Universidad Zaragoza
1542

Trabajo Fin de Grado

Grado en Ingeniería Informática, especialidad de computación

Caracterización de la enfermedad de Alzhéimer utilizando datos genómicos y redes neuronales profundas

Characterization of Alzheimer's disease from genomic data and deep neural networks

Autor

Alejandro Gómez González

Directoras

Elvira Mayordomo Cámara

Mónica Hernández Giménez

ESCUELA DE INGENIERÍA Y ARQUITECTURA
2021

RESUMEN

Actualmente existen distintos procedimientos diagnósticos que pueden ayudar a la identificación de la enfermedad de Alzhéimer (Alzheimer's disease - AD), como pueden ser una evaluación de las capacidades cognitivas del paciente o la atrofia en estructuras cerebrales observada en imágenes por resonancia magnética (Magnetic Resonance Image - MRI). Estas pruebas por norma general son más fiables en etapas más avanzadas de la enfermedad y además, la única prueba que puede confirmar el AD con total seguridad es una biopsia post-mortem del cerebro. Ante esto surge la necesidad de elaborar nuevos métodos diagnósticos útiles en su detección temprana.

La componente genética del AD en su variante más común, el Alzheimer de inicio tardío (late onset Alzheimer's disease - LOAD) se estima del 56–79 %, y para el Alzheimer de aparición temprana (Early Onset Alzheimer's Disease - EOAD) del 90 %. De la componente genética del LOAD se ha conseguido explicar sólo un 33 % mediante estudios de asociación del genoma completo (Genome-wide association study - GWAS), por lo que un 40 % de la varianza fenotípica de la enfermedad está todavía por explicar.

La comunidad científica cree que la metodología detrás de los GWAS tiene ciertas limitaciones debido a que se centra en la causalidad de SNPs de forma individual, y no en grupo, por lo que nuevos métodos basados en aprendizaje automático podrían contribuir a ampliar la identificación de nuevas zonas del genoma relacionadas de alguna forma con la enfermedad. A pesar de que existen antecedentes de otros estudios en los que se combina resultados de MRI con otros biomarcadores como pueden ser la información genética, los que usan exclusivamente información genética son realmente escasos.

El objetivo de este trabajo es explorar la caracterización de la enfermedad mediante información genética cedida por la "Alzheimer's Disease Neuroimaging Initiative" (ADNI) aplicando técnicas de aprendizaje automático como el deep learning. Se ha podido alcanzar una área bajo la curva (AUC - Area Under the Curve) del 96 %.

AGRADECIMIENTOS

Este trabajo de fin de grado pone punto y final a mi paso por el grado de Ingeniería Informática y no es casual que haya conseguido llegar hasta aquí, por lo que me gustaría dar las gracias a todos aquellos y aquellas que hacen que esto haya sido posible.

Gracias a mi familia, por quererme tanto, proporcionarme un hogar, siempre confiar en mí y apoyarme todos los días. Tener un padre informático ya era un "factor de riesgo" para ir por el mismo camino, pero quien sabe si la culpa de todo esto es de mis padres al regalarme aquel Alfa-Berto y Junior... De cualquier forma, si soy como soy y estoy donde estoy es principalmente gracias a vosotros.

Gracias a Talía, por quererme tanto, tener un momento para escucharme siempre, darme ánimos, apoyarme en todo y cuidarme cuando lo he necesitado. Me siento muy afortunado de haberte conocido, no te cambiaría por nada. Juntos podemos con todo, somos un equipo.

Gracias a mis compañeros de la universidad por hacer que ir a la universidad fuera algo más ameno e incluso divertido algunos días. En estos cuatro años he conocido a gente maravillosa, que comenzando como compañeros de clase hemos acabado siendo verdaderos amigos. Hicimos piña y creo que hemos tenido suerte de tenernos los unos a los otros, os aprecio mucho. Maldita pandemia que nos empujó a todos a casa, se nos han quedado pendientes muchas comidas, con sus respectivas jugadas de cartas y los muy necesarios miércoles de montaditos. Me da mucha pena que vaya a dejar de veros tan a menudo.

Gracias a mis amigos, por todos los grandes momentos compartidos, me lo paso genial siempre que estoy con vosotros. Valéis muchísimo.

Gracias a Elvira y a Mónica, las directoras de este trabajo. La idea de hacer esto fue vuestra, y me habéis indicado los pasos a seguir para llevar a cabo este trabajo.

Gracias a los participantes del estudio que prestaron los datos necesarios para este trabajo.

Y por último, no quiero dejar sin mencionar a María, quien fue mi vecina desde que era pequeño, y que lamentablemente padeció de Alzheimer. Me he acordado de ti.

Índice

1. Introducción y objetivos **1**

2. Materiales y métodos **5**

2.1. Control de calidad 6

2.1.1. Diagnósticos no contemplados 7

2.2. Unión de datos de distintos instrumentos 8

2.3. Uso de los datos 9

2.3.1. División entre GWAS y validación cruzada 10

2.4. Estudio de asociación del genoma completo 11

3. Aprendizaje automático **13**

3.1. Red neuronal profunda (Deep Neural Network - DNN) 14

3.2. Otros modelos 16

3.2.1. Máquina de Vector Soporte (Support Vector Machine - SVM) 16

3.2.2. Bosque aleatorio (Random Forest) 17

3.2.3. Potenciación del gradiente (Gradient boosting) 17

4. Resultados **18**

4.1. Resultados con ADNI1/GO/2 - Illumina Omni 2.5M BeadChip 18

4.1.1. Comparativa GWAS 18

4.1.2. Comparativa SNPs de fuentes externas 21

4.1.3. Comparativa curvas ROC 21

4.1.4. Interpretabilidad del modelo 26

4.2. Resultados con unión de datos de distintos chips 32

5. Discusión **33**

Bibliografía **35**

6. Bibliografía **35**

Lista de Figuras **39**

Lista de Tablas **45**

Anexos **46**

A. Uso de PLINK **47**

B. Recopilación de resultados **48**

B.1. Resultados con ADNI1/GO/2 - Illumina Omni 2.5M BeadChip 48

B.1.1. Método de división simple 48

B.1.2. Método de división por k-fold 61

B.1.3. Método de filtración de datos 79

B.1.4. Método de análisis externo 95

B.2. Resultados con unión de datos de distintos chips 100

B.2.1. Método de división simple 100

B.2.2. Método de división por k-fold 109

B.2.3. Método de filtración de datos 118

C. Código desarrollado **127**

1

Introducción y objetivos

La enfermedad de Alzheimer (Alzheimer's disease - AD) es una enfermedad neurodegenerativa e irreversible que afecta a la memoria, las habilidades del pensamiento, lenguaje y al comportamiento del paciente. Esta enfermedad en la mayoría de los casos conlleva en la mayoría de los casos a la muerte del paciente [1]. Este deterioro se produce debido a que las neuronas comienzan a degenerarse y a morir. Estos cambios físicos en el cerebro pueden llegar a comenzar una década o más antes de la aparición de los primeros síntomas. El AD es actualmente la forma más común de demencia, siendo responsable del 60-80 % de los casos de esta, y que puede llegar a afectar a cerca de la mitad de la población mayor de 85 años.

Actualmente existen distintos procedimientos diagnósticos que pueden ayudar a la identificación del AD, como pueden ser una evaluación de las capacidades cognitivas del paciente o el análisis de la imagen de una resonancia magnética (Magnetic Resonance Image - MRI) del área cerebral. De hecho, aunque puede hacerse un diagnóstico con gran fiabilidad con estas pruebas, ya sea aplicando técnicas de aprendizaje automático o no, estas pruebas son más fiables en una etapa avanzada de la enfermedad, en la que ya se han manifestado los síntomas, y por tanto aumenta la gravedad de la enfermedad. Además, pese a estos procedimientos diagnósticos, la única prueba definitiva para confirmar AD es mediante una biopsia post-mortem del cerebro del paciente [2] [3].

Se ha demostrado también, que aunque no exista por el momento un tratamiento reversible o una cura, el diagnóstico temprano del AD favorece el pronóstico de esta enfermedad [4] [5]. Ante esto surge la necesidad de elaborar nuevos métodos diagnósticos útiles para la detección temprana del AD.

Las hipótesis actuales sostienen que el Alzheimer de aparición tardía (Late Onset Alzheimer's Disease - LOAD) puede tener origen tanto por motivos ambientales, como genéticos. En concreto se cree que tanto la enfermedad de Alzheimer de aparición temprana como tardía, tiene cierta componente genética. Se estima que el LOAD tiene una probabilidad de heredarse del 56–79 % [6]. Por otro lado el Alzheimer de aparición temprana (Early Onset Alzheimer's Disease - EOAD) del 90 % [7]. Por lo tanto, la información genética de un paciente podría utilizarse en un diagnóstico de AD.

El ácido desoxirribonucleico (ADN) almacena su información en forma de código, el cual esta compuesto por una cadena de nucleótidos. Cada uno de estos nucleótidos está formado por un glúcido, (desoxirribosa) una base nitrogenada, que puede ser adenina (A), timina (T), citosina (C) o guanina (G) y finalmente un grupo fosfato. Los nucleótidos a lo largo de la cadena se ven diferenciados por la base nitrogenada, por lo que la secuencia del ADN se especifica nombrando solo la secuencia de sus bases. El ADN está organizado en cromosomas y cada cromosoma contiene una serie de bases. El ADN humano consta de unos 3 mil millones de bases distribuidos en 23 pares de cromosomas. De estos 3 mil millones de bases, compartimos cerca del 99.9 % de estas bases. Esto implica que en un 0.1 % del ADN del ser humano se encuentran todas las variantes causantes de todos los rasgos de origen genético. Las posiciones del código genético en las que existe cierta variación se le denominan polimorfismos de nucleótido único (Single Nucleotide Polymorphism - SNP). Estos SNPs son los que implican la expresión de un fenotipo asociado a un determinado gen.

Los estudios de asociación del genoma completo (Genome-wide association study - GWAS) se basan en encontrar asociaciones entre determinados SNPs y rasgos de su portador. La comunidad científica cree que la metodología detrás de los GWAS tiene ciertas limitaciones debido a que los GWAS se centran en la correlación de SNPs de forma individual, y no en conjunto. En el caso de estudiar enfermedades o rasgos que no dependen de un solo gen, sino de una combinación de varios, suele conllevar que cada una de estas mutaciones por separado tiene una relevancia menor en el diagnóstico, por lo que puede llegar a dificultar la tarea de encontrar estas asociaciones. Para el AD, de la componente genética del LOAD se ha conseguido explicar sólo un 33 % mediante GWAS, por lo que un 40 % de la varianza fenotípica

de la enfermedad está todavía por explicar. Se cree además, que el Alzheimer es una enfermedad genéticamente muy heterogénea, en la que no hay un gen que sea claramente responsable de su heredabilidad genética.

En los últimos años se han realizado estudios GWAS sobre el Alzheimer [8] [9] y estos han encontrado distintos SNPs que parecen estar relacionados con la manifestación de esta enfermedad. Sin embargo estos marcadores solo podrían explicar una parte de la componente genética. En este mismo año (2021) se ha elaborado también un estudio [10] en el que se trata de predecir el diagnóstico de AD mediante información genómica y áreas de inerés (Regions Of Interest - ROI) extraídas de imágenes de resonancia magnética (Magnetic Resonance Image - MRI), con los cuales se ha llegado a obtener una AUC del 0.77 mediante información genómica exclusivamente y una AUC del 0.8 combinando información genómica y las ROI.

La combinación de GWAS con técnicas de aprendizaje automático podría resultar en métodos con una mejor capacidad de diagnóstico de la enfermedad. Por eso, en este trabajo va a realizarse un estudio sobre la caracterización de la enfermedad de Alzheimer utilizando los SNPs seleccionados con estudios de asociación de genoma completo (Genome-wide association study - GWAS) y mediante diversas técnicas de aprendizaje automático, como las redes neuronales profundas, con lo cual se crearán modelos predictivos que dado una serie variantes genéticas, proporcione la probabilidad de que el paciente desarrolle AD.

Este trabajo ha sido inspirado por otro estudio previo realizado por Javier de Velasco Oriol et al. “Predicting late-onset Alzheimer’s disease from genomic data using deep neural networks” [11], realizado con el objetivo de predecir, en este caso LOAD también con información genómica. Se va a intentar reproducir los resultados de este trabajo ya que además, el artículo se trata de un pre-print y está en proceso de revisión. Se ha elegido especialmente este artículo porque, además de compartir la meta común de predecir AD usando exclusivamente datos genómicos, también se tienen en común los datos con los que se realiza, provenientes de la “Alzheimer’s Disease Neuroimaging Initiative”. Javier de Velasco Oriol et al. alcanzaron AUC de el 66 % aproximadamente eligiendo los 20 mejores SNPs del estudio de asociación.

Se espera que los modelos de aprendizaje automático establezcan de forma automática las relaciones entre los marcadores genéticos seleccionados en el estudio y el diagnóstico del paciente. Se va a realizar un calculo del valor SHAP [12] de cada SNP para cada modelo, con el cual interpretar importancia de cada SNP en el diagnóstico predicho. Es decir, se van a intentar interpretar el funcionamiento de los modelos realizados y cuantificar el valor que aporta cada SNP en el diagnóstico.

2

Materiales y métodos

Los datos con los cuales se ha realizado el trabajo provienen de la "Alzheimer's Disease Neuroimaging Initiative" (ADNI)[13] la cual, entre los datos que recoge de pacientes, se incluyen sus secuencias del genoma completo (Whole-Genome Sequence - GWS).

Estos datos están disponibles en formato PLINK, el cual se usa de forma habitual en estudios genómicos. Dado este formato, se ha usado la herramienta de PLINK[14] en su version 1.9. En un set de datos en PLINK se puede encontrar la lista de los SNPs recogidos junto a su información, como por ejemplo, su posición en la cadena de ADN y sus alelos en un archivo .bim (PLINK extended MAP file); las llamadas genotípicas de las muestras recogidas en un archivo .bed (PLINK binary biallelic genotype table) ; información de los individuos, como por ejemplo su sexo o su fenotipo.

ADNI lleva a cabo un estudio que en el momento actual consta de 4 fases: ADNI 1, ADNI GO, ADNI2 y ADNI3. Respecto a los datos genómicos publicados de cada una de las fases, estos se encuentran agrupados para su descarga según el instrumento de secuenciamiento utilizado. En la tabla 2.1 se puede observar el numero de SNPs disponibles para analizar.

Fase(s)	**Instrumento de secuenciamiento**	**Numero SNPs**
ADNI1	Illumina Human610-Quad BeadChip	620 901
ADNIGO/2	Illumina HumanOmniExpress BeadChip	730 525
ADNI1/GO/2	Illumina Omni 2.5M BeadChip	2 379 855
ADNI3	Illumina Infinium Global Screening Array v2	759 903

Tabla 2.1: Fases de estudio, instrumento de secuenciamiento y número de SNPs

De los SNPs recogidos por los distintos chips, se encuentra que sólo una pequeña cantidad de estos son comunes a todos ellos. Esto tiene sus consecuencias si se qui-

siese realizar un estudio uniendo los datos, ya que el número de SNPs disponibles constituye un factor limitante a tener en cuenta.

2.1. Control de calidad

Los datos genómicos a ser usados deben de pasar previamente por un control de calidad, con el cual seleccionar las muestras y SNPs que cumplan ciertas condiciones. Para esto, se ha tomado como base, el proceso descrito por Turner et al. [15]. En este proceso se pueden apreciar los siguientes pasos.

Filtrado de muestras:

- Detección de inconsistencias sexuales
- Detección de anomalías cromosómicas
- Parentesco entre las muestras
- Estratificación de la población
- Eficiencia de genotipado

Filtrado de características:

- Eficiencia de genotipado de marcadores - Marker call rates
- Concordancia HapMap
- Frecuencia del alelo menos común (Minor allele frecuency - MAF)
- Equilibrio de Hardy-Weinberg (Hardy-Weinberg Equilibrium - HWE)

Contemplando las necesidades y características de este estudio se ha procedido a adaptar este control de calidad. Para esto, el tutorial escrito por Marees, Andries T. et al. [16] ha sido de gran utilidad para entender y recrear los pasos necesarios para realizar este proceso en PLINK.

En primer lugar, respecto a la estratificación de la población de las muestras, se ha procedido a comprobar en la información demográfica de los participantes de ADNI. Tras observar que la gran mayoría de estos eran de etnia europea, y que el resto de

etnias no formaban una cantidad lo suficientemente grande como para ser analizadas, se ha decidido incluir exclusivamente a participantes de etnia europea. Aquí se encuentra la primera discrepancia con el artículo que se quiere reproducir[11], y es que en el artículo se afirma que en los datos de ADNI1/GO/2 del Illumina Omni 2.5M BeadChip no se encontraron participantes que no fueran de etnia europea y por lo tanto no tuvieron que retirar ninguna muestra del estudio. En la tabla 2.2 se detalla el número de pacientes de cada etnia en los diferentes conjuntos de datos.

Fase(s)	**Nativo americano**	**Asiático**	**Isleño del pacífico**	**Subsahariano**	**Europeo**	**Heterogeneo**
ADNI1	1	12	0	37	704	3
ADNIGO/2	2	12	2	30	734	13
ADNI1/GO/2	2	11	2	28	768	11
ADNI3	1	5	0	7	307	7

Tabla 2.2: Fases de estudio y número de pacientes de cada etnia

Respecto a las anomalías sexuales, se descarta el cromosoma 23, quedándonos con los cromosomas del 1 al 22, ya que no se está interesado en el sexo de los individuos. Por otra parte, se descartan las muestras que puedan estar relacionadas (familiares cercanos) estudiando la similitud entre las muestras con el coeficiente de relación descrito en el articulo de GCTA [17]; si este coeficiente supera cierto umbral, una de las muestras se descarta. En concreto se ha usado un coeficiente del 0.05.

Respecto al filtrado de SNPs, se comienza aplicando un umbral en la eficacia de genotipado para eliminar marcadores que conozcan en menos del 99.9 % de las muestras. También se estudia la frecuencia de los alelos menos común para eliminar aquellos SNPs cuyo alelo menor tenga una frecuencia menor al 1 %. Por último, se aplica también el test de equilibrio de Hardy-Weinberg con un umbral de significancia del 5 % con el cual se esperan detectar posibles errores de genotipado dada la frecuencia esperada de los alelos.

2.1.1. Diagnósticos no contemplados

Los datos genómicos deben de ser asociados a un diagnóstico para generar la información fenotípica correspondiente. La información diagnóstica de los pacientes, actualizada históricamente en cada consulta del estudio, también ha sido proporcio-

nada por ADNI. El criterio diagnóstico seguido por ADNI es el de NINCDS-ADRDA [18].

Para este estudio, se ha decidido no incluir los pacientes con MCI. Esto es debido a que, aunque muchos pacientes con AD son previamente diagnosticados con MCI, esto no quiere decir que los pacientes con MCI acaben desarrollando AD. El diagnóstico de pacientes con MCI no suele ser estable, y esto implica que personas diagnosticadas con MCI puedan evolucionar finalmente a AD o no. Por lo tanto, la inclusión de datos con este diagnóstico, entorpecería el estudio. Además, se tomó la misma decisión en el artículo [11] cuyos resultados se busca reproducir. De otra forma, si se hubiesen incluido estos diagnósticos, o bien podría considerarse un problema de clasificación multiclase entre CN, MCI o AD. Sin embargo, finalmente se va a centrar el estudio en el diagnóstico entre CN y AD, resultando así en un problema de clasificación estrictamente binaria.

En la tabla 2.3 se detalla el número de pacientes cognitivo normales (Cognitive normal - CN), con deterioro cognitivo leve (Mild Cognitive Impairment - MCI), o con la enfermedad del Alzheimer (Alzheimer's disease - AD) para los distintos grupos de datos.

	Pacientes			
Fase(s)	**CN**	**MCI**	**AD**	**Total**
ADNI1	168	192	397	757
ADNIGO/2	257	301	235	793
ADNI1GO/2	243	318	251	812
ADNI3	233	68	26	327

Tabla 2.3: Fases de estudio y número de pacientes con cada diagnóstico

2.2. Unión de datos de distintos instrumentos

Cuando se quiere unir datos genómicos de distintos chips, puede darse la situación de que solo una parte de las variantes recogidas por cada chip sean comunes, y por lo tanto, que se puedan utilizar para el estudio. Esto ocasiona que el número de variables disponibles se pueda reducir drásticamente. La reducción de SNPs disponibles no es un aspecto positivo, pero por otra parte, realizar el estudio con muestras de más individuos podría mejorar los resultados y encontrar más SNPs relevantes [19].

Por ejemplo, en el caso de unir los datos de ADNI1 del Illumina Human610-Quad BeadChip (620 901 SNPs) con los de ADNIGO/2 del Illumina HumanOmniExpress BeadChip (716 503 SNPs), se obtiene que solo 342 603 SNPs son comunes a ambos. Por lo tanto, se ha decidido comprobar la diferencia entre los resultados obtenidos sólo con los provenientes de un solo chip, y los provenientes de distintos chips.

2.3. Uso de los datos

Los datos en este estudio cumplen distintas funcionalidades

- Ser empleados en el GWAS
- Ser usados en la validación cruzada (cross-validation). Dentro de la cual se incluyen:
    - Datos usados en entrenamiento
    - Datos usados en validación.

Respecto a la validación cruzada, estos datos son usados para entrenar el modelo (con los datos de entrenamiento), y posteriormente realizar su evaluación (con los datos de validación). El uso de este método permite descartar que los resultados se vean afectados por sesgos en la división de entrenamiento y test.

En concreto se va a usar una validación cruzada estratificada de k-folds (stratified k-fold cross-validation). Con la validación cruzada estratificada de k-folds, los datos se dividen en k-partes, haciendo que en cada fold se seleccionen k-1 de estas partes para realizar el entrenamiento, y la parte restante para test. La estratificación permite asegurar que en cada una de las partes se mantenga la misma proporción de muestras de cada clase.

2.3.1. División entre GWAS y validación cruzada

Los datos empleados en el GWAS son con los cuales se van a obtener los SNPs más relevantes y se han propuesto distintas formas de elegir los datos con el objetivo de comparar los resultados obtenidos. En primer lugar, se ha optado por un diseño simple en cual, desde un principio se dividen los datos en los que se van a usar en GWAS y los que van a ser usados en la validación cruzada dada una proporción preestablecida. Para este trabajo se ha fijado una proporción de 80 % para el GWAS y un 20 % para la validación cruzada De ahora en adelante, este método se le llamará como el **método de división simple**.

En segundo lugar, se ha diseñado un método con el cual se prioriza la cantidad de muestras para cada uno de los conjuntos de datos. Para esto, en cada un de los folds de la validación cruzada se realiza el GWAS con los datos de entrenamiento y por otra parte, los datos de test se reservan exclusivamente para esta finalidad, sin ser incluidos en el estudio de asociación. Esto tiene como inconveniente que para cada iteración de la validación cruzada, las variables (SNPs) con los que se va a entrenar el modelo son distintos, haciendo más difícil su interpretación. De ahora en adelante, este método se le llamará como el **método de división por k-fold**.

En tercer lugar, se ha decidido no hacer distinción entre los datos destinados al GWAS y a la validación cruzada, de forma que todos los datos se usan tanto para el GWAS como para la validación cruzada. Esto entra en contradicción con uno de los principios básicos del aprendizaje automático, y es que los datos de test no deben usarse nunca en otro contexto que no sea el del propio test. De ahora en adelante, este método se le llamará como el **método de filtración de datos** (data leakage).

Por último, se ha decidido prescindir de realizar un GWAS, y usar todos los datos para la validación cruzada. Para la elección de los SNPs que se van a emplear los ya reconocidos por previas investigaciones como relevantes para el AD. De ahora en adelante, este método se le llamará como el **método de análisis externo**.

2.4. Estudio de asociación del genoma completo

Como se ha comentado anteriormente, se va a realizar un pequeño GWAS propio. PLINK proporciona las herramientas necesarias para hacer este estudio posible. La prueba de asociación básica para una enfermedad se basa en la comparación de frecuencias alélicas entre casos y controles. Se puede realizar un análisis caso/control estándar, PLINK usa por defecto la prueba de chi-cuadrado, pero también se puede utilizar el test exacto de Fisher, que calcula un p-valor más exacto a expensar de un mayor consumo computacional. Debido a que actualmente el tiempo de computación extra que supone usar Fisher es mínimo (para el tamaño del conjunto de datos que se va a analizar en el trabajo), y por tanto es la que se va a usar.

De esta prueba se obtiene un p-valor para cada uno de los SNPs. Este p-valor indica como de posible es que los alelos de dicho SNP estén relacionados con el rasgo (en este trabajo AD). Cuanto menor sea este p-valor, menor será la probabilidad de que la distribución encontrada de los alelos haya sido fruto del azar, y por tanto, de una mayor probabilidad de correlación.

Posteriormente se aplica un agrupamiento de desequilibrio de ligamiento (Linkage Disequilibrium Clumping - LD Clumping), este procedimiento es necesario debido a que habitualmente los SNPs de menor p-valor se encuentran agrupados en posiciones cercanas físicamente, lo cual es debido al agrupamiento de desequilibrio.

De forma habitual, se considera que una hipótesis es válida cuando tiene un p-valor inferior a 0.05, es decir, que con un 5 % de probabilidad o menos esta hipótesis considerada verdadera se pueda tratarse de un falso un positivo. Sin embargo cuando esto se realiza una gran cantidad de SNPs, un 5 % de falsos positivos equivale a una gran cantidad de falsos positivos. Es entonces cuando de forma habitual se aplica la corrección de Bonferroni, asignando un umbral al p-valor de $\frac{0{,}05}{SNPs}$, lo cual es igual a $0{,}5 * 10^{-8}$.

Lo que se quiere comprobar es si al relajar este umbral, considerando así más SNPs como relevantes, el modelo predictivo pueda obtener mejores resultados, o si por el contrario, funciona mejor con el umbral tradicional. Por una parte, al incluir más SNPs se pueden tener en cuenta más variables en la predicción y el rendimiento del

modelo podría mejorar. Sin embargo, si estas variables solo representan ruido y no tienen relación con el valor a predecir el rendimiento del modelo podría empeorar, o simplemente no aportar nada.

3

Aprendizaje automático

El aprendizaje automático se basa en que una serie de muestras, con sus respectivos atributos, sean introducidos a un modelo para realizar predicciones. En el aprendizaje automático se distinguen dos categorías principales, el aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado se conoce el verdadero valor del estado a predecir, por lo que se entrena el modelo con un subconjunto de los datos para intentar predecir sus valores para posteriormente realizar la comprobación de la fiabilidad del modelo con datos de test, los cuales el modelo no había visto antes nunca ni ha entrenado con ellos. Por otra parte, en el aprendizaje no supervisado no se conoce el verdadero valor a predecir, en este caso se buscan patrones en los datos con el objetivo de agrupar distintos tipos de muestras, reducir su dimensionalidad o investigar la estructura y distribución de los datos.

En este trabajo, se va a tener como datos de entrada una serie de pacientes, cuyos atributos son ciertos valores genotípicos de una serie de SNPs, y el valor a predecir, es su diagnóstico. Es decir, en base a información genética, mediante un modelo predecir el AD, tal y como se ve en la figura 3.1.

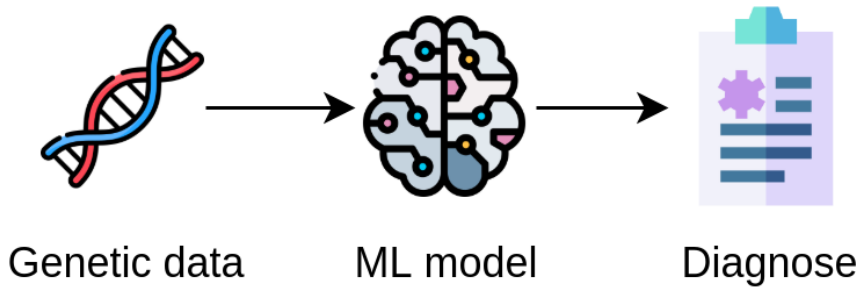

Figura 3.1: Esquema conceptual del trabajo (iconos obtenidos en Flaticon.com)

En este trabajo, se cuenta además con el valor real a predecir, por lo tanto, entra

dentro del campo del aprendizaje supervisado.

Para leer los datos en formato PLINK y poder ser interpretados en Python se ha utilizado la librería de pandas-plink[20].

Se han contemplado en este apartado una serie de modelos de clasificación. El que es protagonista de este trabajo está formado por una red neuronal, pero también se van a evaluar otros modelos que se verán a continuación.

Dentro del campo de las redes neuronales, se ha desestimado el uso de una red neuronal convolucional ya que a priori, la información genética disponible carece de dimensionalidad. Esto es debido a que en la información genética a usar se encuentra en una lista, en la que para cada posición se encuentra un SNP, el cual no tiene por que guardar ninguna relación con los SNP adyacentes en la lista. No obstante, existen herramientas como DeepInsight con la cual se pueden convertir este tipo de datos en imágenes que pueda usar una red neuronal convolucional [21] [22]. En este tipo de arquitectura, se intenta explotar la dimensionalidad de los datos y la relación que tienen los datos que se encuentran cerca de una misma posición, como por ejemplo, en los píxeles de una imagen.

3.1. Red neuronal profunda (Deep Neural Network - DNN)

El modelo principal a analizar en este trabajo consiste en una red neuronal profunda, cuyas capas de neuronas están totalmente conectadas. Es decir, en la que cada neurona de una capa conecta con todas las neuronas de la siguiente capa. Se espera que con una red neuronal profunda se puedan detectar estructuras complejas de los datos de entrada, con los cuales se pueda establecer un diagnóstico.

En la implementación, se ha usado TensorFlow[23] para el backend con la GPU, y Keras[24] como framework de las redes neuronales. La función de activación elegida en las capas de neuronas ha sido la de “ReLU”, ya que con esta función se espera que

el gradiente se propague de manera eficiente y la activación de los nodos será pequeña, resultando en que la tarea de decisión se divida a lo largo de la red. Para evitar el sobre-ajuste, se ha decidido incluir un dropout en las salidas del 40 %. La función de pérdida elegida ha sido la entropía cruzada binaria (binary cross-entropy), y se ha elegido Adam como algoritmo de optimización. Además, se ha incluido como criterio de parada temprana para detener las épocas de entrenamiento se ha decidido monitorizar la función de pérdida en busca de encontrar la época cuyo valor sea mínimo para los datos de validación.

En comparación con la red del trabajo de Javier de Velasco Oriol et al. [11], se ha cambiado la regularización L2 por la L1, aumentar el dropout y reducir el ruido gausiano después de la primera capa densa. Con esta arquitectura se ha podido comprobar que el modelo obtenía de forma general mejores métricas que con la red original. La arquitectura de la red neuronal profunda de este trabajo está formada por las siguientes capas:

- **Entrada**: Plana de dimensión igual a los SNPs elegidos
- Densa con un número de neuronas igual al doble del número de SNPs de entrada, activación ReLU, regularización L1 al 0.02 e inicialización He
- Ruido Gaussiano al 20 %, normalización Batch y dropout del 40 % de las salidas
- Densa con 1024 neuronas, activación ReLU, regularización L1 al 0.01 e inicialización He
- Normalización Batch y dropout del 40 % de las salidas
- Densa con 512 neuronas, activación ReLU, regularización L1 al 0.01 e inicialización He
- Normalización Batch y dropout del 40 % de las salidas
- Densa con 256 neuronas, activación ReLU, regularización L1 al 0.01 e inicialización He
- Normalización Batch y dropout del 40 % de las salidas
- Densa con 64 neuronas, activación ReLU, regularización L1 al 0.01 e inicialización He
- Normalización Batch y dropout del 40 % de las salidas

- **Salida**: Una neurona de activación sigmoide, predicción de la probabilidad de padecer de AD

Dada la naturaleza de las capas densas, en las que cada una de sus entradas esta conectada a todas las neuronas, hace que el número de conexiones aumente de forma cuadrática respecto al tamaño de la entrada. Es por esto que reducir los SNPs de entrada es importante no solo para intentar deducir el ruido de SNPs que no tengan relación con el AD, sino también por el costo computacional que tendría entrenar un modelo con una gran cantidad de SNPs.

3.2. Otros modelos

Por otra parte, aunque en este trabajo el modelo principal a diseñar y evaluar es la red neuronal profunda, se ha optado por probar otros modelos que se han considerado relevantes, que son usados de forma habitual en el aprendizaje automático. Estos modelos a pesar de ser relativamente más sencillos que una red neuronal profunda, asientan una base y aportan un punto con el cual comparar el rendimiento de la red neuronal. Además, suelen generar predicciones bastante razonables con una configuración mínima.

3.2.1. Máquina de Vector Soporte (Support Vector Machine - SVM)

Uno de los modelos incluidos es un clasificador basado en una maquina de vector. La SVM se basa en encontrar un hiperplano que separe las clases a clasificar tratando de maximizar el margen de separación entre clases.

En ocasiones, con la dimensionalidad disponible no es posible separar linealmente las clases. En estos casos, el uso de un kernel que añada una nueva dimensión puede solventar este problema. En concreto, en este trabajo se va a usar un kernel de función de base radial.[25]

3.2.2. Bosque aleatorio (Random Forest)

Por otra parte, también se ha decidido probar un clasificador basado en bosques aleatorios (Random Forests).Random forest es un conjunto de árboles de decisión combinados mediante bagging. Con bagging, cada uno de estos árboles de decisión dispone de distintos datos de entrenamiento de manera que ninguno de los arboles entrena con todos los datos disponibles y los distintos árboles de decisión ven distintos datos. De esta forma, al combinar los resultados de los árboles de decisión, los errores de uno se compensan con los de otros y se obtiene un modelo que generaliza mejor.[26]

3.2.3. Potenciación del gradiente (Gradient boosting)

Finalmente, también se ha incluido un clasificador basado en la potenciación del gradiente.

De forma resumida, este modelo se basa en mejorar un modelo base, como por ejemplo, un árbol de decisión e intentar mejorar centrándose en los casos en los que no ha rendido correctamente. [27]

4

Resultados

En este estudio se van a comparar los resultados de los SNPs elegidos con umbrales del p-valor típicamente usados en estudios GWAS de $5 * 10^{-8}$, los cuales se aseguran una baja tasa de falsos positivos, con umbrales más permisivos, de por ejemplo $1 * 10^{-5}$ o $1 * 10^{-4}$ que devuelvan más SNPs pero con mayor tasa de falsos positivos. Se espera que con el uso de técnicas de aprendizaje automático, que el modelo no le de importancia estos SNPs que se traten de falsos positivos. Es decir, se espera que el modelo pueda eliminar el ruido de los SNPs no relacionados de los datos y pueda seguir usando el resto de SNPs que sí que le son útiles.

A el umbral del p-valor de $5 * 10^{-8}$ va hacer referencia como el umbral **tradicional**, al de $1 * 10^{-5}$, umbral **sugestivo**, y al de $1 * 10^{-4}$ umbral **relajado**

4.1. Resultados con ADNI1/GO/2 - Illumina Omni 2.5M BeadChip

Se ha decidido usar en primer lugar los datos del del Illumina Omni 2.5M BeadChip ya que es en este conjunto en el que mejor balance de muestras y SNPs hay. Tras aplicar a los datos el control de calidad descrito, se obtiene que se cuentan con 902 728 SNPs y 458 muestras, de las cuales, 219 son controles y 239 casos.

4.1.1. Comparativa GWAS

Los resultados del estudio GWAS se van a mostrar en gráficos de Manhattan. En un gráfico de Manhattan se pueden apreciar los p-valores calculados para cada SNP en el estudio de asociación. En el eje horizontal se representa la posición del SNP, y dividido en colores, su cromosoma. En el eje vertical se representa el p-valor calculado

para cada SNP. En los gráficos de Manhattan de este trabajo, se observa una linea roja horizontal a la altura del umbral usado más habitualmente en GWAS en color rojo. También se puede apreciar más abajo una linea en color azul a la altura del umbral sugestivo, y finalmente, una linea en color púrpura a la altura del umbral sugestivo.

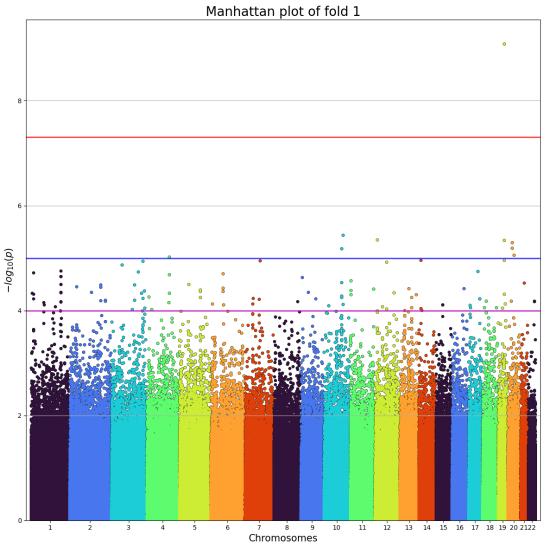

Figura 4.1: Gráfico de Manhattan para el método de división simple

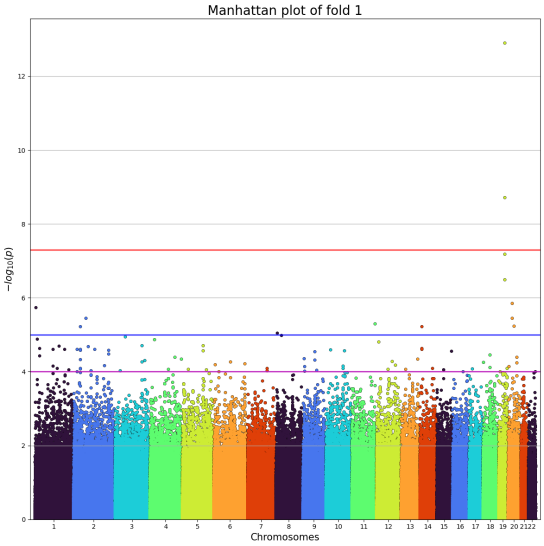

Figura 4.2: Gráfico de Manhattan para el método de no división (filtración de datos)

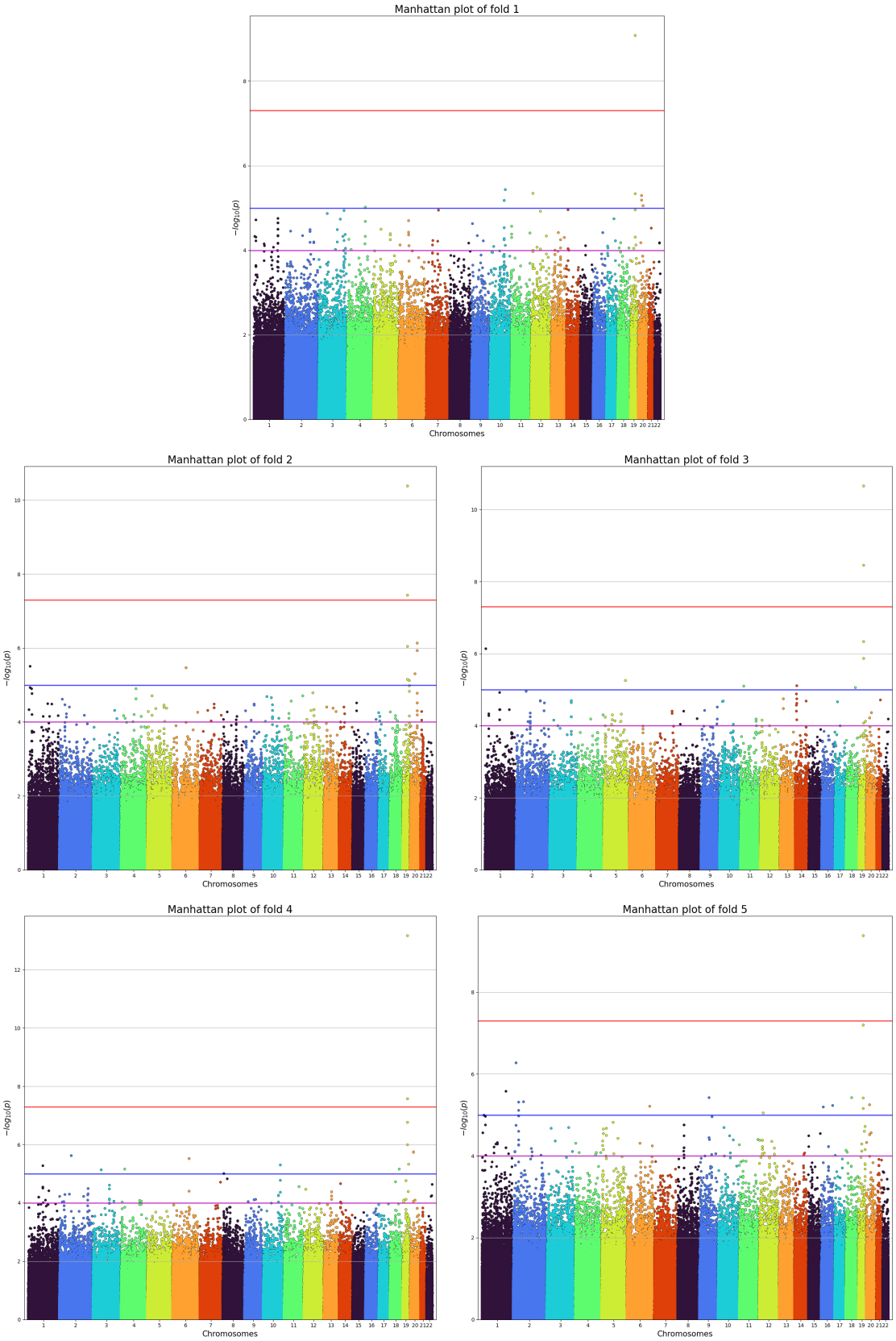

Figura 4.3: Gráficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip

Los p-valores calculados de los SNPs se pueden encontrar en el anexo B.

Aparece en primera posición para todos los GWAS el SNP rs769449 [28], del gen APOE, asociado con el riesgo de padecer AD. Este SNP siempre es elegido, ya que es el que se encuentra en primera posición y tiene el p-valor menor. Cuando se empieza a bajar el umbral aparece el SNP rs157582 [29], del gen TOMM40, asociado con la capacidad de memoria. Para el resto de SNPs no se han encontrado investigaciones que los relacionen con el AD.

4.1.2. Comparativa SNPs de fuentes externas

Se han contemplado 3 listas de las cuales se han obtenido los SNPs con los cuales realizar los experimentos.

En la primera lista, se han recogido SNPs que han sido considerado relevantes respecto al AD de un estudio realizado por Mohan Giri, Man Zhang, y Yang Lü [30]. Partiendo de 47 SNPs recogidos, en los datos disponibles se han encontrado 13. En concreto, se van a usar los SNPs rs3764650, rs3818361, rs3865444, rs3826656, rs610932, rs4938933, rs9296559, rs3851179, rs11771145, rs190982, rs16979934, rs8093731 y rs3129882.

En la segunda lista, se han recogido SNPs que han sido considerado relevantes respecto al AD de la lista publicada en SNPedia [31]. Partiendo de 35 SNPs recogidos, en los datos disponibles se han encontrado 9. En concreto, se van a usar los SNPs rs3818361, rs3851179, rs10519262, rs2227564, rs669, rs10868366, rs7019241, rs6265 y rs4878104

En la tercera lista, se han recogido SNPs que han sido considerado relevantes respecto al AD del Trabajo de Fin de Grado de Eduardo Alonso Monge [32]. Partiendo de 7 SNPs recogidos, en los datos disponibles se han encontrado 5. En conreto, se van a utilizar los SNPs rs439401, rs10779339, rs9314349, rs4726618 y rs2242601.

4.1.3. Comparativa curvas ROC

Calculando la media de las distintas iteraciones de la validación cruzada, se obtienen las siguientes curvas ROC.

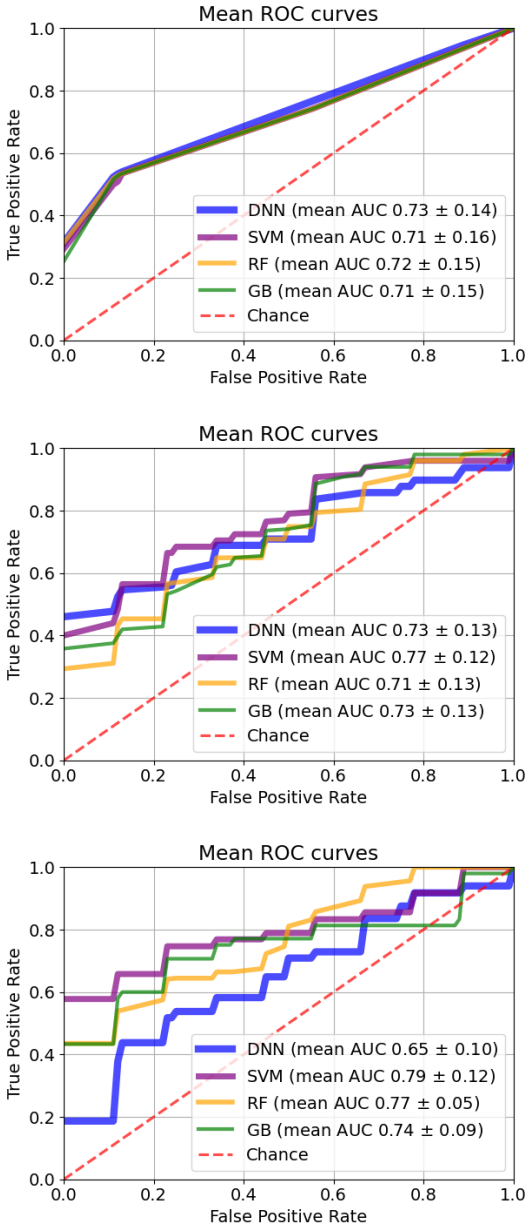

Figura 4.4: Curvas ROC, método de división simple, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo.

