Universitat Oberta de Catalunya uoc.edu

Predicción de pacientes diabéticos, insulina-sensibles o insulina-resistentes aplicando técnicas de Inteligencia Artificial sobre genes obtenidos de un análisis de expresión diferencial

Jesús María González Martín
Máster Universitario en Bioinformática y Bioestadística
Área 4

Nombre Consultor/a: **Romina Astrid Rebrij**
Nombre Profesor/a responsable de la asignatura: **Ferran Briansó Castilla**

12/2021

Jesús Mª González Martín

Agradecimientos

En primer lugar, me gustaría agradecer al Dr. Francisco J. Rodríguez Esparragón por ayudarme con las dudas surgidas, no solo en este TFM, sino con las asignaturas asociadas a la biología cursadas en este máster.

En segundo lugar, a mi aita, ya fallecido y a mi ama, por educarnos en principios y valores. A mis hermanos, José y Carmelo, y Natalia por el apoyo recibido en los momentos difíciles. A mi hijo Asier, por lo mucho que aprendo de él cada día.

Jesús Mª González Martín

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción de pacientes diabéticos, insulina-sensibles o insulina-resistentes aplicando técnicas de Inteligencia Artificial sobre genes obtenidos de un análisis de expresión diferencial.*
Nombre del autor:	*Jesús Mª González Martín*
Nombre del consultor/a:	*Romina Astrid Rebrij*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 4*
Idioma del trabajo:	Español
Número de créditos:	15
Palabras clave	*Diabetes, Insulina, microarrays, machine learning, Bioconductor, R.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En la actualidad, 463 millones de adultos tienen diabetes y 374 millones presentan intolerancia a la glucosa. La insulina es una potente hormona pleiotrópica que afecta a los procesos como el crecimiento celular, el gasto energético y el metabolismo de carbohidratos, lípidos y proteínas. Por otra parte, el músculo esquelético es el sitio principal para la eliminación de la glucosa insulinodependiente. Los mecanismos moleculares por los que la insulina regula el metabolismo muscular y los defectos subyacentes que causan la resistencia a la insulina no se han dilucidado por completo. El objetivo de este estudio es realizar un análisis de datos de microarrays para encontrar genes diferencialmente expresados. El análisis se ha basado en los datos de un estudio depositado en Gene Expresion Omnibus (GEO) con identificador “GSE22309” y cuyo título es “Expression data from human skeletal muscle”. Los datos seleccionados contienen muestras de tres tipos de pacientes después de realizar un tratamiento con insulina: pacientes con diabetes (DB), pacientes sensibles a la insulina (IS) y pacientes resistentes a la insulina (IR). Una vez obtenidos los 20 genes expresados de forma diferencial entre las tres comparaciones posibles (DB vs IS, DB vs IR y IS vs IR), se ha utilizado este conjunto de datos para elaborar modelos predictivos a través de técnicas de Machine Learning para clasificar a los pacientes respecto de las tres categorías comentadas previamente. Todas las técnicas utilizadas

presentan una exactitud superior al 80%, alcanzando casi el 90% al unificar las categorías IR y DB.

Abstract (in English, 250 words or less):

Currently, 463 million adults have diabetes and 374 million have impaired glucose tolerance. Insulin is a powerful pleiotropic hormone that affects processes such as cell growth, energy expenditure, and carbohydrate, lipid, and protein metabolism. On the other hand, skeletal muscle is the main site for insulin-dependent glucose excretion. The molecular mechanisms by which insulin regulates muscle metabolism and the underlying defects that cause insulin resistance have not been fully elucidated. The objective of this study is to perform an analysis of microarray data to find differentially expressed genes. The analysis has been based on the data of a study deposited in Gene Expression Omnibus (GEO) with identifier "GSE22309" and whose title is "Human skeletal muscle expression data". The selected data contains samples from three types of patients after taking insulin treatment: patients with diabetes (DB), patients with insulin sensitivity (IS), and patients with insulin resistance (IR). Once the 20 genes expressed differentially between the three possible comparisons were obtained (DB vs IS, DB vs IR and IS vs IR), this data set has been used to develop predictive models through Machine Learning techniques to classify patients with respect to the three categories mentioned previously. All the techniques used present an accuracy superior to 80%, reaching almost 90% when unifying the categories IR and DB.

