consumo de tabaco. Las siguientes 10 variables más relevantes son:

- **V121**: ¿Fuma actualmente? (variable categórica)
- **S110**: Peso en kg (variable numérica)
- **CLASE_PR**: Clase social basada en la ocupación de la persona de referencia (variable categórica)
- **NIVEST**: Nivel de estudios (variable categórica)
- **CCAA**: Comunidad Autónoma de residencia (variable categórica)
- **F19a_2**: Ocupación, profesión u oficio actual (código CNO2011, 3 dígitos) (variable categórica)
- **S109**: Altura en cm (variable numérica)
- **F19b_2**: Última ocupación, profesión u oficio (código CNO2011, 3 dígitos) (variable categórica)

- **F18a_2**: Actividad de la empresa en la que trabaja (código CNAE2009, 3 dígitos) (variable categórica)
- **T112**: Frecuencia con la que realiza alguna actividad física en su tiempo libre (variable categórica)



2.45. Figura: *Feature significance* del modelo predictivo Random Forest de *alcoholm* (entrenado con todas las variables de ENSE 2017)

Conclusiones

Entre las variables predictoras más determinantes para las predicciones de consumo de tabaco y alcohol aparecen características que ya habíamos tenido en cuenta y que coinciden con las que dispone el conjunto Perinatal. Ejemplos de estas variables son la edad, el lugar de residencia (comunidad autónoma), el nivel de estudios o el tipo de trabajo o actividad económica. Sin embargo, características no recogidas en el *dataset* Perinatal también parecen ser útiles para estos modelos predictivos. Estas nuevas variables serían, entre otras, el peso y la estatura, la actividad física diaria o el tiempo sin movimiento, y los hábitos alimenticios como la frecuencia de consumo de frutas o refrescos azucarados. También se observa que el consumo de tabaco es un gran predictor del consumo de alcohol, y viceversa.

En conclusión, la edad es la variable más determinante en las dos predicciones, consumo de tabaco y alcohol, seguida de variables socioeconómicas y los hábitos de alimentación y de actividad física mencionados, así como el peso y la estatura. Agregar estas nuevas variables a las características de las madres y padres del conjunto de datos Perinatal podría incluso mejorar las predicciones de peso de los recién nacidos.

Bibliografía

- [1] X Bao, Y Wang, S Zhang, L Yang, G Liu, Y Yang, X Li, D Hao, A Chen, X Liu, and J Shao. Establishment of a personalized fetal growth curve model. *Technol Health Care.*, 29:311–317, 2021.
- [2] J Gardosi, A Williams, O Hugh, and A Francis. Grow – customised weight centiles. *Gestation Network.*, Dec 2020.