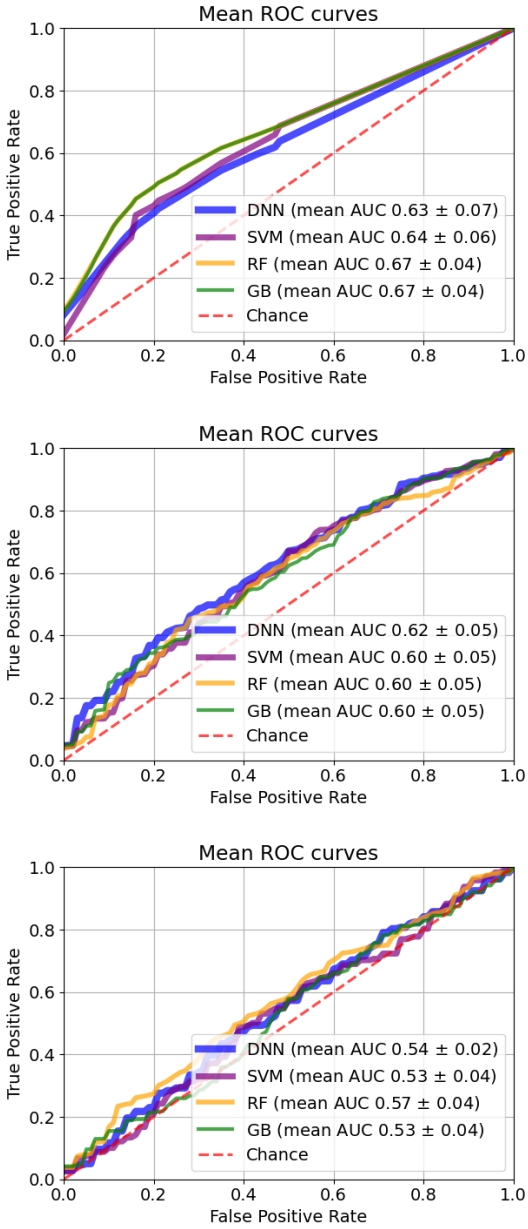

Figura 4.5: Curvas ROC, método de división por k fold, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo.

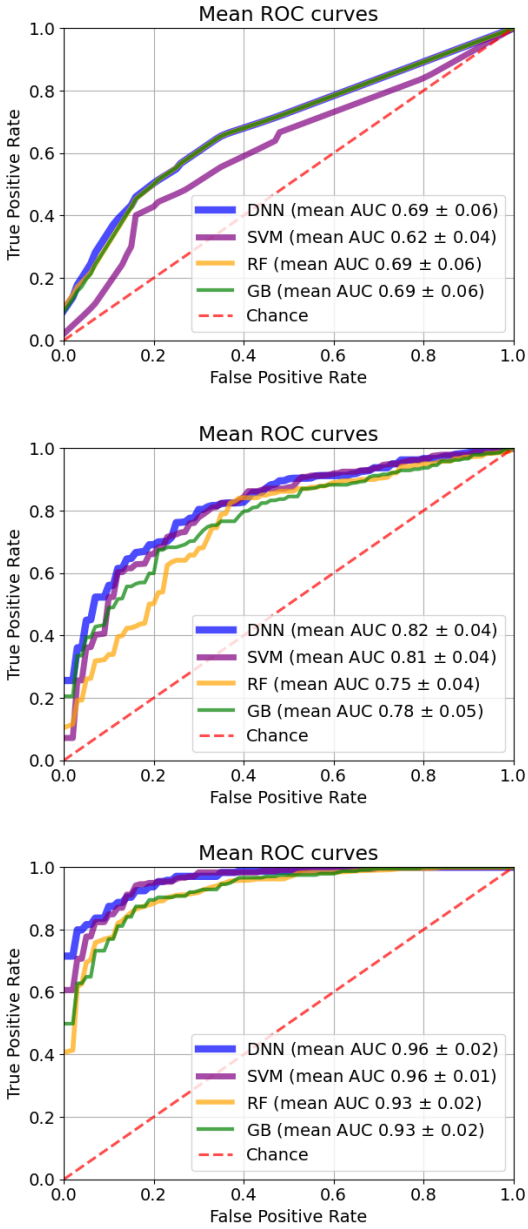

Figura 4.6: Curvas ROC, método de data leakage, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo.

Tanto para el método de división simple como para el de filtración de datos, la AUC aumenta cuando se reduce el umbral del p-valor. Sin embargo, para la división por k-fold empeora. El mejor resultado se obtiene con el método de filtración de datos y el umbral relajado.

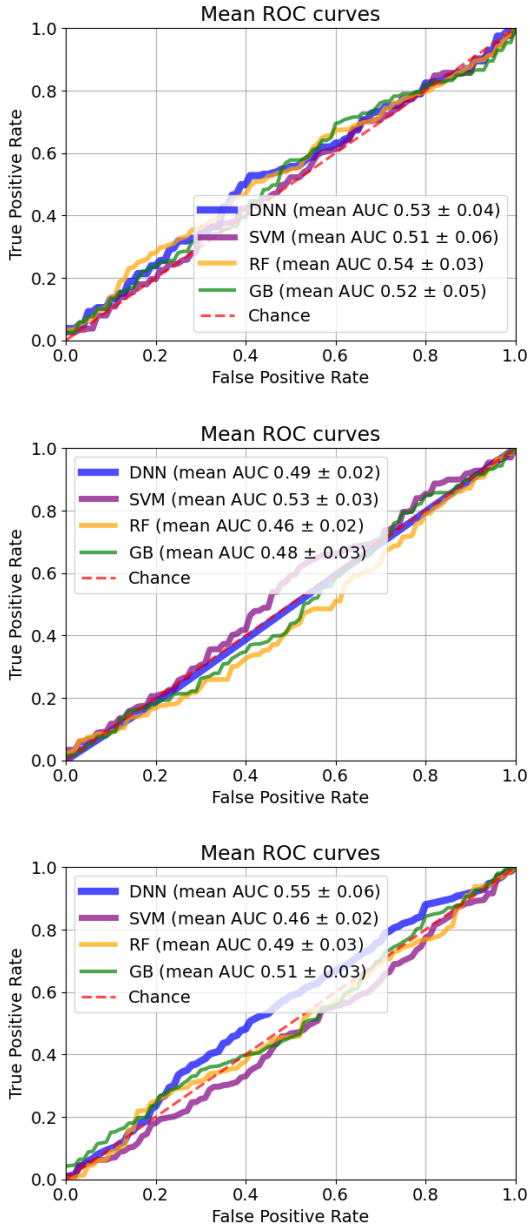

Figura 4.7: Curvas ROC, SNPs de lista 1 arriba, lista 2 en medio, lista 3 abajo.

Los tres modelos con SNPs extraídos de fuentes externas (figura 4.7) tienen una fiabilidad pésima, casi tanto como elegir el diagnóstico al azar.

4.1.4. Interpretabilidad del modelo

El calculo de los valores SHAP [12] nos permiten saber más acerca de cómo de importantes son cada una de las variables en el modelo de predicción. En este trabajo, con los valores SHAP podemos ver la correlación de los valores de los SNPs con el diagnóstico predicho, pero se debe recordar que la correlación no implica causalidad. Para poder decir que un SNP incrementa o reducir la probabilidad de tener AD no es suficiente con este valor, pero puede ser un punto de partida para futuros experimentos.

Se van a analizar los valores SHAP calculados para el mejor modelo, el obtenido con el método de data leakage y umbral p-valor relajado. Notase que en las gráficas solo se van a mostrar los 20 SNPs con mayor importancia en el modelo (según el valor SHAP), en realidad se están tomando en cuenta 82 SNPs en la entrada.

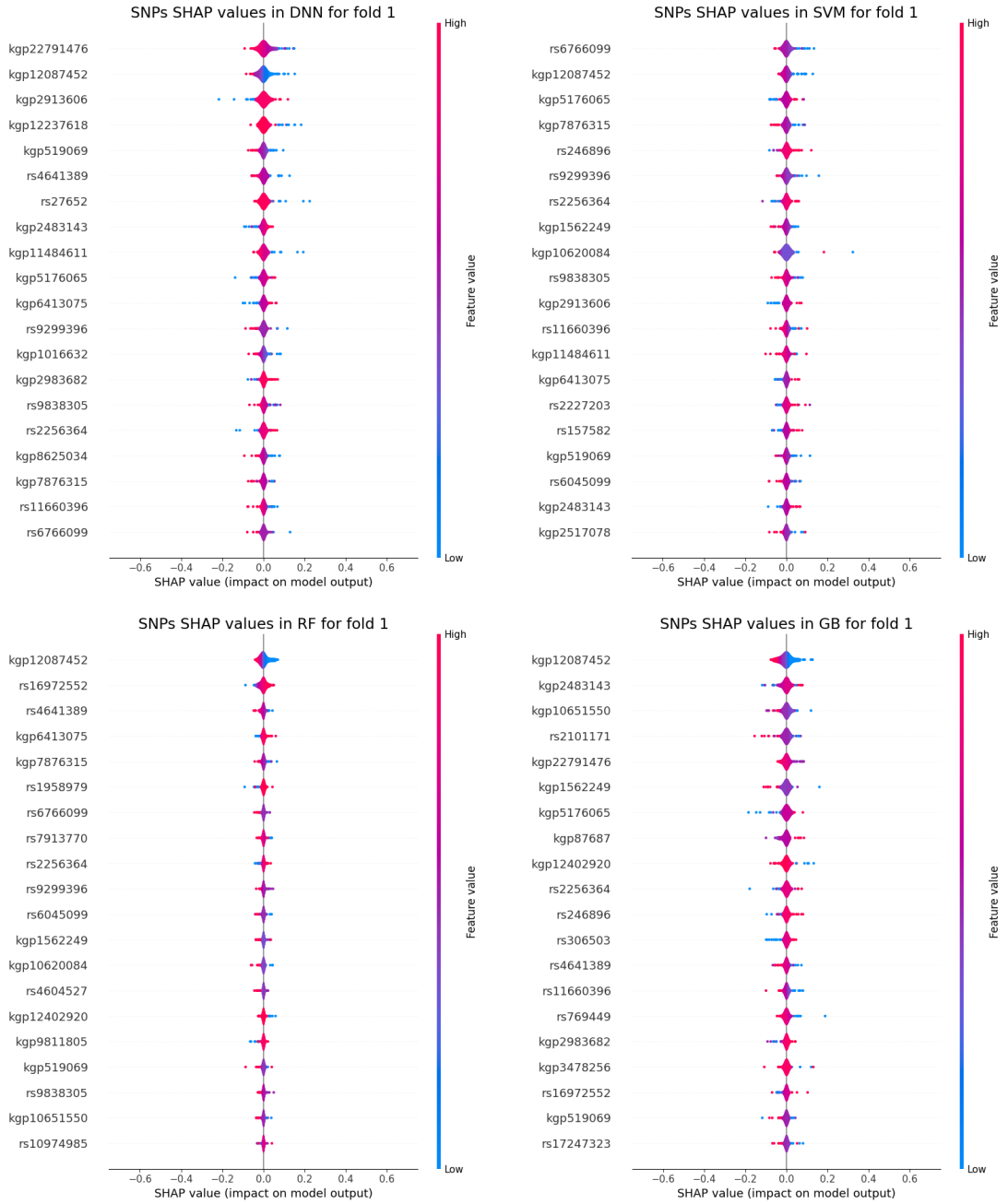

Figura 4.8: Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado

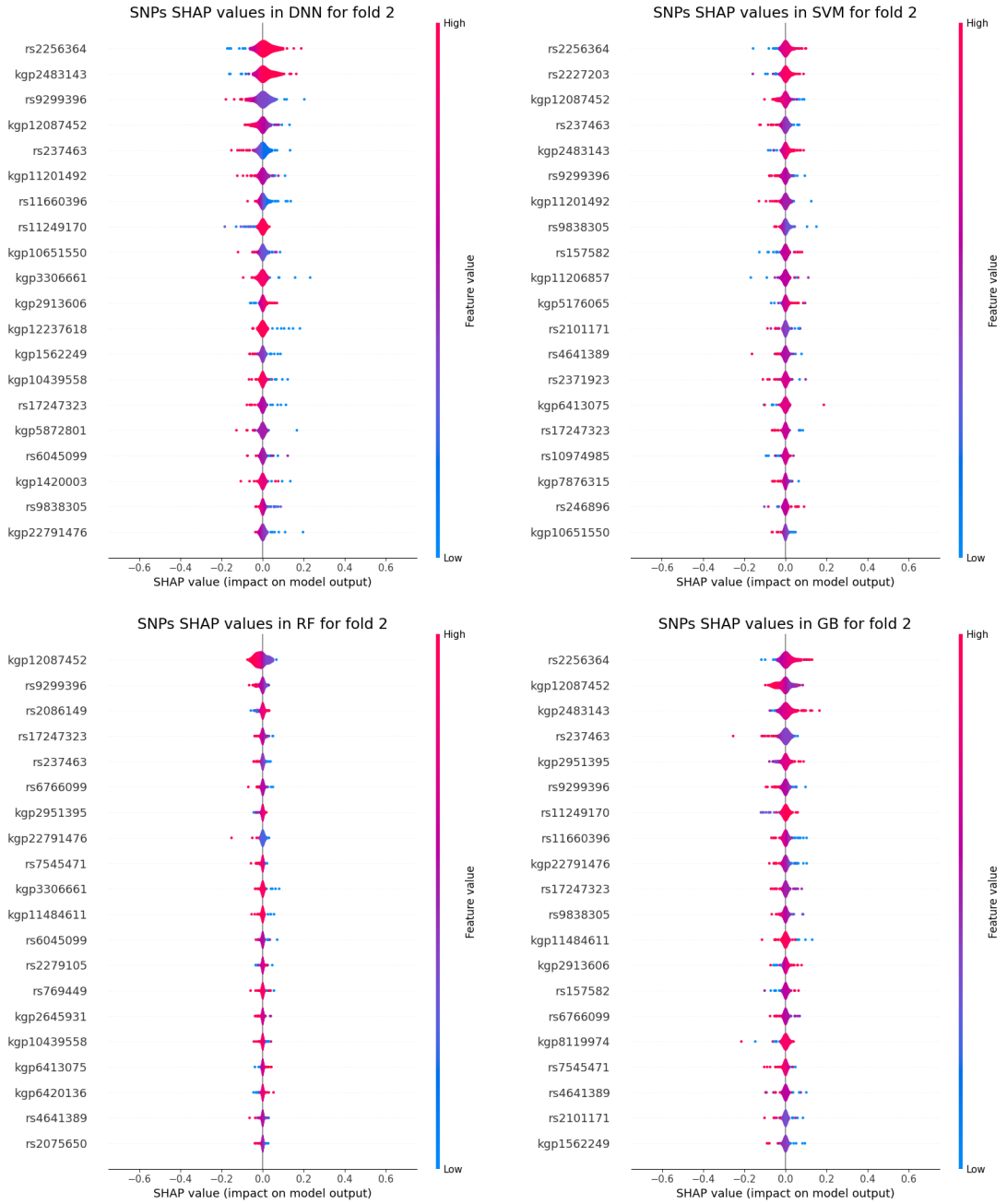

Figura 4.9: Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado

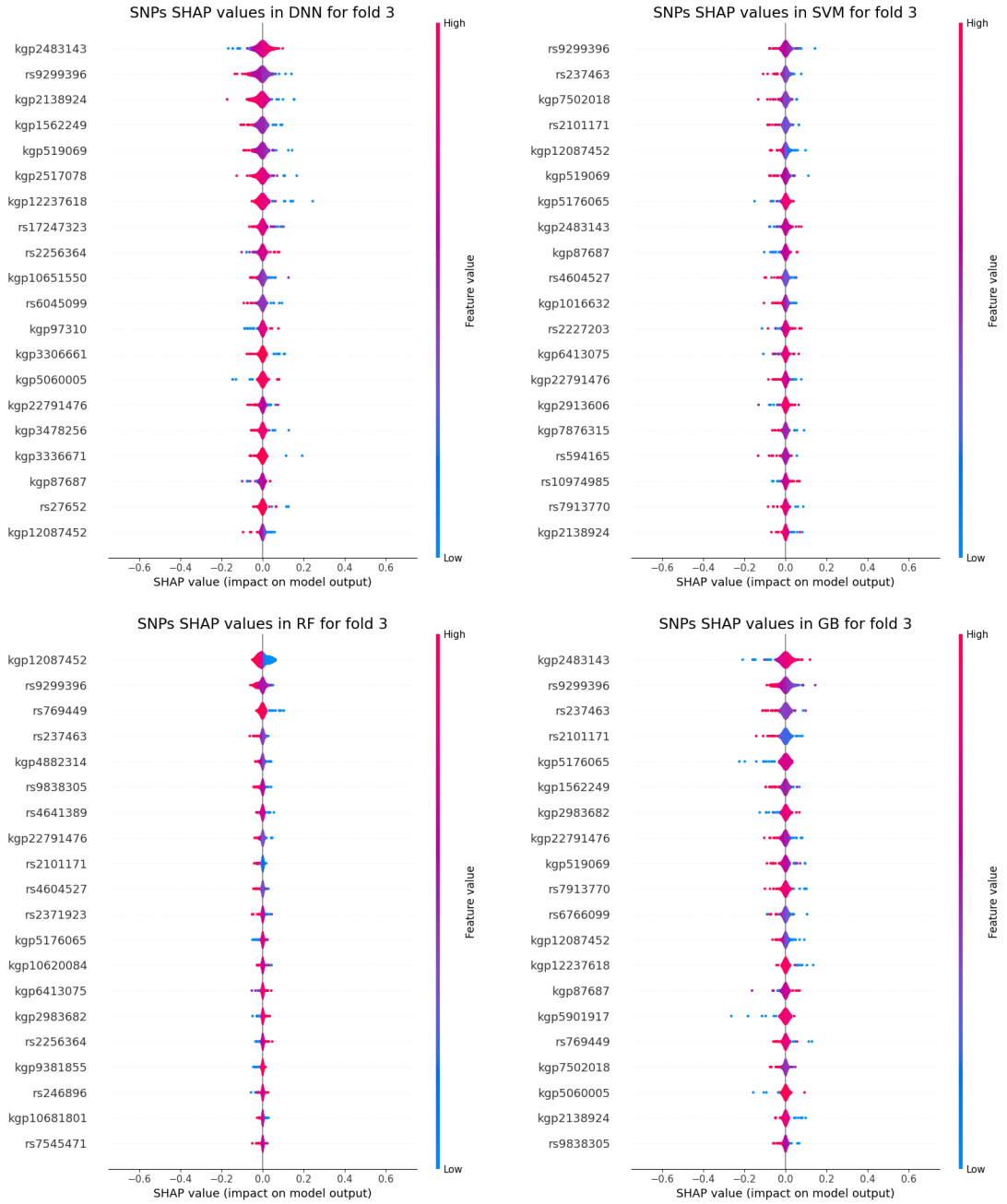

Figura 4.10: Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado

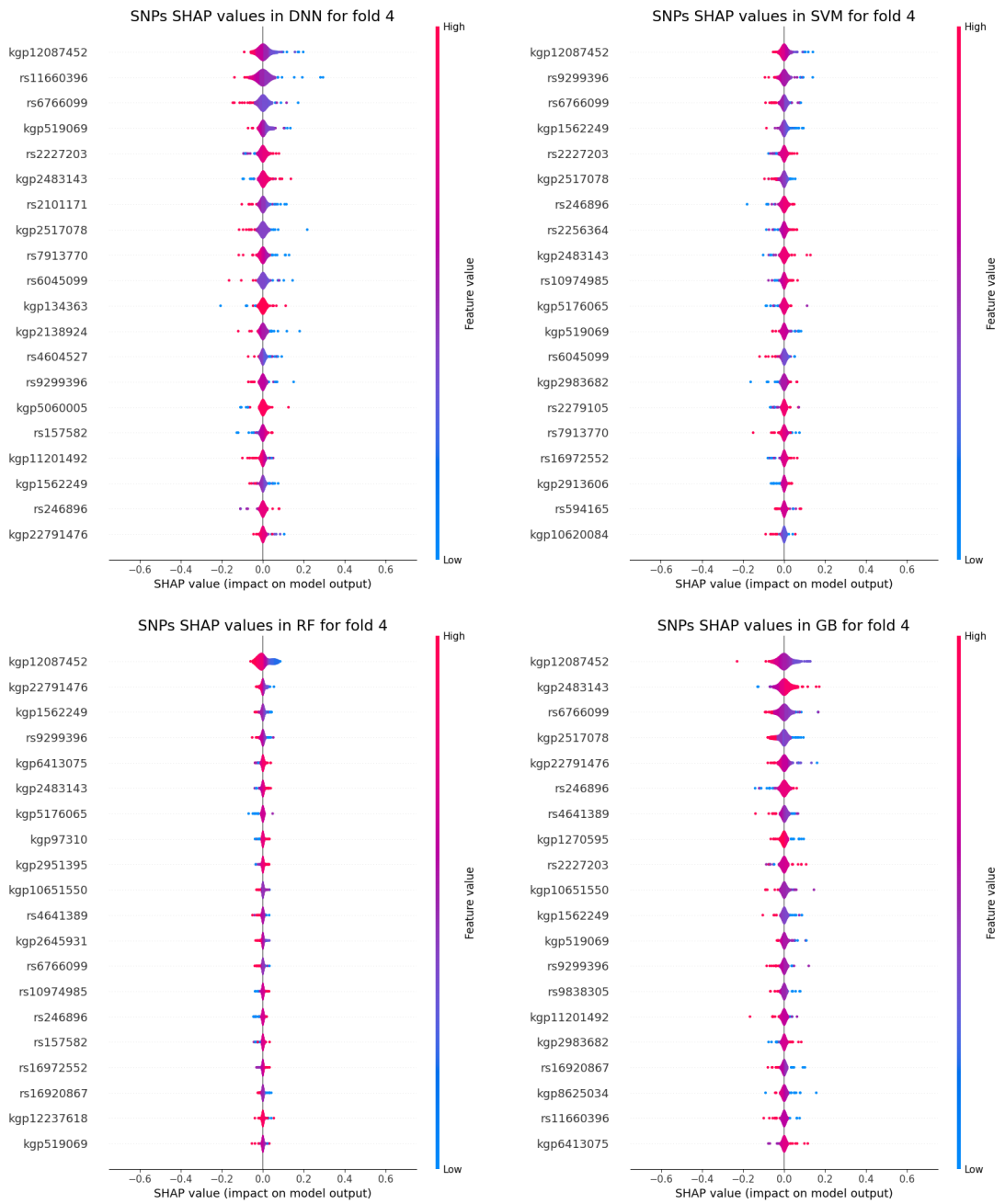

Figura 4.11: Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado

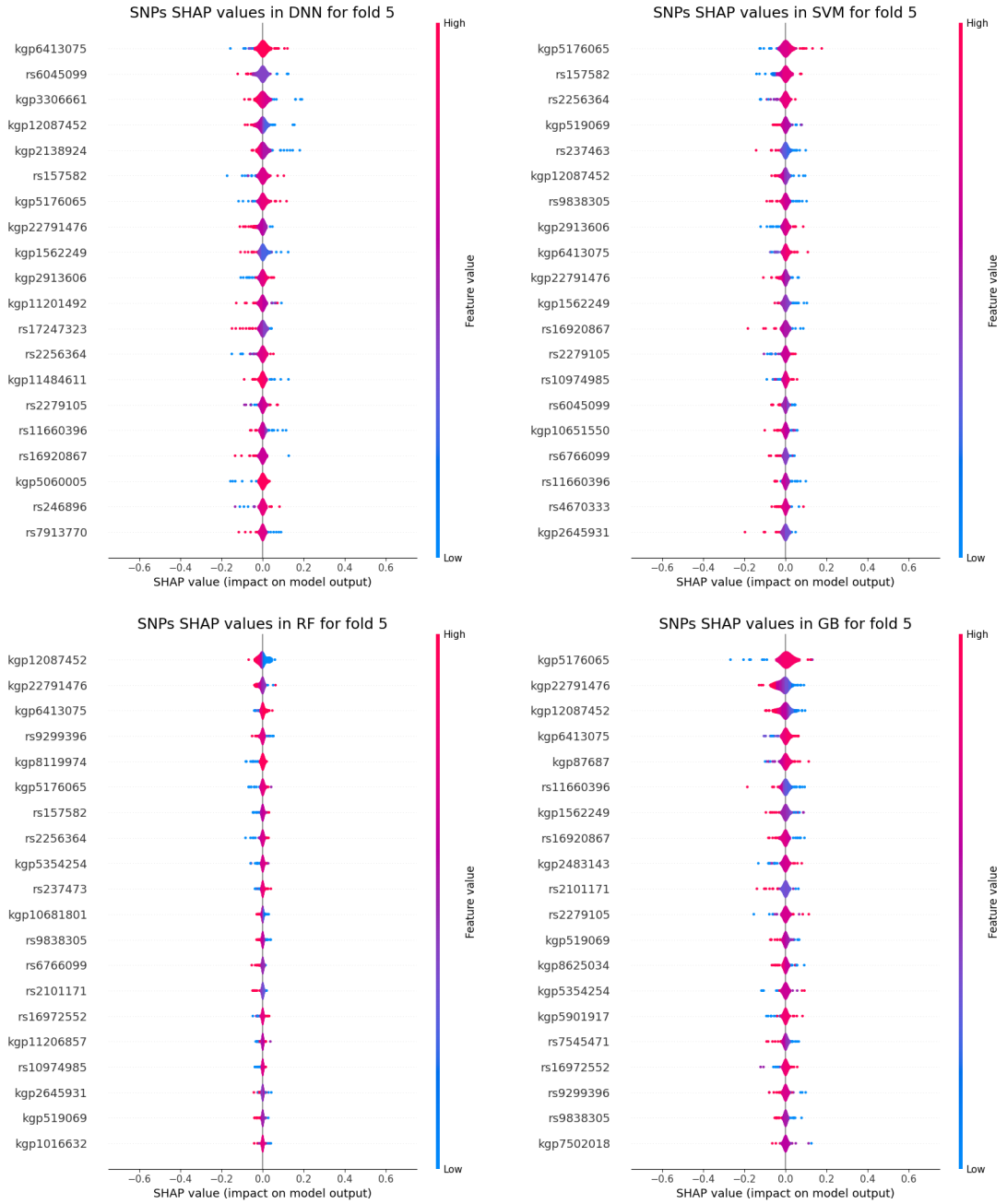

Figura 4.12: Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado

No parece haber un consenso ni entre iteraciones ni entre modelos a la hora de calcular los valores SHAP de los SNPs (figuras 4.8, 4.9, 4.10, 4.11 y 4.12). Alcanzar un consenso absoluto es prácticamente imposible, pero sí que se debería esperar la repetición de los SNPs para las posiciones más altas (las que tienen mayor valor SHAP). lo cual es un indicador más de la dificultad de encontrar relación entre información genética y el AD incluso para algoritmos de ML.

En otros contextos, los valores SHAP calculados sí son consistentes, sobre todo para los variables con valores SHAP más altos, como es el trabajo de Tomohisa Seki et al. “Machine learning-based prediction of in-hospital mortality using admission laboratory data” [33], o en el de Mónica Hernandez et al. “Explainable AI towards understanding the performance of the top-three Tadpole challenge methods in the forecast of Alzheimer’s disease diagnosis” [34]. Teniendo ambos trabajos finalidad clínica y además, en el segundo mencionado, con el pronóstico del AD.

4.2. Resultados con unión de datos de distintos chips

Se han unido los datos de los chips Illumina Human610-Quad BeadChip, Illumina HumanOmniExpress BeadChip y el Illumina Infinium Global Screening Array v2. Esto se hace con el objetivo de experimentar con un conjunto de datos en el cual se dispone de datos de más participantes, a expensas de, como se ha comentado anteriormente en el punto 2.2, se vaya a partir de menos SNPs disponibles.

Tras aplicar el control de calidad se obtienen 33710 SNPs y 1212 muestras, de las cuales 593 son controles y 619 son casos. Contar con tan pocos SNPs ha tenido un impacto negativo en los resultados obtenidos, por lo que se han obtenido peores resultados que con los daos del Illumina Omni 2.5M BeadChip.

Los resultados se encuentran recogidos en el anexo B.

5

Discusión

El objetivo de este trabajo es explorar la caracterización de la enfermedad mediante información genética cedida por la "Alzheimer's Disease Neuroimaging Initiative" (ADNI) aplicando técnicas de aprendizaje automático como el deep learning. Se ha tenido como punto de partida el trabajo realizado por Javier de Velasco Oriol, Edgar E. Vallejo, Karol Estrada y ADNI [11].

Los resultados obtenidos pueden considerarse superiores a los que se querían reproducir del artículo seleccionado. Para ser exactos en el trabajo realizado se ha alcanzado una AUC del 96 % con el umbral del p-valor tradicional, respecto al 66 % alcanzado en el articulo mencionado eligiendo los 20 mejores SNPs del estudio de asociación

Los modelos de aprendizaje automático necesitan de una selección previa de los SNPs ya que no sería lógico alimentar a un modelo con, por ejemplo, 2 millones de SNPs cuando realmente solo unos pocos van tener relación con el AD. Es aquí donde un GWAS previo permite hacer un filtrado y obtener una lista reducida de SNPs estadísticamente significativos que presentan cierta correlación con la enfermedad (dadas las muestras analizadas). Haciendo un uso combinado de GWAS con aprendizaje automático se ha logrado obtener sistemas que mejoran ambos.

Se han obtenido los mejores resultados haciendo el GWAS con todos los datos, incluidos los de test (leakage). Se deja abierto a un futuro debate si esto invalida o no el procedimiento, ya que en el GWAS sólo se realiza una preselección de los SNPs a usar, y posteriormente se realiza una validación cruzada, en la cual los datos de test no son usados durante el entrenamiento del modelo. De ser un método válido, se estaría obteniendo en este trabajo una AUC del del 0.96 %, lo cual demuestra que esta prueba podría utilizarse con una finalidad clínica.

Resultan sorprendentes los resultados tan pésimos utilizando SNPs de estudios externos, cuando estos SNPs están respaldados por otros estudios y se asume que

están relacionados con el AD. Este desajuste puede ser atribuido al hecho de que estos estudios pueden haberse realizado con poblaciones distintas e instrumentos de secuenciamiento (recogiendo distintos SNPs).

La interpretabilidad del modelo con el calculo de los valores SHAP tiene sus limitaciones, ya que solo cuantifica la correlación de las variables con el diagnóstico calculado. Es un punto de partida pero para poder confirmar la correlación entre SNPs y AD se necesitarían más pruebas que lo demuestren.

En un futuro se plantea seguir explorando esta línea de investigación, con por ejemplo, repetir los experimentos descritos cuando se disponga de más muestras. De esta forma se esperaría obtener mejores resultados.

Esta prueba tiene el beneficio de que su resultado es invariante respecto al estado de la enfermedad y por lo tanto podría realizarse con la misma efectividad antes de que la enfermedad pueda llegar a desarrollarse, ayudando ayudando así a hacer un diagnóstico temprano.

Ya que este modelo devuelve la probabilidad de que un paciente desarrolle eventualmente AD, se podría usar para detectar personas con mayor probabilidad de desarrollar AD, y por tanto, que deberían ser monitorizadas con mayor frecuencia y detectar lo antes posible los cambios físicos ocasionados por la enfermedad.

Finalmente, se plantea la posibilidad de extrapolar esta técnica a otras enfermedades que tengan, al menos en parte, causalidad genética. Se considera que una prueba que confirme cierta predisposición genética a une enfermedad puede ayudar a facilitar su detección temprana. Esto es debido a que los datos usados en esta prueba son marcadores genómicos, los cuales se mantienen constantes a lo largo de la vida del paciente.

6

Bibliografía

[1] I. Usma, "Deterioro Cognitivo Leve y Enfermedad de Alzheimer: Revisión de conceptos," *1.*, vol. 5, pp. 53–82, Aug 2017.

[2] M. A. Fearing, E. D. Bigler, M. Norton, J. A. Tschanz, C. Hulette, C. Leslie, K. Welsh-Bohmer, and C. C. Investigators, "Autopsy-confirmed Alzheimer's disease versus clinically diagnosed Alzheimer's disease in the Cache County Study on Memory and Aging: a comparison of quantitative MRI and neuropsychological findings," *J. Clin. Exp. Neuropsychol.*, vol. 29, pp. 553–560, Jul 2007.

[3] A. Kautzky, R. Seiger, A. Hahn, P. Fischer, W. Krampla, S. Kasper, G. G. Kovacs, and R. Lanzenberger, "Prediction of Autopsy Verified Neuropathological Change of Alzheimer's Disease Using Machine Learning and MRI," *Front. Aging Neurosci.*, vol. 10, Dec 2018.

[4] J. Sevigny, P. Chiao, T. Bussière, P. H. Weinreb, L. Williams, M. Maier, R. Dunstan, S. Salloway, T. Chen, Y. Ling, *et al.*, "The antibody aducanumab reduces Aβ plaques in Alzheimer's disease," *Nature*, vol. 537, pp. 50–56, Sep 2016.

[5] E. D. Roberson and L. Mucke, "100 Years and Counting: Prospects for Defeating Alzheimer's Disease," *Science (New York, N.Y.)*, vol. 314, p. 781, Nov 2006.

[6] M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, and N. L. Pedersen, "Role of genes and environments for explaining Alzheimer disease," *Arch. Gen. Psychiatry*, vol. 63, pp. 168–174, Feb 2006.

[7] T. S. Wingo, J. J. Lah, A. I. Levey, and D. J. Cutler, "Autosomal recessive causes likely in early-onset Alzheimer disease," *Arch. Neurol.*, vol. 69, pp. 59–64, Jan 2012.

[8] H. Hu, H. Li, J. Li, J. Yu, L. Tan, and A. D. N. Initiative, "Genome-wide association study identified ATP6V1H locus influencing cerebrospinal fluid BACE activity," *BMC Med. Genet.*, vol. 19, p. 75., May 2018.

[9] T. Zhou, K.-H. Thung, M. Liu, and D. Shen, "Brain-Wide Genome-Wide Association Study for Alzheimer's Disease via Joint Projection Learning and Sparse Regression Model," *IEEE Trans. Biomed. Eng.*, vol. 66, pp. 165–175, Jan 2019.

[10] J. Zhou, L. Hu, Y. Jiang, and L. Liu, "A Correlation Analysis between SNPs and ROIs of Alzheimer's Disease Based on Deep Learning," *Biomed Res. Int.*, vol. 2021, 2021.

[11] J. de Velasco Oriol, E. E. Vallejo, K. Estrada, and , "Predicting late-onset Alzheimer's disease from genomic data using deep neural networks," *bioRxiv*, p. 629402, May 2019.

[12] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions. arxiv," *arXiv preprint arXiv:1705.07874*, vol. 25, 2017.

[13] "ADNI | Alzheimer's Disease Neuroimaging Initiative," May 2021. [Online; accessed 17. May 2021].

[14] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, *et al.*, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *Am. J. Hum. Genet.*, vol. 81, p. 559, Sep 2007.

[15] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. de Andrade, K. F. Doheny, J. L. Haines, G. Hayes, *et al.*, "Quality Control Procedures for Genome Wide Association Studies," *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, vol. CHAPTER, p. Unit1.19, Jan 2011.

[16] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *Int. J. Methods Psychiatr. Res.*, vol. 27, Jun 2018.

[17] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: A Tool for Genome-wide Complex Trait Analysis," *Am. J. Hum. Genet.*, vol. 88, p. 76, Jan 2011.

[18] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group∗ under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology*, vol. 34, p. 939, Jul 1984.

[19] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nat. Rev. Genet.*, vol. 20, pp. 467–484, Aug 2019.

[20] D. Horta, "pandas-plink," 2020.

[21] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Sci. Rep.*, vol. 9, pp. 1–7, Aug 2019.

[22] A. Sharma, A. Lysenko, B. K. A., E. Vans, and T. Tsunoda, "DeepInsight-FS: Selecting features for non-image data using convolutional neural network," Sep 2020. [Online; accessed 28. Jun. 2021].

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[24] F. Chollet *et al.*, "Keras," 2015.

[25] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and Applications," *ResearchGate*, vol. 2049, pp. 249–257, Jan 2001.

[26] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *arXiv*, Mar 2016.

[28] "rs769449 - SNPedia," Nov 2021. [Online; accessed 6. Nov. 2021].

[29] "rs157582 - SNPedia," Nov 2021. [Online; accessed 21. Nov. 2021].

[30] M. Giri, M. Zhang, and Y. Lü, "Genes associated with Alzheimer's disease: an overview and current status," *Clin. Interv. Aging*, vol. 11, p. 665, 2016.

[31] "Alzheimer's disease - SNPedia," Nov 2021. [Online; accessed 1. Nov. 2021].

[32] E. Alonso Monge, E. Mayordomo Cámara, and M. Hernández Giménez, "Un estudio de asociación genómica basado en aprendizaje automático para la caracterización de la enfermedad de Alzheimer," 2020.

[33] T. Seki, Y. Kawazoe, and K. Ohe, "Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data," *PLoS One*, vol. 16, no. 2, 2021.

[34] M. Hernandez, U. Ramon-Julvez, F. Ferraz, and ADNI, "Explainable ai towards understanding the performance of the top-three tadpole challenge methods in the forecast of alzheimer's disease diagnosis." submitted to PLoS One, 2021.

Lista de Figuras

3.1. Esquema conceptual del trabajo (iconos obtenidos en Flaticon.com) . 13

4.1. Gráfico de Manhattan para el método de división simple 19
4.2. Gráfico de Manhattan para el método de no división (filtración de datos) 19
4.3. Gráficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip 20
4.4. Curvas ROC, método de división simple, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo. 22
4.5. Curvas ROC, método de división por k fold, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo. 23
4.6. Curvas ROC, método de data leakage, umbral p-valor tradicional arribo, sugestivo en medio y relajado abajo. 24
4.7. Curvas ROC, SNPs de lista 1 arriba, lista 2 en medio, lista 3 abajo. . 25
4.8. Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado 27
4.9. Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado 28
4.10. Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado 29
4.11. Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado 30
4.12. Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado 31

B.1. Gráfico de Manhattan para el método de división simple 49
B.2. Matrices de confusión, método de división simple, umbral p-valor tradicional 50

B.3. Curvas ROC, método de división simple, umbral p-valor tradicional . 51
B.4. Valores SHAP fold 1, método de división simple, umbral p-valor sugestivo 51
B.5. Valores SHAP fold 2, método de división simple, umbral p-valor sugestivo 52
B.6. Valores SHAP fold 3, método de división simple, umbral p-valor sugestivo 52
B.7. Valores SHAP fold 4, método de división simple, umbral p-valor sugestivo 53
B.8. Valores SHAP fold 5, método de división simple, umbral p-valor sugestivo 53
B.9. Matrices de confusión, método de división simple, umbral p-valor sugestivo 54
B.10. Curvas ROC, método de división simple, umbral p-valor sugestivo . . 54
B.11. Valores SHAP fold 1, método de división simple, umbral p-valor relajado 55
B.12. Valores SHAP fold 2, método de división simple, umbral p-valor relajado 56
B.13. Valores SHAP fold 3, método de división simple, umbral p-valor relajado 57
B.14. Valores SHAP fold 4, método de división simple, umbral p-valor relajado 58
B.15. Valores SHAP fold 5, método de división simple, umbral p-valor relajado 59
B.16. Matrices de confusión, método de división simple, umbral p-valor relajado 60
B.17. Curvas ROC, método de división simple, umbral p-valor relajado . . . 60
B.18. Graficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip 66
B.19. Matrices de confusión, método de división por k-fold, umbral p-valor tradicional 67
B.20. Curvas ROC, método de división por k-fold, umbral p-valor tradicional 68
B.21. Valores SHAP fold 1, método de división por k-fold, umbral p-valor sugestivo 68
B.22. Valores SHAP fold 2, método de división por k-fold, umbral p-valor sugestivo 69
B.23. Valores SHAP fold 3, método de división por k-fold, umbral p-valor sugestivo 69
B.24. Valores SHAP fold 4, método de división por k-fold, umbral p-valor sugestivo 70

B.25. Valores SHAP fold 5, método de división por k-fold, umbral p-valor sugestivo 71
B.26. Matrices de confusión, método de división por k-fold, umbral p-valor sugestivo 72
B.27. Curvas ROC, método de división por k-fold, umbral p-valor sugestivo 72
B.28. Valores SHAP fold 1, método de división por k fold, umbral p-valor relajado 73
B.29. Valores SHAP fold 2, método de división por k fold, umbral p-valor relajado 74
B.30. Valores SHAP fold 3, método de división por k fold, umbral p-valor relajado 75
B.31. Valores SHAP fold 4, método de división por k fold, umbral p-valor relajado 76
B.32. Valores SHAP fold 5, método de división por k fold, umbral p-valor relajado 77
B.33. Matrices de confusión, método de división por k fold, umbral p-valor relajado 78
B.34. Curvas ROC, método de división por k fold, umbral p-valor relajado . 78
B.35. Gráfico de Manhattan para el método de no división (filtración de datos) 80
B.36. Matrices de confusión, método de no división (filtración de datos), umbral p-valor tradicional 81
B.37. Curvas ROC, método de no división (filtración de datos), umbral p-valor tradicional 82
B.38. Valores SHAP fold 1, método de no división (filtración de datos), umbral p-valor sugestivo 83
B.39. Valores SHAP fold 2, método de no división (filtración de datos), umbral p-valor sugestivo 84
B.40. Valores SHAP fold 3, método de no división (filtración de datos), umbral p-valor sugestivo 85
B.41. Valores SHAP fold 4, método de no división (filtración de datos), umbral p-valor sugestivo 86
B.42. Valores SHAP fold 5, método de no división (filtración de datos), umbral p-valor sugestivo 87

B.43. Matrices de confusión, método de no división (filtración de datos), umbral p-valor sugestivo 88
B.44. Curvas ROC, método de no división (filtración de datos), umbral p-valor sugestivo 88
B.45. Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado 89
B.46. Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado 90
B.47. Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado 91
B.48. Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado 92
B.49. Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado 93
B.50. Matrices de confusión, método de filtración de datos, umbral p-valor relajado 94
B.51. Curvas ROC, método de filtración de datos, umbral p-valor relajado . 94
B.52. Matrices de confusión, lista 1 SNPs 95
B.53. Curvas ROC, SNPs de lista 1 96
B.54. Matrices de confusión, SNPs de lista 2 97
B.55. Curvas ROC, SNPs de lista 2 98
B.56. Matrices de confusión, SNPs de lista 3 99
B.57. Curvas ROC, SNPs de lista 3 100
B.58. Gráfico de Manhattan para el método de división simple 101
B.59. Matrices de confusión, método de división simple, umbral p-valor tradicional 102
B.60. Curvas ROC, método de división simple, umbral p-valor tradicional . 103
B.61. Matrices de confusión, método de división simple, umbral p-valor sugestivo 104
B.62. Curvas ROC, método de división simple, umbral p-valor sugestivo . . 105
B.63. Valores SHAP fold 1, método de división simple, umbral p-valor relajado 105
B.64. Valores SHAP fold 2, método de división simple, umbral p-valor relajado 106
B.65. Valores SHAP fold 3, método de división simple, umbral p-valor relajado 106
B.66. Valores SHAP fold 4, método de división simple, umbral p-valor relajado 107
B.67. Valores SHAP fold 5, método de división simple, umbral p-valor relajado 107

B.68. Matrices de confusión, método de división simple, umbral p-valor relajado . . . 108
B.69. Curvas ROC, método de división simple, umbral p-valor relajado . . . 108
B.70. Graficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip . . . 110
B.71. Matrices de confusión, método de división por k-fold, umbral p-valor tradicional . . . 111
B.72. Curvas ROC, método de división por k-fold, umbral p-valor tradicional 112
B.73. Matrices de confusión, método de división por k-fold, umbral p-valor sugestivo . . . 113
B.74. Curvas ROC, método de división por k-fold, umbral p-valor sugestivo 114
B.75. Valores SHAP fold 1, método de división por k fold, umbral p-valor relajado . . . 114
B.76. Valores SHAP fold 2, método de división por k fold, umbral p-valor relajado . . . 115
B.77. Valores SHAP fold 3, método de división por k fold, umbral p-valor relajado . . . 115
B.78. Valores SHAP fold 4, método de división por k fold, umbral p-valor relajado . . . 116
B.79. Valores SHAP fold 5, método de división por k fold, umbral p-valor relajado . . . 116
B.80. Matrices de confusión, método de división por k fold, umbral p-valor relajado . . . 117
B.81. Curvas ROC, método de división por k fold, umbral p-valor relajado . 117
B.82. Gráfico de Manhattan para el método de no división (filtración de datos) . . . 118
B.83. Matrices de confusión, método de no división (filtración de datos), umbral p-valor tradicional . . . 119
B.84. Curvas ROC, método de no división (filtración de datos), umbral p-valor tradicional . . . 120
B.85. Valores SHAP fold 1, método de no división (filtración de datos), umbral p-valor sugestivo . . . 120
B.86. Valores SHAP fold 2, método de no división (filtración de datos), umbral p-valor sugestivo . . . 121

B.87. Valores SHAP fold 3, método de no división (filtración de datos), umbral p-valor sugestivo 121
B.88. Valores SHAP fold 4, método de no división (filtración de datos), umbral p-valor sugestivo 121
B.89. Valores SHAP fold 5, método de no división (filtración de datos), umbral p-valor sugestivo 122
B.90. Matrices de confusión, método de no división (filtración de datos), umbral p-valor sugestivo 122
B.91. Curvas ROC, método de no división (filtración de datos), umbral p-valor sugestivo 123
B.92. Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado 123
B.93. Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado 124
B.94. Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado 124
B.95. Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado 125
B.96. Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado 125
B.97. Matrices de confusión, método de filtración de datos, umbral p-valor relajado 126
B.98. Curvas ROC, método de filtración de datos, umbral p-valor relajado . 126

Lista de Tablas

2.1. Fases de estudio, instrumento de secuenciamiento y número de SNPs 5
2.2. Fases de estudio y número de pacientes de cada etnia 7
2.3. Fases de estudio y número de pacientes con cada diagnóstico 8

Anexos

A

Uso de PLINK

En la realización de este trabajo, una de las herramientas especialmente relacionadas del campo de la biología y genética ha sido PLINK. Dicha herramienta permite el tratamiento de datos genómicos de forma compacta.

El uso de PLINK ha simplificado la tarea de aplicar el control de calidad a los datos y realización del estudio de asociación del genoma completo.

Partes del control de calidad han tenido que realizarse haciendo uso de programas desarrollados en Python. Por ejemplo, a la hora de analizar la estratificación de la población, en el que se hace uso de los datos demográficos proporcionados por ADNI que se tuvieron que cargar desde un fichero externo. Este script puede ejecutarse de la siguiente forma:

```
python3 exclude_non_white.py plink_file diagnosis_file
   ↪ demographic_file out_file
```

El resto del control de calidad se ha podido realizar con el siguiente comando en PLINK:

```
plink --bfile fichero_entrada --make-bed --missing-
   ↪ genotype N --chr 1-22 --maf 0.01 --geno 0.001 --hwe
   ↪ 0.05 --rel-cutoff 0.25 --out fichero_salida
```

Finalmente para realizar los k-folds de la validación cruzada nuevamente se ha diseñado un script en Python para realizar de forma automática esta tarea, el cual puede ejecutarse de la siguiente forma:

```
python3 split_data.py plink_file splits out_file_root
```

B

Recopilación de resultados

B.1. Resultados con ADNI1/GO/2 - Illumina Omni 2.5M BeadChip

Tras aplicar a los datos el control de calidad descrito, se obtiene que se cuentan con 902 728 SNPs y 458 muestras, de las cuales, 219 son controles y 239 casos.