Índice

Lista de figuras

Lista de tablas

1 Resumen

1.1 Antecedentes

En la actualidad, alrededor de 463 millones de adultos entre 20 y 79 años tienen diabetes. Esto representa el 9.3% de la población mundial en este grupo de edad. Se prevé que la cantidad total aumente a 578 millones (10.2%) para 2030 y a 700 millones (10.9%) para 2045[1].

La diabetes tipo 2 (DM2), que representa el 90-95% de todas las diabetes, se caracteriza esencialmente por una disfunción de las células β pancreáticas y resistencia a la insulina, por lo que se requiere un aumento de la secreción de insulina para compensar[2]. El descubrimiento de la insulina en 1921 fue un Big Bang del que ha surgido un vasto y creciente universo de investigación sobre la acción y la resistencia de la insulina[3]. Ahora es posible una evaluación completa de la expresión diferencial en respuesta a la insulina utilizando la tecnología de microarrays. Este conocimiento podría mejorar nuestra comprensión de la acción de la insulina y cómo se integran las respuestas para mediar en el espectro de efectos hormonales.

A pesar de que actualmente hay técnicas más sofisticadas como Next Generation Sequencing (NGS), el análisis de datos de microarrays ha sido uno de los éxitos más importantes en la interacción entre la estadística y la bioinformática en las últimas dos décadas[4,5]. El análisis de datos de microarrays se puede realizar de diferentes formas utilizando diferentes herramientas, sin embargo, en este análisis se va a realizar el flujo de trabajo típico para analizar datos de microarrays utilizando paquetes de R y Bioconductor.

Por otra parte, en los últimos años, se han desarrollado técnicas de predicción asociadas a Machine Learning (K-nearest neighbors, Redes neuronales, Support Vector Machine, Random Forest) y técnicas de última generación como Deep Learning (Multilayer Perceptron) para elaborar modelos predictivos con una alta exactitud. En este análisis se pretenden utilizar estas técnicas con la finalidad de predecir a futuros pacientes que sean sensibles a la insulina, resistentes a la insulina o que terminen siendo diabéticos.

Entonces, debido a la importancia cada vez mayor de la prevalencia de diabetes en el mundo, y en particular en las islas Canarias, se pretende mediante técnicas avanzadas, obtener genes expresados de forma diferencial y posteriormente comprobar si estos genes son capaces de realizar buenas predicciones que puedan servir para clasificar a los pacientes de forma correcta.

1.2 Método

El presente trabajo se puede desglosar en dos partes. En la primera parte se realiza el análisis clásico de datos de microarrays a través de la plataforma de Bioconductor y el programa R, es decir, se parte de los archivos ".cel" depositados en Gene Expresion Omnibus (GEO) con identificador "GSE22309" y cuyo título es "Expression data from human skeletal muscle" (datos de expresión del músculo esquelético humano). Estos datos contienen muestras de tres tipos de pacientes (insulina sensibles (IS), insulina resistentes (IR) y diabéticos (DB)) antes y después de un tratamiento con insulina, aunque para este análisis únicamente se han seleccionado las 55 muestras tomadas después del tratamiento con insulina. Las 55 muestras se desglosan de la siguiente forma:

- 20 muestras de pacientes que son sensibles a la insulina
- 20 muestras de pacientes que son resistentes a la insulina
- 15 muestras de pacientes diabéticos

El flujo de trabajo comienza con la lectura del archivo “targets” que contiene las características de las 55 muestras y la lectura de los 55 archivos “.cel”. Posteriormente se realizan los pasos correspondientes al análisis de datos de microarray como: control de calidad, normalización, filtrado, selección de genes diferencialmente expresados, comparación de listas seleccionadas y análisis de importancia biológica. Dentro del apartado selección de genes diferencialmente expresados, se han seleccionado los 20 genes con presentaban un p-valor ajustado más pequeño en las comparaciones dos a dos (DB vs IR, DB vs IS y IR vs IS), es decir, en total 60 genes.

En la segunda parte del análisis, se ha tomado este conjunto de datos formado por 60 genes, junto la variable tipo de paciente (DB, IR o IS), que ha sido la variable objetivo y se han elaborado distintos modelos predictivos utilizando técnicas de Machine Learning (ML) para clasificar a los pacientes en las tres categorías descritas previamente.