B.1.1. Método de división simple

Resultados estudio GWAS (hasta p-valor relajado): rs769449 (8.19e-10), kgp11201492 (3.61e-06), kgp12193771 (4.51e-06), rs157582 (4.60e-06), kgp2517078 (5.08e-06), rs6045099 (6.48e-06), rs7913770 (6.62e-06), kgp9381855 (8.80e-06), kgp9008013 (9.43e-06), rs2075650 (1.10e-05), kgp87687 (1.10e-05), rs7799610 (1.12e-05), rs872013 (1.14e-05), rs17247323 (1.19e-05), rs6766099 (1.34e-05), kgp5049911 (1.74e-05), kgp9042226 (1.79e-05), kgp10548801 (1.83e-05), rs2999899 (1.90e-05), rs870032 (1.96e-05), rs12505461 (2.05e-05), rs10449303 (2.23e-05), rs859706 (2.25e-05), kgp6482496 (2.25e-05), rs16920867 (2.34e-05), kgp11630376 (2.71e-05), kgp9002459 (2.93e-05), rs6517046 (3.00e-05), kgp12394039 (3.17e-05), rs2724387 (3.19e-05), kgp6404324 (3.19e-05), rs1196184 (3.22e-05), rs819086 (3.23e-05), rs4670333 (3.50e-05), kgp1897704 (3.57e-05), rs1929762 (3.73e-05), kgp9443084 (3.78e-05), rs4883820 (3.80e-05), kgp5176065 (3.87e-05), rs246896 (4.04e-05), rs926101 (4.08e-05), kgp9016601 (4.20e-05), rs246887 (4.29e-05), rs12196881 (4.32e-05), kgp9283537 (4.44e-05), kgp8119974 (4.44e-05), kgp5837647 (4.50e-05), rs2745980 (4.50e-05), kgp12402920 (4.54e-05), kgp3377949 (4.60e-05), kgp22791476 (4.62e-05), kgp2983682 (4.62e-05), rs12509463 (4.68e-05), kgp22729362 (4.81e-05), kgp7502018 (4.86e-05), kgp1560994 (4.92e-05), kgp4882314 (4.94e-05), rs4641389 (5.20e-05), kgp12419671 (5.49e-05), rs27652 (5.54e-05), rs1360631 (5.54e-05), rs7899065 (5.64e-05), kgp3504413 (5.78e-

05), rs4912515 (5.80e-05), rs2131606 (5.89e-05), kgp8631591 (5.89e-05), kgp7046169 (5.97e-05), rs1416623 (5.99e-05), rs917424 (6.07e-05), rs9883988 (6.25e-05), rs16854581 (6.28e-05), rs1196166 (6.46e-05), kgp9209674 (6.56e-05), rs4800353 (6.56e-05), kgp2083150 (6.58e-05), kgp8625034 (6.60e-05), kgp679032 (6.60e-05), kgp2913606 (6.66e-05), rs7017572 (6.67e-05), kgp840298 (6.73e-05), rs7541230 (6.79e-05), rs7658630 (6.93e-05), rs6665401 (7.02e-05), rs10281903 (7.43e-05), kgp7876315 (7.44e-05), rs279447 (7.44e-05), rs17819035 (7.51e-05), kgp2645931 (7.67e-05), kgp10681801 (7.67e-05), rs11571943 (7.79e-05), rs166642 (7.80e-05), kgp9921316 (7.91e-05), rs4082518 (8.12e-05), rs2227203 (8.41e-05), rs7973722 (8.43e-05), rs4555239 (8.69e-05), rs1360632 (8.73e-05), rs2072321 (8.83e-05), rs4606813 (8.83e-05), rs9873487 (8.86e-05), kgp2579012 (8.89e-05), kgp3306661 (9.18e-05), kgp6357006 (9.36e-05), kgp7187035 (9.45e-05), kgp5354254 (9.45e-05), rs4981220 (9.83e-05), rs7337312 (9.92e-05).

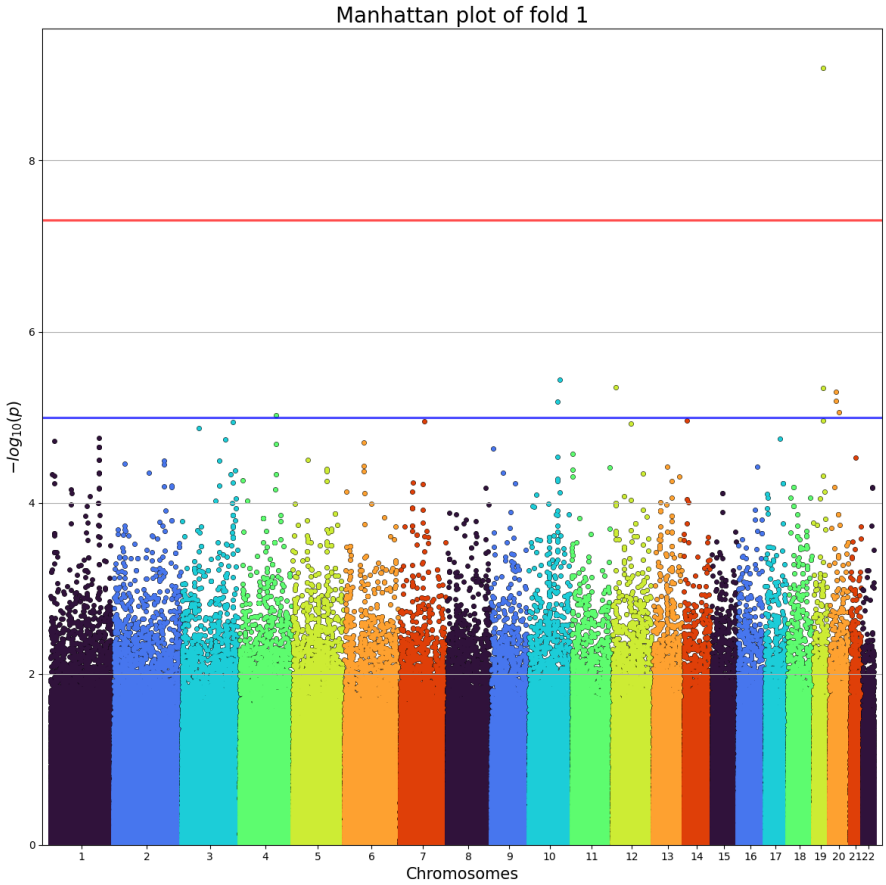

Figura B.1: Gráfico de Manhattan para el método de división simple

Umbral clásico

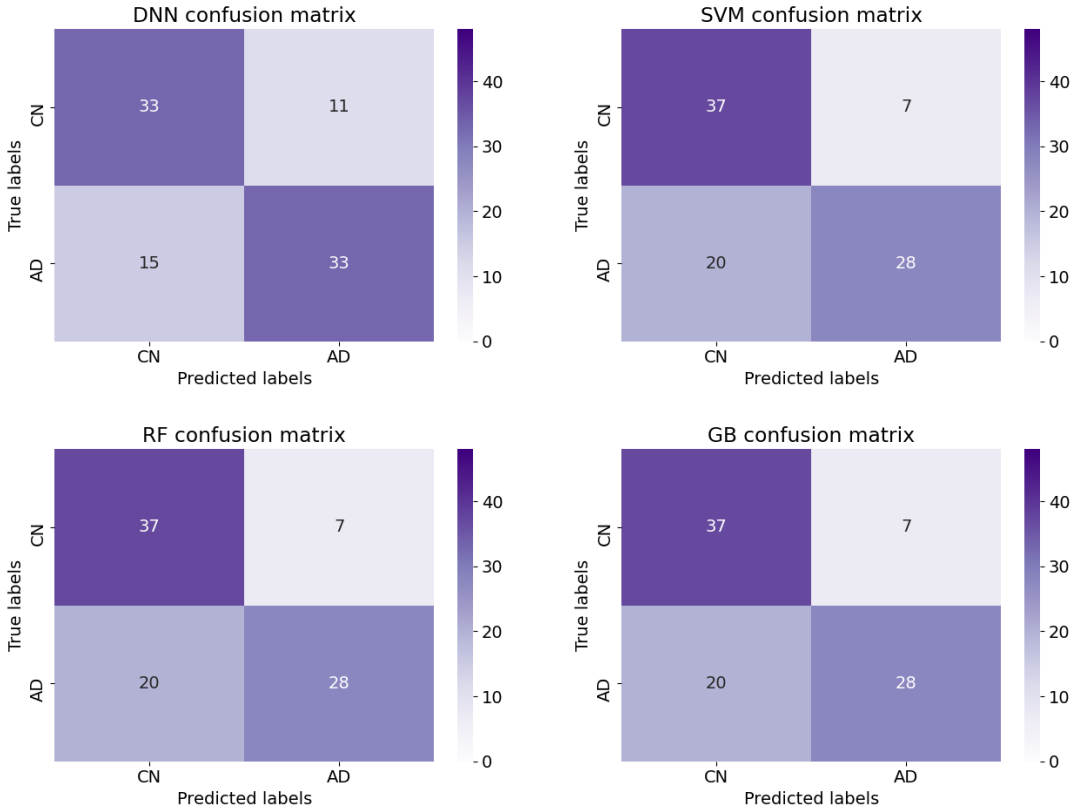

Figura B.2: Matrices de confusión, método de división simple, umbral p-valor tradicional

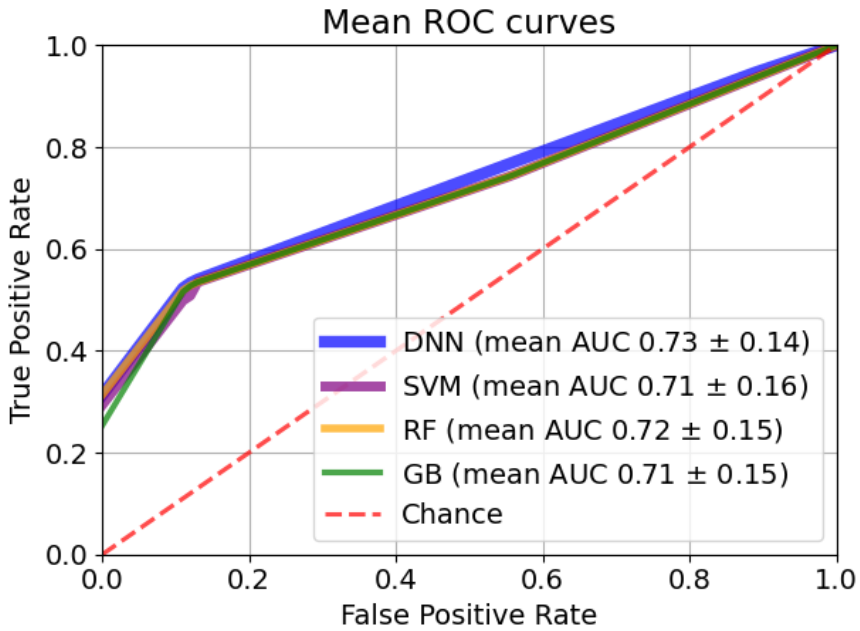

Figura B.3: Curvas ROC, método de división simple, umbral p-valor tradicional

Umbral sugestivo

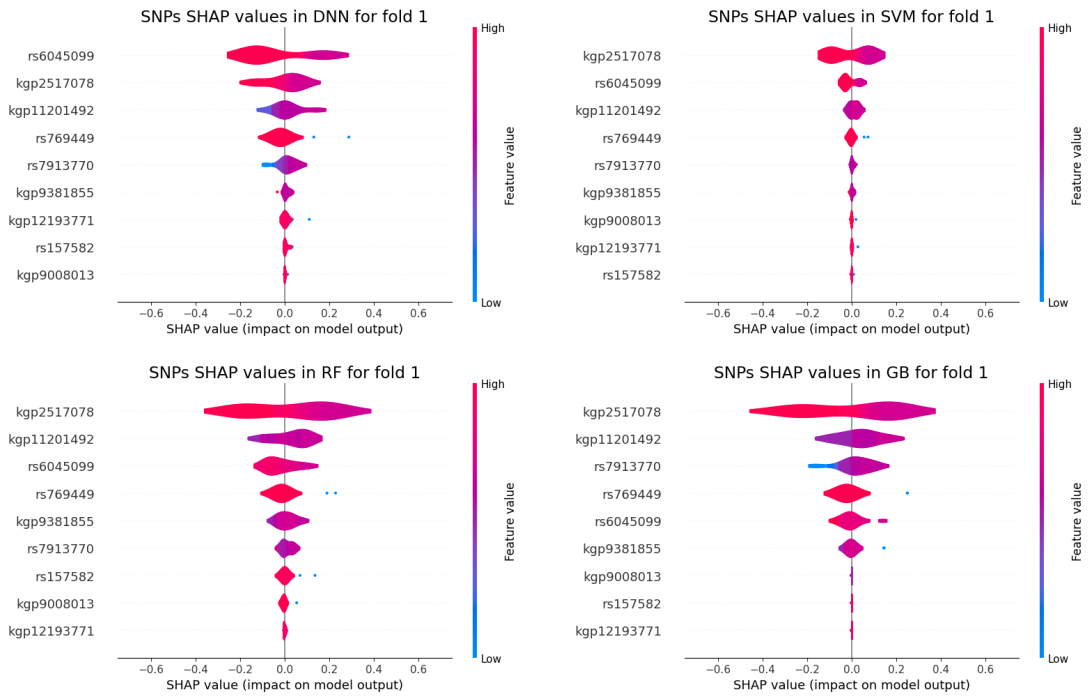

Figura B.4: Valores SHAP fold 1, método de división simple, umbral p-valor sugestivo

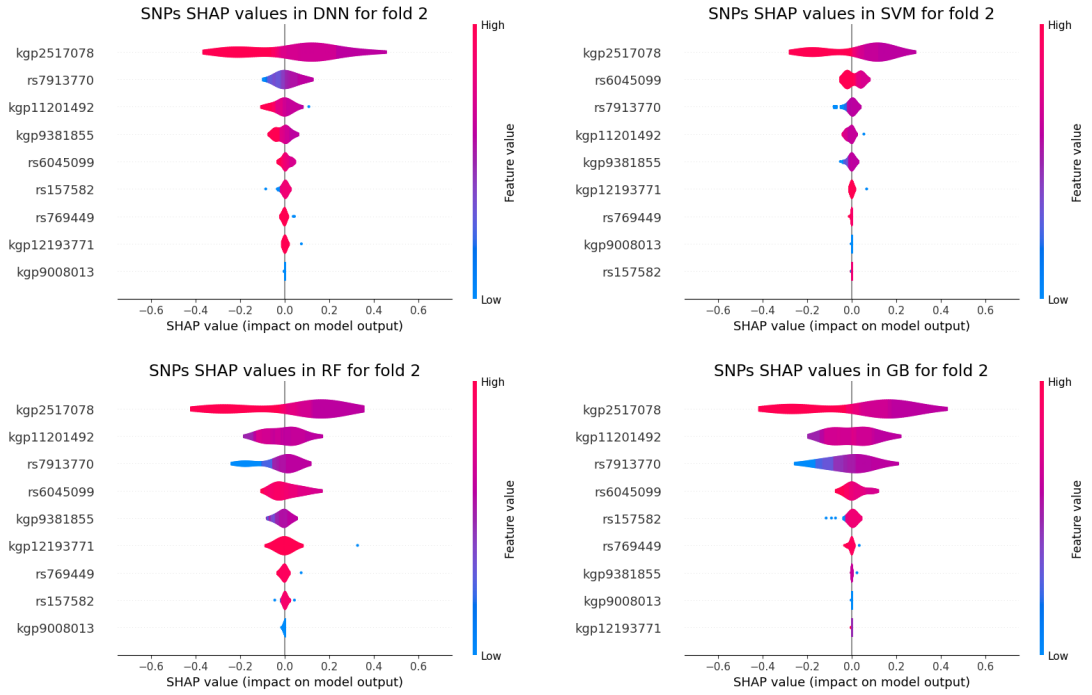

Figura B.5: Valores SHAP fold 2, método de división simple, umbral p-valor sugestivo

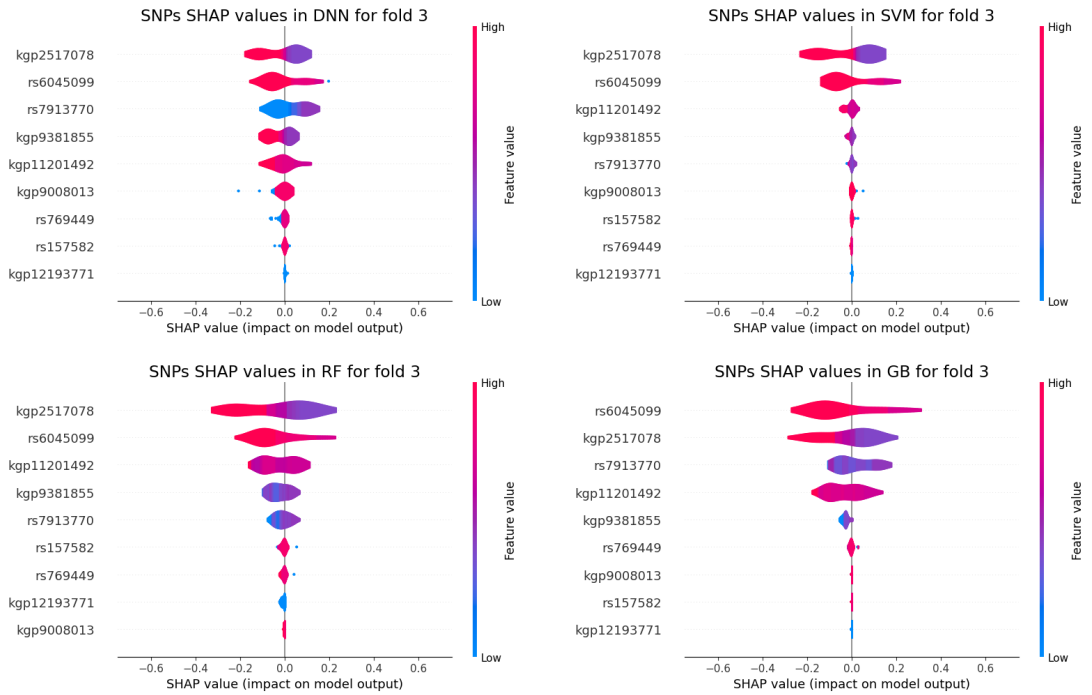

Figura B.6: Valores SHAP fold 3, método de división simple, umbral p-valor sugestivo

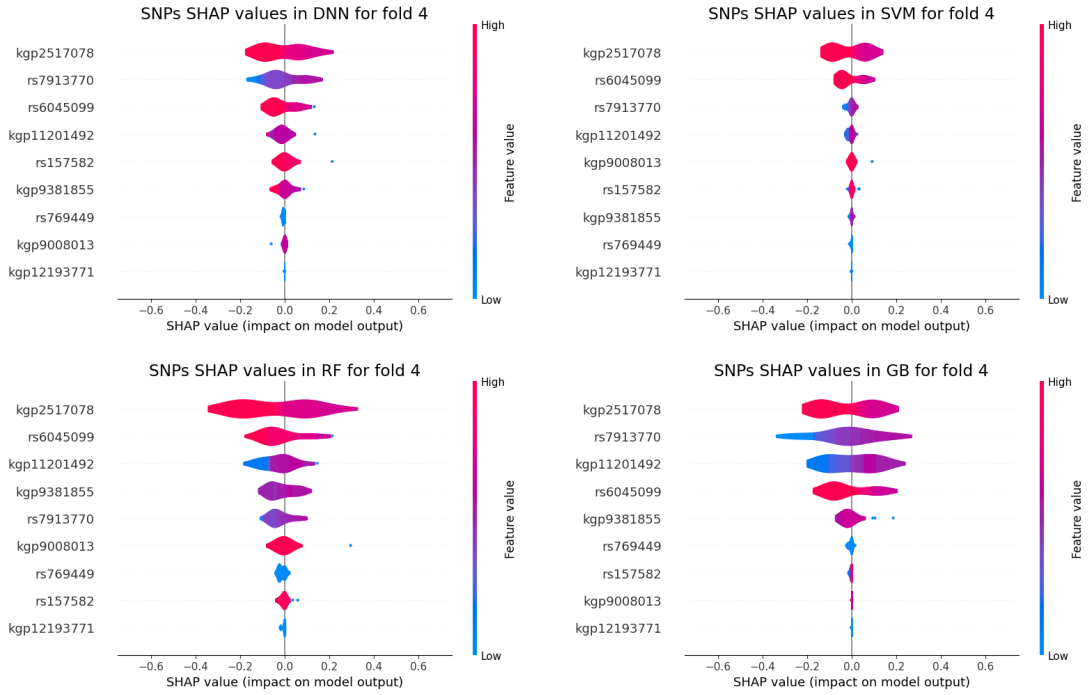

Figura B.7: Valores SHAP fold 4, método de división simple, umbral p-valor sugestivo

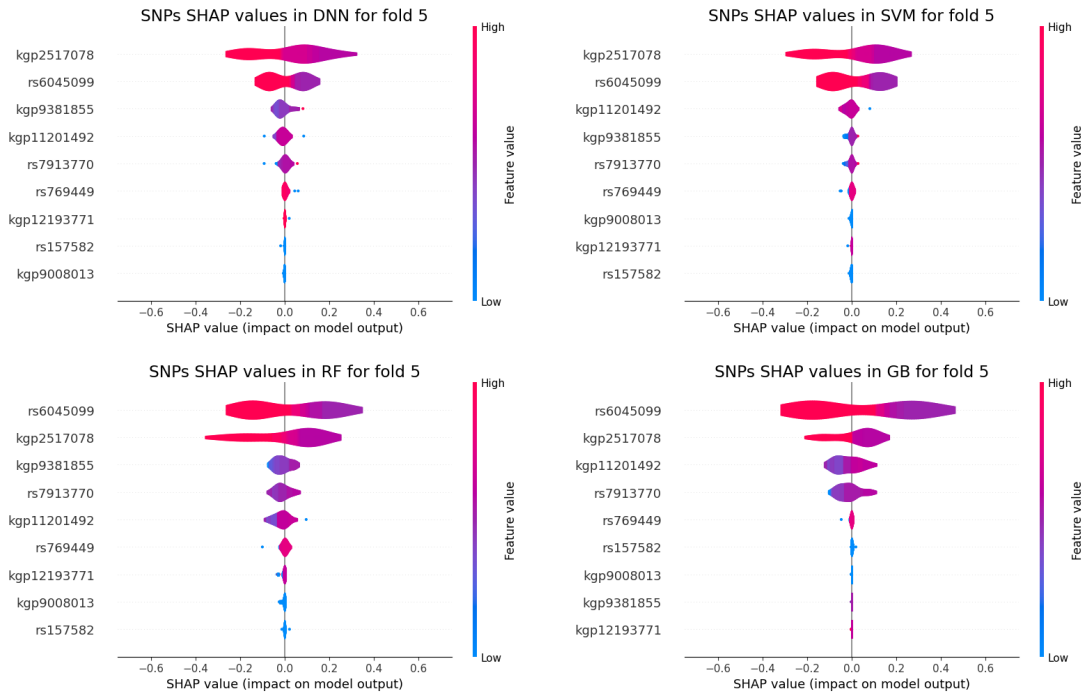

Figura B.8: Valores SHAP fold 5, método de división simple, umbral p-valor sugestivo

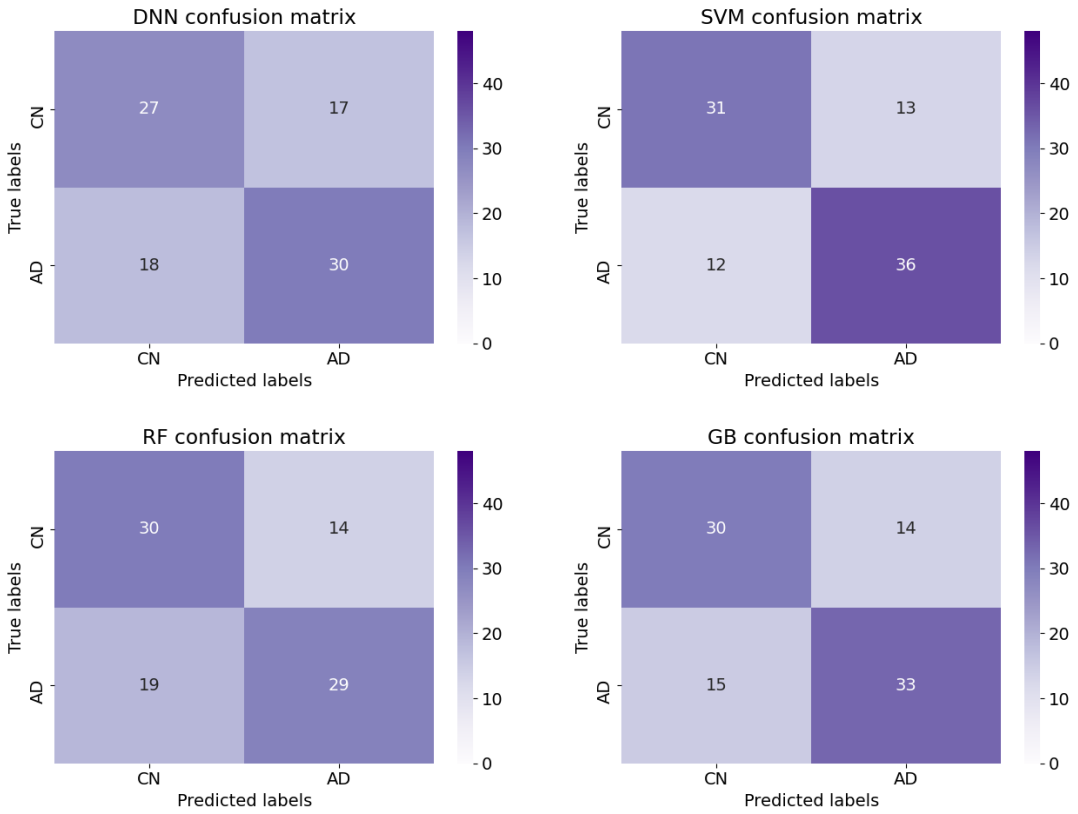

Figura B.9: Matrices de confusión, método de división simple, umbral p-valor sugestivo

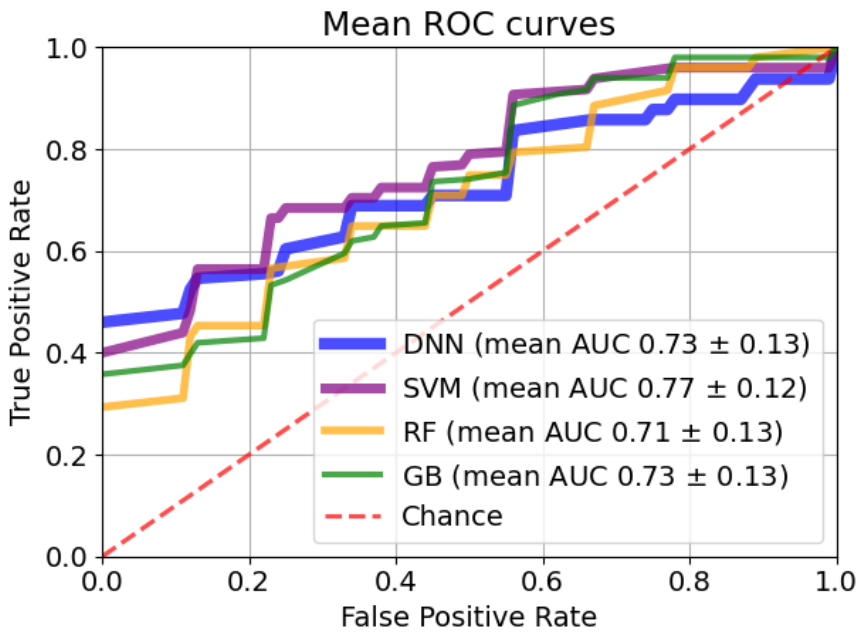

Figura B.10: Curvas ROC, método de división simple, umbral p-valor sugestivo

Umbral relajado

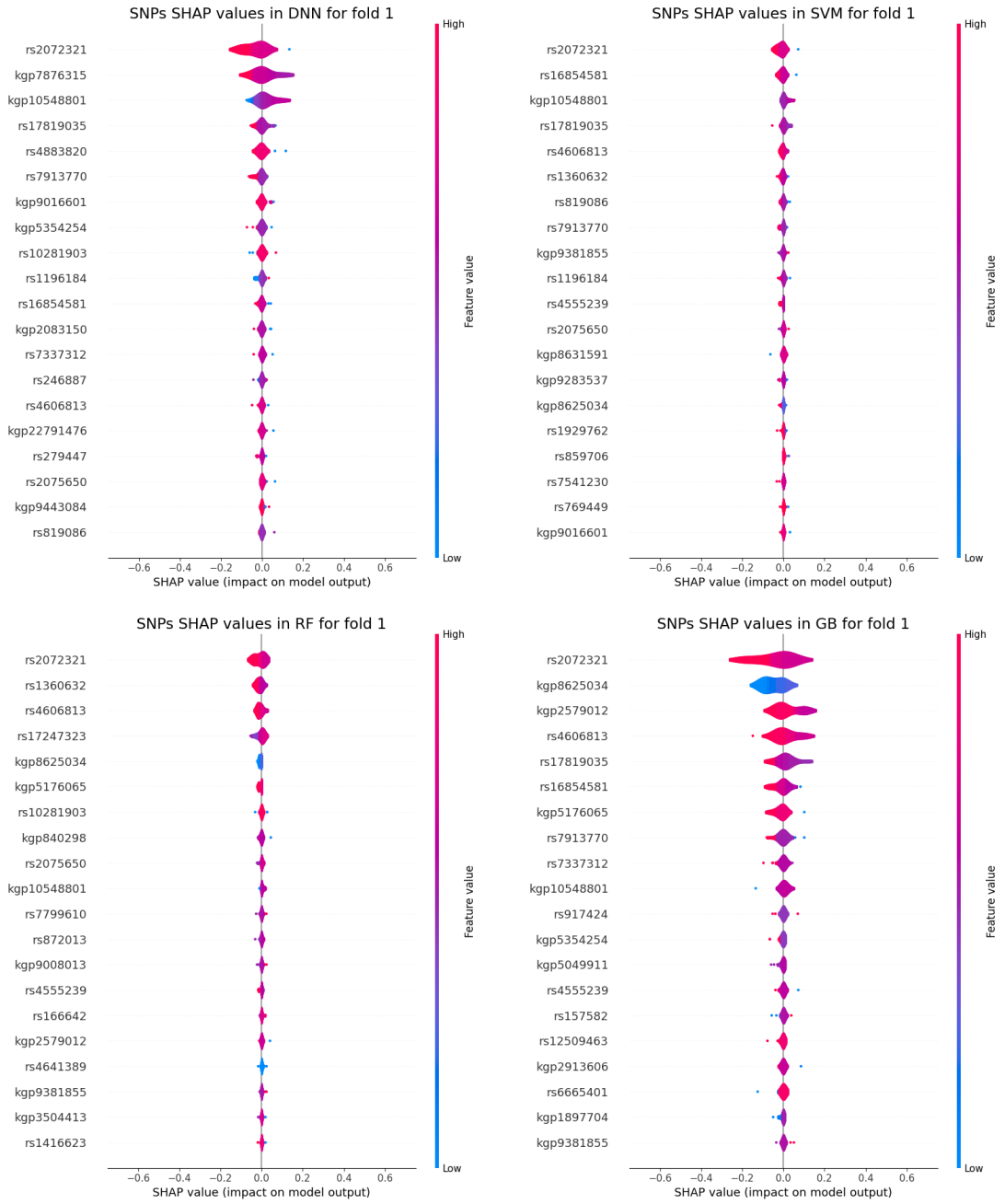

Figura B.11: Valores SHAP fold 1, método de división simple, umbral p-valor relajado

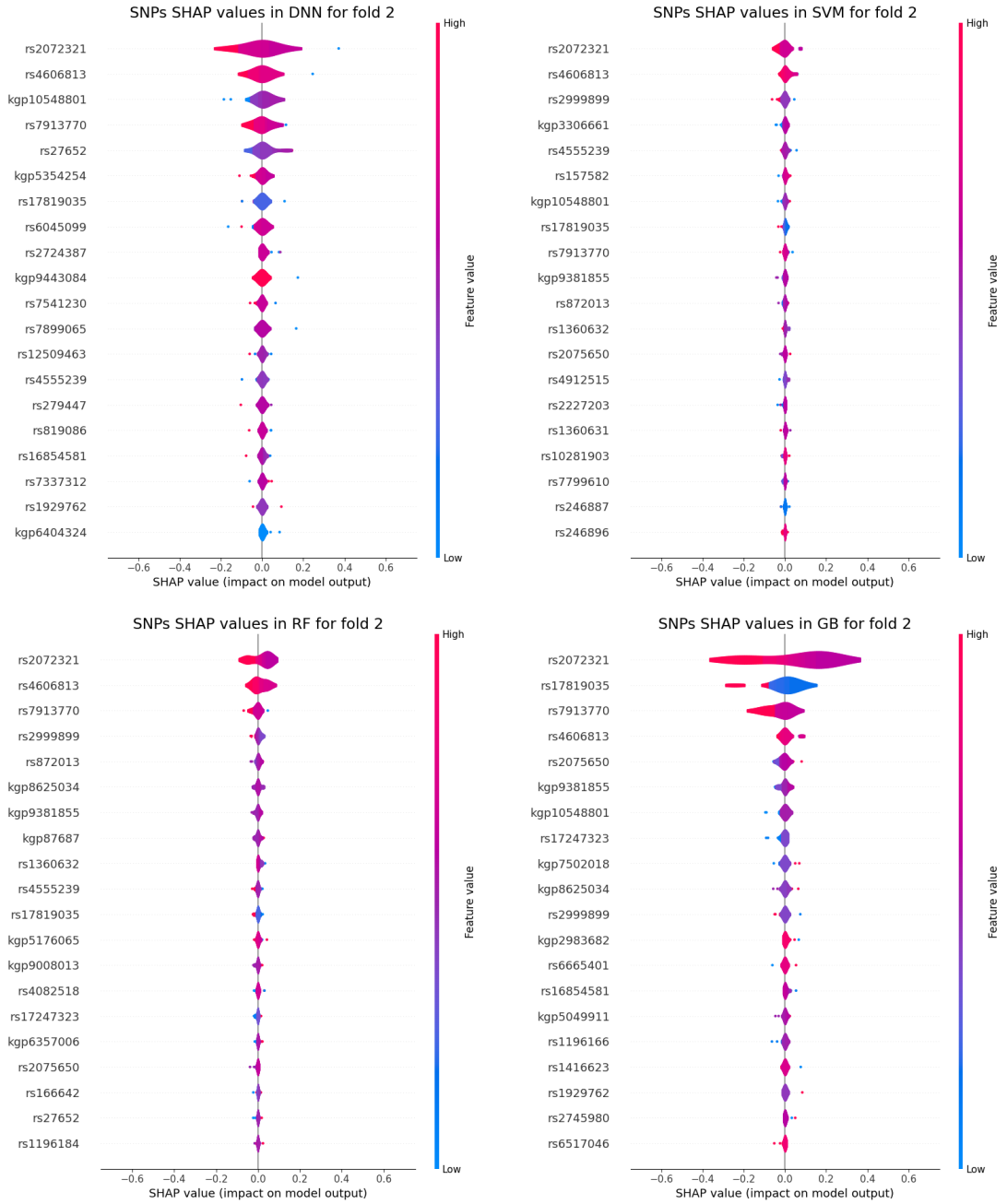

Figura B.12: Valores SHAP fold 2, método de división simple, umbral p-valor relajado

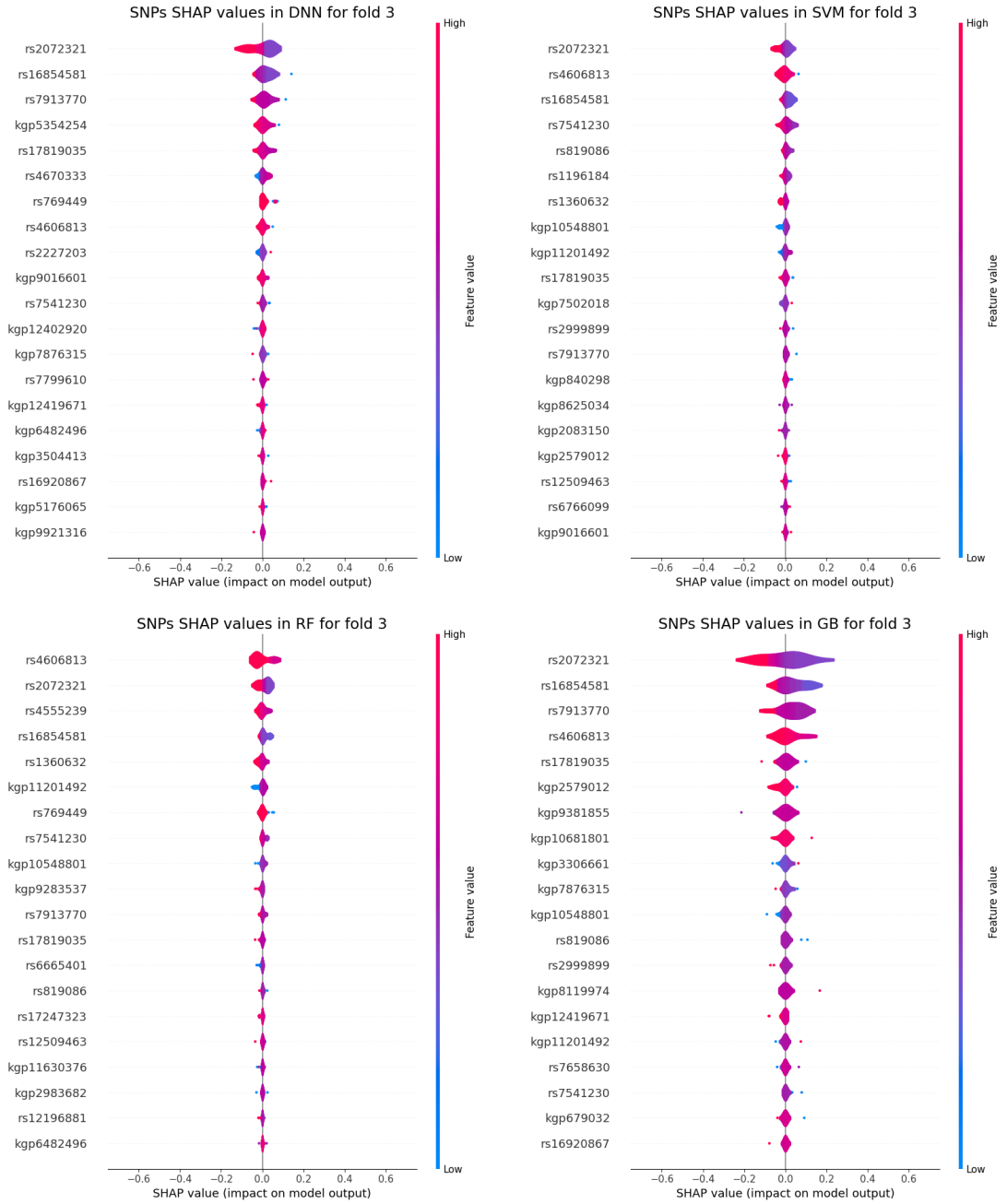

Figura B.13: Valores SHAP fold 3, método de división simple, umbral p-valor relajado

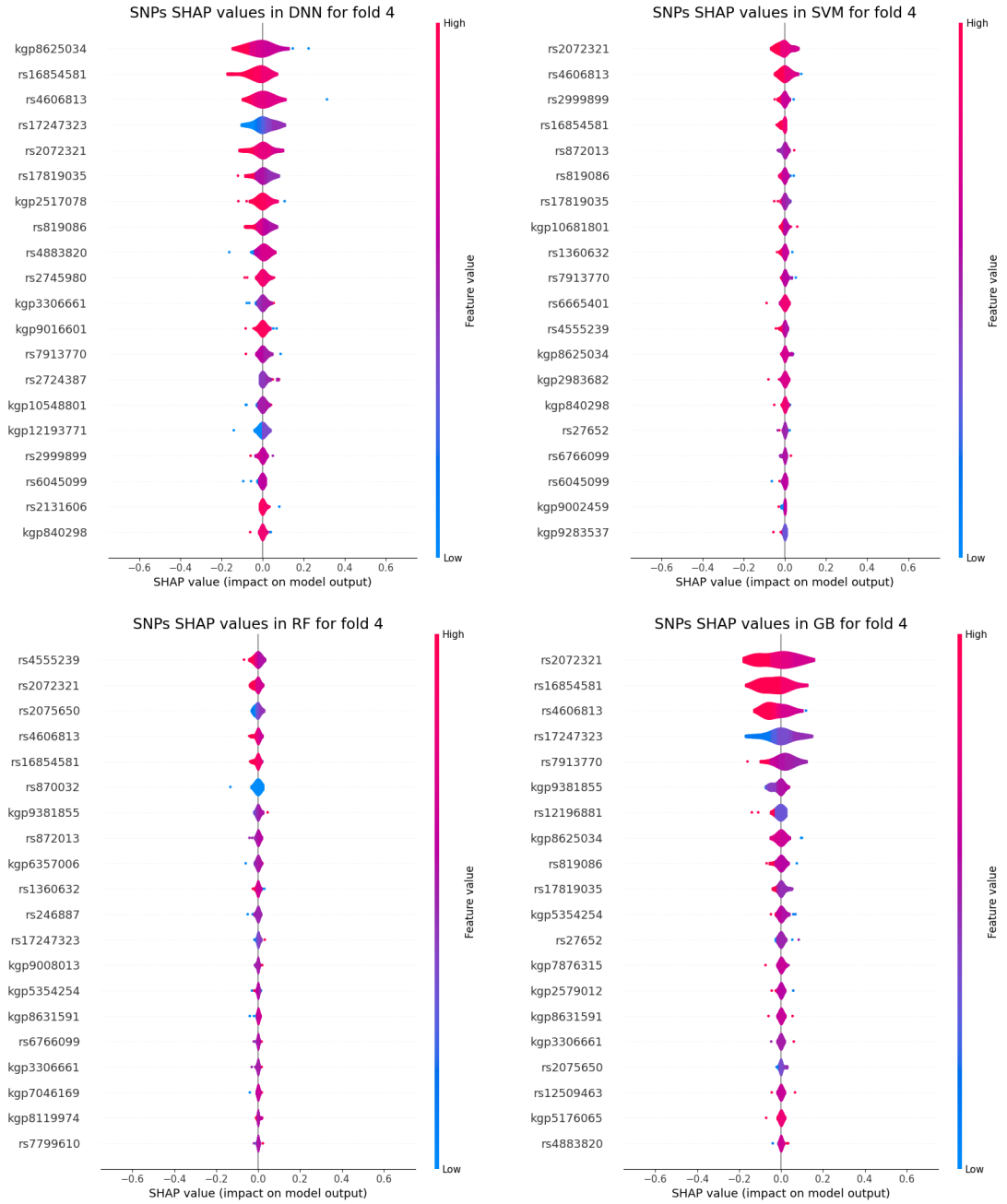

Figura B.14: Valores SHAP fold 4, método de división simple, umbral p-valor relajado

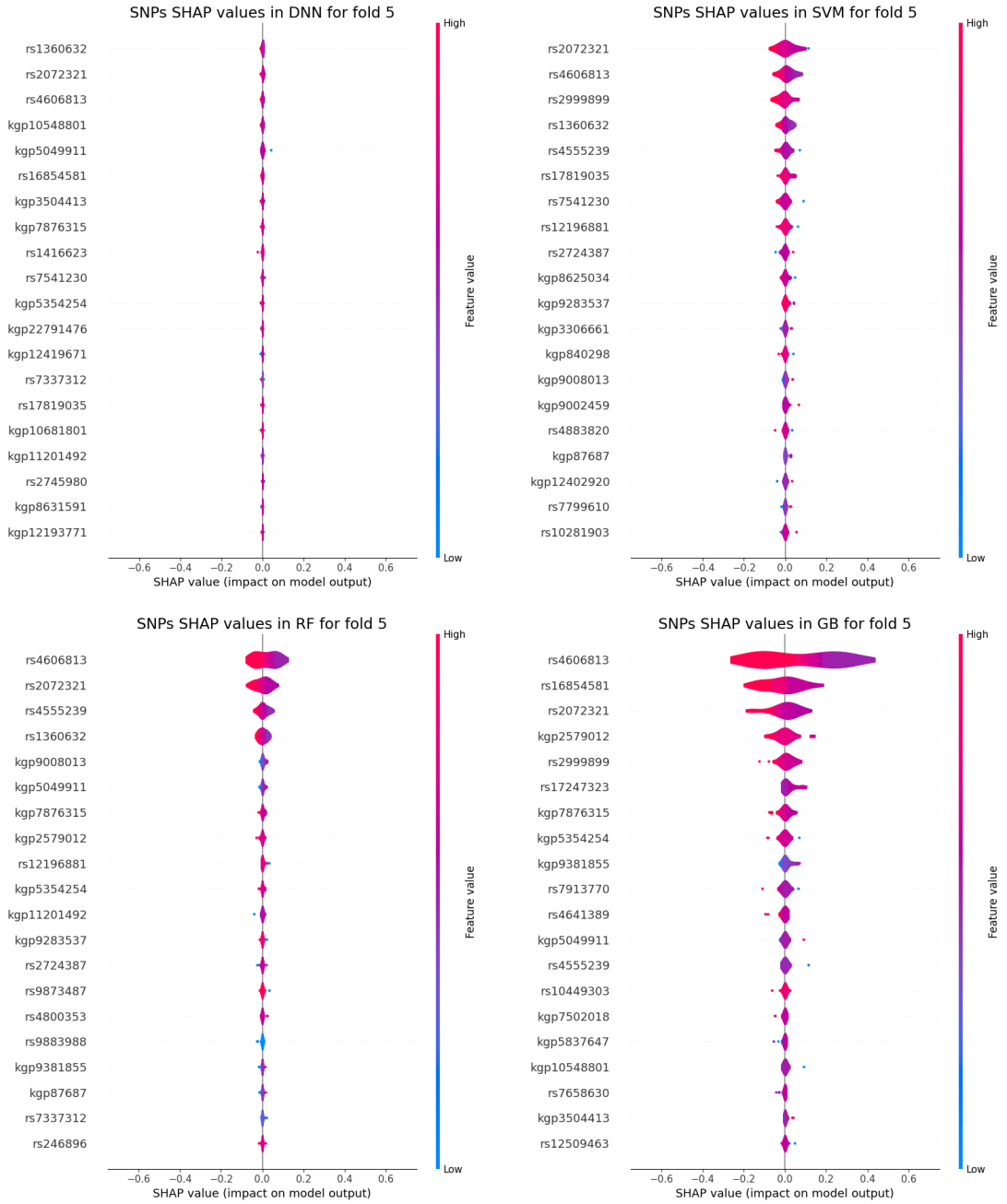

Figura B.15: Valores SHAP fold 5, método de división simple, umbral p-valor relajado