1.3 Resultados

Tras la lectura de los archivos “.cel”, se parte de una matriz de datos de 409.600 genes de 55 pacientes, aunque tras el proceso de normalización de los datos, el conjunto de datos contiene 12.626 genes de los 55 pacientes pertenecientes a los tres grupos descritos con anterioridad (IS, IR y DB) y se realiza un modelo lineal para encontrar los genes que se expresan de forma diferencial entre las 3 comparaciones posibles:

- Se han obtenido 294 genes que se expresan de forma diferencial entre las categorías DB e IR.
    - Hay 139 genes poco expresados (down)
    - Hay 155 genes sobreexpresados (up)

- Se han obtenido 1.843 genes que se expresan de forma diferencial entre las categorías DB e IS.
    - Hay 731 genes poco expresados (down)
    - Hay 1112 genes sobreexpresados (up)

- Se han obtenido 863 genes que se expresan de forma diferencial entre las categorías DB e IS.
    - Hay 511 genes poco expresados (down)
    - Hay 352 genes sobreexpresados (up)

Tras realizar un análisis de significación biológica (gene enrichment analysis) con la herramienta Reactome, encontramos las siguientes relaciones a nivel de redes:

- En la comparación entre DB vs IR no se han encontrado redes en relación a los genes expresados de forma diferencial. Esto puede ser posible por el reducido número de genes expresados de forma diferencial, que eran 294.
- En la comparación entre DB vs IS se han encontrado las siguientes redes:

    - Signaling by ROBO receptors
    - Regulation of expresion of SLITs and ROBOs
    - Eukaryotic Translation Initiation
    - Cap−dependent Translation Initiation
    - L13a−mediated translational silencing of Ceruloplasmin expression

- En la comparación entre IR vs IS se han encontrado las siguientes redes:
    - Eukaryotic Translation Initiation
    - Cap−dependent Translation Initiation
    - GTP hydrolysis and joining of the 60S ribosomal subunit
    - L13a−mediated translational silencing of Ceruloplasmin expression
    - Regulation of expresion of SLITs and ROBOs

Resulta curioso que de las 5 redes que aparecen en ambas comparaciones, 4 de ellas coinciden en ambas.

Una vez terminado en análisis de datos ómicos, se han seleccionado los 20 genes con menor p-valor ajustado en las tres comparaciones realizadas (DB vs IR, DB vs IS y IR vs IS) y estos 60 genes se han utilizado como fichero de entrada para clasificar a los pacientes en su relación con la insulina o la diabetes tras aplicar diversas técnicas de ML. Todas las técnicas utilizadas presentan una exactitud (accuracy) superior al 80% y en particular, las técnicas de MLP y SVM-lineal presentan una exactitud superior al 90%. Tras unificar las categorías DB y IR y de esta forma tener una matriz de confusión de dimensiones 2*2, la exactitud es superior al 90 % en todas las técnicas exceptuando RF que se queda en el 89.39%.

1.4 Conclusiones

Se ha demostrado la validez de las herramientas utilizadas en el análisis de datos ómicos a través de Bioconductor y R, y las técnicas de ML para, por una parte, encontrar aquellos genes expresados de forma diferencial entre 3 grupos y posteriormente la buena predisposición de los 20 genes expresados de forma diferencial en cada una de las 3 comparaciones para elaborar modelos predictivos capaces de alcanzar una exactitud superior al 90%.

Es decir, partiendo de 60 genes, se podrá predecir con una exactitud superior al 90% si el paciente va a ser sensible a la insulina, resistente a la insulina o diabético.

Otra conclusión que se puede sacar del estudio, y viendo el número de genes expresados de forma diferencial en cada comparación y observando los "*heatmaps*" es la cercanía entre los grupos insulina-resistentes y diabéticos, es decir, aquellos pacientes que sean resistentes a la insulina, pues es muy probable que terminen siendo diabéticos.

1.5 Aportación

Por una parte, el presente trabajo pretende aportar una serie de genes (60 genes en particular) que se expresen de forma diferencial entre pacientes que sean sensibles a la insulina, resistentes a la insulina o diabéticos tras utilizar el análisis de datos de microarrays a través de la plataforma de Bioconductor y el programa R. Por otra parte, con estos 60 genes se pretende crear modelos predictivos utilizando técnicas de ML

capaces de predecir con una exactitud superior al 80% (superior al 90% si se tiene en cuenta la técnica de MLP) en cuál de las tres categorías comentadas previamente (IS, IR o DB) puede ser clasificado un nuevo paciente. Con el análisis de estos 60 genes, es posible predecir la clase a la que va a pertenecer el paciente con una exactitud superior al 90%.