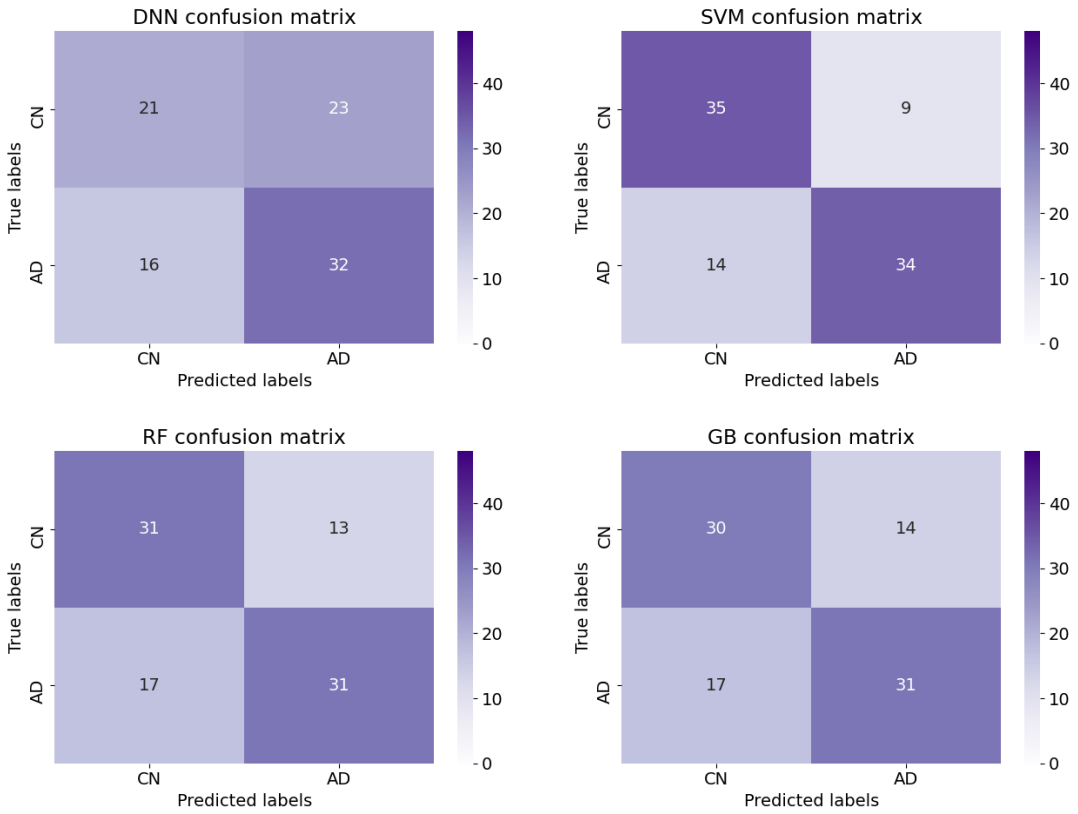

Figura B.16: Matrices de confusión, método de división simple, umbral p-valor relajado

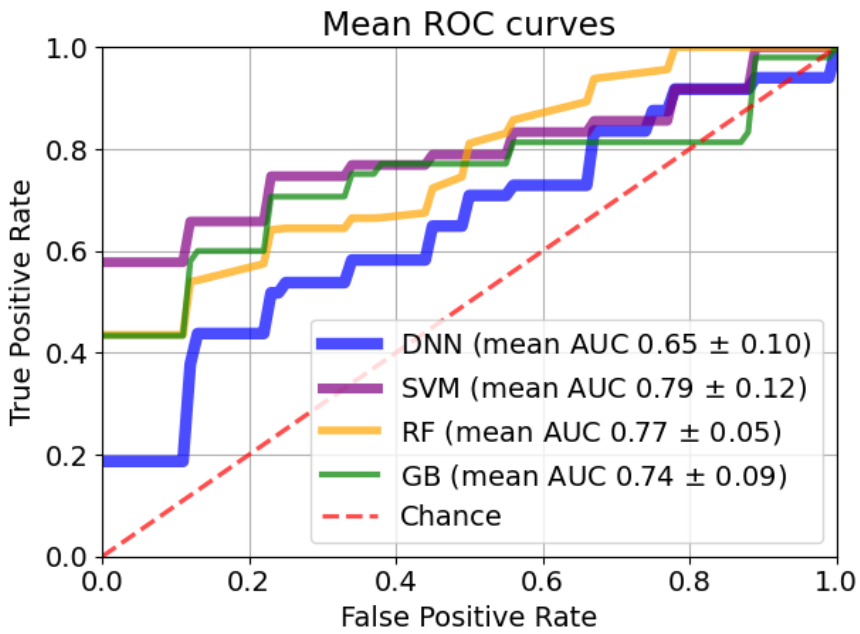

Figura B.17: Curvas ROC, método de división simple, umbral p-valor relajado

B.1.2. Método de división por k-fold

Resultados estudio GWAS (hasta p-valor relajado) para fold 1: rs769449 (8.19e-10), kgp11201492 (3.61e-06), kgp12193771 (4.51e-06), rs157582 (4.60e-06), kgp2517078 (5.08e-06), rs6045099 (6.48e-06), rs7913770 (6.62e-06), kgp9381855 (8.80e-06), kgp9008013 (9.43e-06), rs2075650 (1.10e-05), kgp87687 (1.10e-05), rs7799610 (1.12e-05), rs872013 (1.14e-05), rs17247323 (1.19e-05), rs6766099 (1.34e-05), kgp5049911 (1.74e-05), kgp9042226 (1.79e-05), kgp10548801 (1.83e-05), rs2999899 (1.90e-05), rs870032 (1.96e-05), rs12505461 (2.05e-05), rs10449303 (2.23e-05), rs859706 (2.25e-05), kgp6482496 (2.25e-05), rs16920867 (2.34e-05), kgp11630376 (2.71e-05), kgp9002459 (2.93e-05), rs6517046 (3.00e-05), kgp12394039 (3.17e-05), rs2724387 (3.19e-05), kgp6404324 (3.19e-05), rs1196184 (3.22e-05), rs819086 (3.23e-05), rs4670333 (3.50e-05), kgp1897704 (3.57e-05), rs1929762 (3.73e-05), kgp9443084 (3.78e-05), rs4883820 (3.80e-05), kgp5176065 (3.87e-05), rs246896 (4.04e-05), rs926101 (4.08e-05), kgp9016601 (4.20e-05), rs246887 (4.29e-05), rs12196881 (4.32e-05), kgp9283537 (4.44e-05), kgp8119974 (4.44e-05), kgp5837647 (4.50e-05), rs2745980 (4.50e-05), kgp12402920 (4.54e-05), kgp3377949 (4.60e-05), kgp22791476 (4.62e-05), kgp2983682 (4.62e-05), rs12509463 (4.68e-05), kgp22729362 (4.81e-05), kgp7502018 (4.86e-05), kgp1560994 (4.92e-05), kgp4882314 (4.94e-05), rs4641389 (5.20e-05), kgp12419671 (5.49e-05), rs27652 (5.54e-05), rs1360631 (5.54e-05), rs7899065 (5.64e-05), kgp3504413 (5.78e-05), rs4912515 (5.80e-05), rs2131606 (5.89e-05), kgp8631591 (5.89e-05), kgp7046169 (5.97e-05), rs1416623 (5.99e-05), rs917424 (6.07e-05), rs9883988 (6.25e-05), rs16854581 (6.28e-05), rs1196166 (6.46e-05), kgp9209674 (6.56e-05), rs4800353 (6.56e-05), kgp2083150 (6.58e-05), kgp8625034 (6.60e-05), kgp679032 (6.60e-05), kgp2913606 (6.66e-05), rs7017572 (6.67e-05), kgp840298 (6.73e-05), rs7541230 (6.79e-05), rs7658630 (6.93e-05), rs6665401 (7.02e-05), rs10281903 (7.43e-05), kgp7876315 (7.44e-05), rs279447 (7.44e-05), rs17819035 (7.51e-05), kgp2645931 (7.67e-05), kgp10681801 (7.67e-05), rs11571943 (7.79e-05), rs166642 (7.80e-05), kgp9921316 (7.91e-05), rs4082518 (8.12e-05), rs2227203 (8.41e-05), rs7973722 (8.43e-05), rs4555239 (8.69e-05), rs1360632 (8.73e-05), rs2072321 (8.83e-05), rs4606813 (8.83e-05), rs9873487 (8.86e-05), kgp2579012 (8.89e-05), kgp3306661 (9.18e-05), kgp6357006 (9.36e-05), kgp7187035 (9.45e-05), kgp5354254 (9.45e-05), rs4981220 (9.83e-05), rs7337312 (9.92e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 2: rs769449 (4.06e-11), rs157582 (3.66e-08), rs237473 (7.25e-07), rs2075650 (8.86e-07), rs237463 (1.17e-06), rs2999899 (3.12e-06), rs2256364 (3.40e-06), kgp9381855 (4.95e-06), kgp7502018 (6.89e-06), rs10418480 (7.38e-06), rs10413037 (1.04e-05), kgp5394102 (1.16e-05),

kgp5901917 (1.26e-05), kgp15687332 (1.26e-05), rs10421632 (1.46e-05), kgp3061255 (1.60e-05), kgp22829069 (1.67e-05), kgp9366075 (1.69e-05), kgp5176065 (1.92e-05), rs12515837 (1.95e-05), rs473015 (2.01e-05), kgp9650943 (2.16e-05), kgp3601525 (2.34e-05), rs29413 (2.34e-05), kgp364516 (2.34e-05), kgp3478256 (2.41e-05), kgp10139756 (2.67e-05), rs237468 (3.00e-05), rs237462 (3.00e-05), rs7179777 (3.01e-05), rs2227203 (3.13e-05), rs2782524 (3.25e-05), kgp12138439 (3.27e-05), kgp10439558 (3.27e-05), rs7087608 (3.35e-05), rs4670333 (3.42e-05), rs4604527 (3.46e-05), rs246896 (3.47e-05), rs4672367 (3.87e-05), rs1253679 (3.88e-05), kgp6152748 (3.93e-05), rs27652 (4.04e-05), rs3757737 (4.11e-05), kgp4059590 (4.11e-05), kgp433319 (4.11e-05), kgp945991 (4.17e-05), rs1094506 (4.17e-05), kgp6420136 (4.19e-05), rs432999 (4.30e-05), kgp22791476 (4.59e-05), kgp3658675 (4.64e-05), kgp9406725 (4.83e-05), kgp11360138 (4.83e-05), kgp6318914 (4.83e-05), rs17224674 (4.83e-05), rs740094 (4.83e-05), kgp4159738 (4.84e-05), kgp3791913 (4.85e-05), rs16920867 (4.94e-05), kgp5572870 (4.94e-05), rs572845 (4.94e-05), rs2699316 (5.11e-05), kgp3175471 (5.16e-05), rs707645 (5.20e-05), rs2588252 (5.26e-05), rs11660396 (5.27e-05), kgp5872801 (5.42e-05), kgp9481735 (5.49e-05), kgp3634366 (5.85e-05), rs17030368 (5.89e-05), kgp3897901 (5.99e-05), rs11249170 (6.17e-05), kgp11278477 (6.25e-05), rs237454 (6.34e-05), kgp5161509 (6.46e-05), rs1196184 (6.46e-05), kgp3251588 (6.56e-05), rs7311350 (6.60e-05), rs7692432 (6.67e-05), kgp5170780 (6.69e-05), kgp4786128 (6.70e-05), kgp10509397 (6.89e-05), kgp8304730 (6.92e-05), kgp4473224 (6.94e-05), rs4796566 (6.94e-05), rs6990507 (6.94e-05), kgp4964336 (6.94e-05), kgp7857731 (7.02e-05), rs4133345 (7.16e-05), kgp6810456 (7.24e-05), kgp5129569 (7.79e-05), kgp1426673 (7.79e-05), rs522089 (7.83e-05), kgp3627688 (7.93e-05), kgp11289786 (8.02e-05), kgp2983682 (8.10e-05), kgp1135012 (8.10e-05), rs7838972 (8.42e-05), kgp6433360 (8.43e-05), kgp12487058 (8.50e-05), kgp11680544 (8.53e-05), rs1155463 (8.79e-05), kgp757360 (8.89e-05), kgp1873920 (9.17e-05), rs10852888 (9.30e-05), rs7242356 (9.30e-05), rs9948496 (9.30e-05), kgp1255352 (9.50e-05), rs2762637 (9.56e-05), kgp8331440 (9.59e-05), rs772492 (9.59e-05), kgp5825387 (9.83e-05), kgp1521332 (9.83e-05), rs1958979 (9.95e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 3: rs769449 (2.18e-11), rs157582 (3.50e-09), rs2075650 (4.53e-07), kgp22791476 (7.12e-07), kgp7502018 (1.33e-06), kgp1106764 (5.43e-06), kgp87687 (7.62e-06), rs2625312 (7.87e-06), kgp9129240 (8.66e-06), rs4672367 (1.10e-05), rs4970834 (1.18e-05), rs1012146 (1.29e-05), kgp3434986 (1.75e-05), kgp3306661 (1.79e-05), kgp6818987 (1.92e-05), kgp6566063 (1.98e-05), kgp3791913 (2.00e-05), rs6710432 (2.03e-05), rs473015 (2.04e-05), kgp6629063 (2.06e-05), kgp10260701 (2.10e-05), kgp6413075 (2.13e-05), rs9838305 (2.28e-05), kgp8625034

(2.30e-05), rs1958979 (2.36e-05), kgp5142166 (2.63e-05), kgp4350374 (3.03e-05), rs608039 (3.29e-05), kgp8108355 (3.52e-05), kgp7296220 (3.52e-05), rs646776 (3.52e-05), kgp7857731 (3.59e-05), rs10810541 (3.75e-05), kgp7219319 (3.81e-05), kgp10620084 (3.89e-05), rs17110902 (3.89e-05), rs2588252 (3.91e-05), kgp10998429 (4.34e-05), rs3171927 (4.42e-05), kgp12087452 (4.49e-05), rs7545471 (4.60e-05), rs27986 (4.75e-05), kgp2483143 (4.94e-05), kgp9278422 (5.00e-05), kgp8501386 (5.00e-05), rs11249170 (5.16e-05), rs3731935 (5.54e-05), kgp8932683 (5.55e-05), kgp11766293 (5.55e-05), rs9823723 (5.55e-05), rs932207 (5.55e-05), kgp7054990 (5.55e-05), kgp22747896 (5.72e-05), kgp329470 (5.72e-05), kgp2517078 (6.03e-05), rs1906492 (6.26e-05), kgp11278477 (6.46e-05), rs551141 (6.46e-05), kgp8918725 (6.46e-05), kgp11239964 (6.46e-05), kgp4179269 (6.82e-05), rs660240 (6.84e-05), kgp307686 (6.84e-05), rs1363785 (6.84e-05), kgp12193771 (6.91e-05), kgp11484611 (6.92e-05), kgp6813909 (7.18e-05), rs17420101 (7.36e-05), rs17479466 (7.36e-05), kgp1362893 (7.39e-05), rs10413037 (7.42e-05), kgp1884643 (7.51e-05), rs246896 (7.56e-05), rs27652 (7.56e-05), rs3214058 (7.56e-05), kgp12237618 (7.66e-05), kgp9081044 (8.45e-05), rs2278563 (8.50e-05), kgp2995630 (8.89e-05), rs230844 (8.90e-05), kgp11206857 (8.90e-05), rs7913770 (8.96e-05), rs6766099 (9.82e-05), rs9912576 (9.83e-05), rs2256111 (9.92e-05), rs17247323 (9.94e-05), kgp2025242 (9.95e-05), rs12522444 (9.95e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 4: rs769449 (6.66e-14), rs157582 (2.63e-08), rs2075650 (1.71e-07), kgp7502018 (9.95e-07), rs6045099 (1.74e-06), kgp2517078 (1.79e-06), kgp12237618 (2.35e-06), kgp124656 (3.00e-06), kgp3556207 (4.65e-06), kgp5230206 (4.89e-06), rs594165 (5.23e-06), kgp5354254 (6.77e-06), rs1943440 (6.86e-06), rs6766099 (7.29e-06), kgp11206857 (9.78e-06), rs2588252 (1.46e-05), rs4752300 (1.65e-05), kgp6927229 (1.68e-05), rs11082405 (1.86e-05), kgp2568973 (1.90e-05), kgp87687 (2.15e-05), kgp2951395 (2.29e-05), rs4687827 (2.42e-05), rs1974953 (2.75e-05), kgp12571859 (2.79e-05), rs4970834 (3.08e-05), rs6758001 (3.17e-05), kgp12295062 (3.27e-05), kgp12193771 (3.32e-05), kgp7775481 (3.74e-05), kgp8959836 (3.87e-05), kgp1331883 (4.00e-05), rs803764 (4.19e-05), kgp466423 (4.19e-05), kgp5581882 (4.19e-05), kgp9148496 (4.19e-05), rs4623810 (4.29e-05), kgp7949809 (4.85e-05), rs4672367 (4.95e-05), rs4670333 (5.07e-05), rs16854581 (5.22e-05), kgp7244105 (5.36e-05), kgp11917342 (5.48e-05), rs4823717 (5.56e-05), rs9317985 (5.66e-05), kgp2558556 (5.80e-05), kgp9827807 (5.82e-05), kgp703792 (5.94e-05), rs627873 (6.34e-05), rs6665401 (6.34e-05), kgp8640098 (7.30e-05), rs1423066 (7.38e-05), rs9299396 (7.44e-05), rs4884931 (7.44e-05), rs2684311 (7.70e-05), kgp10846942 (7.74e-05), rs11999426 (7.74e-05), kgp3910028 (7.83e-05), rs608527 (7.85e-05), kgp11358179 (7.93e-05), kgp5253828 (7.99e-05), kgp6411744

(8.05e-05), rs9810250 (8.56e-05), rs2618648 (8.72e-05), kgp7553176 (8.73e-05), rs2304678 (8.91e-05), rs1571907 (8.97e-05), kgp7555974 (9.06e-05), kgp5008088 (9.06e-05), rs10884402 (9.08e-05), kgp8598717 (9.20e-05), kgp3478256 (9.30e-05), kgp1455176 (9.59e-05), rs798577 (9.99e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 5: rs769449 (4.07e-10), rs157582 (6.27e-08), rs6723770 (5.30e-07), kgp2913606 (2.60e-06), kgp134363 (3.72e-06), kgp10439558 (3.76e-06), rs2075650 (3.83e-06), rs4672367 (4.75e-06), kgp9811805 (4.80e-06), kgp2517078 (5.58e-06), rs8061054 (5.77e-06), kgp9057945 (6.09e-06), kgp11623754 (6.38e-06), kgp7502018 (6.84e-06), kgp1420003 (7.60e-06), rs10505904 (8.93e-06), rs2999899 (9.94e-06), kgp1562249 (1.04e-05), rs11249170 (1.06e-05), rs16922059 (1.10e-05), rs3797655 (1.51e-05), kgp9917186 (1.73e-05), kgp5310431 (1.73e-05), rs7545471 (1.75e-05), kgp1220081 (1.79e-05), rs306503 (1.91e-05), rs9838305 (2.00e-05), kgp8928990 (2.00e-05), rs4867289 (2.08e-05), kgp11484611 (2.10e-05), rs4702595 (2.19e-05), rs2739430 (2.49e-05), kgp11357236 (2.49e-05), rs4670333 (2.49e-05), kgp22791476 (2.68e-05), kgp9381855 (2.68e-05), rs2101171 (2.79e-05), rs7702790 (2.84e-05), rs6045099 (3.00e-05), rs17603912 (3.13e-05), kgp1832421 (3.17e-05), kgp9573706 (3.23e-05), kgp6918617 (3.23e-05), rs4604527 (3.58e-05), kgp6598539 (3.67e-05), kgp9886473 (3.91e-05), kgp5176065 (3.91e-05), rs10174698 (4.03e-05), kgp11201492 (4.05e-05), kgp4002280 (4.10e-05), rs2239065 (4.13e-05), rs7727470 (4.28e-05), rs6766099 (4.31e-05), kgp8331440 (4.38e-05), kgp5060005 (4.62e-05), rs6537822 (4.64e-05), rs2748324 (4.84e-05), rs1468309 (4.85e-05), rs12030629 (5.12e-05), rs11589933 (5.12e-05), rs896199 (5.12e-05), rs12459777 (5.12e-05), rs536715 (5.28e-05), rs4670335 (5.32e-05), rs2371923 (5.42e-05), rs7725455 (5.50e-05), kgp5872801 (5.64e-05), rs2719889 (5.87e-05), kgp7246956 (5.90e-05), kgp7925672 (5.90e-05), rs1559331 (5.90e-05), rs2816881 (6.06e-05), kgp7238095 (6.32e-05), kgp5353951 (6.32e-05), kgp8765231 (6.32e-05), kgp11598172 (6.32e-05), rs7299801 (6.47e-05), kgp5967232 (6.54e-05), rs12312266 (6.75e-05), rs2117900 (7.03e-05), kgp8490417 (7.03e-05), kgp9715015 (7.03e-05), rs11660396 (7.31e-05), rs436760 (7.55e-05), kgp1378739 (7.59e-05), kgp2320463 (7.75e-05), kgp4474389 (7.78e-05), rs10928585 (7.88e-05), kgp2668302 (7.95e-05), kgp6912392 (8.11e-05), rs17475331 (8.18e-05), kgp5354254 (8.20e-05), kgp3493035 (8.20e-05), kgp2483143 (8.20e-05), rs10922744 (8.36e-05), kgp3668512 (8.42e-05), rs3904537 (8.42e-05), rs1874797 (8.42e-05), kgp1016632 (8.49e-05), kgp2995630 (8.75e-05), rs2290044 (8.76e-05), rs169822 (8.79e-05), rs2099380 (8.86e-05), rs12297204 (9.02e-05), rs3818337 (9.09e-05), rs17631987 (9.10e-05), rs11044136 (9.39e-05), kgp8184155 (9.40e-05), kgp519069 (9.42e-05), rs9912659 (9.48e-05), kgp11437041 (9.60e-05), kgp9366075 (9.62e-05), rs1160769 (9.90e-05), kgp111318 (9.90e-05), rs1368056

(9.94e-05), rs13406300 (9.94e-05), rs1932713 (9.94e-05).

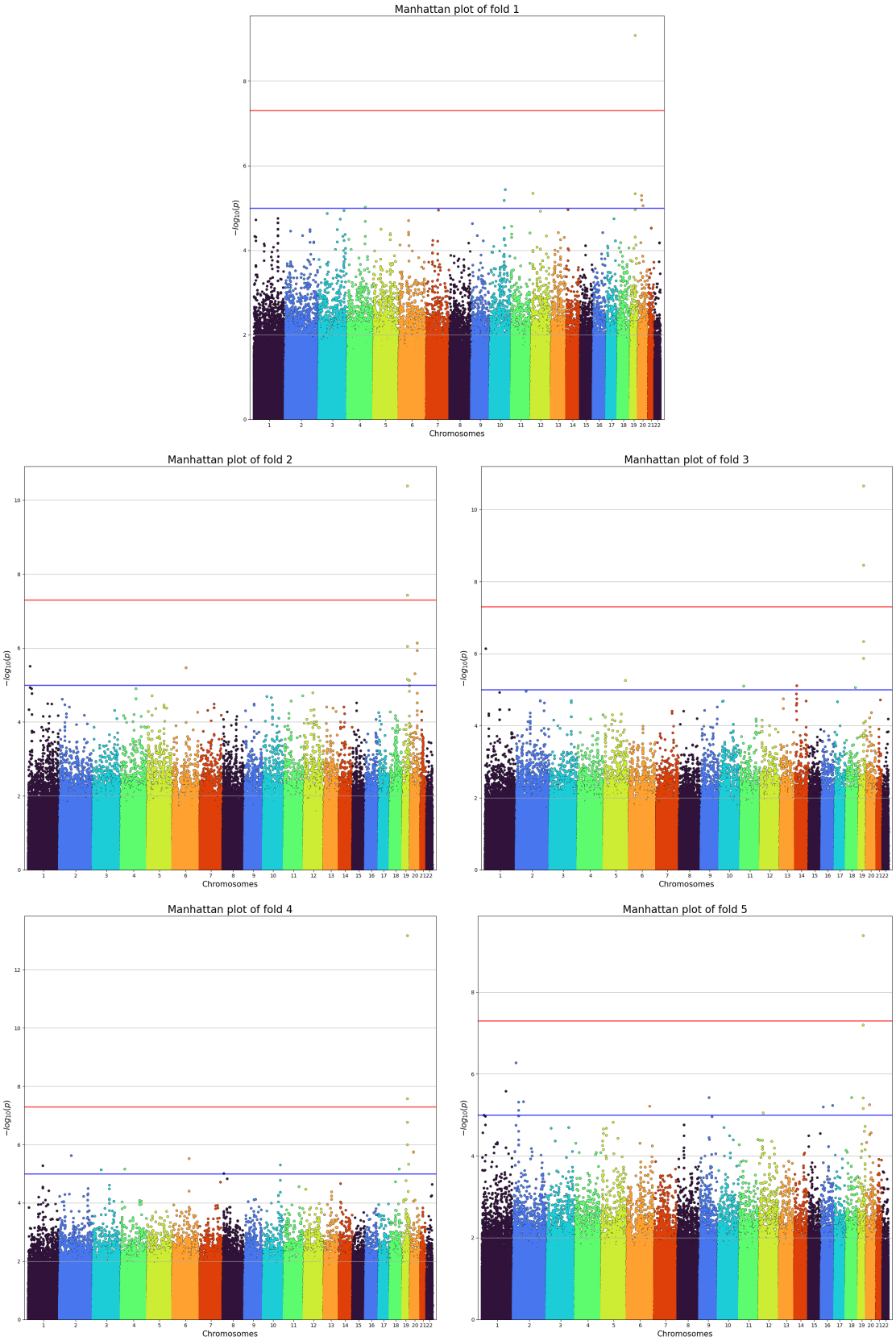

Figura B.18: Graficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip

Umbral clásico

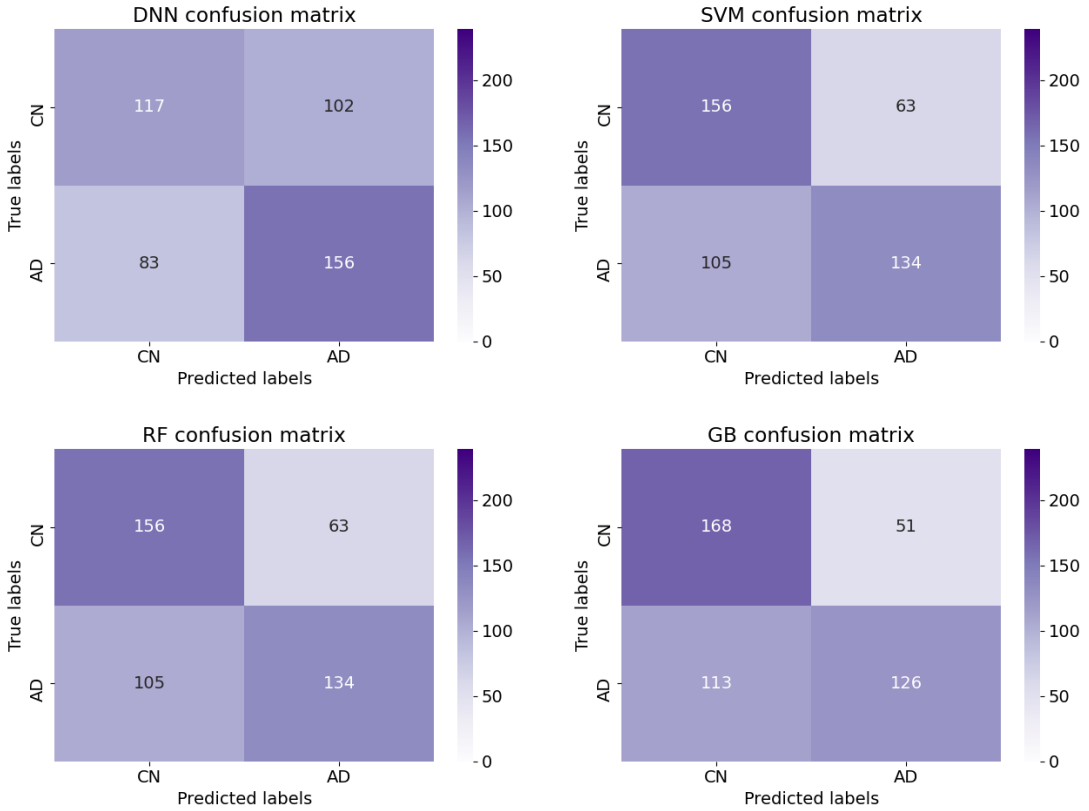

Figura B.19: Matrices de confusión, método de división por k-fold, umbral p-valor tradicional

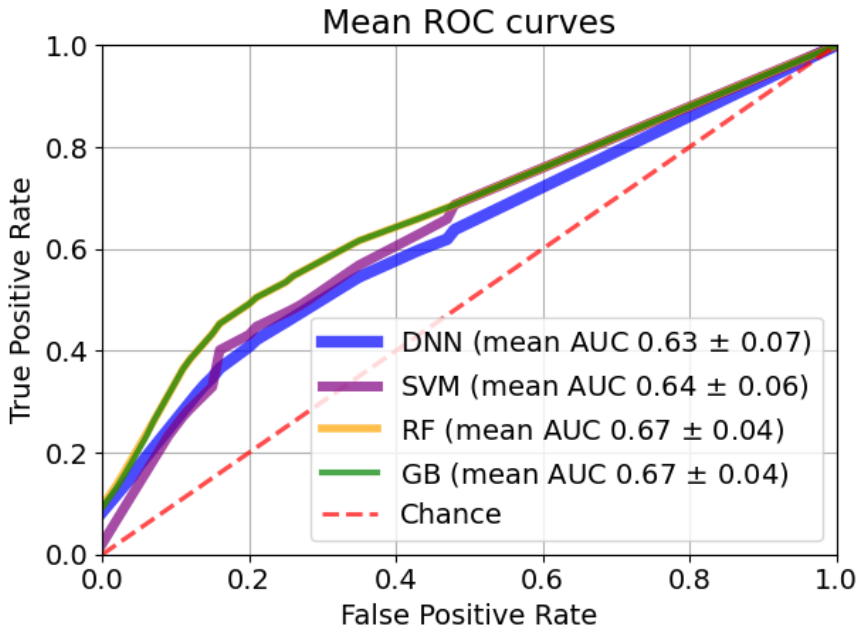

Figura B.20: Curvas ROC, método de división por k-fold, umbral p-valor tradicional

Umbral sugestivo

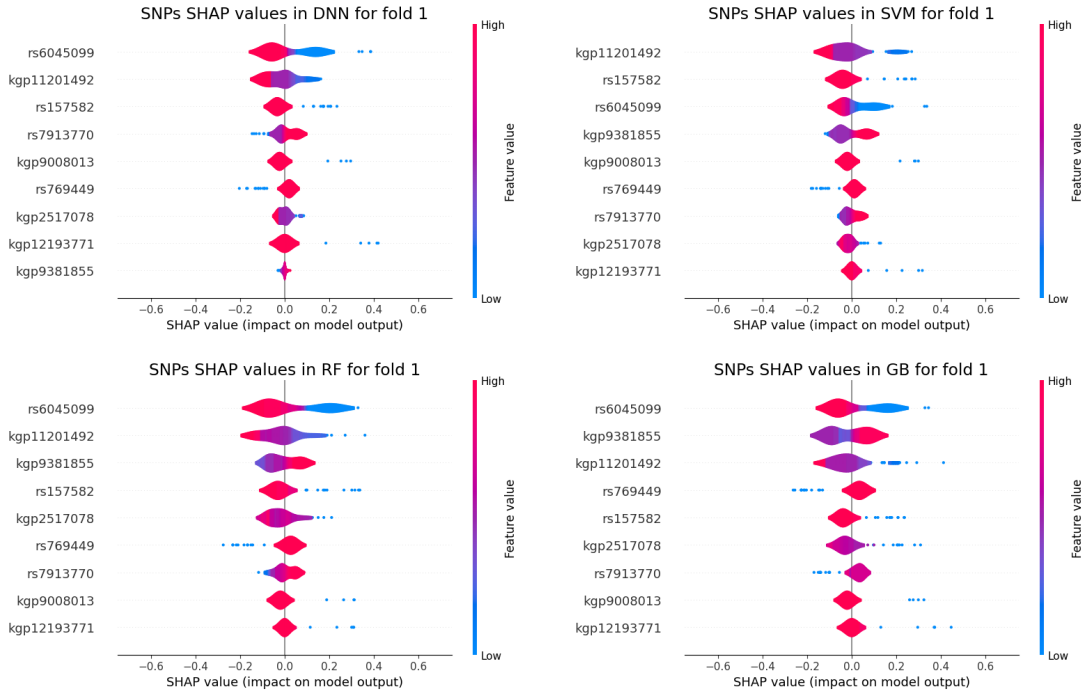

Figura B.21: Valores SHAP fold 1, método de división por k-fold, umbral p-valor sugestivo

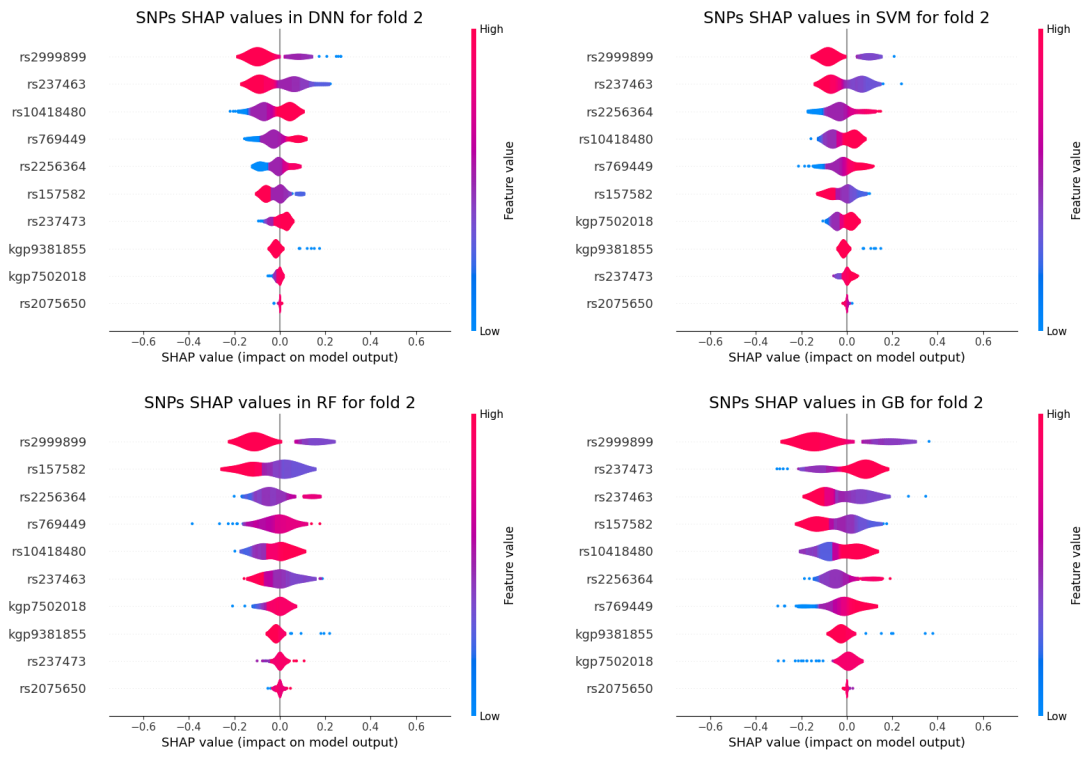

Figura B.22: Valores SHAP fold 2, método de división por k-fold, umbral p-valor sugestivo

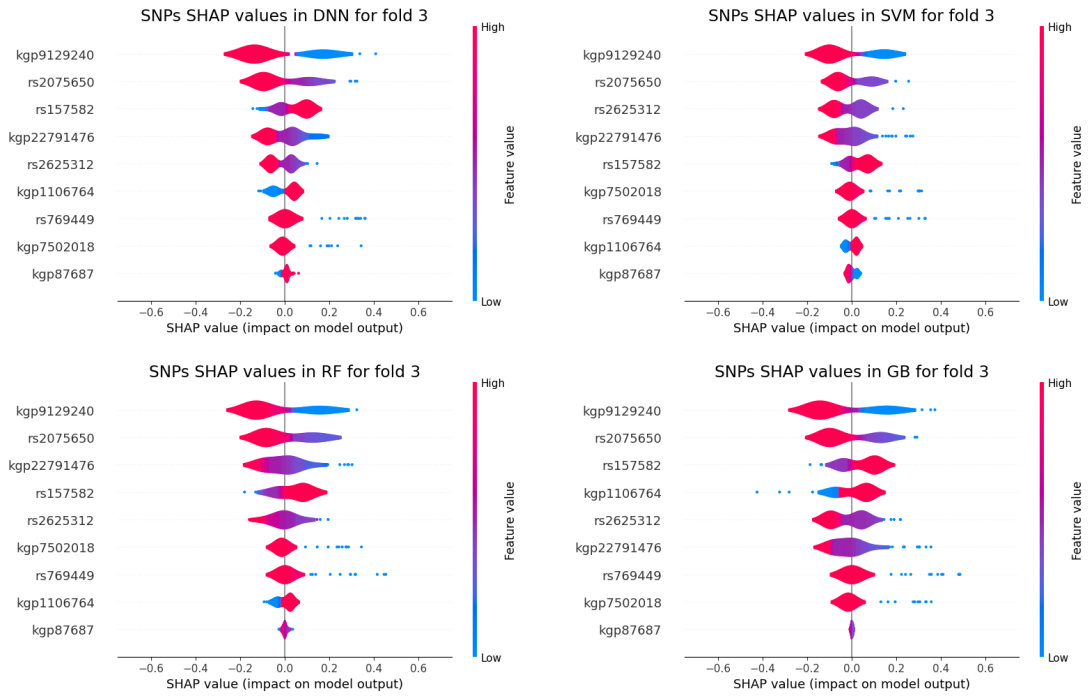

Figura B.23: Valores SHAP fold 3, método de división por k-fold, umbral p-valor sugestivo

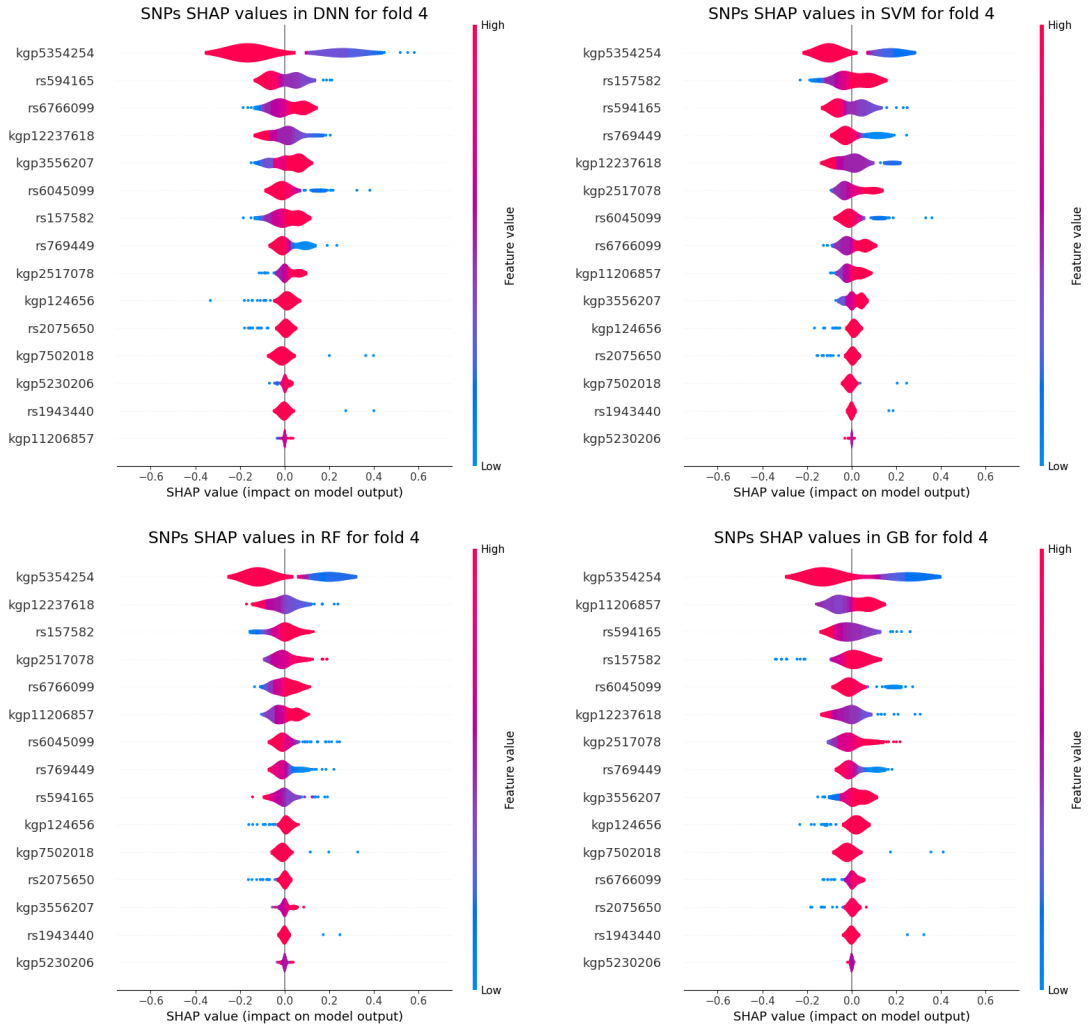

Figura B.24: Valores SHAP fold 4, método de división por k-fold, umbral p-valor sugestivo

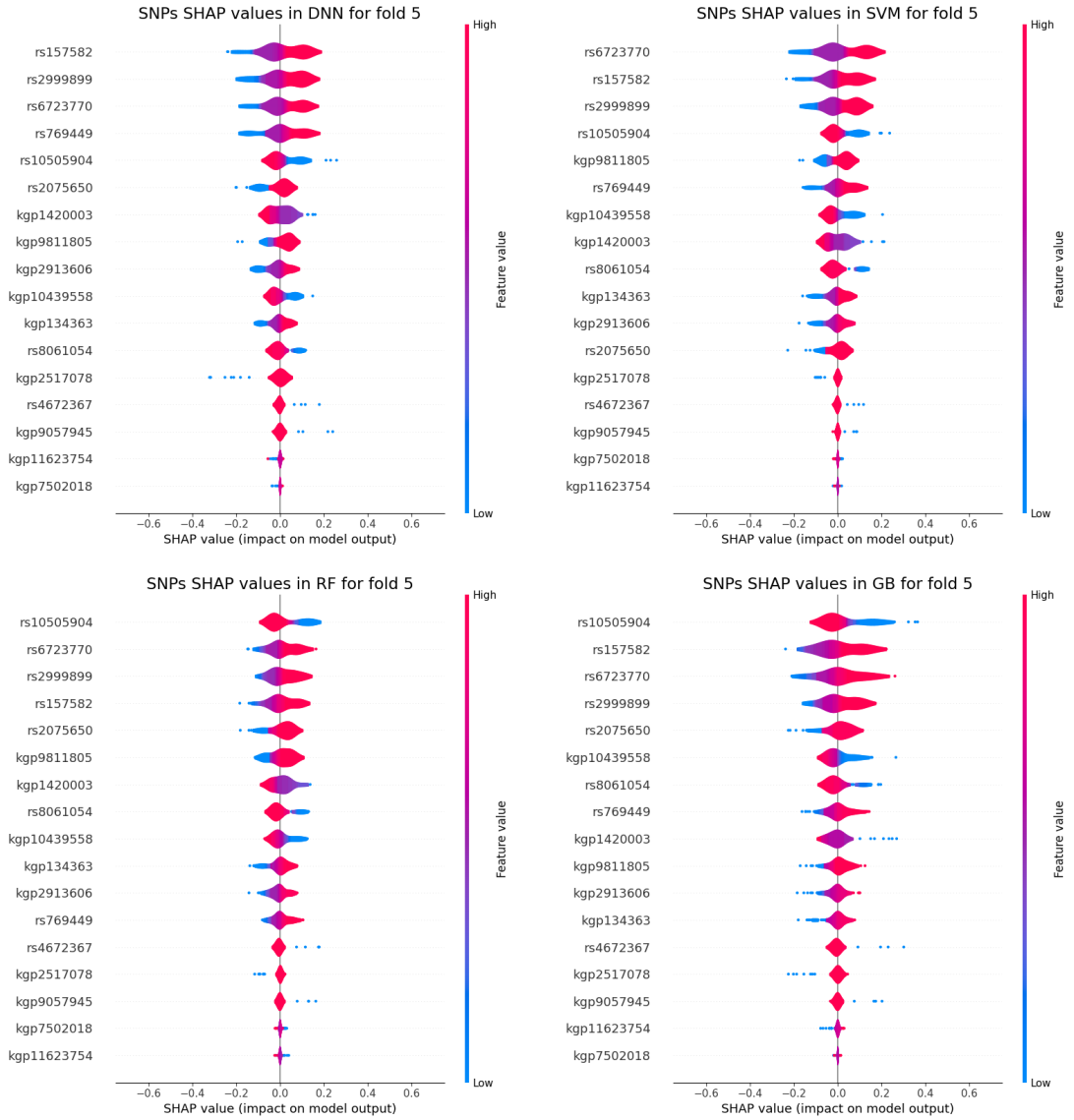

Figura B.25: Valores SHAP fold 5, método de división por k-fold, umbral p-valor sugestivo