Es decir, se pretende validar el proceso de búsqueda de genes diferencialmente expresados y que posteriormente sirvan para predecir una variable objetivo con una alta validez a través de técnicas de ML.

Por último, merece la pena mencionar, que los artículos publicados en pubmed asociados a estos datos, plantean otras hipótesis distintas a las planteadas en este análisis. En el artículo “The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle” únicamente se tienen en cuenta a los pacientes que son sensibles a la insulina antes y después de una infusión de insulina. Por otra parte, en el artículo “Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning” se tienen en cuenta a los pacientes que son sensibles a la insulina y resistentes a la insulina, pero no tiene en cuenta a los pacientes diabéticos. Por ello, este análisis se considera una nueva aportación respecto de los artículos seminales realizados con estos datos.

2 Introducción

2.1 Contexto y justificación del Trabajo

La Diabetes mellitus es una enfermedad crónica que se caracteriza por presentar un conjunto de trastornos metabólicos relativos a la aparición de hiperglucemia crónica, así como alteraciones en el metabolismo de los hidratos de carbono, las grasas y las proteínas, debido a la existencia de problemas en la secreción y/o acción de la insulina[6]. Esto se debe a que el páncreas no produce insulina en una cantidad suficiente o bien a que el propio organismo no es capaz de utilizar de forma adecuada la insulina que genera, y consecuentemente la glucosa no es asimilada por las células y permanece en la sangre, donde se produce un incremento de su concentración[7].

La Diabetes Mellitus tiene una elevada prevalencia e incidencia en la población, y puede ocasionar importantes problemas de salud, dando lugar a serias complicaciones como son la enfermedad cardiovascular, el ictus, la ceguera, la amputación de miembros inferiores, entre otros. En el caso de la diabetes gestacional, puede ocasionar problemas en el embarazo, tanto para la madre como para el feto o el recién nacido.

Algunas de las complicaciones que produce la diabetes se pueden evitar o demorar mediante el desarrollo de acciones preventivas y un buen control de la misma. En este sentido, existen diversos factores de riesgo que predisponen a las personas a desarrollar la enfermedad, y específicamente a las mujeres la DMG. Algunos de estos factores de riesgo no se pueden modificar, pero sobre otros es posible desarrollar acciones de prevención que eviten o retrasen su impacto en el desarrollo de la patología.

En la actualidad, la prevalencia de la diabetes, una enfermedad potencialmente mortal, está aumentando en todo el mundo, alrededor de 463 millones de adultos de entre 20 y 79 años tienen diabetes. Esto representa el 9,3% de la población mundial en este grupo de edad. Se prevé que la cantidad total aumente a 578 millones (10,2%) para 2030 y a 700 millones (10,9%) para 2045. El número calculado de adultos de entre 20 y 79 años con tolerancia anormal a la glucosa (TAG) es de 374 millones (7,5% de la población mundial en este grupo de edad). Se prevé que esta cifra aumente a 454 millones (8,0%) para 2030 y a 548 millones (8,6%) para 2045. Se calcula que 1,1 millones de niños y adolescentes (que no superan los 20 años de edad) tienen diabetes tipo 1 (no se disponen datos en relación a la cantidad de niños y adolescentes que tienen diabetes tipo 2). Se calcula que el gasto anual en salud a nivel mundial destinado a la diabetes será de 760 mil millones de USD. Se pronostica que para los años 2030 y 2045, el gasto alcanzará los 825 mil y 845 mil millones de USD, respectivamente[1].

En Europa, se estima que el número de adultos de entre 20 y 79 años con diabetes es de 59,3 millones, lo que representa el 8,9% de la población regional en este grupo de edad. Esto incluye a 24,2 millones de adultos con diabetes no diagnosticada. Se estima que otros 36,6 millones de adultos de entre 20 y 79 años, el 5,5% de la población regional en este grupo de edad, tiene TAG en 2019. En 2030, se pronostica que habrá 66 millones de adultos con diabetes y 39,7 millones con TAG en la región. Las predicciones para 2045 sugieren que esta cifra aumentará hasta 68,1 millones de personas con diabetes y 40,3 millones con TAG.

En España, unos 3.6 millones de adultos de entre 20 y 79 años presentaban diabetes en 2019, lo que representa el 10.5% de la población. Por otra parte, la prevalencia declarada de la diabetes en Canarias en la población de 15 y más años, es la más alta

de España, habiendo mantenido una tendencia creciente en los últimos años. En el año 2017, la prevalencia de la diabetes en Canarias asciende casi al 11%, 3.19 puntos porcentuales superior a la media española (7.82%.)[7].

Debido a ello, se pretende analizar los genes relacionados con las 3 posibilidades respecto del binomio insulina-diabetes:

- Insulina-sensibles (IS), la cual sería la situación perfecta. Es decir, cuando el páncreas solo se necesita segregar un poco de insulina para hacer descender los niveles de glucosa en sangre.
- Insulina-resistentes (IR), es decir, que cada vez se va necesitando más cantidad de insulina para que ésta haga su función correctamente ya que las células no hacen caso a las señales y no quieren absorber glucosa. Se puede considerar un estadio pre-diabético. Tener la insulina alta hace aumentar el almacenamiento de grasa, lo cual a la vez te hace más resistente a la insulina. Por eso la resistencia a la insulina es un factor de riesgo para desarrollar obesidad, enfermedades cardiovasculares y diabetes tipo 2.
- Diabéticos (DB), sería el peor estadio. La diabetes es una enfermedad en la que los niveles de glucosa de la sangre están muy altos. La glucosa proviene de los alimentos que consume. La insulina es una hormona que ayuda a que la glucosa entre a las células para suministrarles energía. En la diabetes tipo 1, el cuerpo no produce insulina. En la diabetes tipo 2, la más común, el cuerpo no produce o no usa la insulina de manera adecuada y sin suficiente insulina, la glucosa permanece en la sangre.

Realizando las comparaciones dos a dos entre estás 3 categorías (DBvsIR, DBsvIS y IRvsIS), se podría encontrar los genes que caracterizan cada una de las 3 situaciones. Una vez encontrados los genes que se expresan de forma diferencial, se pretende dar un enfoque práctico y por ello se pretende elaborar modelos predictivos mediante técnicas de Machine Learning (ML) con la finalidad de poder clasificar a los pacientes con la mayor antelación y validez posible sobre su posible estadio respecto de la insulina y diabetes.

2.2 Objetivos del Trabajo

Objetivo general:

Obtener un modelo predictivo utilizando técnicas de Inteligencia Artificial (IA) que permita clasificar a los pacientes en insulina-sensibles, insulina-resistentes o diabéticos a través de datos de expresión génica expresados de forma diferencial.

Objetivos específicos:

a) Obtener los genes expresados de forma diferencial entre las 3 categorías propuestas (DB, IS, IR).
b) Relacionar genes con la diabetes.
c) Crear un modelo predictivo partiendo de genes para clasificar a los pacientes entre las 3 categorías propuestas (DB, IS, IR). Para comprobar la bondad de los modelo se utilizaran matrices de confusión y la exactitud (accuracy).
d) Tras evaluar los resultados preliminares, y observando la similitud entre las categorías IR y DB, se decidió unificar ambas categorías y calcular una matriz de confusión 2*2, donde además de calcular la exactitud, también se ha

calculado la sensibilidad, especificidad, valor predictivo positivo (vpn) y valor predictivo negativo (vpn)

2.3 Enfoque y método seguido

Este proyecto consta de dos partes, por una parte, se ha realizado un análisis clásico de microarrays a través de la plataforma de Bioconductor[8] (https://www.bioconductor.org/) de R[9]. Los pasos seguidos en el proceso de análisis de microarrays han sido los siguientes: lectura del archivo con las características de las muestras (targets) y lectura de los 55 archivos tipo ".cel", control de calidad de los datos crudos, normalización de los datos mediante el método "rma", control de calidad de los datos normalizados, identificación de los genes diferencialmente expresados, filtraje, selección de genes diferencialmente expresados (DB vs IR; DB vs IS y IR vs IS), volcano-plots de los genes más relevantes de cada comparación. Los genes seleccionados como diferencialmente expresados se agruparon para buscar patrones comunes de expresión entre condiciones experimentales mediante "*mapas de calor*" o "*heatmaps*". Las listas de genes diferencialmente expresados se anotaron en diversas bases de datos (Entrez, Unigene, Gene Ontology, KEGG, etc.) utilizando los paquetes de anotación para microarrays de affymetrix disponibles en el proyecto Bioconductor. Para contribuir a la interpretación biológica de los resultados se realizó dos tipos de análisis de enriquecimiento[10,11] o "gene set analysis" que busca establecer si las categorías funcionales de los genes seleccionados aparecen entre estos genes con mayor o menor frecuencia que entre todos los del genoma. De ser así, se indica que la lista de genes se encuentra "enriquecida" en estas funcionalidades, o lo que es lo mismo que los procesos afectados por las diferencias son éstos.