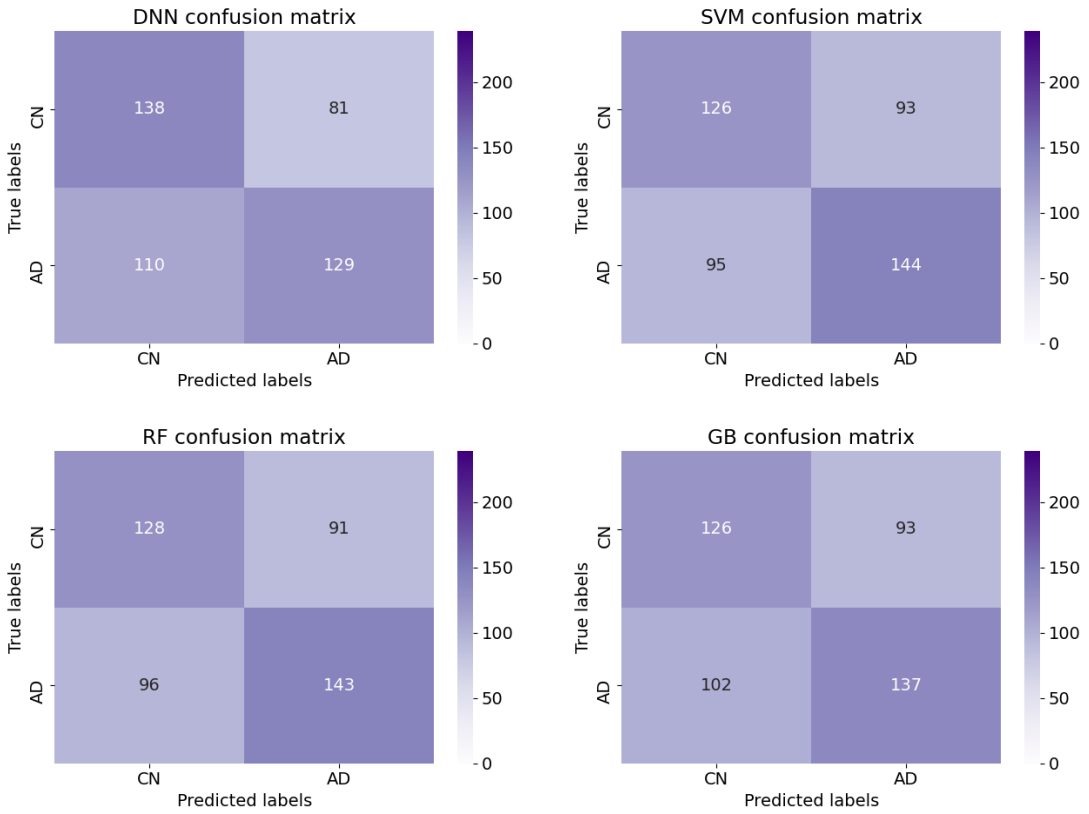

Figura B.26: Matrices de confusión, método de división por k-fold, umbral p-valor sugestivo

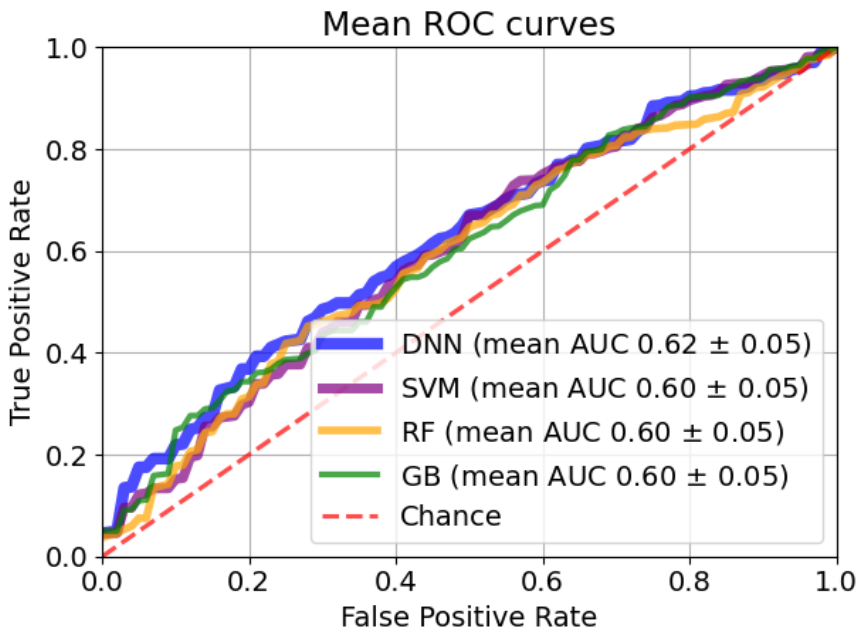

Figura B.27: Curvas ROC, método de división por k-fold, umbral p-valor sugestivo

Umbral relajado

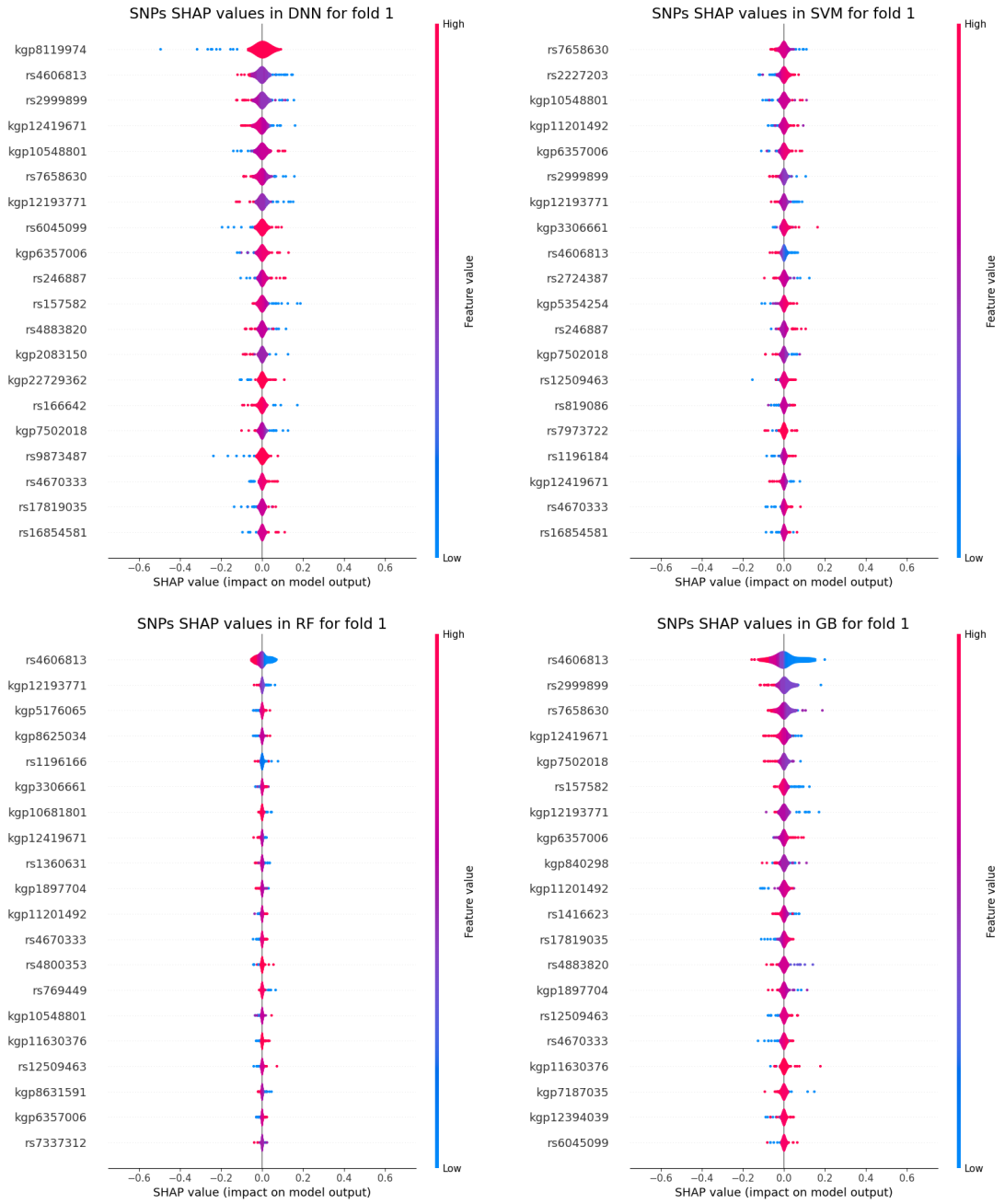

Figura B.28: Valores SHAP fold 1, método de división por k fold, umbral p-valor relajado

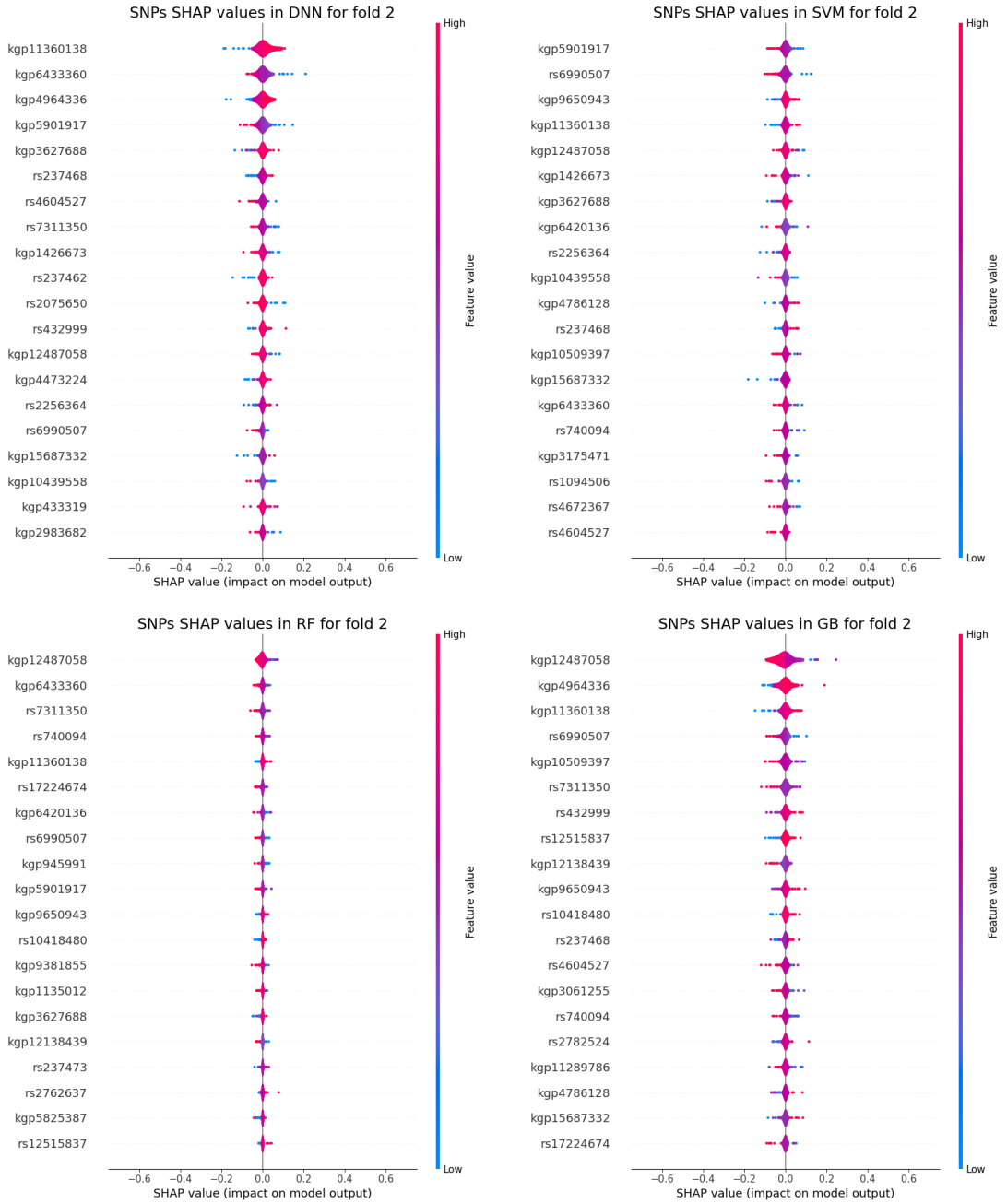

Figura B.29: Valores SHAP fold 2, método de división por k fold, umbral p-valor relajado

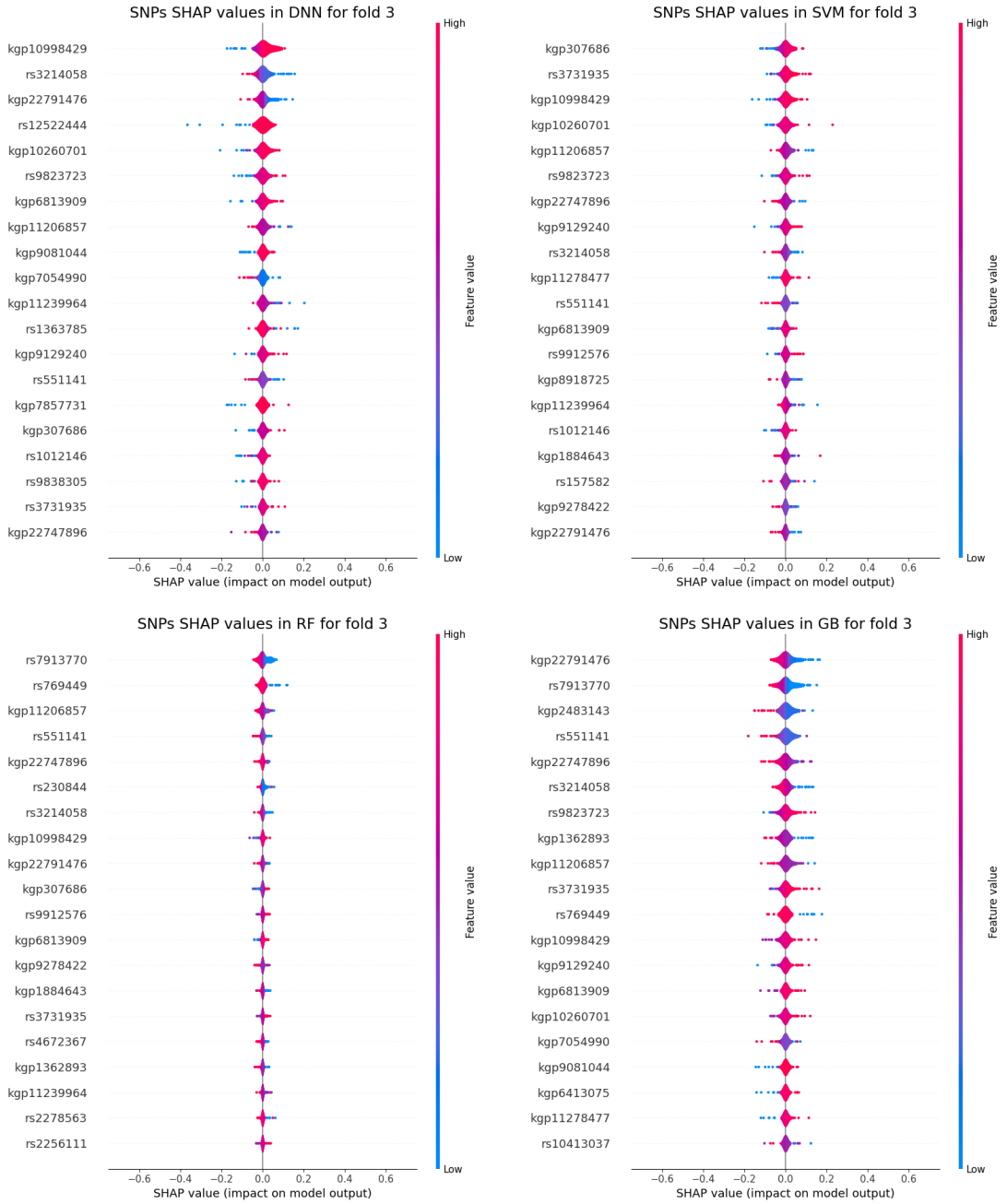

Figura B.30: Valores SHAP fold 3, método de división por k fold, umbral p-valor relajado

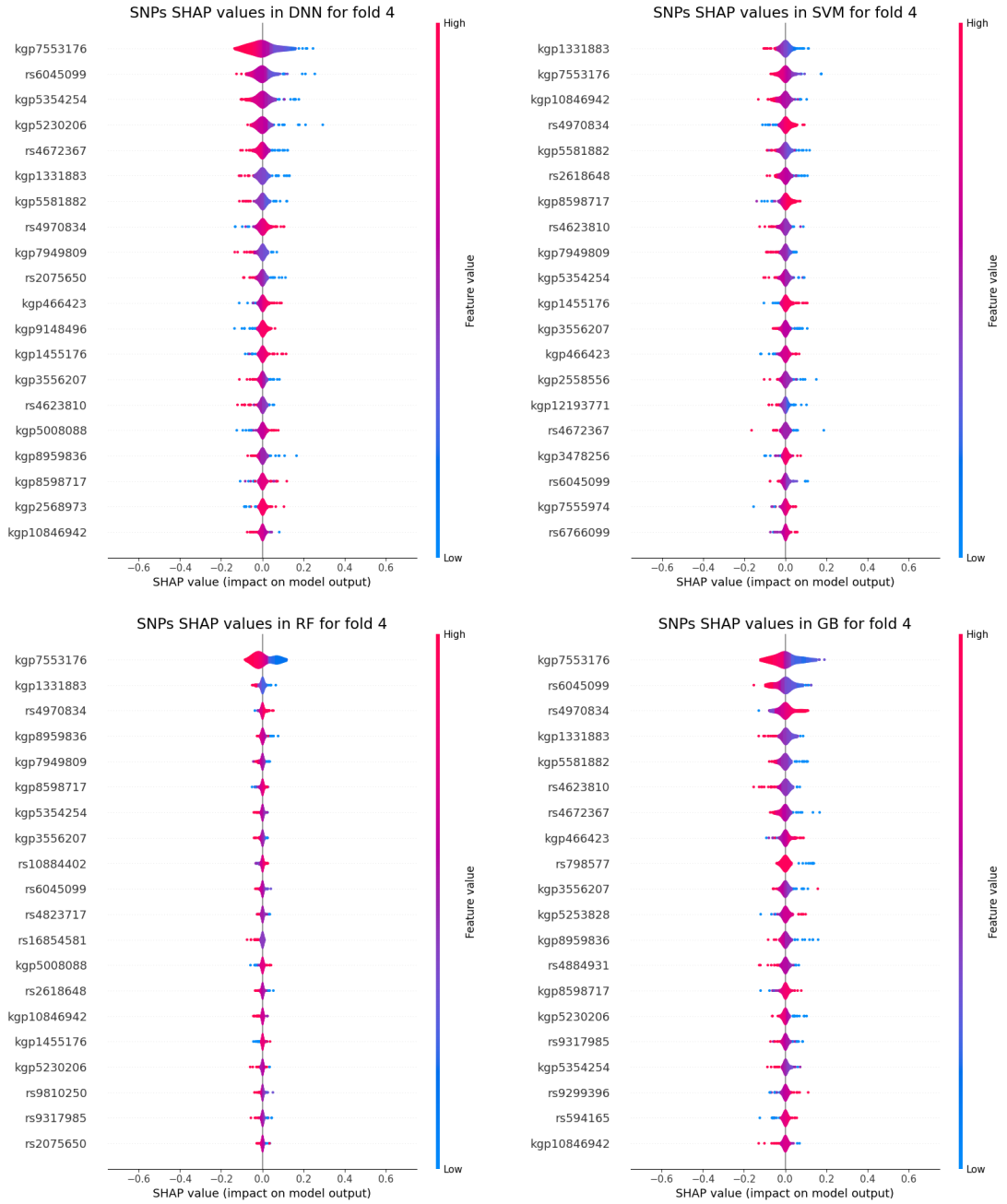

Figura B.31: Valores SHAP fold 4, método de división por k fold, umbral p-valor relajado

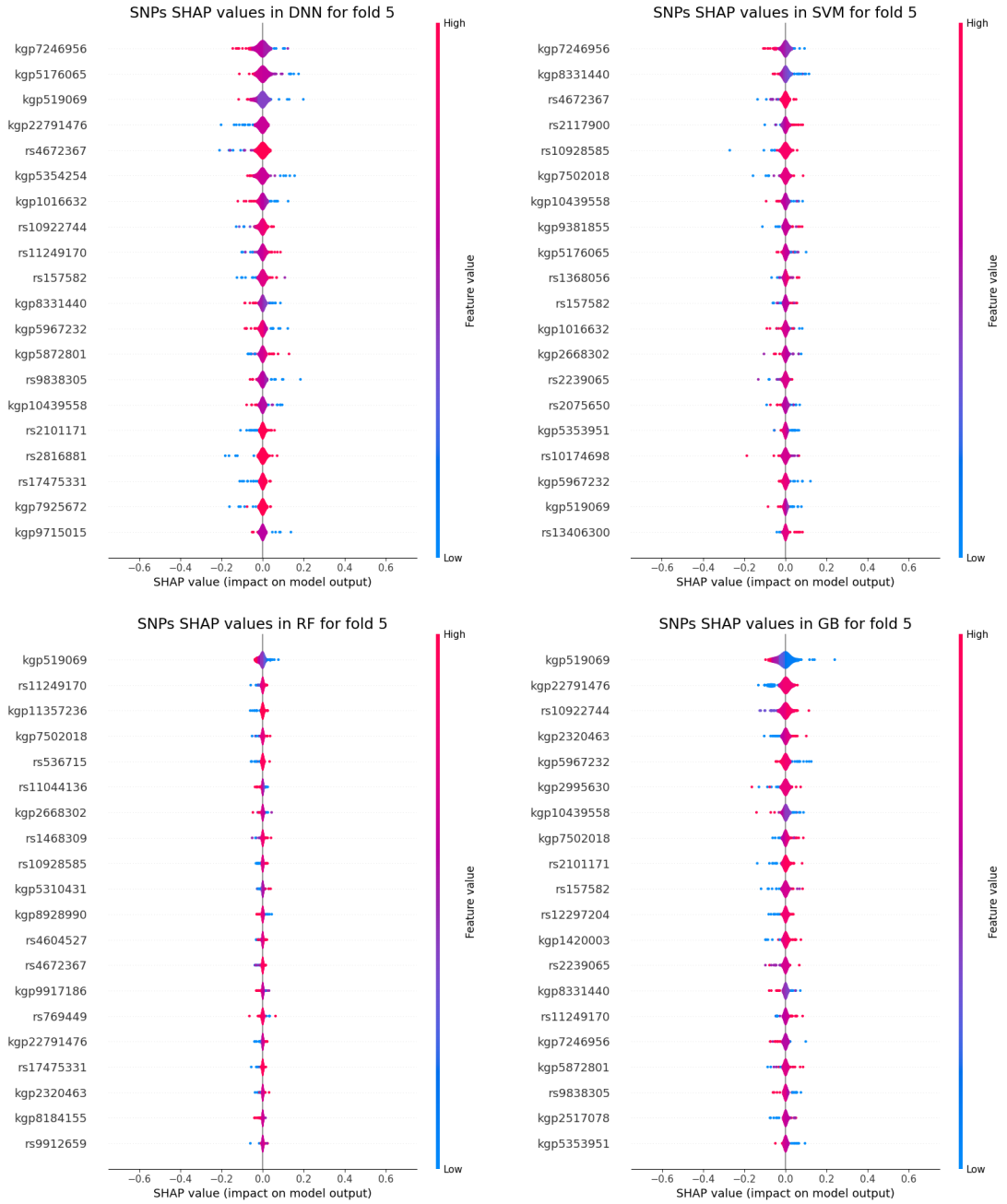

Figura B.32: Valores SHAP fold 5, método de división por k fold, umbral p-valor relajado

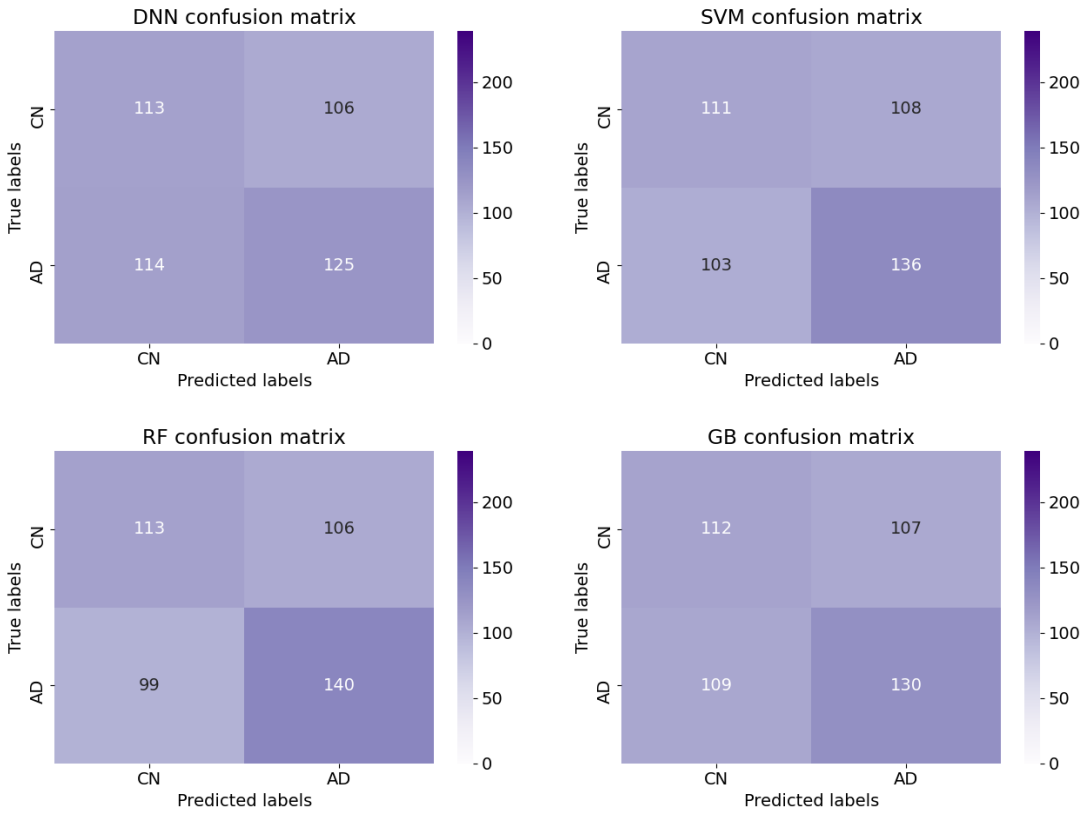

Figura B.33: Matrices de confusión, método de división por k fold, umbral p-valor relajado

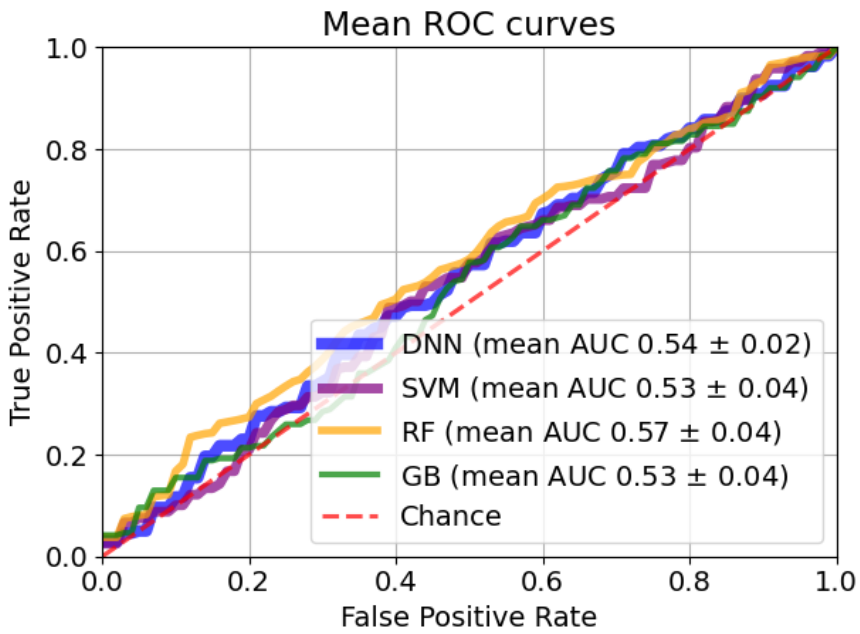

Figura B.34: Curvas ROC, método de división por k fold, umbral p-valor relajado

B.1.3. Método de filtración de datos

Resultados estudio GWAS (hasta p-valor relajado): rs769449 (1.23e-13), rs157582 (1.89e-09), rs2075650 (6.45e-08), kgp7502018 (3.24e-07), kgp2517078 (1.39e-06), kgp22791476 (1.81e-06), rs4672367 (3.51e-06), rs6045099 (3.57e-06), kgp5176065 (4.94e-06), kgp9381855 (5.80e-06), kgp87687 (5.92e-06), rs4670333 (6.01e-06), kgp11206857 (8.84e-06), rs2588252 (1.03e-05), rs6766099 (1.12e-05), rs2999899 (1.31e-05), kgp5354254 (1.34e-05), kgp12193771 (1.56e-05), rs9838305 (1.94e-05), rs246896 (1.98e-05), rs2227203 (2.00e-05), kgp12237618 (2.09e-05), rs1958979 (2.32e-05), rs11249170 (2.34e-05), kgp9811805 (2.44e-05), kgp8119974 (2.46e-05), rs4970834 (2.47e-05), kgp2913606 (2.48e-05), rs6723770 (2.48e-05), kgp3306661 (2.48e-05), rs473015 (2.55e-05), kgp1420003 (2.64e-05), kgp8625034 (2.64e-05), rs27652 (2.66e-05), kgp11201492 (2.69e-05), rs2101171 (2.79e-05), rs4604527 (2.87e-05), rs16854581 (3.25e-05), kgp134363 (3.47e-05), rs7545471 (3.72e-05), kgp519069 (4.06e-05), rs237473 (4.08e-05), rs16920867 (4.38e-05), kgp4882314 (4.55e-05), kgp2483143 (4.55e-05), kgp1562249 (4.58e-05), kgp10439558 (4.80e-05), kgp2983682 (4.87e-05), kgp8331440 (5.27e-05), kgp3791913 (5.31e-05), rs2256364 (5.45e-05), rs11660396 (5.57e-05), rs237463 (5.72e-05), kgp5872801 (6.02e-05), kgp7876315 (6.45e-05), kgp12402920 (6.56e-05), rs1196184 (6.76e-05), rs7913770 (7.03e-05), rs594165 (7.06e-05), kgp97310 (7.12e-05), kgp5060005 (7.13e-05), kgp10651550 (7.55e-05), kgp3336671 (7.99e-05), rs4641389 (8.04e-05), rs2371923 (8.05e-05), kgp10620084 (8.06e-05), kgp1016632 (8.06e-05), rs306503 (8.06e-05), rs10174698 (8.21e-05), kgp6413075 (8.32e-05), kgp5901917 (8.53e-05), kgp2138924 (8.57e-05), rs17247323 (8.62e-05), rs4943768 (8.63e-05), rs4506448 (8.68e-05), kgp6420136 (8.84e-05), kgp10681801 (8.85e-05), kgp2645931 (8.85e-05), rs9299396 (9.15e-05), kgp12087452 (9.29e-05), rs2279105 (9.40e-05), kgp11484611 (9.44e-05), rs10974985 (9.69e-05), rs16972552 (9.85e-05), kgp3478256 (9.88e-05), rs2086149 (9.89e-05), kgp2951395 (9.94e-05), kgp1270595 (9.99e-05),

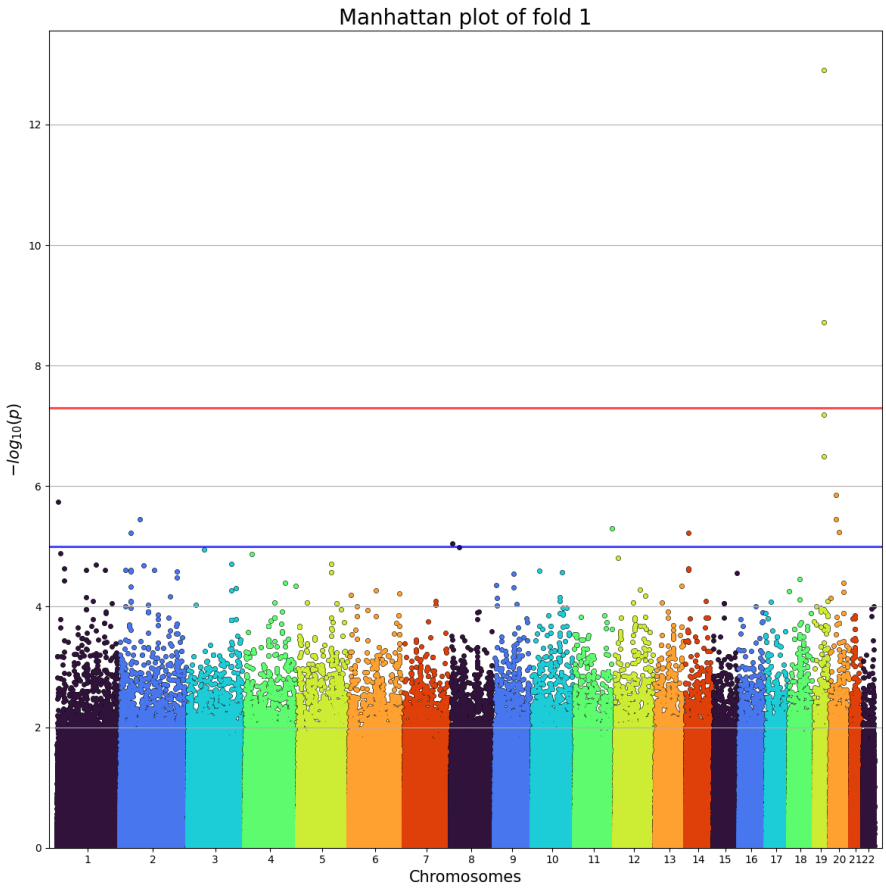

Figura B.35: Gráfico de Manhattan para el método de no división (filtración de datos)

Umbral clásico

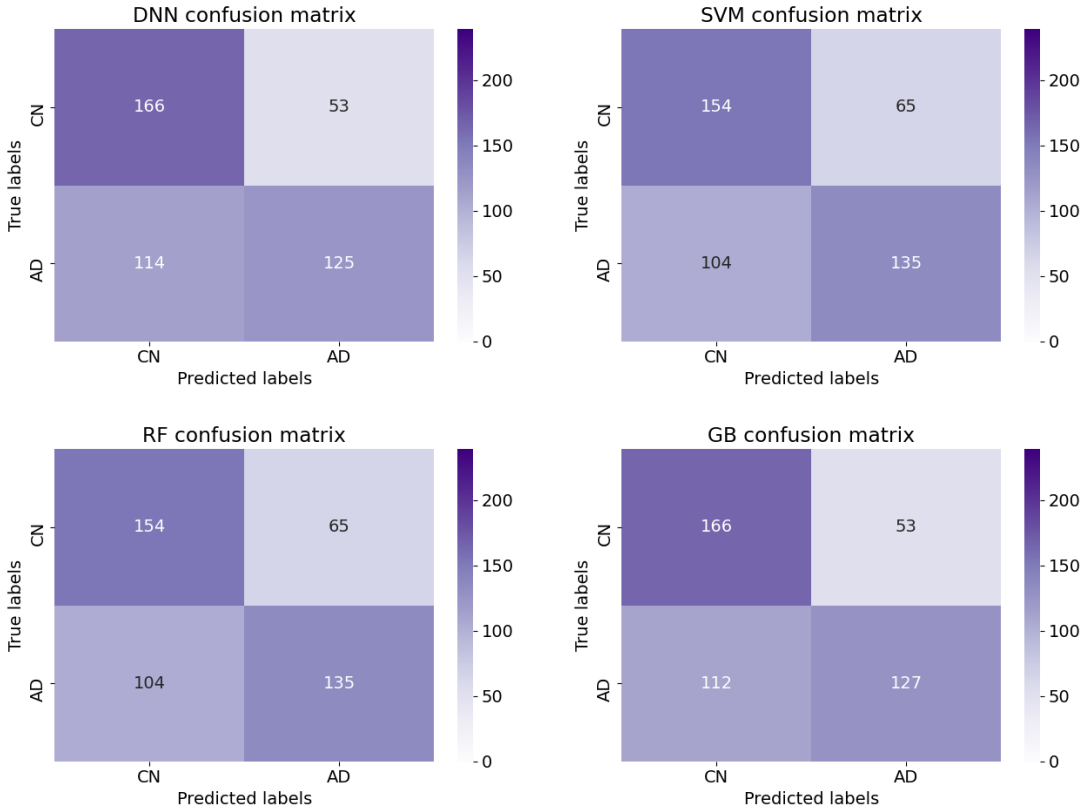

Figura B.36: Matrices de confusión, método de no división (filtración de datos), umbral p-valor tradicional

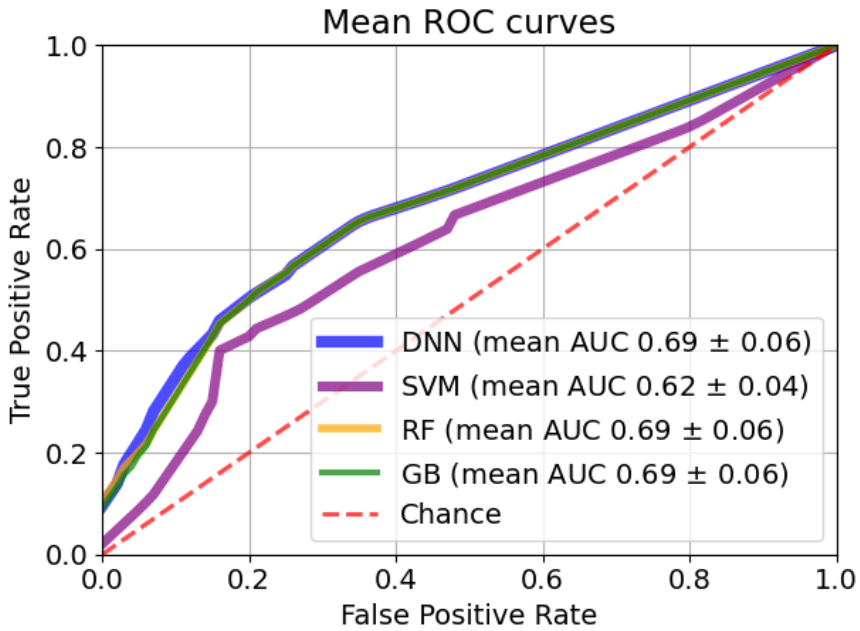

Figura B.37: Curvas ROC, método de no división (filtración de datos), umbral p-valor tradicional

Umbral sugestivo

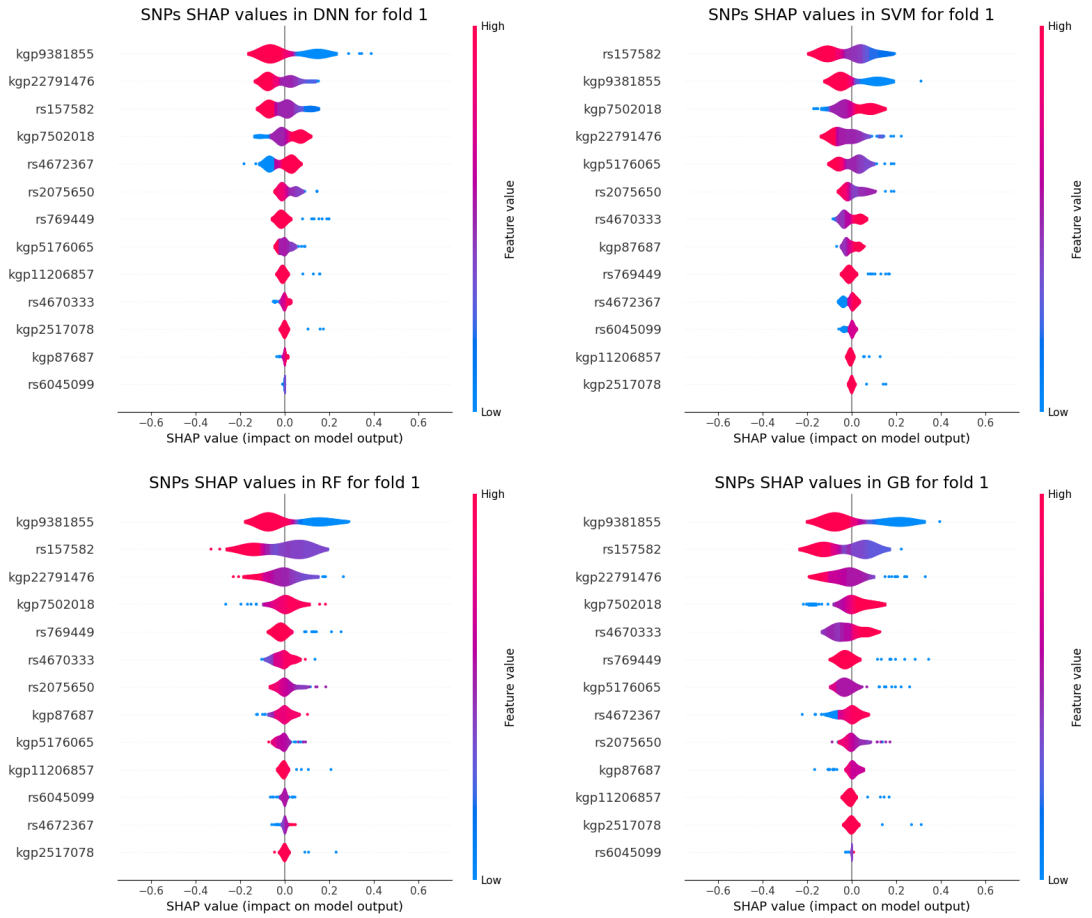

Figura B.38: Valores SHAP fold 1, método de no división (filtración de datos), umbral p-valor sugestivo

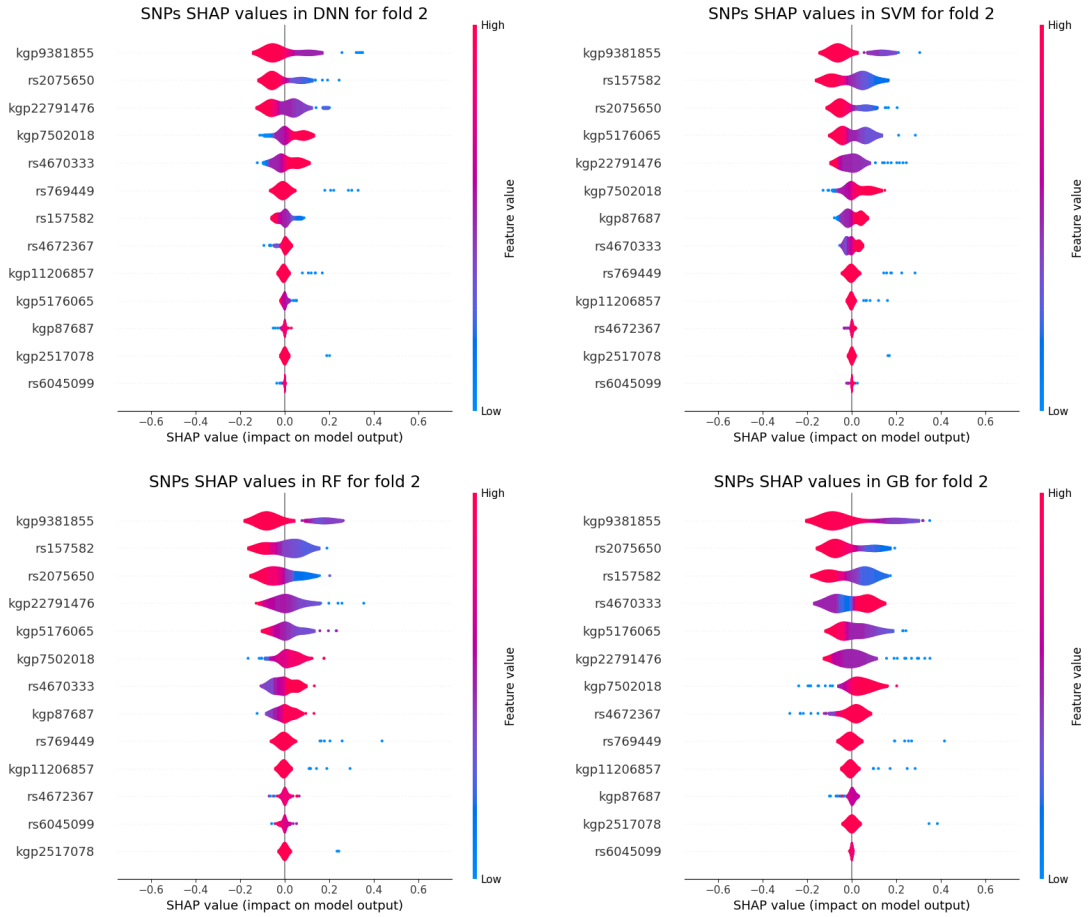

Figura B.39: Valores SHAP fold 2, método de no división (filtración de datos), umbral p-valor sugestivo