Por una parte, se utiliza el análisis básico de enriquecimiento tal y como se describe en los trabajos de Falcon y Gentleman[10] implementado en el paquete GOstats[10] de Bioconductor. Los análisis de este tipo necesitan un número mínimo de genes para resultar fiables por lo que se incluirán en todos los genes con p–valores ajustados inferiores a 0.05 (sin filtrar por mínimo "fold–change"). Por otra parte, también se realiza el análisis básico de enriquecimiento implementado en el paquete ReactomePA[11] de Bioconductor. En este caso, y dado el reducido número de genes expresados de forma diferencial entre las categorías DB e IR, se introducen todos los genes en el análisis para estas dos categorías. En las otras dos comparaciones, DB vs IS e IR vs IS, se han incluido aquellos genes que presentaban un p-valor ajustado inferior a 0.05.

Además, en relación con el apartado "selección de genes diferencialmente expresados", se ha creado un conjunto de datos nuevo con los veinte genes que presentaban un p-valor ajustado más pequeño en cada una de las 3 comparaciones posibles (DB vs IR, DB vs IR e IR vs IS). A este conjunto de datos se la ha añadido la categoría de cada uno de los 55 pacientes (DB, IR o IS) y de esta forma se ha creado un conjunto de datos de 55 pacientes con 60 genes (variables independientes) y una variable que describe estado del paciente (DB, IR o IS) y que se convierte en la variable objetivo en la segunda parte del análisis.

En relación a este segunda parte del análisis, la relacionada con la predicción de la variable objetivo mediantes técnicas de ML, se ha seguido el procedimiento clásico. Se parte de un conjunto de datos formado por 55 filas (15 pacientes DB, 20 pacientes IR y 20 pacientes IS) y 60 variables independientes (los genes expresados de forma diferencial) y una variable objetivo. En un primer momento, se realiza un análisis de componentes principales (ACP) de las 60 variables para ver como se agrupan. Y se

toman las siete variables obtenidas en este ACP que presentan un autovalor igual o mayor a uno y que se utilizarán posteriormente como conjunto de datos de entrada para predecir también la variable objetivo a través de una red neuronal.

Como se van a utilizar técnicas de ML, el conjunto de datos de entrada se divide en datos de entrenamiento (train), formado por el 70% de los 55 registros, y el 30% restante, se utiliza como datos de validación (test). Las técnicas de ML utilizadas han sido:

- Multilayer Preceptron (MLP) con dos capas ocultas
- K-nearest Neighbour (KNN)
- Redes neuronales con dos capas ocultas (ANN)
- Redes neuronales con las 7 variables obtenidas en el ACP con dos capas ocultas
- Support Vector Machine Lineal (SVM-lineal)
- Support Vector Machine Gaussiano (SVM-gaussiano)
- Random forest (RF) con 500 árboles y 7 variables en cada división

Además, para obtener unos resultados robustos y no debidos al azar, el proceso de partición de los datos en datos de entrenamiento y datos de validación y ejecución de las distintas técnicas se repite 1.000 veces y se almacenan los resultados de las 1.000 ejecuciones.

Una vez realizadas las 1.000 ejecuciones, se calculan las medidas de rendimiento para comprobar la validez de las distintas técnicas de dos formas posibles:

1. Acumulación de los resultados de cada una de las 1000 ejecuciones:
    a. En cada una de las 1.000 ejecuciones se han ido acumulando los datos de la matriz de confusión de cada una de las ejecuciones y por lo tanto se obtiene una matriz de confusión final de dimensiones 3*3 (IS, IR y DB) que contiene 17.000 registros (1.000 * 17 datos de validación). Sobre esta matriz acumulada de 3*3 se calcula la exactitud (accuracy).
    b. Partiendo de la matriz acumulada del apartado anterior, se unifican las categorías DB y IR, dada su proximidad y se construye una matriz de confusión de 17.000 registros, pero de dimensiones 2*2. En este matriz de 2*2 y teniendo como resultado positivo la clase “IR-SB” se calculan las medidas de rendimiento exactitud, sensibilidad, especificidad, vpp y vpn.
2. Media de los resultados de cada una de las 1.000 ejecuciones:
    a. En cada una de las 1.000 ejecuciones se ha calculado la exactitud de la matriz 3*3 (IS, IR y DB) y la exactitud, sensibilidad, especificidad, vpp y vpn para la matriz de confusión 2*2 tras unificar las categorías IR y DB. Entonces, tras las 1.000 ejecuciones se calcula la exactitud para la matriz de confusión 3*3 como la media de las 1.000 ejecuciones y para la matriz 2*2, se calculan las medidas de rendimiento de igual forma, como la medias de los resultados de las 1.000 ejecuciones.

Se detalla una figura resumen de los procesos ejecutados:

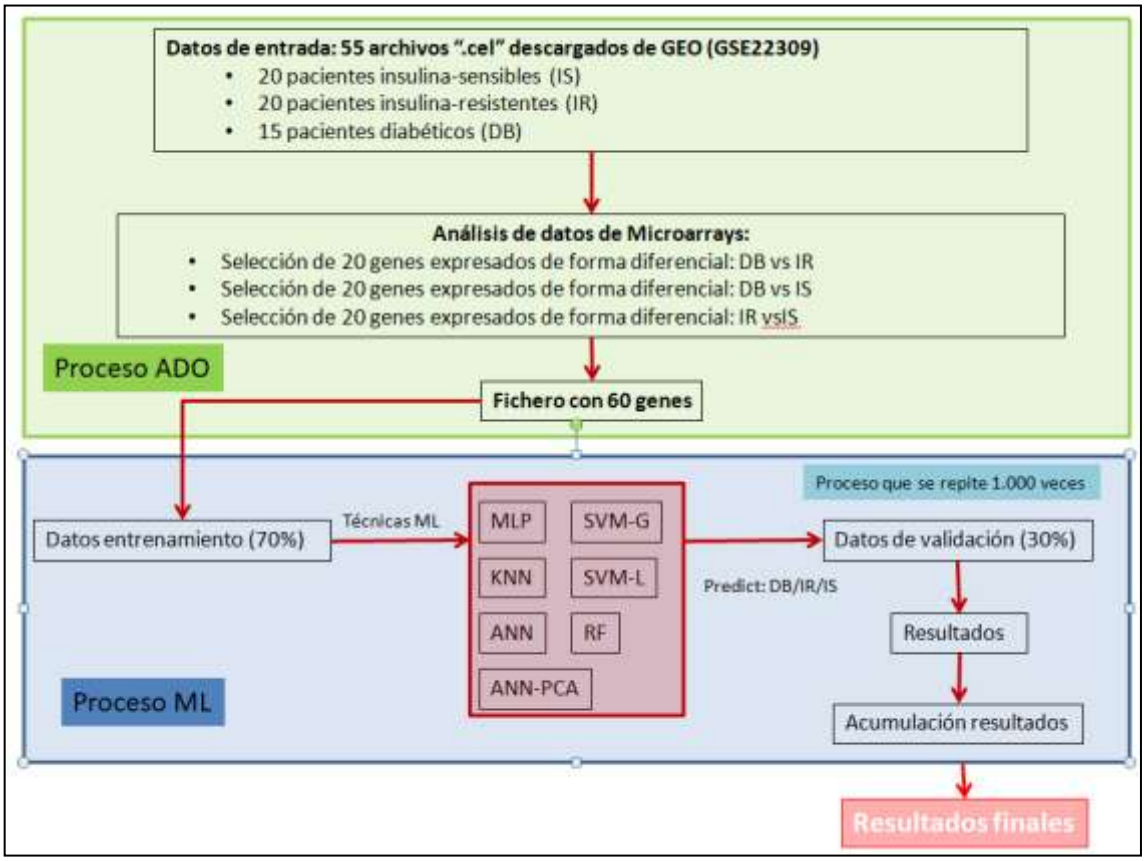

Figura 1. Resumen del proceso ejecutado.