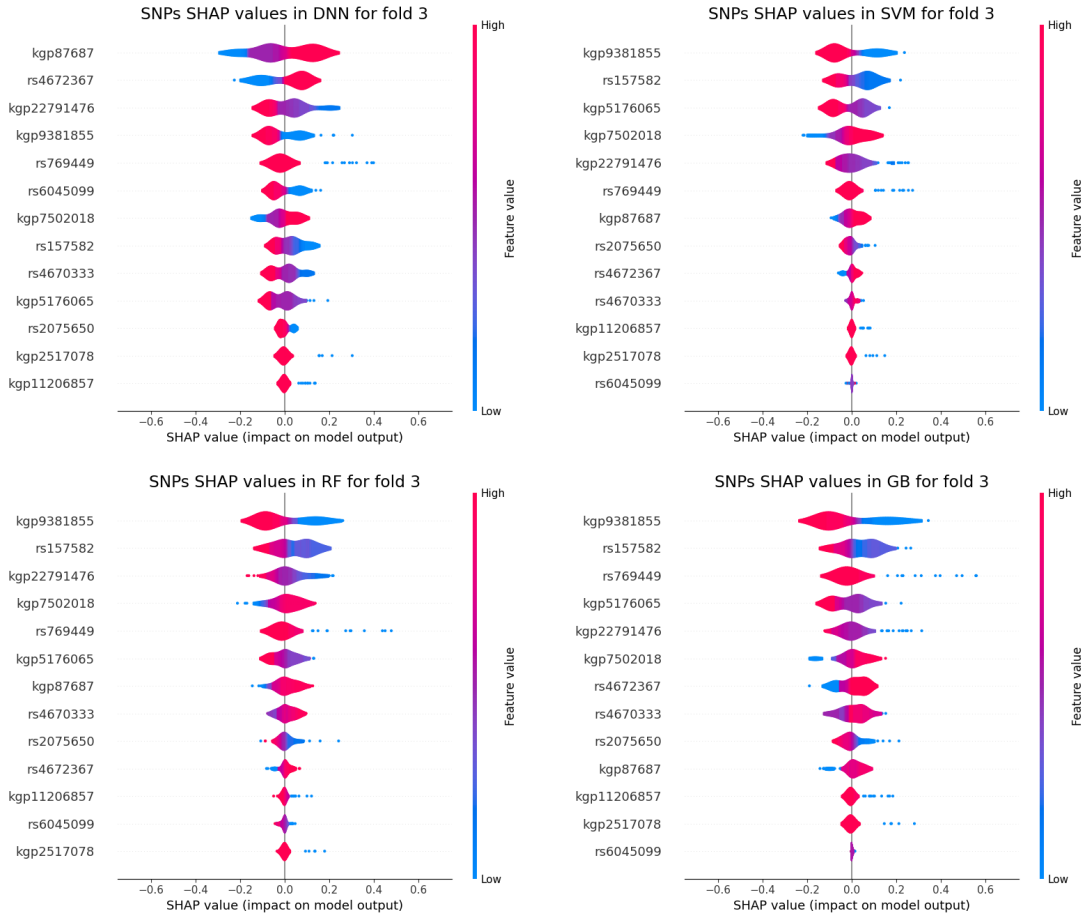

Figura B.40: Valores SHAP fold 3, método de no división (filtración de datos), umbral p-valor sugestivo

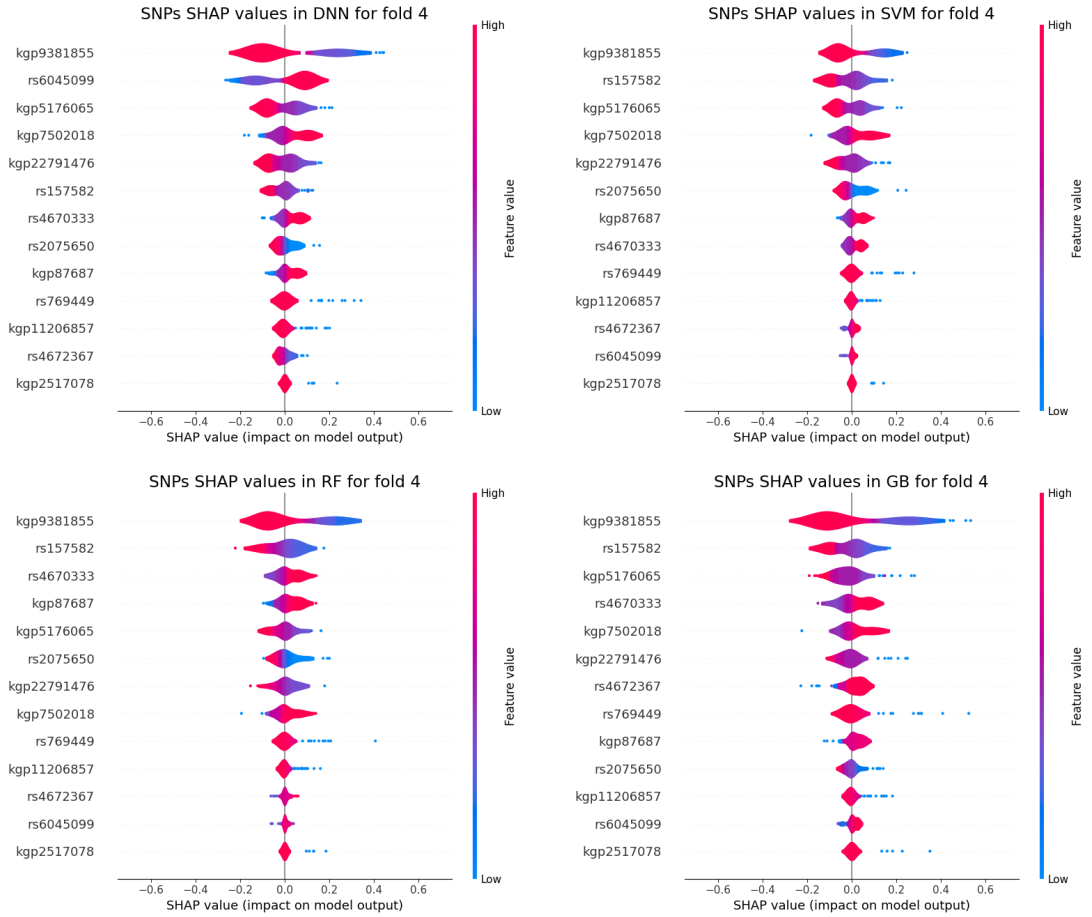

Figura B.41: Valores SHAP fold 4, método de no división (filtración de datos), umbral p-valor sugestivo

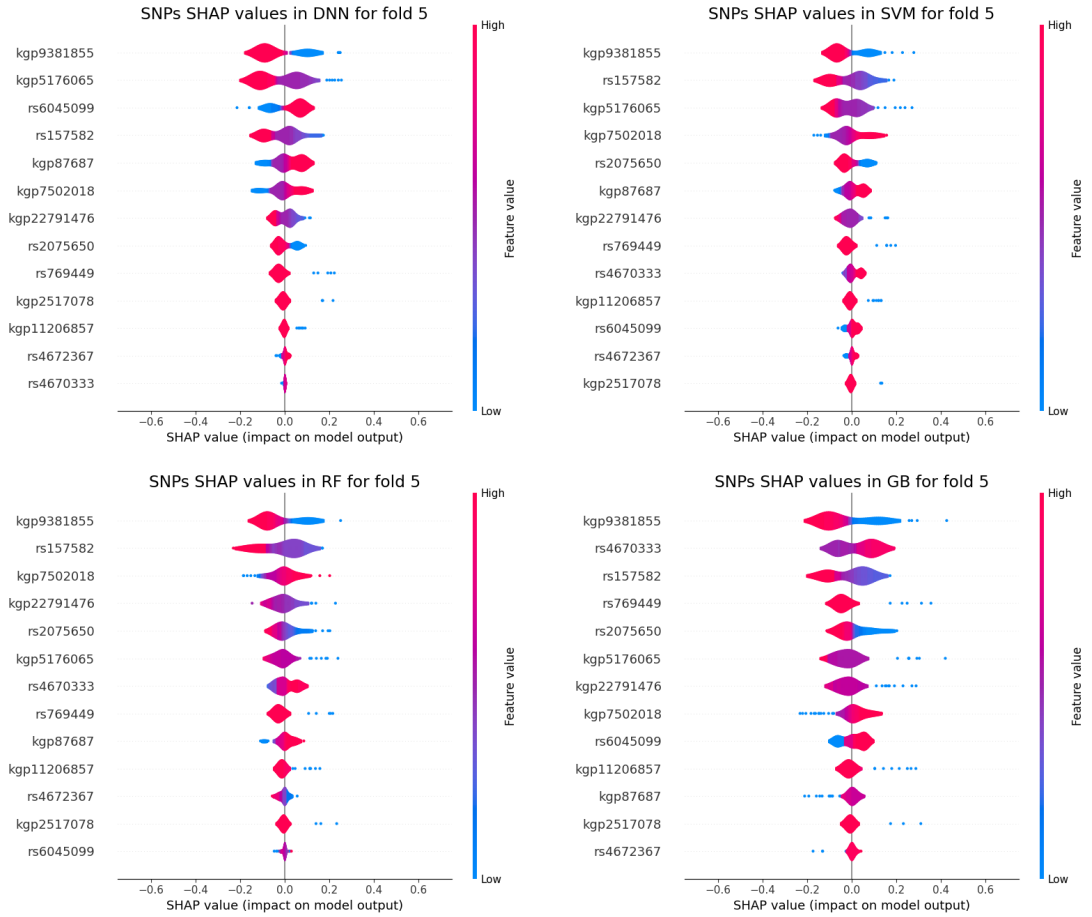

Figura B.42: Valores SHAP fold 5, método de no división (filtración de datos), umbral p-valor sugestivo

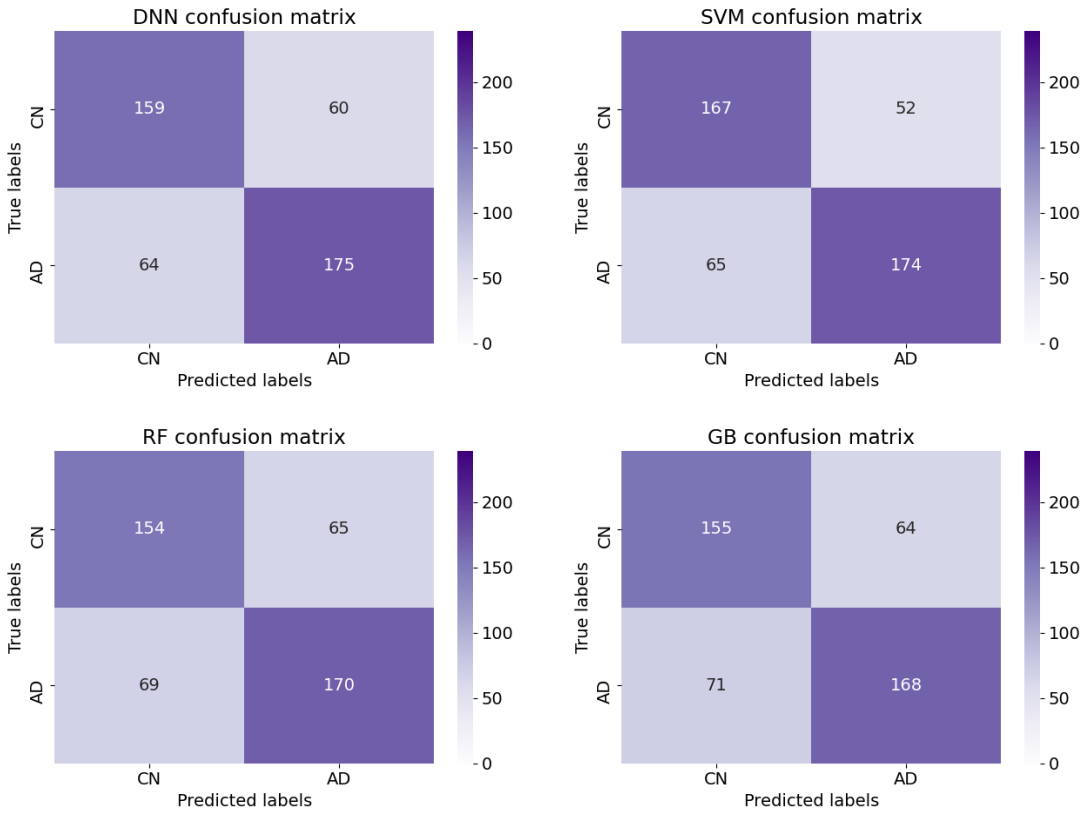

Figura B.43: Matrices de confusión, método de no división (filtración de datos), umbral p-valor sugestivo

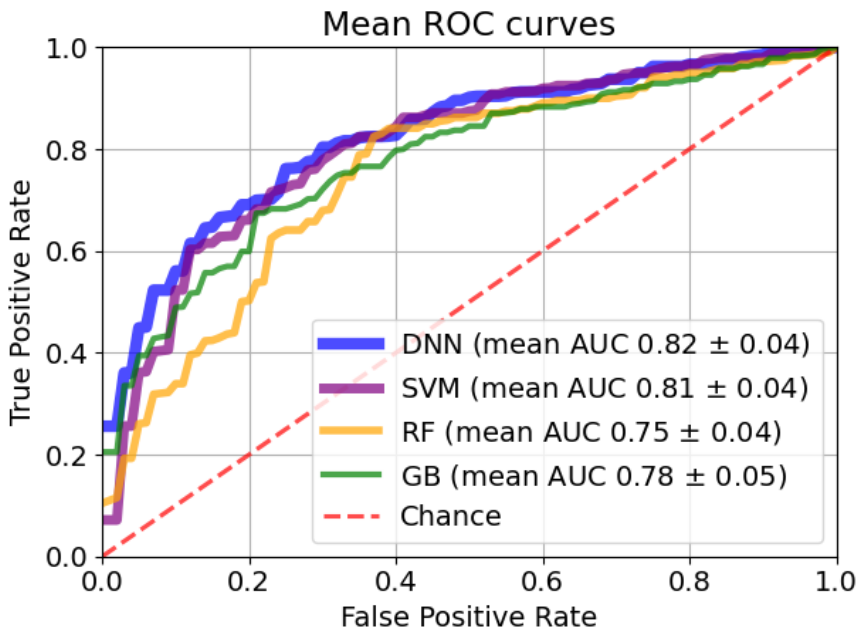

Figura B.44: Curvas ROC, método de no división (filtración de datos), umbral p-valor sugestivo

Umbral relajado

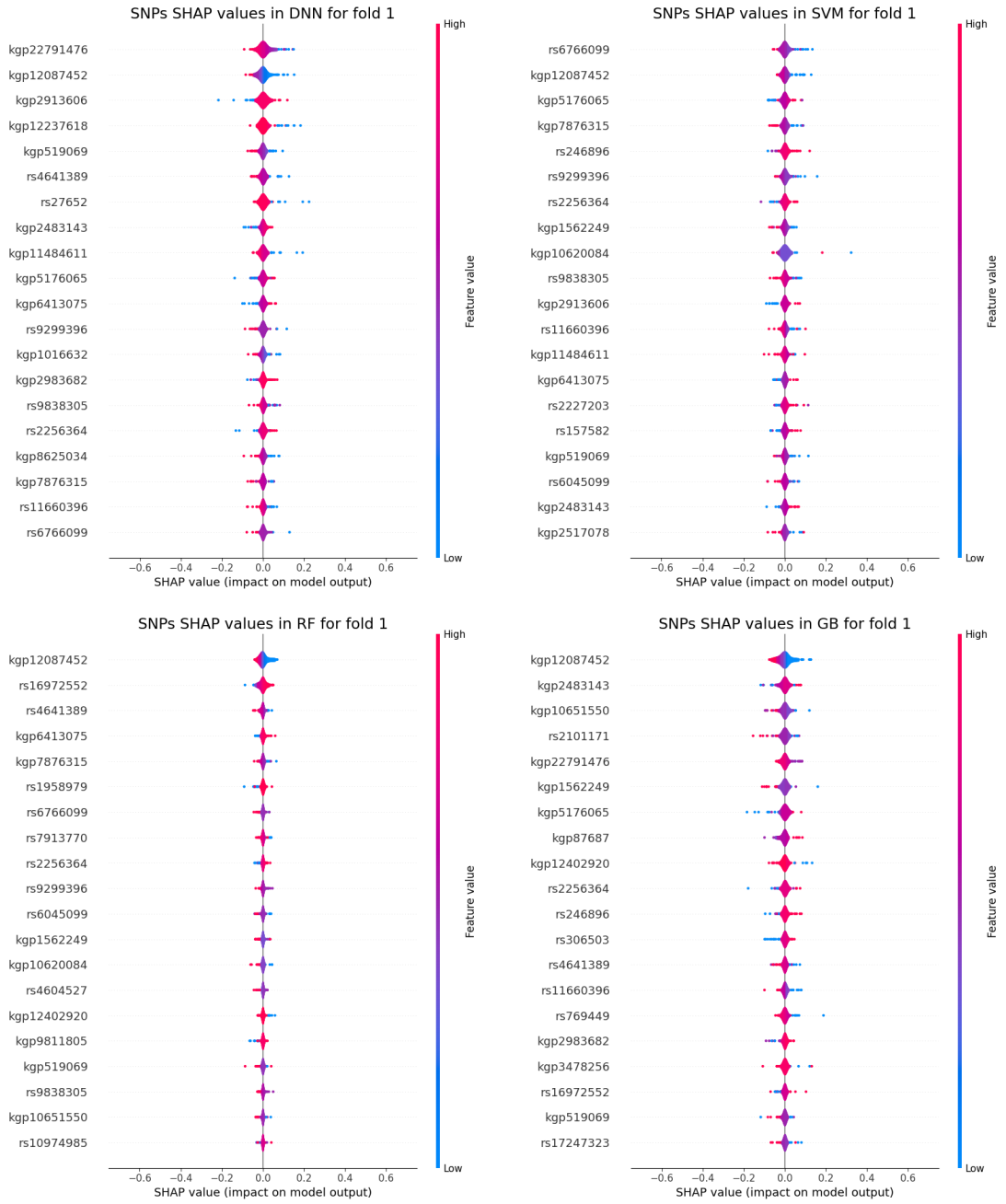

Figura B.45: Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado

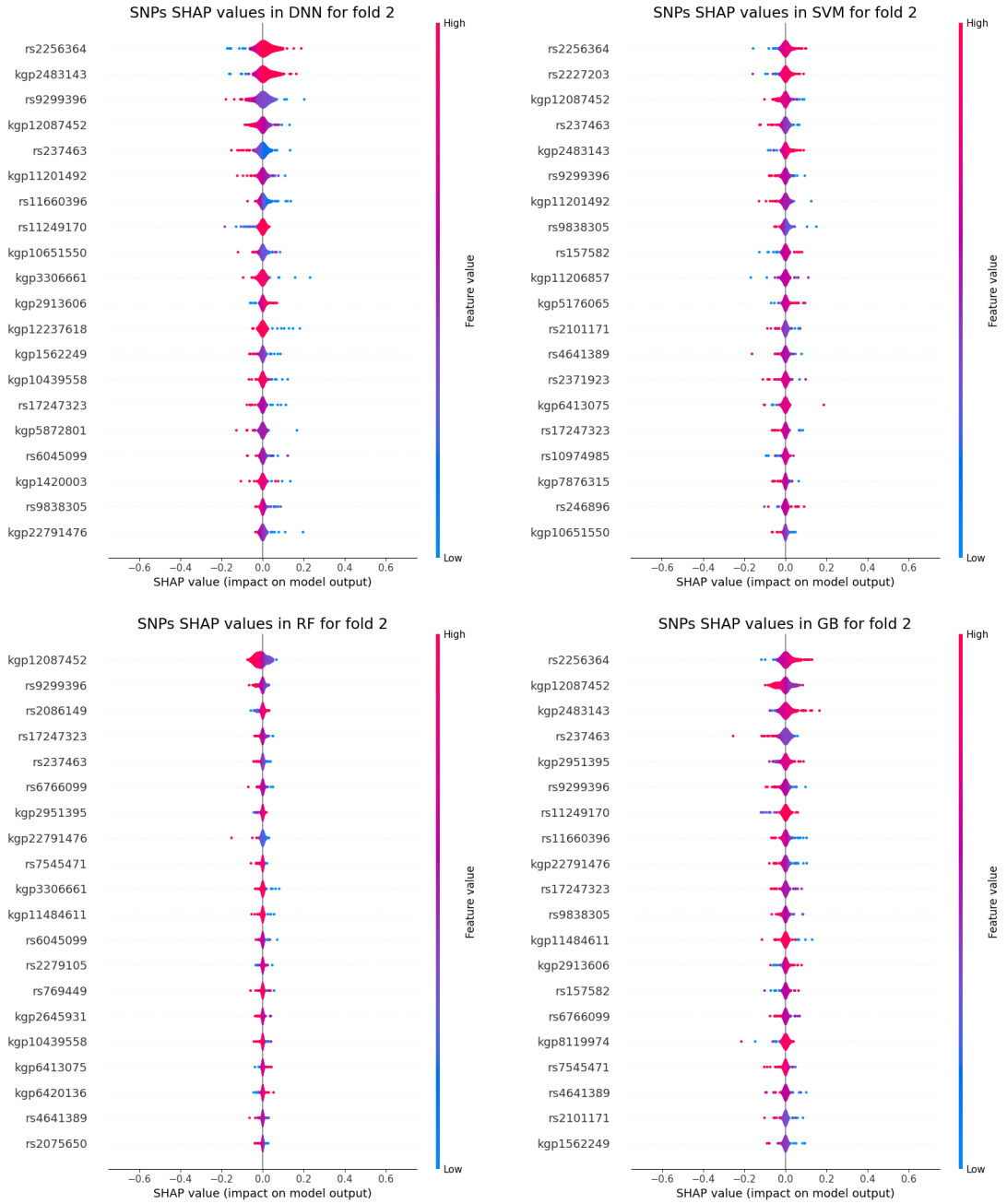

Figura B.46: Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado

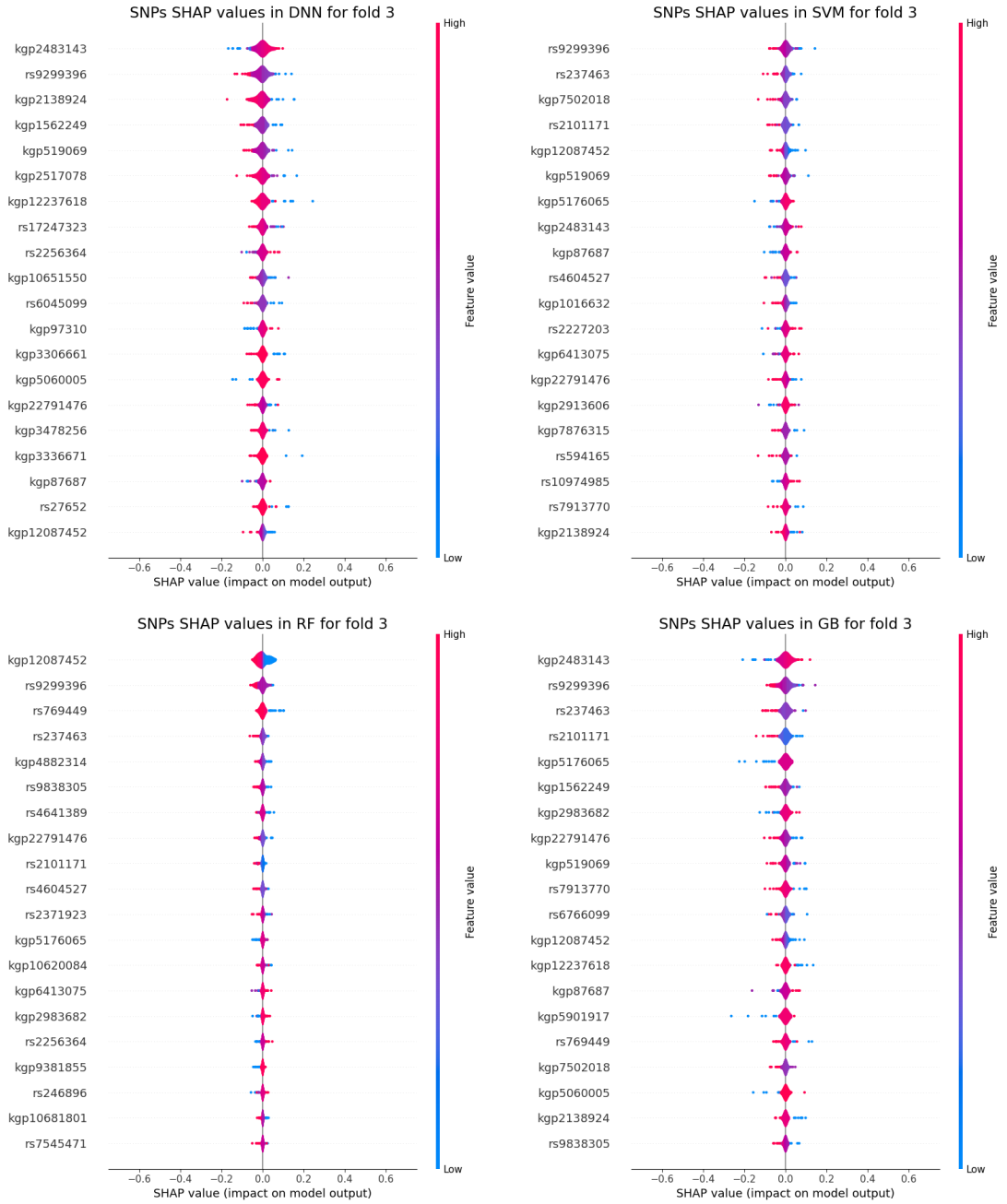

Figura B.47: Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado

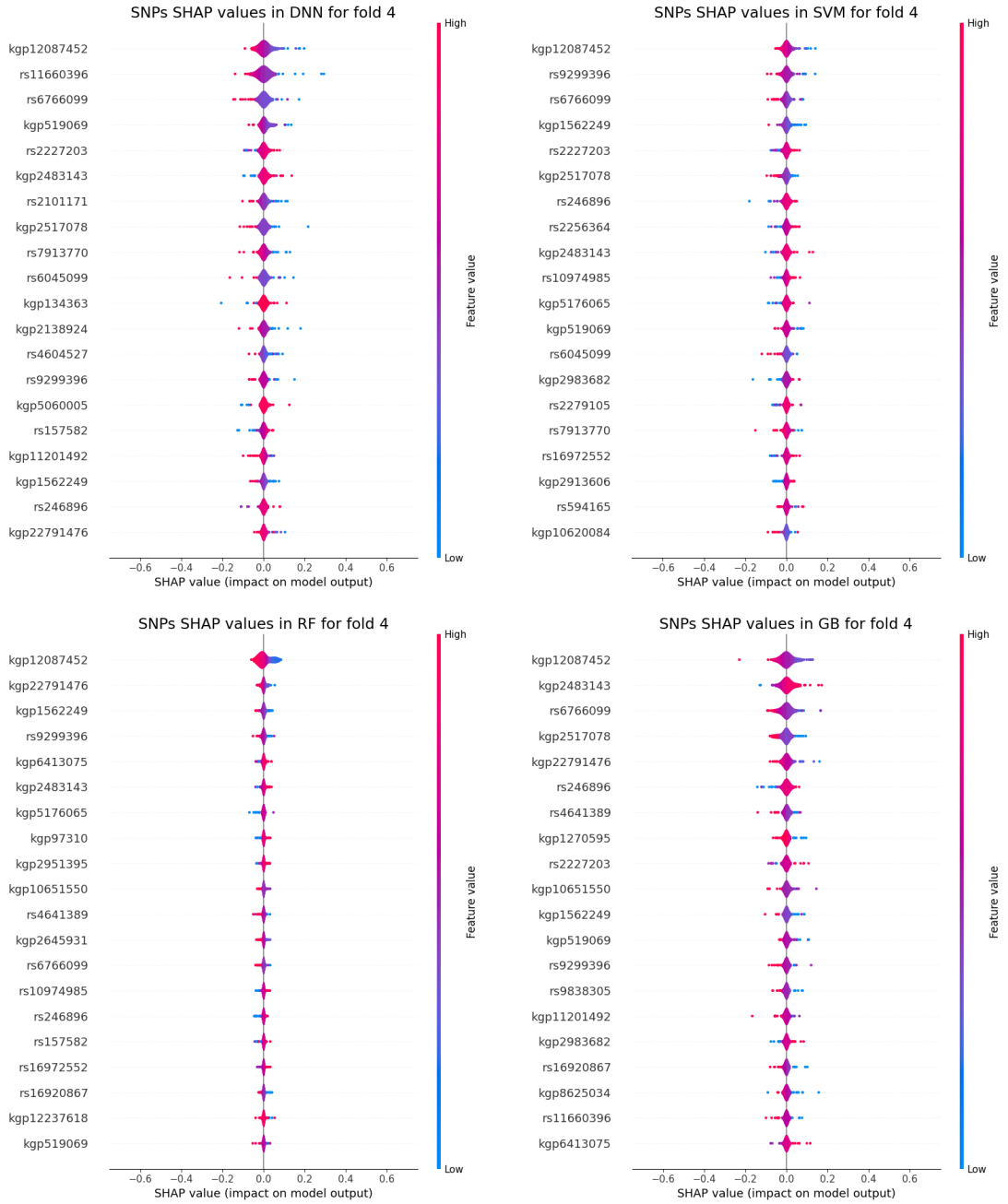

Figura B.48: Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado

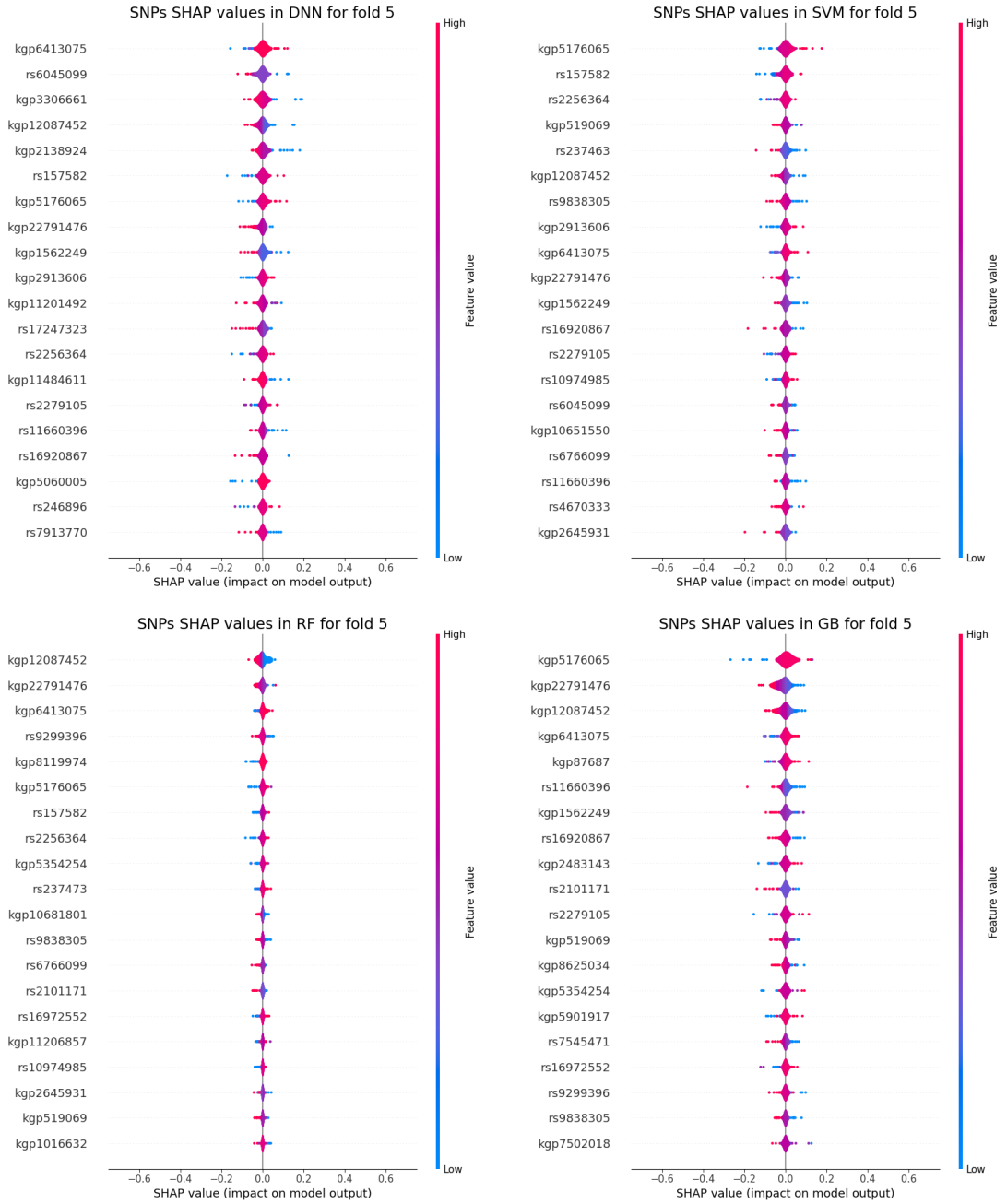

Figura B.49: Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado

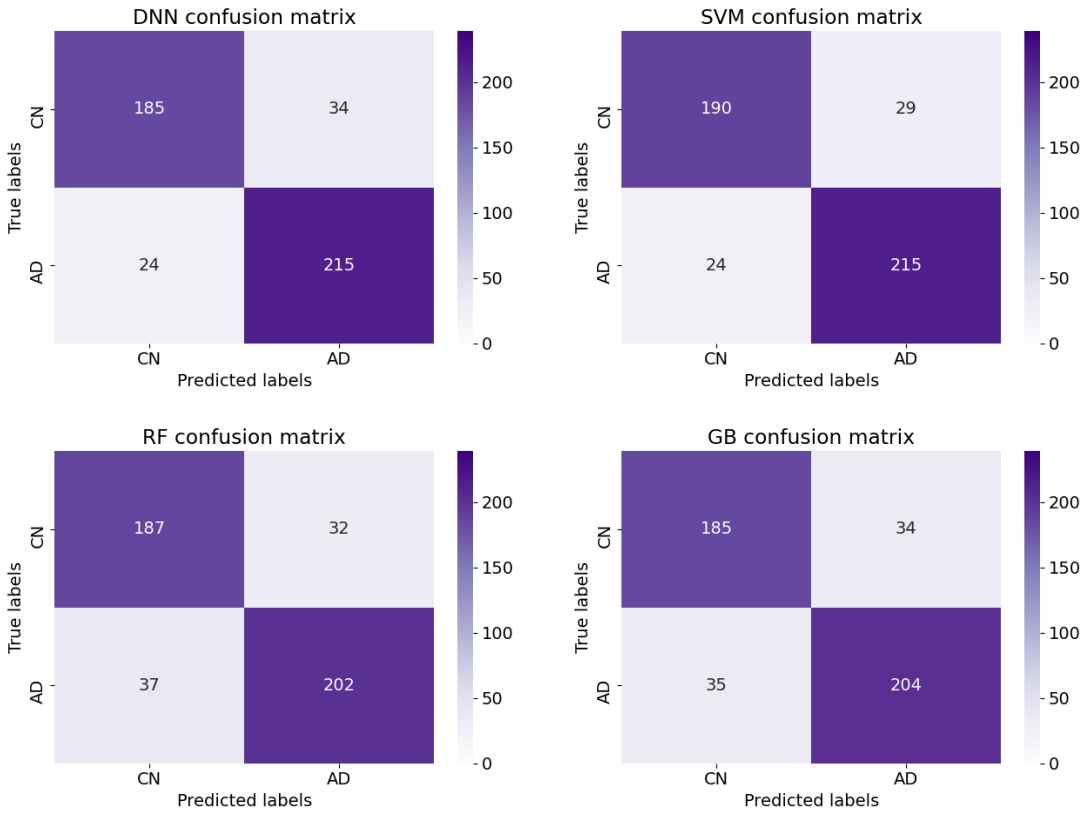

Figura B.50: Matrices de confusión, método de filtración de datos, umbral p-valor relajado

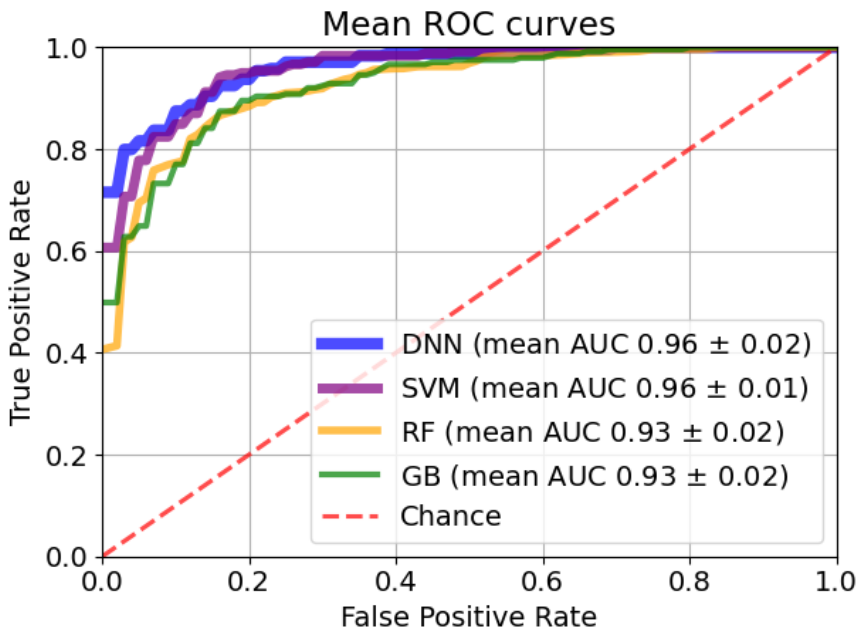

Figura B.51: Curvas ROC, método de filtración de datos, umbral p-valor relajado

B.1.4. Método de análisis externo

Lista 1

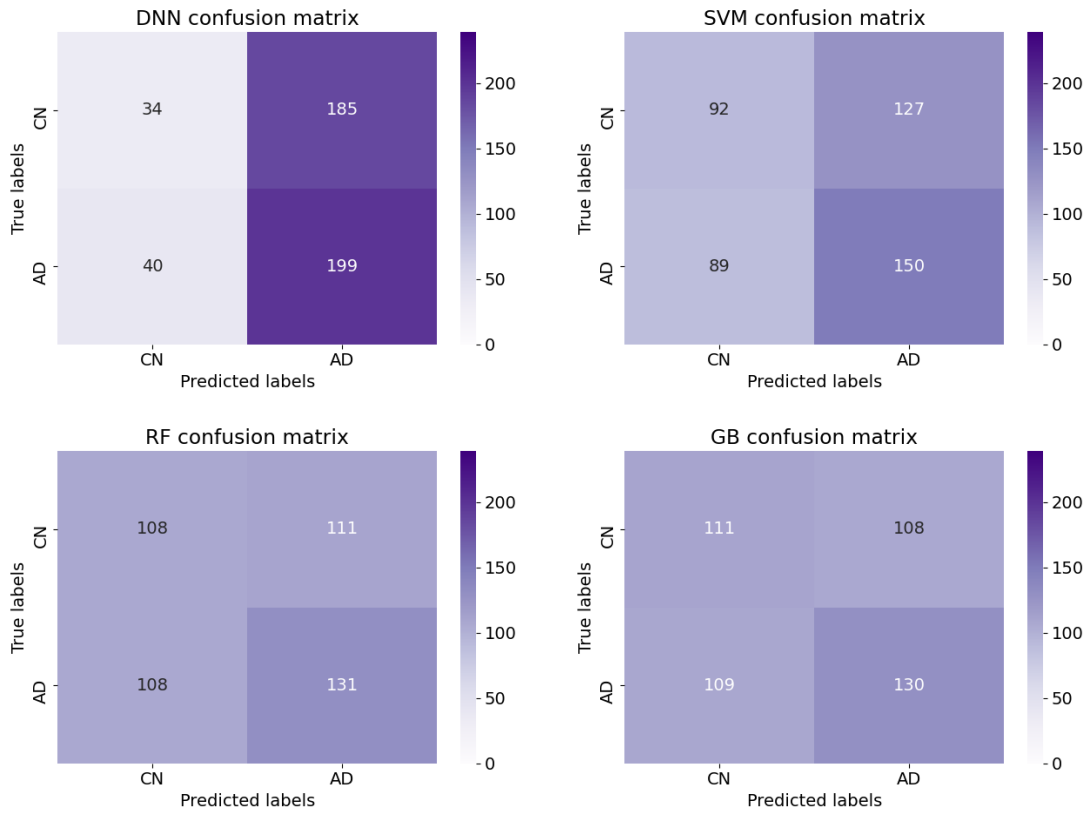

Figura B.52: Matrices de confusión, lista 1 SNPs

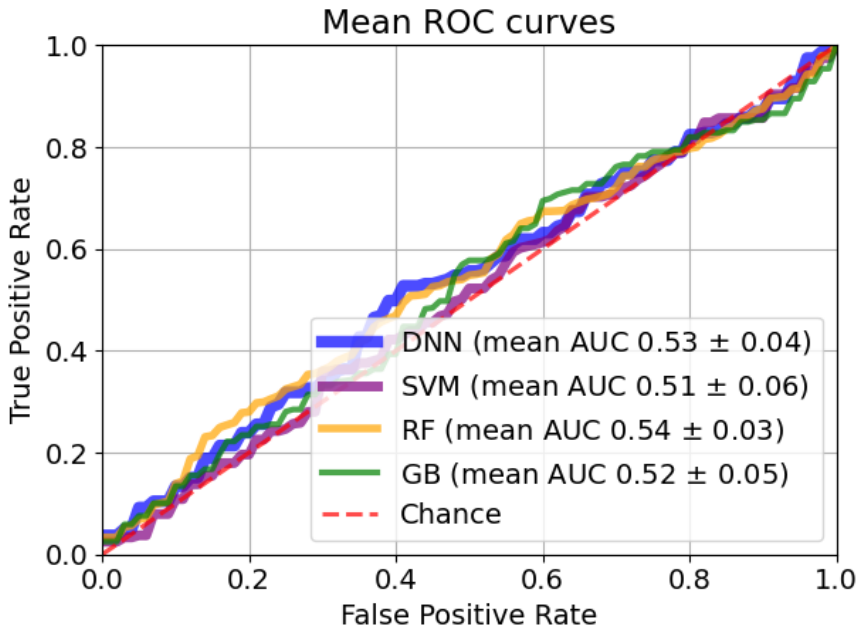

Figura B.53: Curvas ROC, SNPs de lista 1

Lista 2

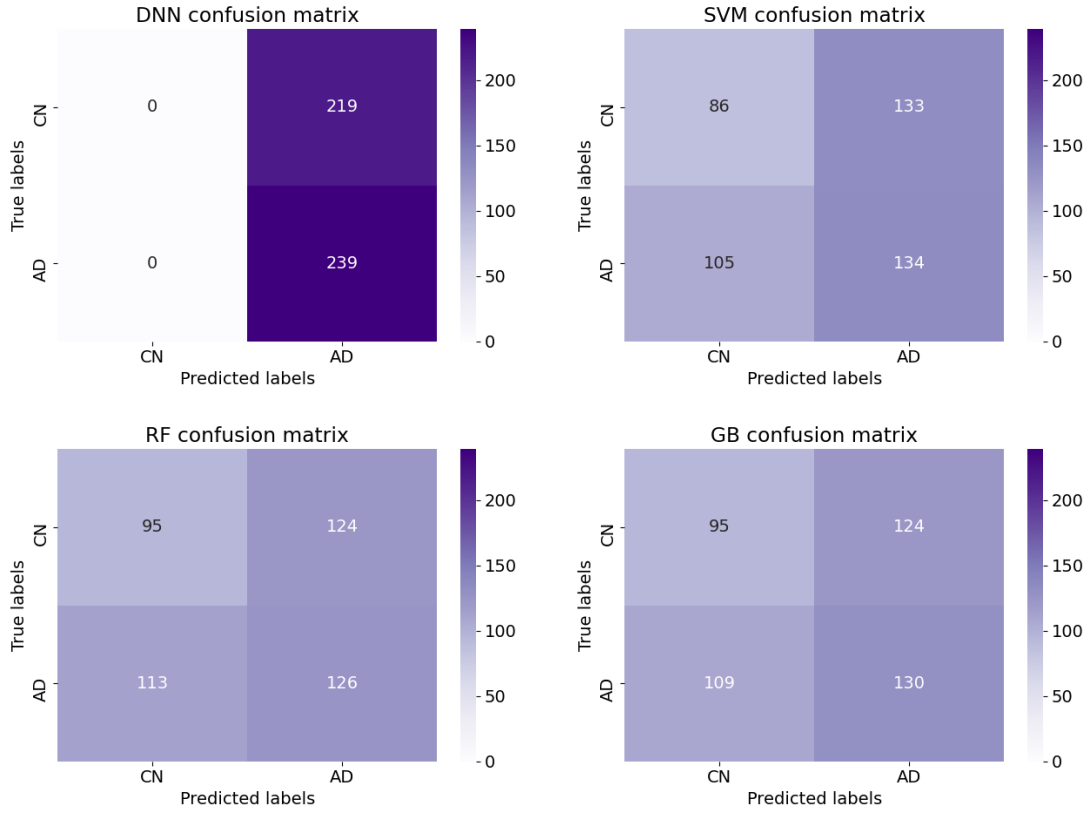

Figura B.54: Matrices de confusión, SNPs de lista 2

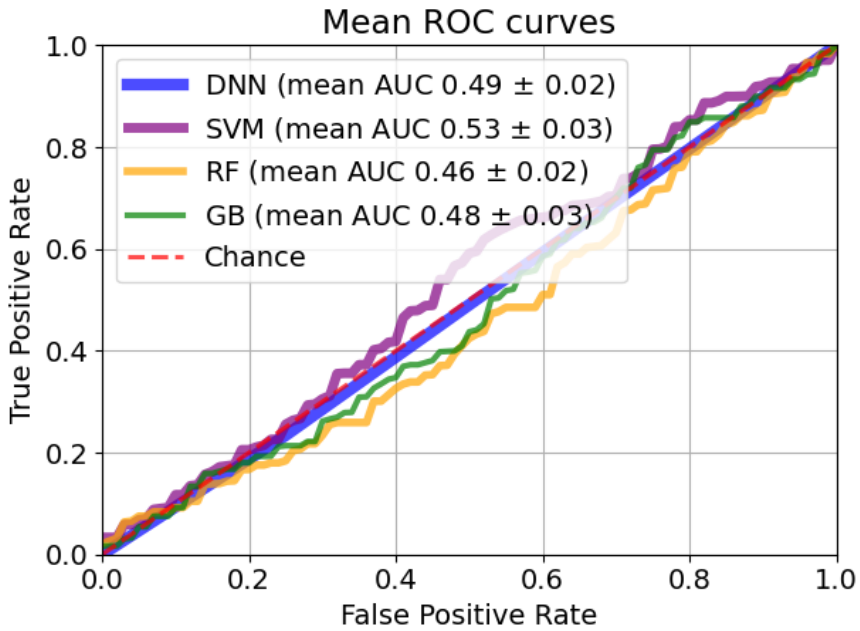

Figura B.55: Curvas ROC, SNPs de lista 2

Lista 3

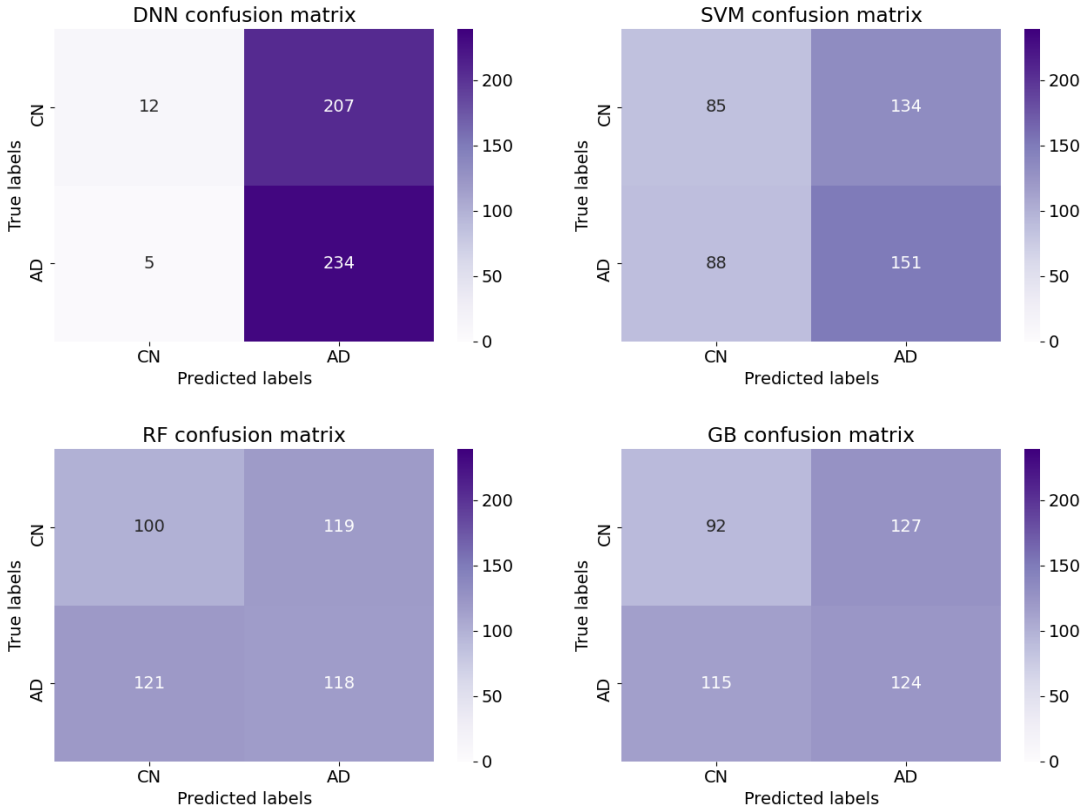

Figura B.56: Matrices de confusión, SNPs de lista 3

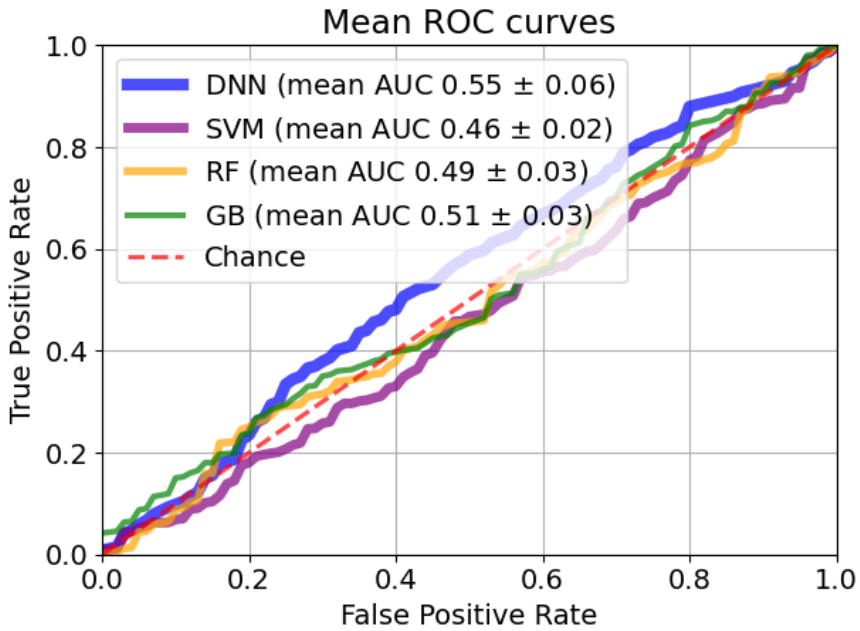

Figura B.57: Curvas ROC, SNPs de lista 3

B.2. Resultados con unión de datos de distintos chips

Tras aplicar el control de calidad se obtienen 33710 SNPs y 1212 muestras, de las cuales 593 son controles y 619 son casos.

B.2.1. Método de división simple

Resultados estudio GWAS (hasta p-valor relajado): rs2075650 (2.14e-15), rs6859 (6.33e-07), rs439401 (2.04e-05), rs486557 (6.11e-05), rs1343030 (6.20e-05),

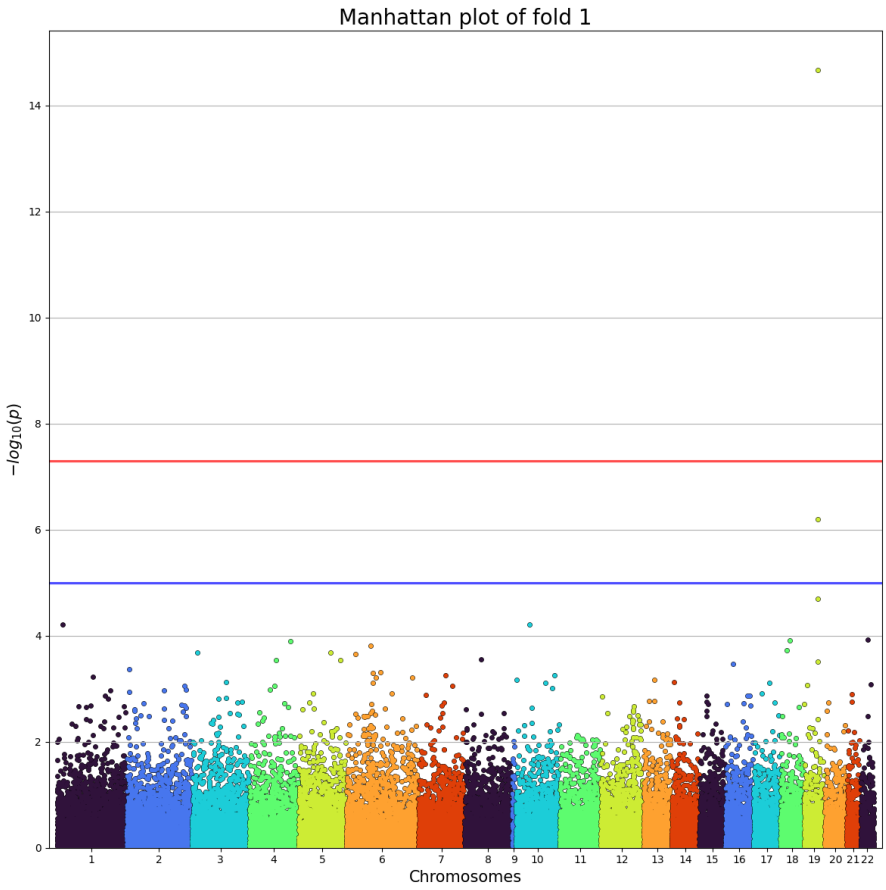

Figura B.58: Gráfico de Manhattan para el método de división simple

Umbral clásico

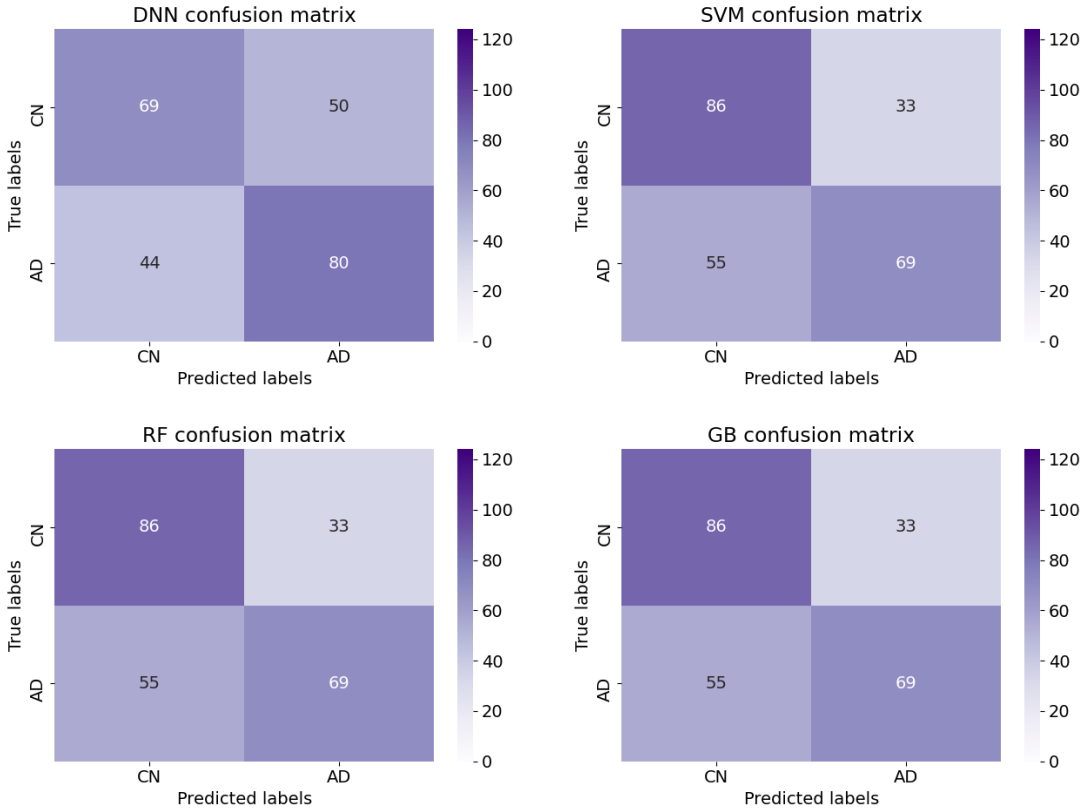

Figura B.59: Matrices de confusión, método de división simple, umbral p-valor tradicional

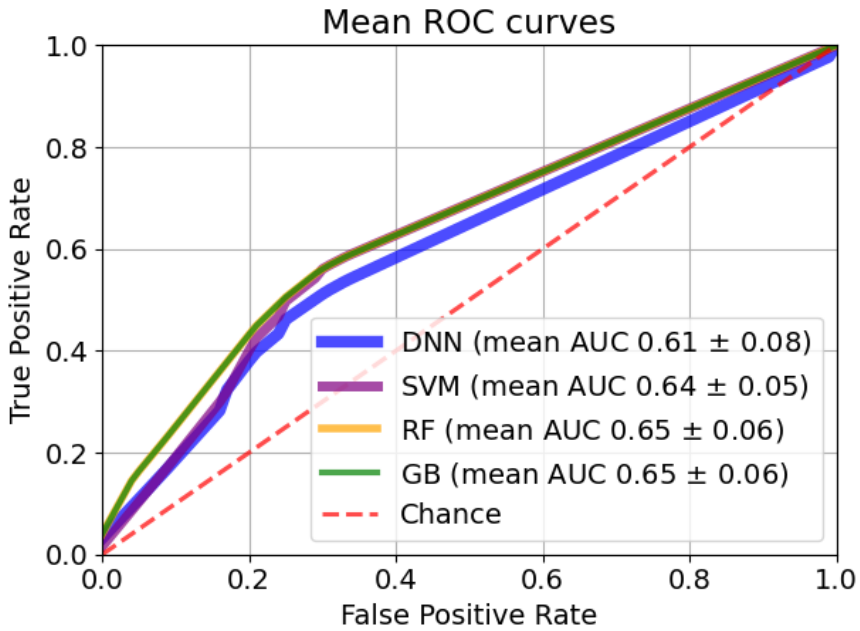

Figura B.60: Curvas ROC, método de división simple, umbral p-valor tradicional

Umbral sugestivo

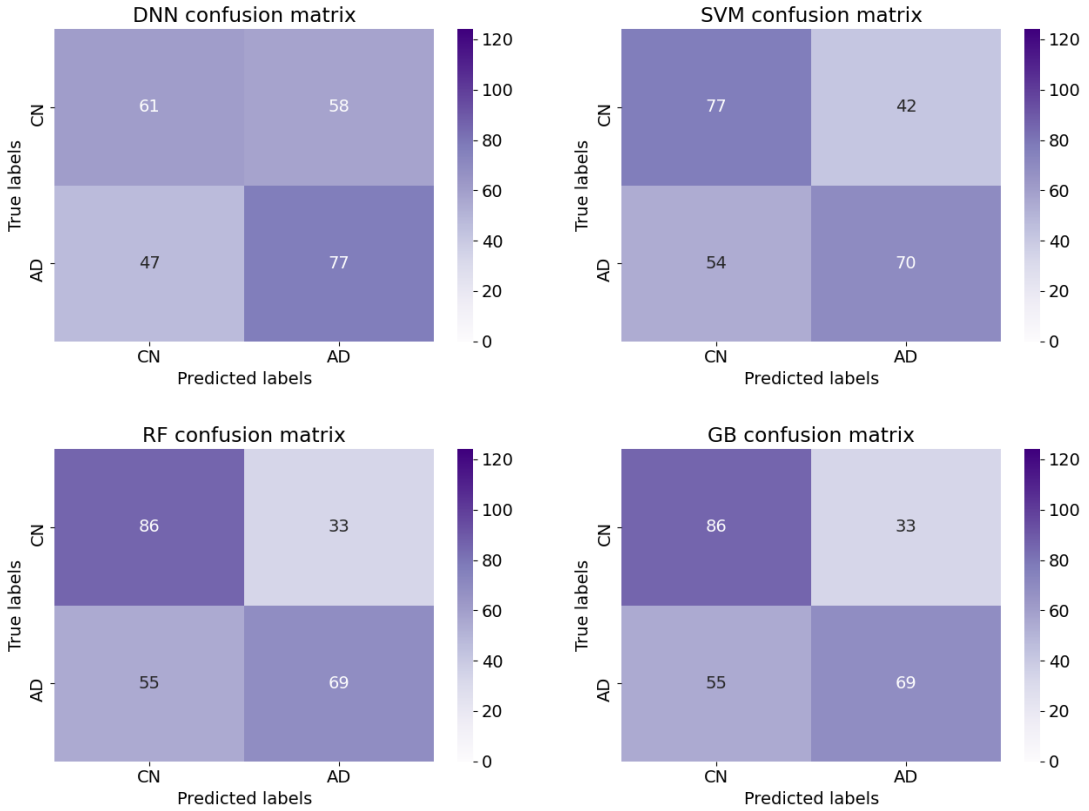

Figura B.61: Matrices de confusión, método de división simple, umbral p-valor sugestivo

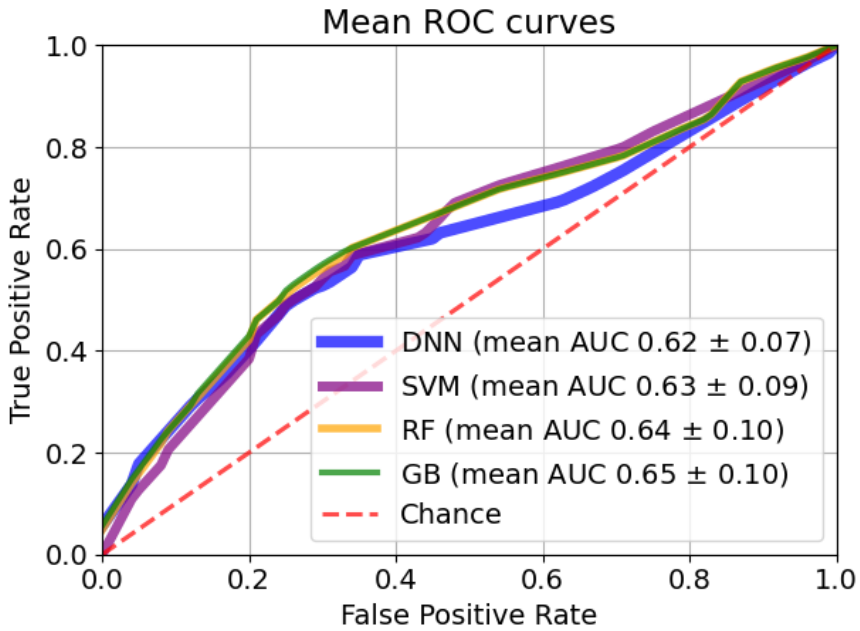

Figura B.62: Curvas ROC, método de división simple, umbral p-valor sugestivo

Umbral relajado

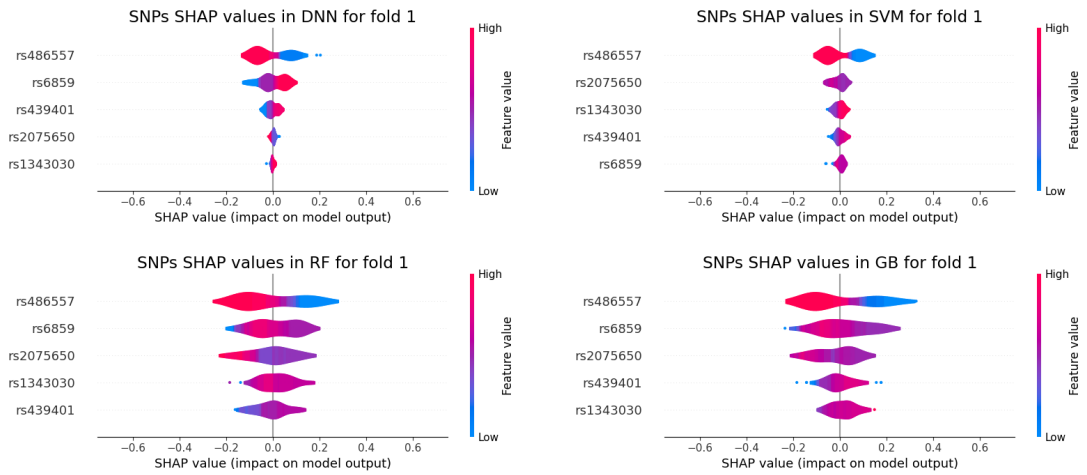

Figura B.63: Valores SHAP fold 1, método de división simple, umbral p-valor relajado

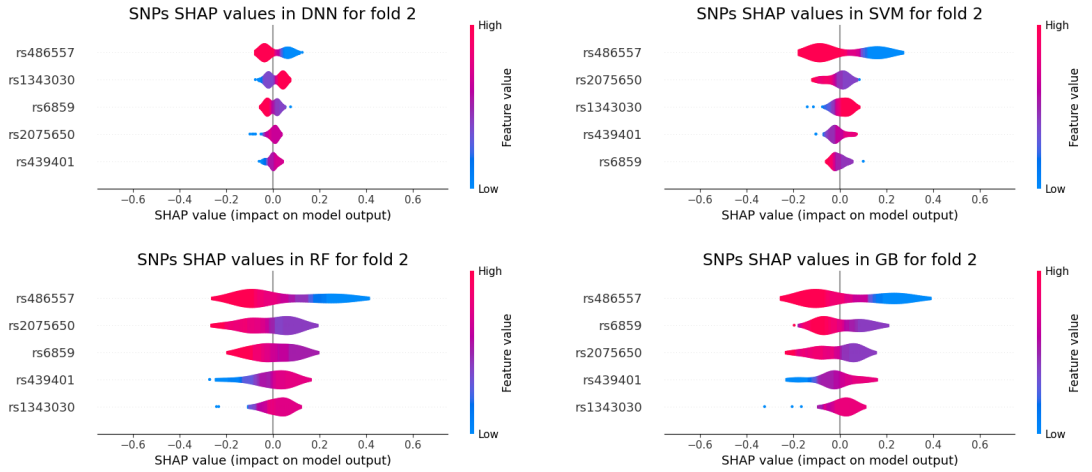

Figura B.64: Valores SHAP fold 2, método de división simple, umbral p-valor relajado

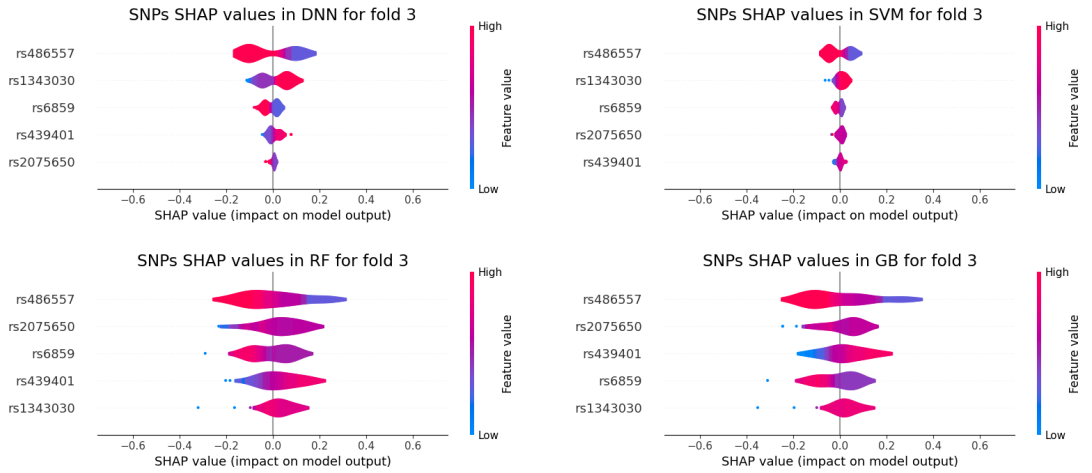

Figura B.65: Valores SHAP fold 3, método de división simple, umbral p-valor relajado

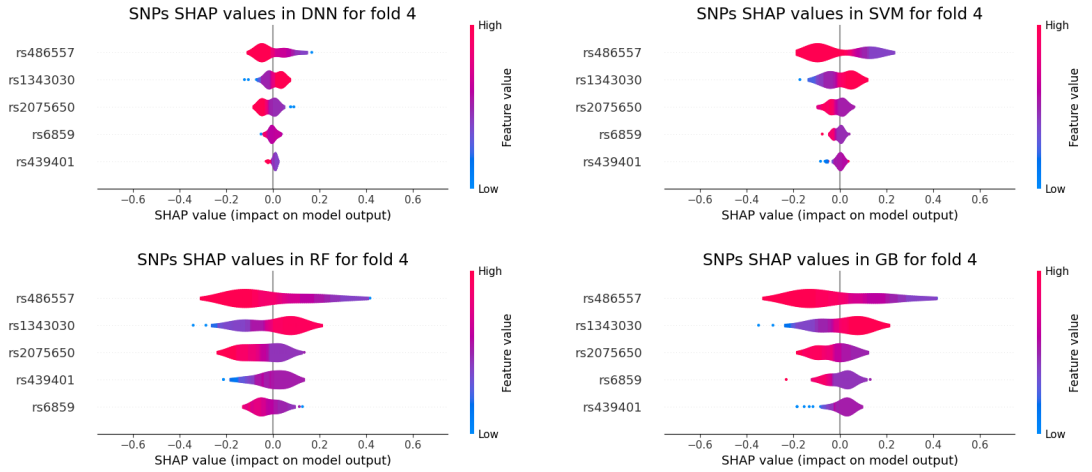

Figura B.66: Valores SHAP fold 4, método de división simple, umbral p-valor relajado

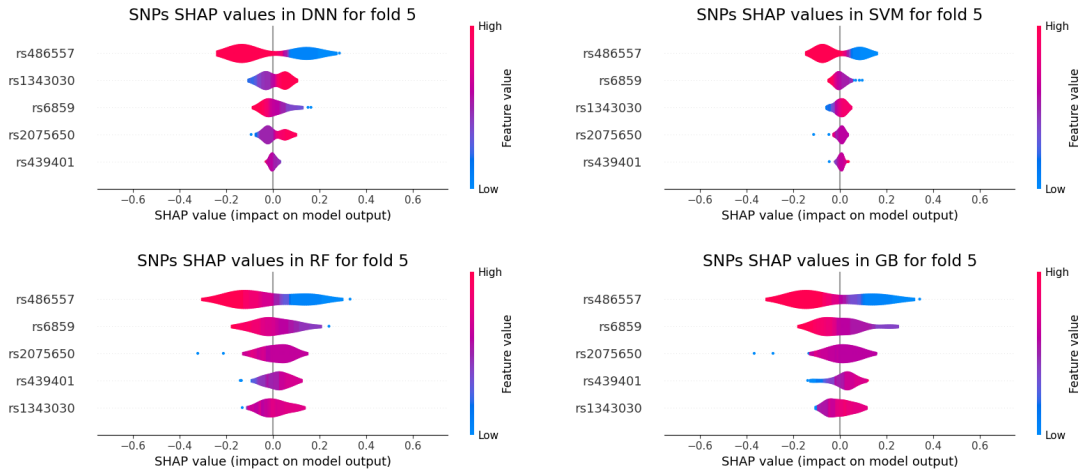

Figura B.67: Valores SHAP fold 5, método de división simple, umbral p-valor relajado

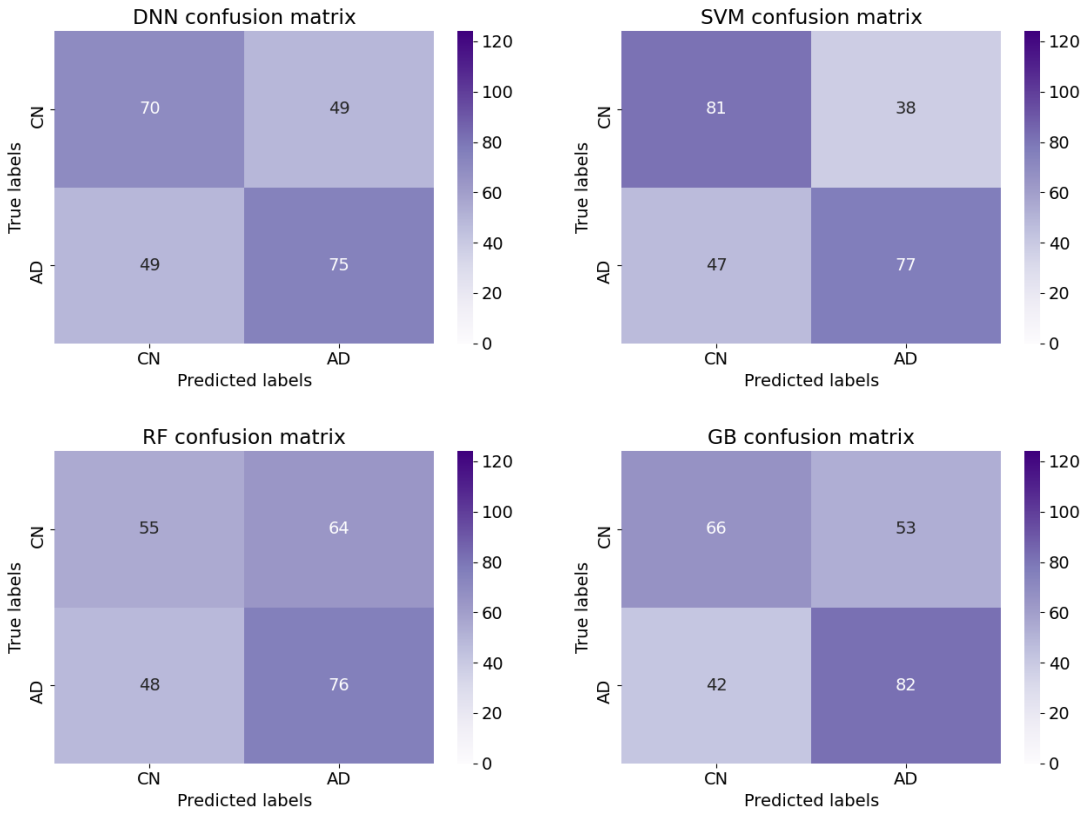

Figura B.68: Matrices de confusión, método de división simple, umbral p-valor relajado

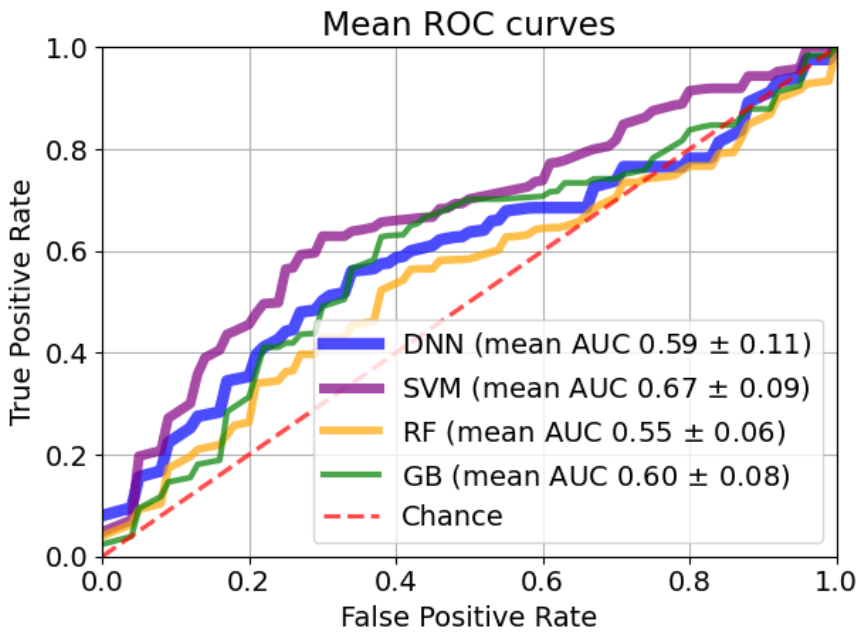

Figura B.69: Curvas ROC, método de división simple, umbral p-valor relajado

B.2.2. Método de división por k-fold

Resultados estudio GWAS (hasta p-valor relajado) para fold 1: rs2075650 (2.14e-15), rs6859 (6.33e-07), rs439401 (2.04e-05), rs486557 (6.11e-05), rs1343030 (6.20e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 2: rs2075650 (1.15e-16), rs439401 (2.03e-06), rs7919352 (1.22e-05), rs2419398 (7.47e-05), rs727919 (8.64e-05), rs4389538 (8.78e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 3: rs2075650 (1.03e-14), rs439401 (9.40e-07), rs8106922 (1.17e-05), rs480092 (3.09e-05), rs12592749 (4.06e-05), rs2051788 (4.41e-05), rs7532949 (8.60e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 4: rs2075650 (2.02e-18), rs6859 (1.77e-06), rs8106922 (1.33e-05), rs439401 (1.56e-05), rs227725 (2.70e-05), rs405509 (3.10e-05), rs4389538 (3.86e-05).

Resultados estudio GWAS (hasta p-valor relajado) para fold 5: rs2075650 (1.00e-15), rs8106922 (5.26e-06), rs439401 (4.67e-05), rs1462060 (8.08e-05).

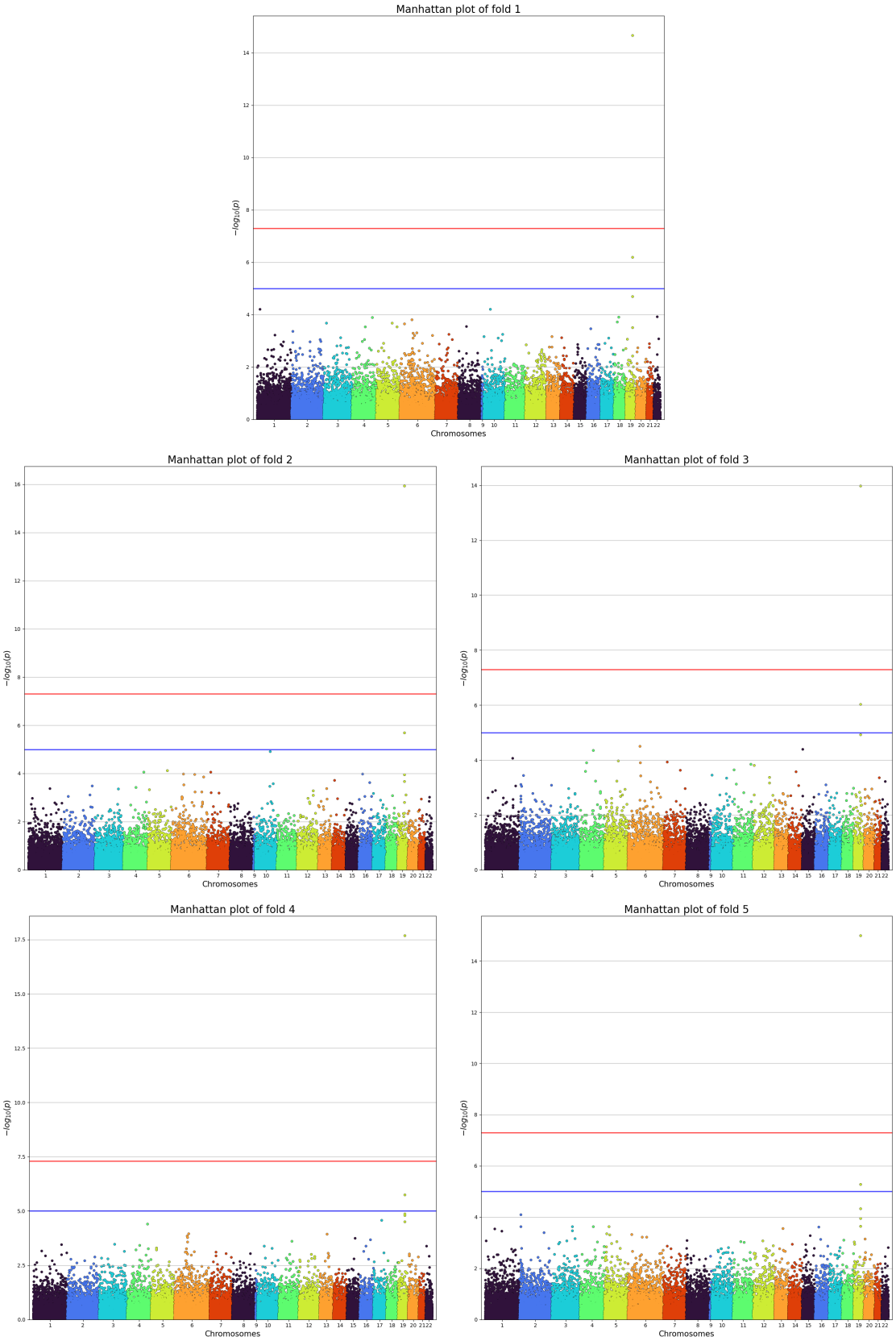

Figura B.70: Graficos de Manhattan obtenidos en cada fold, método de división por k-fold - Illumina Omni 2.5M BeadChip

Umbral clásico

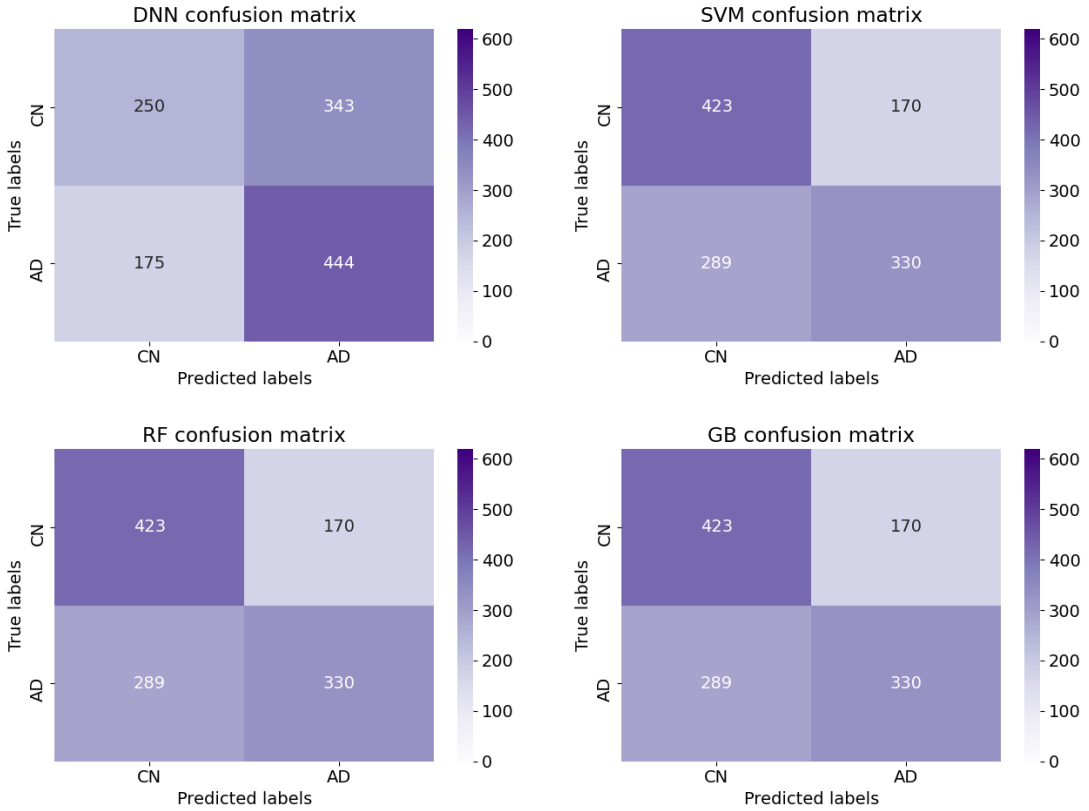

Figura B.71: Matrices de confusión, método de división por k-fold, umbral p-valor tradicional

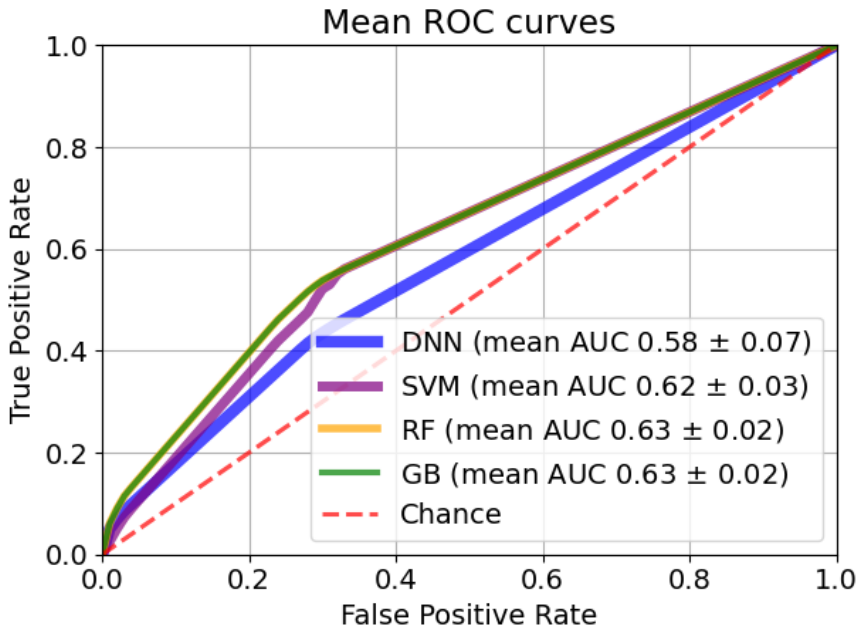

Figura B.72: Curvas ROC, método de división por k-fold, umbral p-valor tradicional

Umbral sugestivo

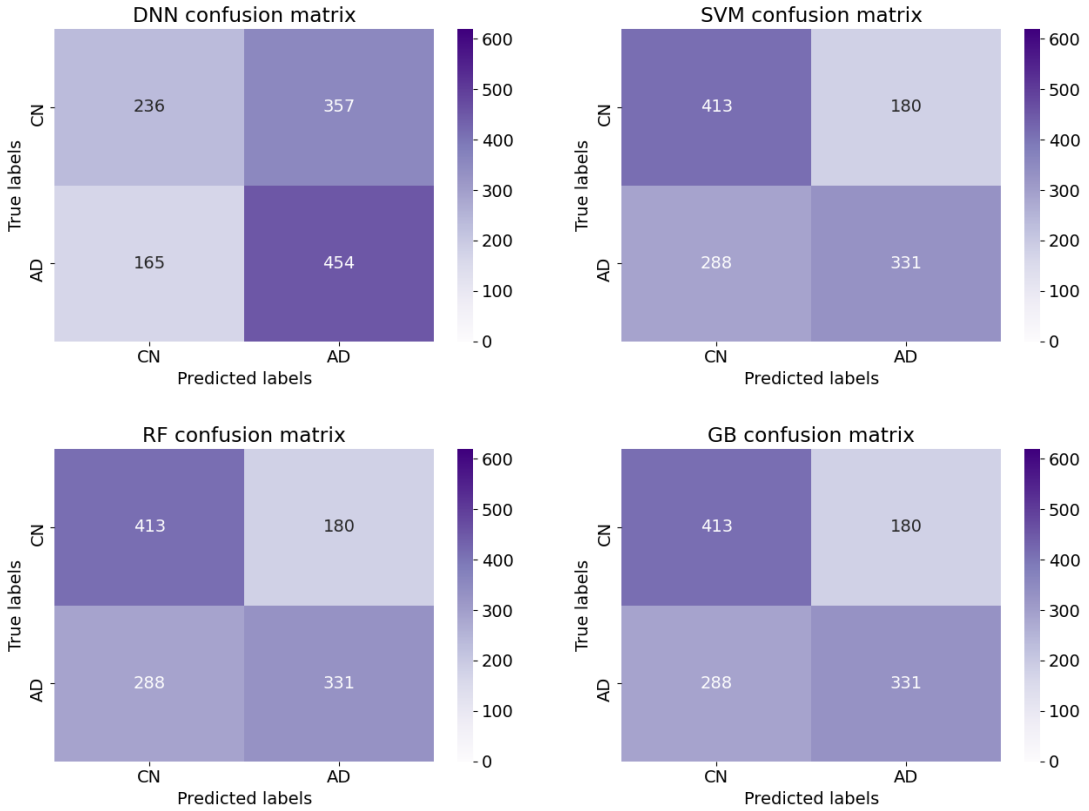

Figura B.73: Matrices de confusión, método de división por k-fold, umbral p-valor sugestivo

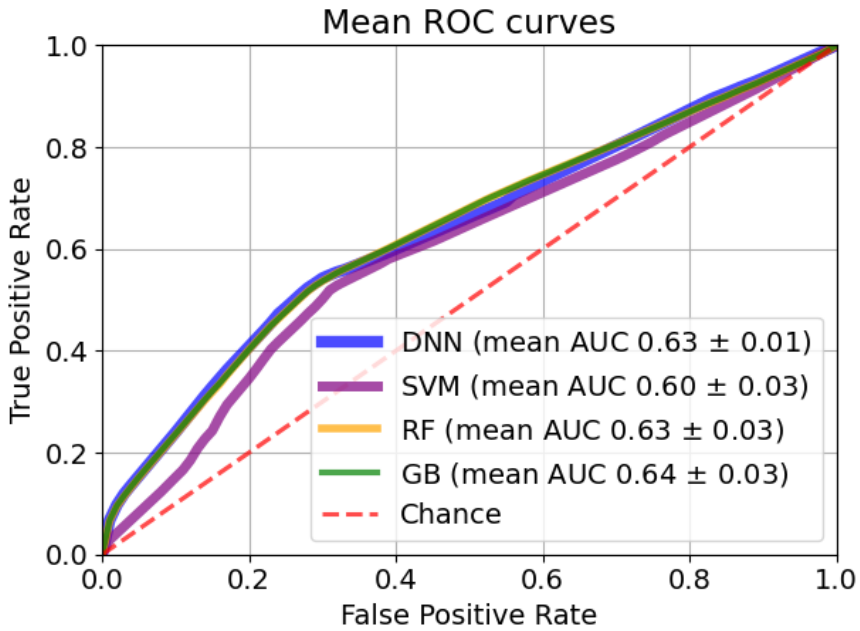

Figura B.74: Curvas ROC, método de división por k-fold, umbral p-valor sugestivo

Umbral relajado

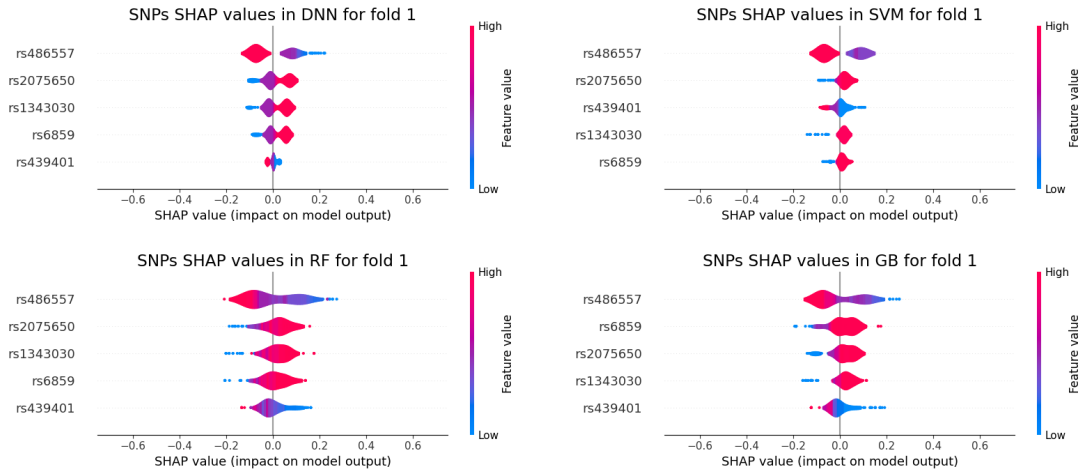

Figura B.75: Valores SHAP fold 1, método de división por k fold, umbral p-valor relajado

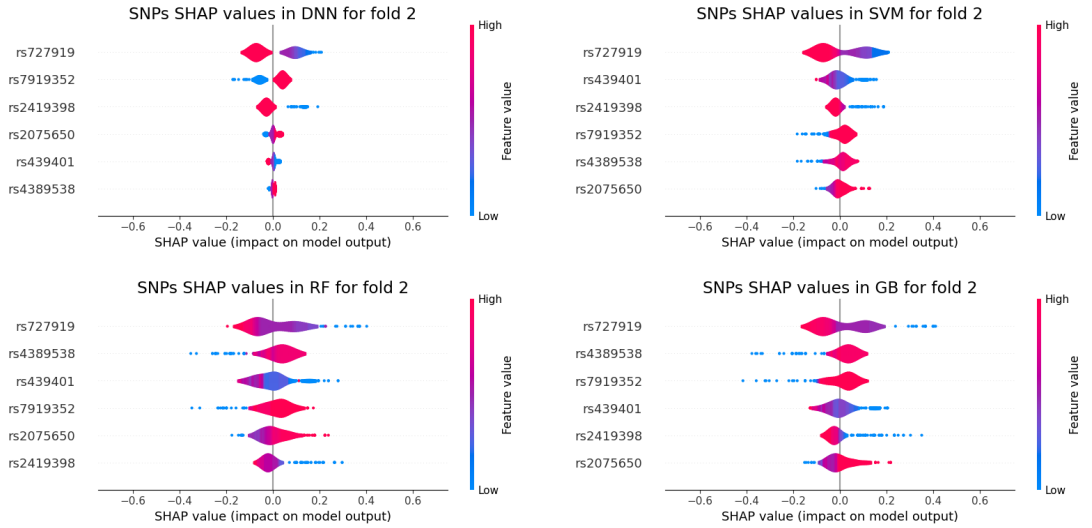

Figura B.76: Valores SHAP fold 2, método de división por k fold, umbral p-valor relajado

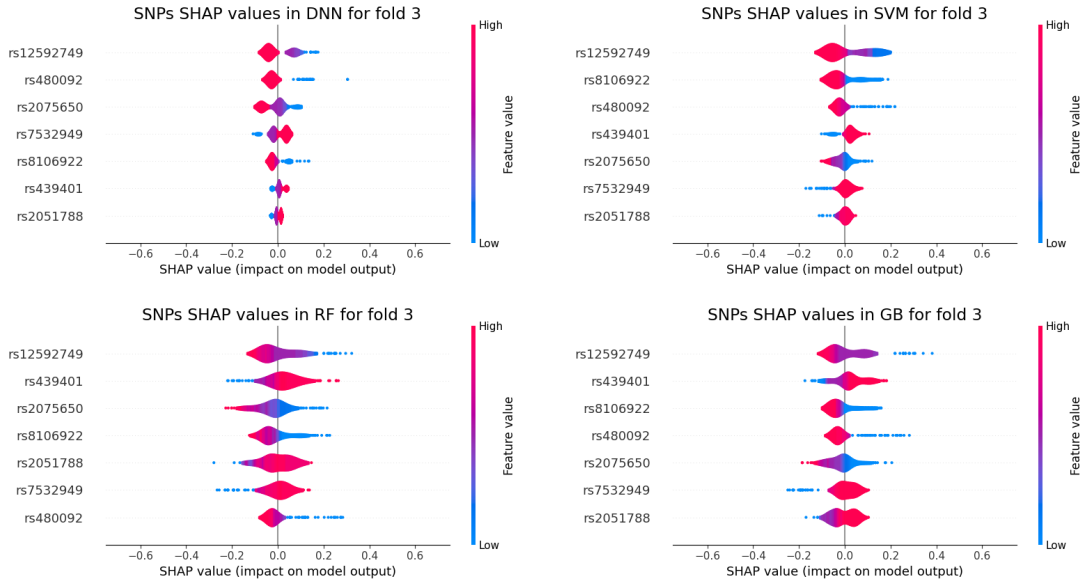

Figura B.77: Valores SHAP fold 3, método de división por k fold, umbral p-valor relajado

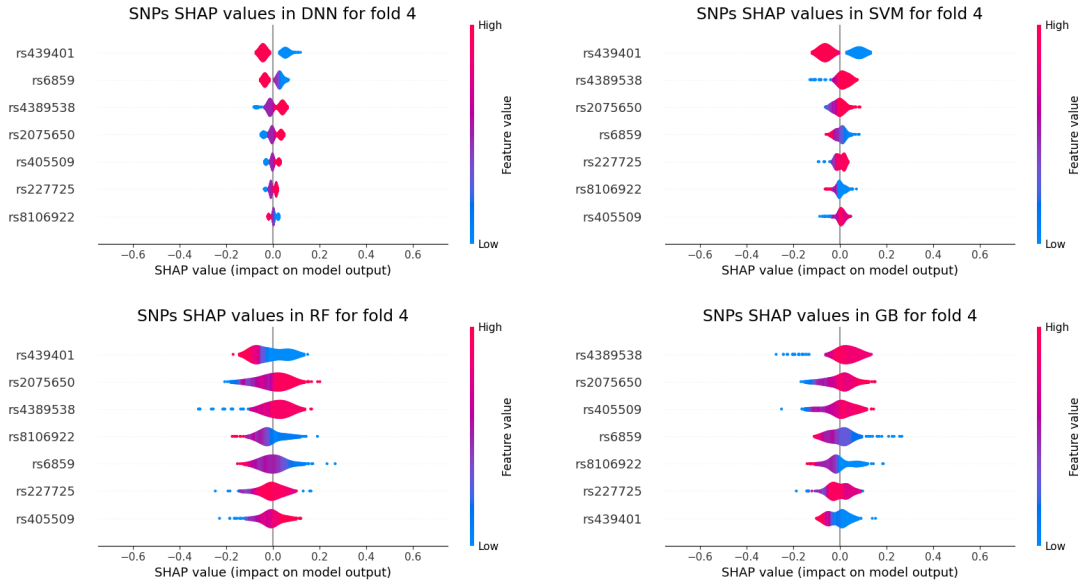

Figura B.78: Valores SHAP fold 4, método de división por k fold, umbral p-valor relajado

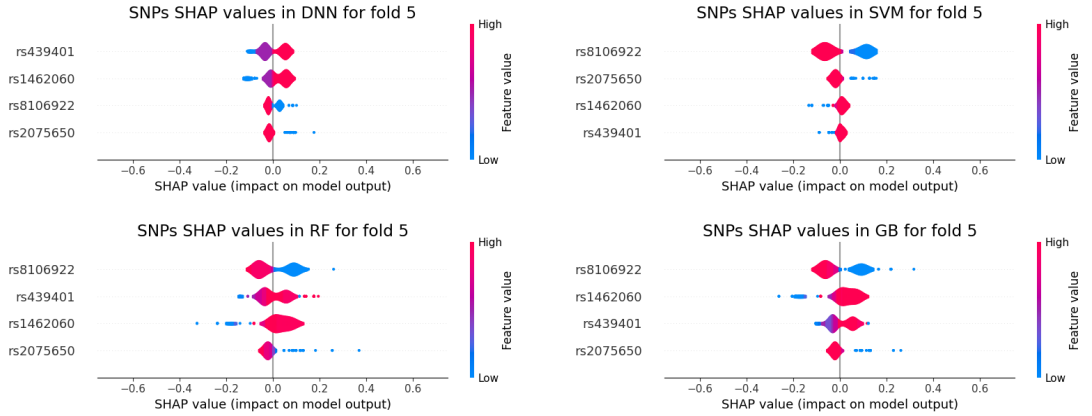

Figura B.79: Valores SHAP fold 5, método de división por k fold, umbral p-valor relajado

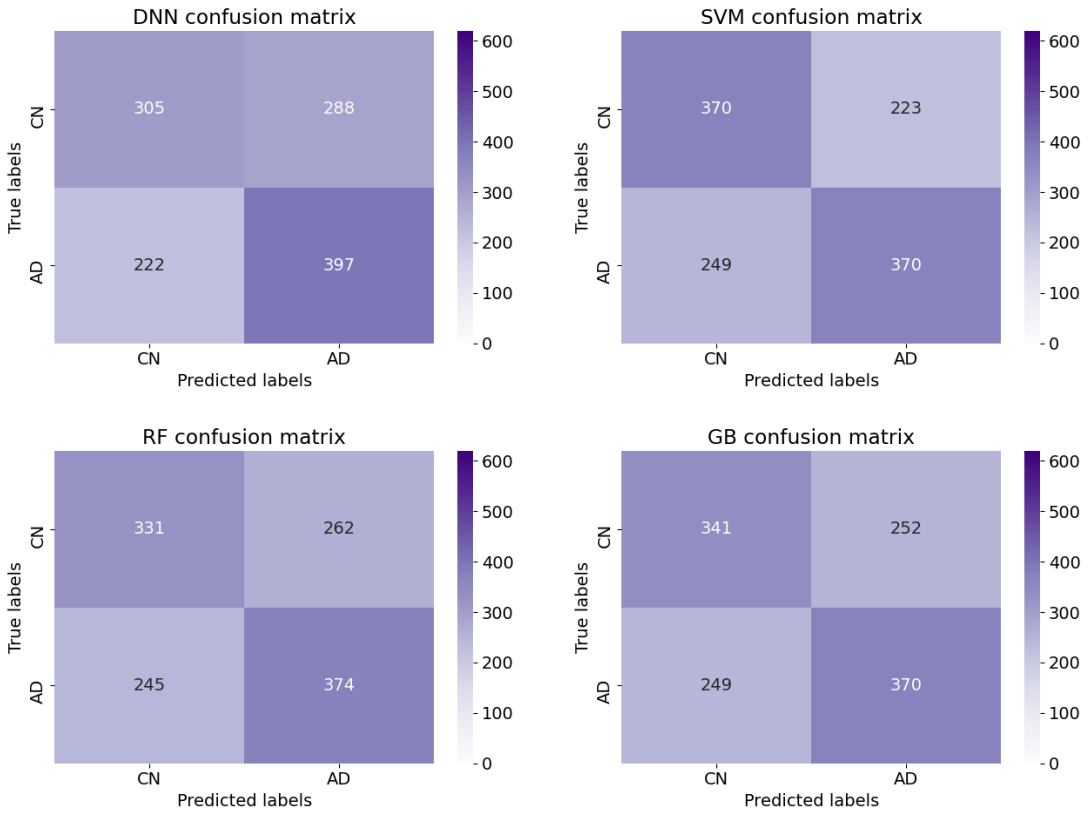

Figura B.80: Matrices de confusión, método de división por k fold, umbral p-valor relajado

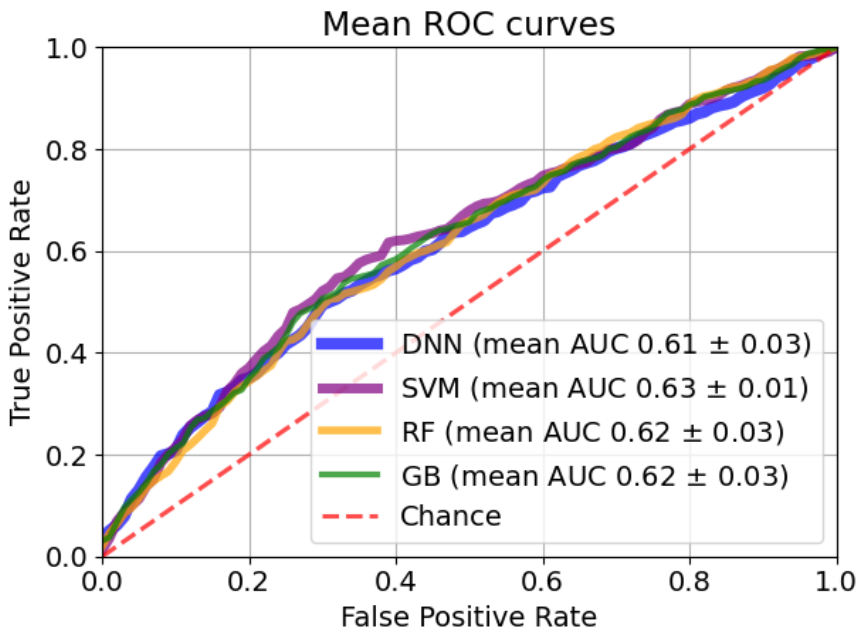

Figura B.81: Curvas ROC, método de división por k fold, umbral p-valor relajado

B.2.3. Método de filtración de datos

Resultados estudio GWAS (hasta p-valor relajado): rs2075650 (7.56e-20), rs439401 (6.27e-07), rs8106922 (3.74e-06), rs6859 (5.57e-06), rs480092 (4.30e-05), rs4389538 (7.72e-05), rs12920708 (8.05e-05),

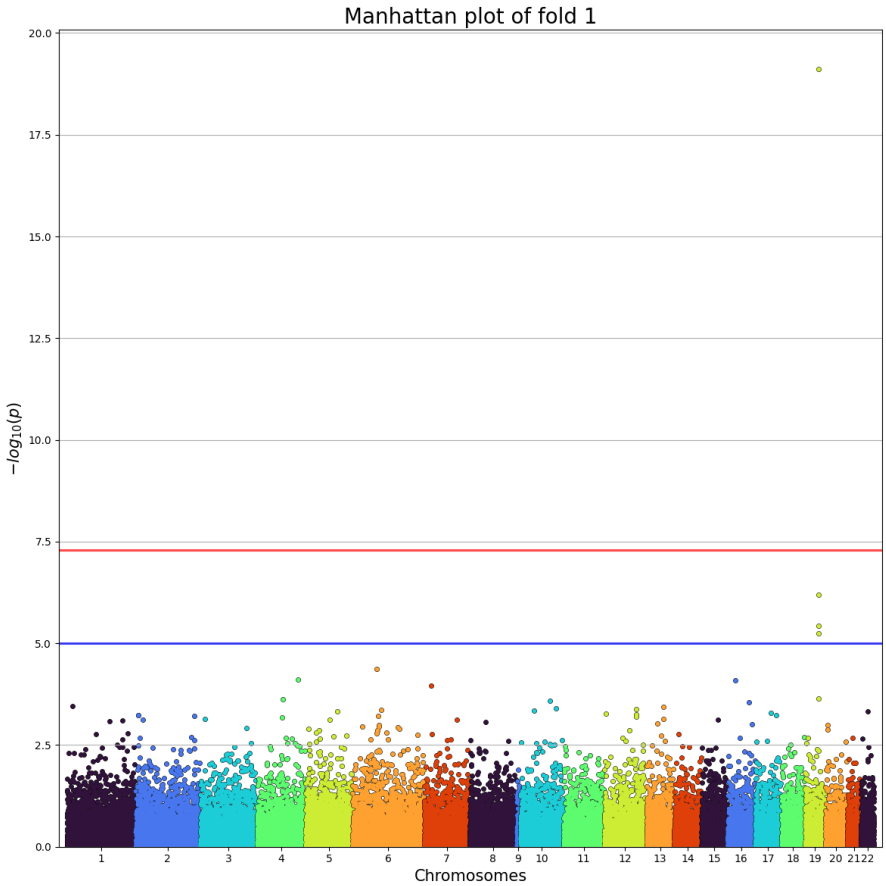

Figura B.82: Gráfico de Manhattan para el método de no división (filtración de datos)

Umbral clásico

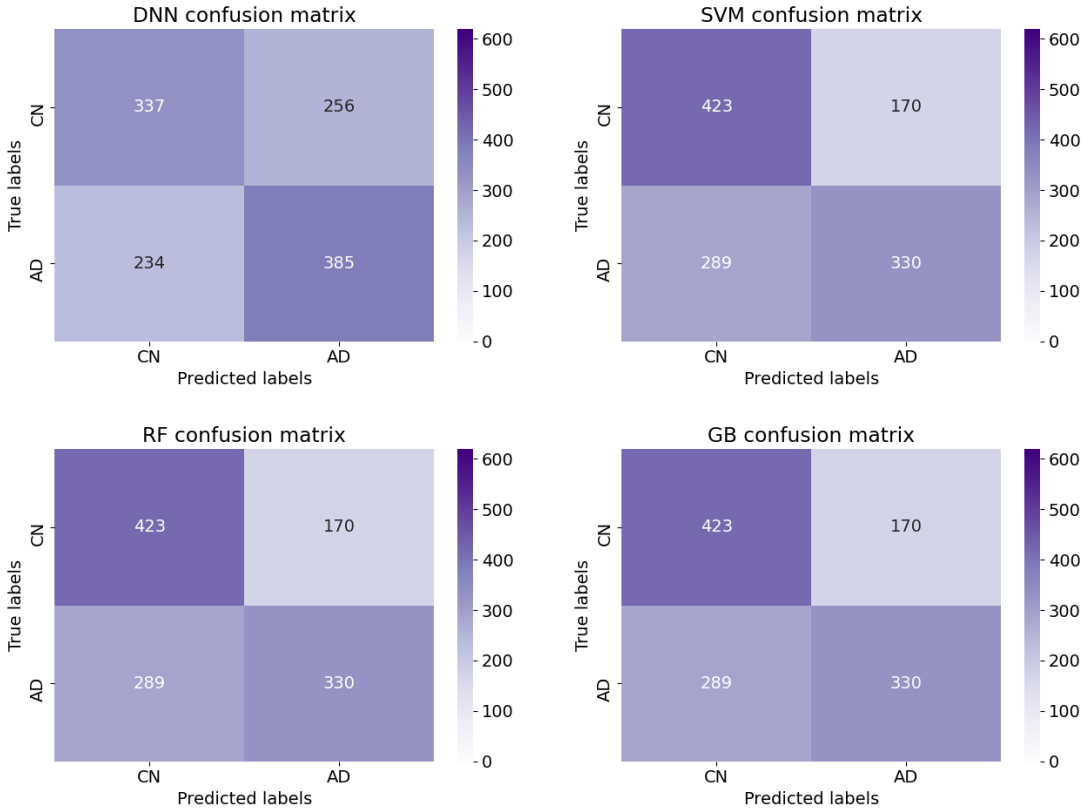

Figura B.83: Matrices de confusión, método de no división (filtración de datos), umbral p-valor tradicional

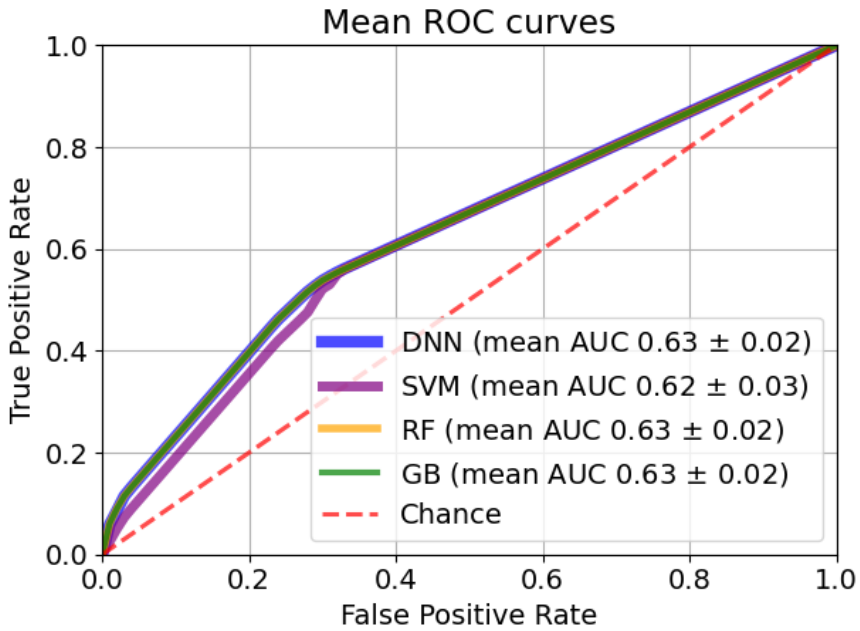

Figura B.84: Curvas ROC, método de no división (filtración de datos), umbral p-valor tradicional

Umbral sugestivo

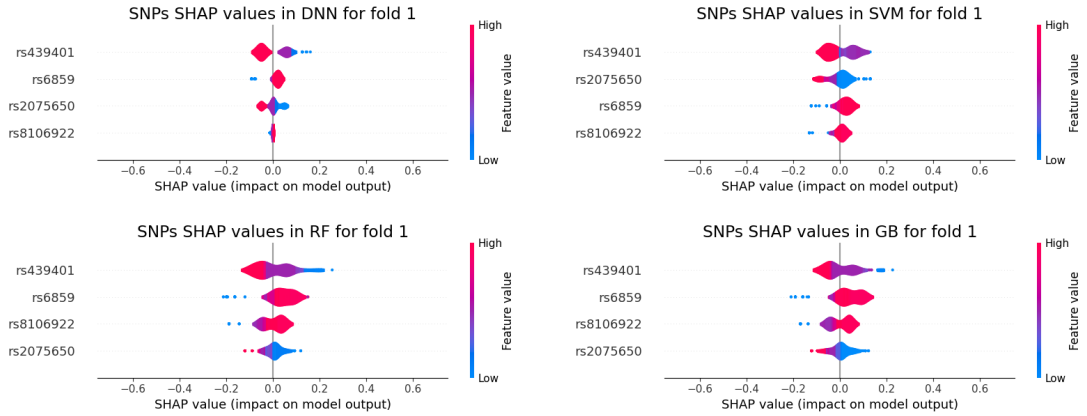

Figura B.85: Valores SHAP fold 1, método de no división (filtración de datos), umbral p-valor sugestivo

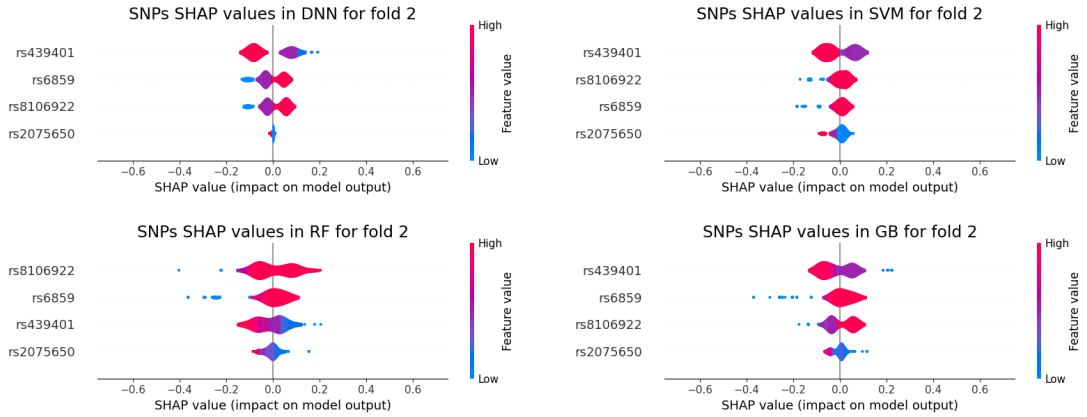

Figura B.86: Valores SHAP fold 2, método de no división (filtración de datos), umbral p-valor sugestivo

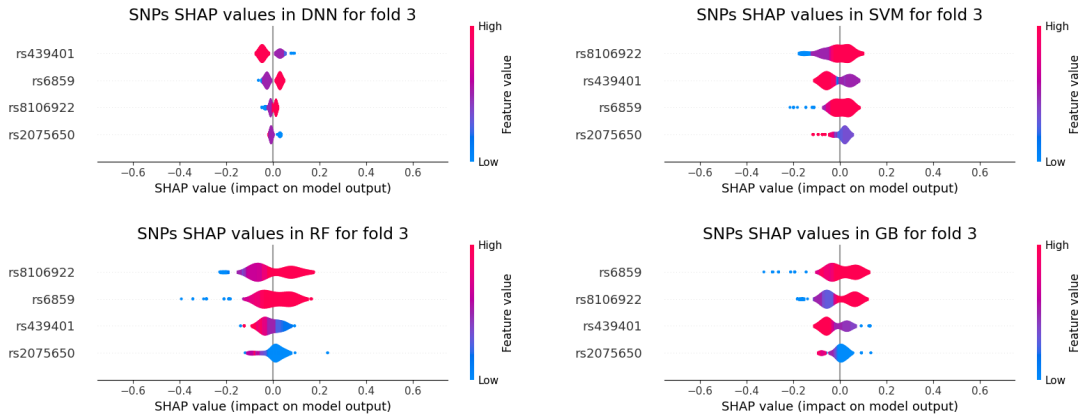

Figura B.87: Valores SHAP fold 3, método de no división (filtración de datos), umbral p-valor sugestivo

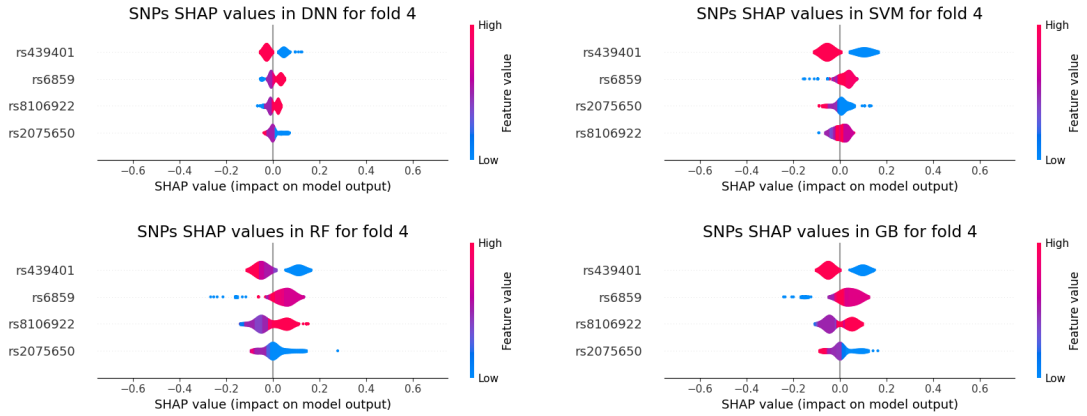

Figura B.88: Valores SHAP fold 4, método de no división (filtración de datos), umbral p-valor sugestivo

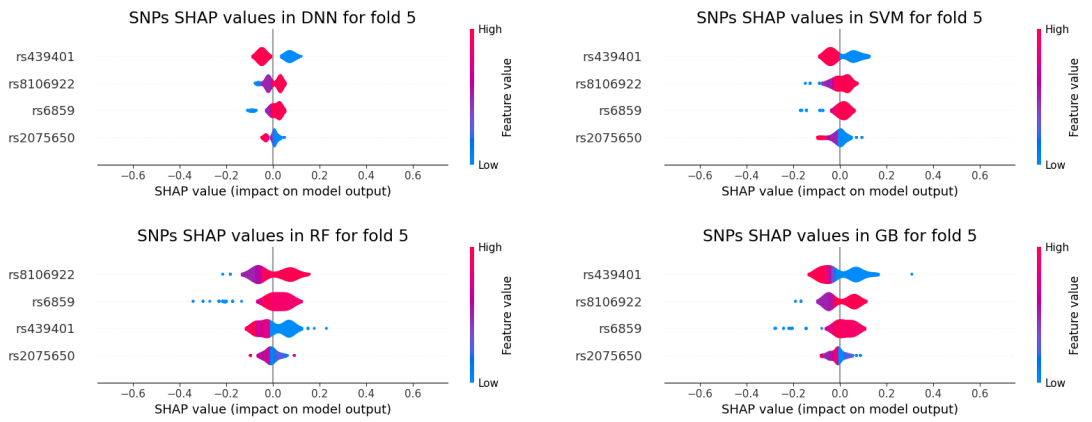

Figura B.89: Valores SHAP fold 5, método de no división (filtración de datos), umbral p-valor sugestivo

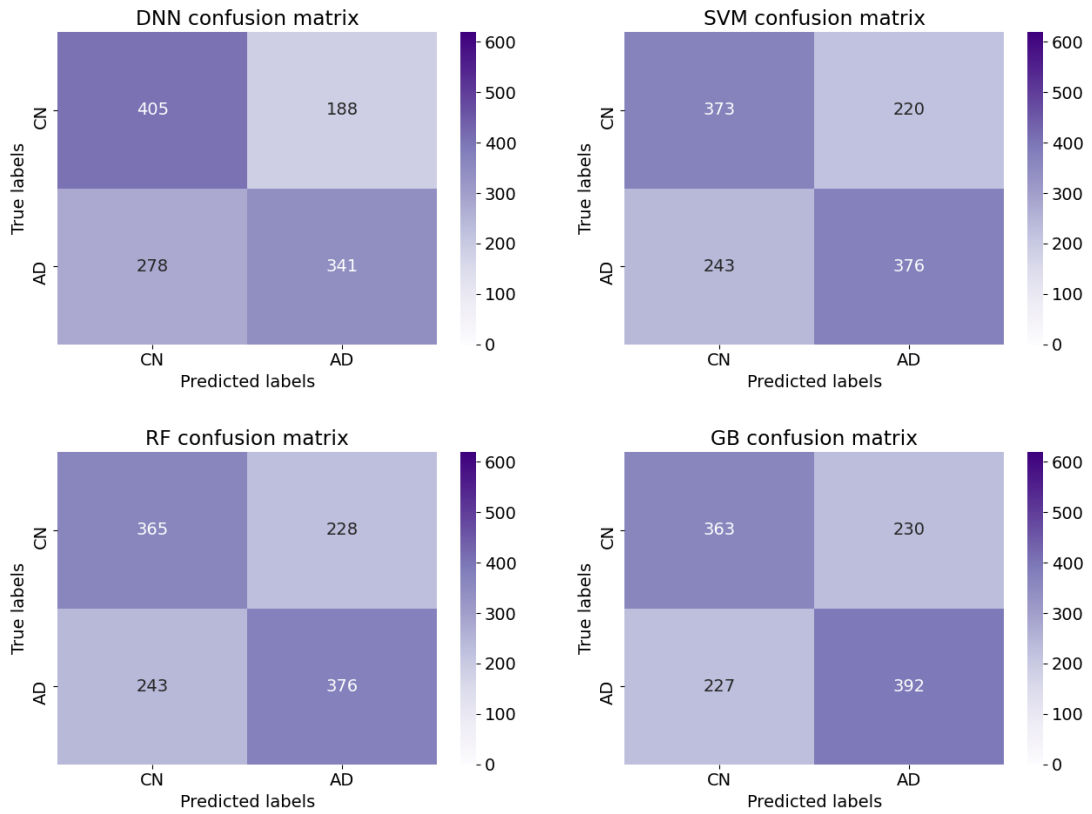

Figura B.90: Matrices de confusión, método de no división (filtración de datos), umbral p-valor sugestivo

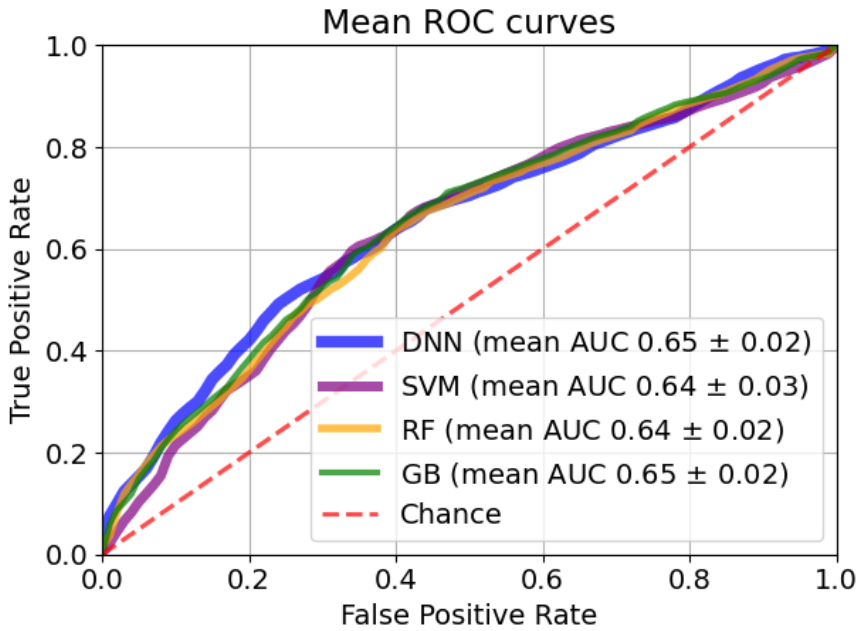

Figura B.91: Curvas ROC, método de no división (filtración de datos), umbral p-valor sugestivo

Umbral relajado

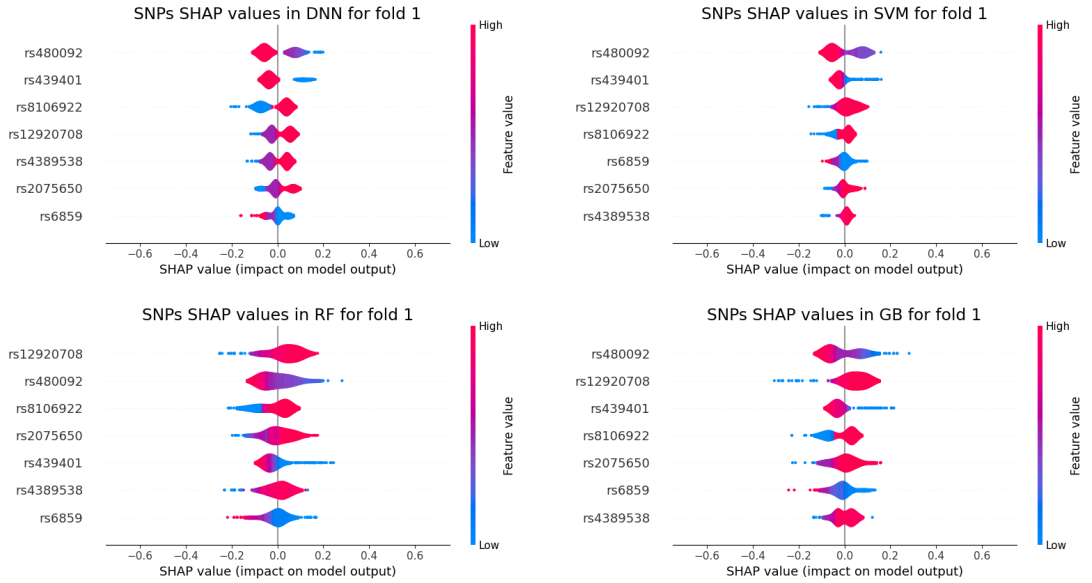

Figura B.92: Valores SHAP fold 1, método de filtración de datos, umbral p-valor relajado

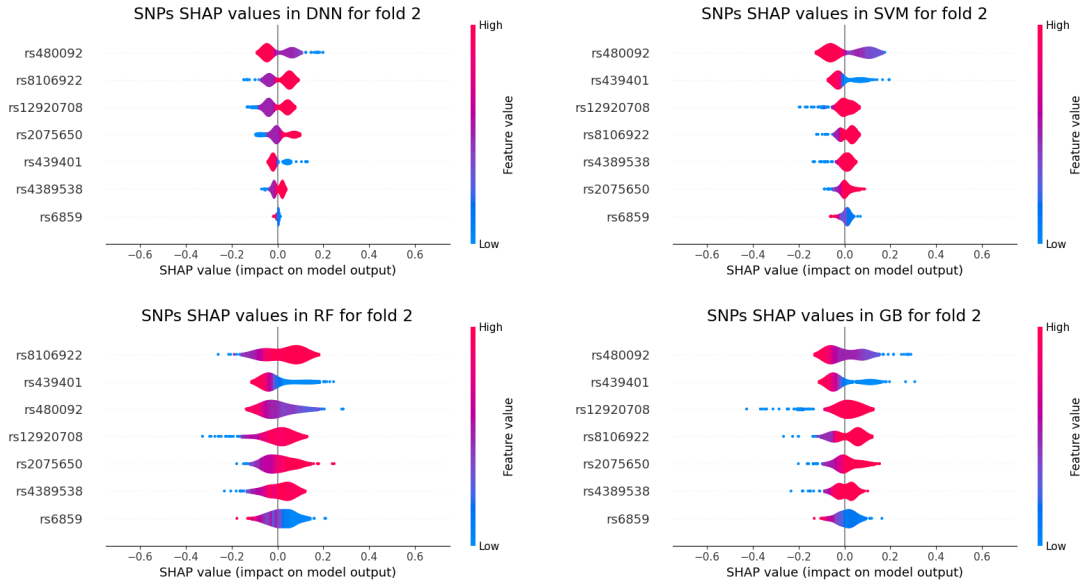

Figura B.93: Valores SHAP fold 2, método de filtración de datos, umbral p-valor relajado

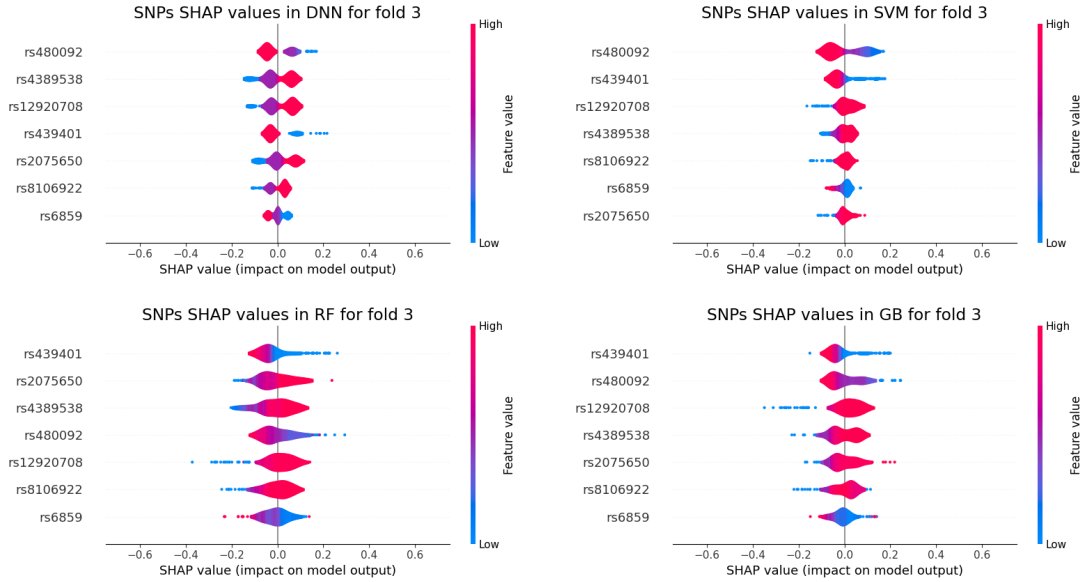

Figura B.94: Valores SHAP fold 3, método de filtración de datos, umbral p-valor relajado

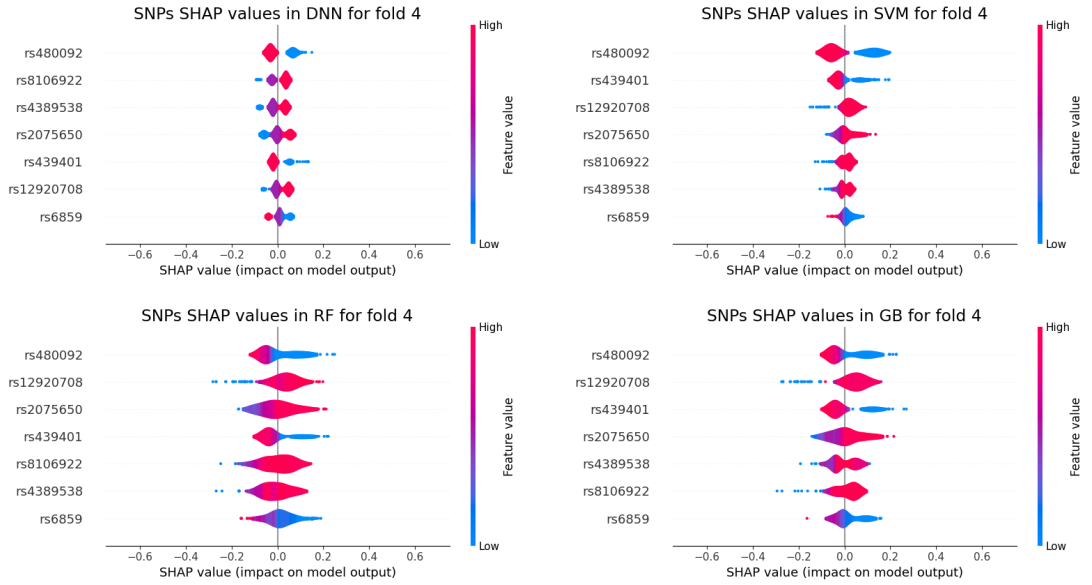

Figura B.95: Valores SHAP fold 4, método de filtración de datos, umbral p-valor relajado

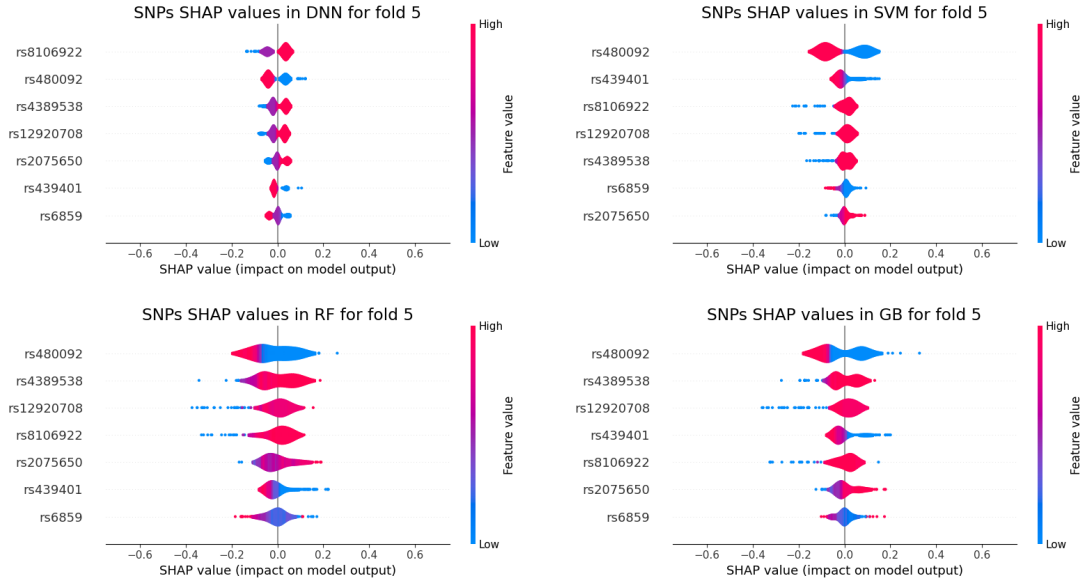

Figura B.96: Valores SHAP fold 5, método de filtración de datos, umbral p-valor relajado

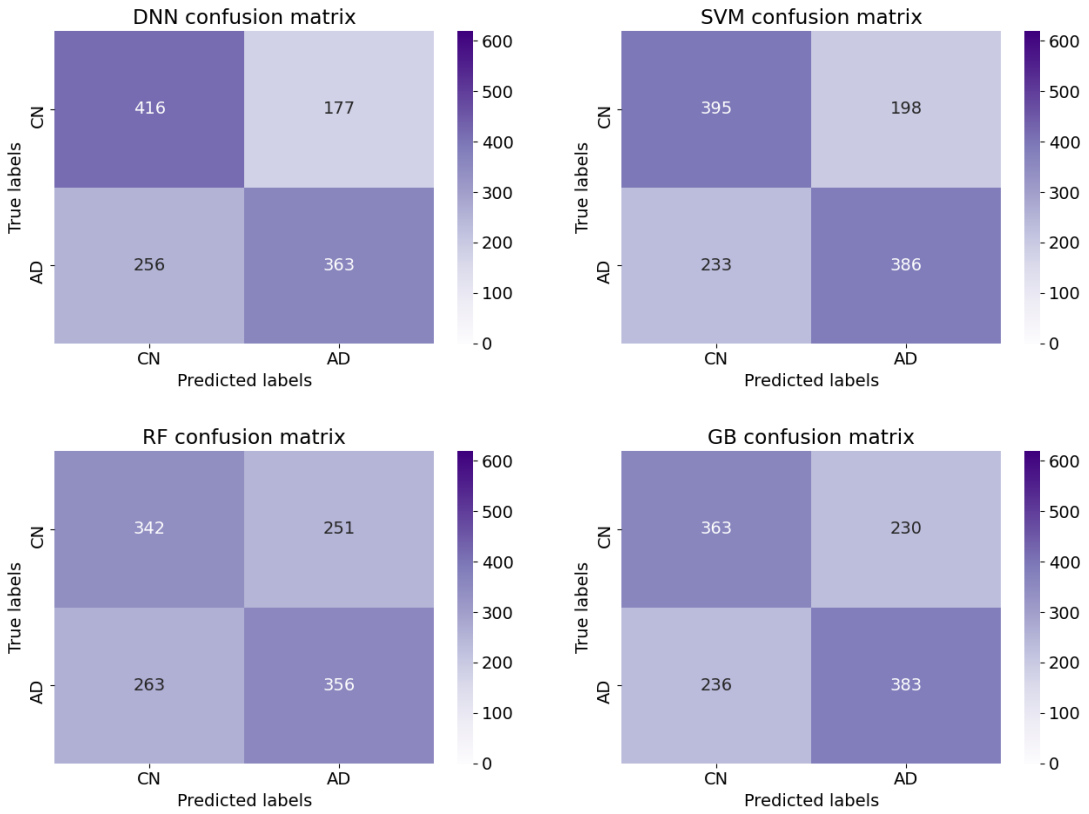

Figura B.97: Matrices de confusión, método de filtración de datos, umbral p-valor relajado

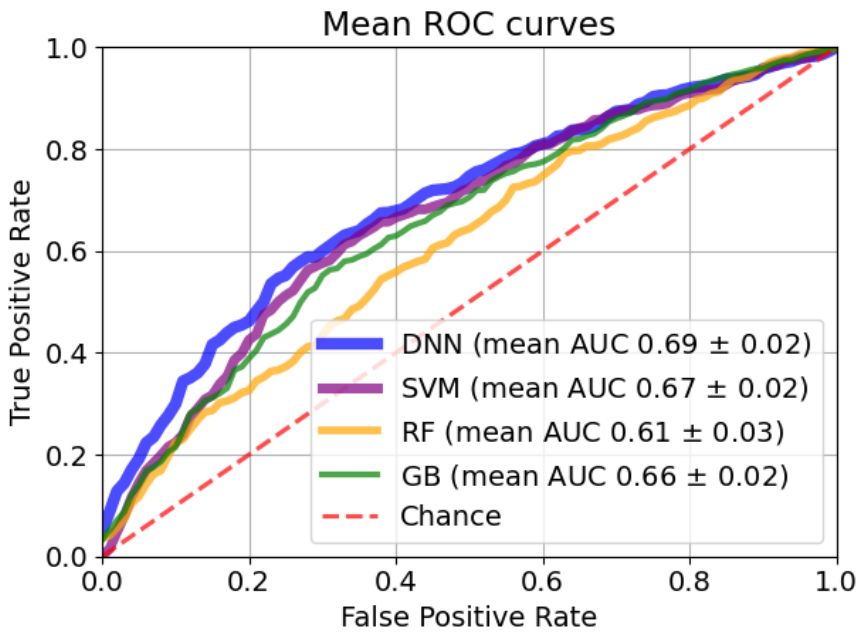

Figura B.98: Curvas ROC, método de filtración de datos, umbral p-valor relajado

C

Código desarrollado

Todo el código desarrollado para este trabajo se encuentra publicado en el portal de GitHub en el repositorio `https://github.com/alexx99gg/ADDLCGD`.