2.4 Planificación del Trabajo

Para la realización de este trabajo se ha utilizado un ordenador cuyo sistema operativo es Windows 10 Pro 64 bits. En relación al software, se han utilizado los programas de software libre R (versión 4.1.2), R Markdown a través de RStudio IDE (versión 2021.09.1+372) y la plataforma asociada Bioconductor (versión 3.13). Sin olvidarnos de una conexión a Internet.

Los paquetes relacionados con Bioconductor han sido:

- Biobase (versión 2.52.0)
- pd.hg.u95a (versión 3.12.0)
- oligoClasses (versión 1.54.0)
- oligo (versión 1.56.0)
- affyio (versión 1.62.0)
- arrayQualityMetrics (versión 3.48.0)
- genefilter (versión 1.74.1)
- limma (versión 3.48.3)
- annotate (1.70.0)
- GOstats (versión 2.58.0)
- GO.db (versión 3.13.0)
- DO.db (versión 2.9)
- ReactomePA (versión 1.36.0)
- org.Hs.eg.db (versión 3.13)

Los paquetes relacionados con Machine Learning han sido:

- Keras (versión 2.7.0)
- Class (versión 7.3-19)
- Neuralnet (versión 1.44.2)
- Kernlab (versión 0.9-29)
- randomForest (versión 4.6-14)
- caret (6.0-90)
- ggplot2 (3.3.5)

Las tareas realizadas han sido:

- Búsqueda de la información a analizar, es decir, selección del trabajo a realizar
- Definición de los objetivos del trabajo
- Plan de trabajo
- Desarrollo del trabajo:
    - Parte de **Análisis de datos ómicos (ADO):**
        - Encontrar genes:
            - Descargar los datos en formato .CEL de la plataforma GEO[12].
            - Descripción de los archivos .CEL según aparece en GEO.
            - Unificar el archivo targets que contiene las características de los datos con los ficheros .CEL.
            - Realizar una exploración y control de calidad de los datos crudos.
            - Realizar una exploración y control de calidad de los datos normalizados.
            - Identificación de los genes diferencialmente expresados.
            - Filtraje. Selección de los genes que presentan variabilidad.
            - Selección de genes diferencialmente expresados en las comparaciones realizadas dos a dos:
                - DB vs IR
                - DB vs IS
                - IR vs IS
            - Volcano plots.
        - Relación genes-diabetes:
            - Anotación de los genes diferencialmente expresados.
            - Comparaciones múltiples.
            - Visualización de los perfiles de expresión.
            - Análisis de significación biológica (gene ontology y Reactome).

    - Parte de **Machine Learning:**
        - Modelo predictivo:

- Lectura del conjunto de datos obtenido en el proceso de ADO y factorización de la variable objetivo (3 categorías: insulina sensible, insulina resistente y diabéticos).
- Análisis gráficos de las variables originales.
- Análisis de componentes principales del conjunto de datos de entrada. Selección de 7 variables del ACP que presentaban un autovalor mayor o igual a 1.
- Partición del conjunto de datos en datos de entrenamiento y datos de validación.
- Ejecución de los datos de entrenamiento con las siguientes técnicas de ML:
    - Multilayer perceptron (MLP)
    - k-nearest neighbors (KNN)
    - Artificial neural network (ANN)
    - Artificial neural network (ANN) junto con análisis de componentes principales (ACP)
    - Support Vector Machine (SVM)
    - Random Forest (RF)
- Predicción de las respectivas técnicas con los datos de validación.
- Creación de las respectivas matrices de confusión con los datos de validación y las predicciones de las respectivas técnicas.
- Calculo de la exactitud (accuracy) de las respectivas técnicas para una matriz de confusión 3*3 y calculo de una matriz de confusión 2*2 unificando dos categorías de la matriz original (IR-DB) y posterior cálculo de la exactitud, sensibilidad, especificidad, VPP y VPN.
- Con la intención de obtener unos resultados validos, el proceso de partición de los datos y ejecución de los siguientes pasos se repite 1000 veces. Es decir, se han realizado 1000 ejecuciones.
- Conclusiones de los resultados.

- Imprevistos y actividades no previstas
- Documentar memoria
- Elaboración de la presentación
- Defensa pública

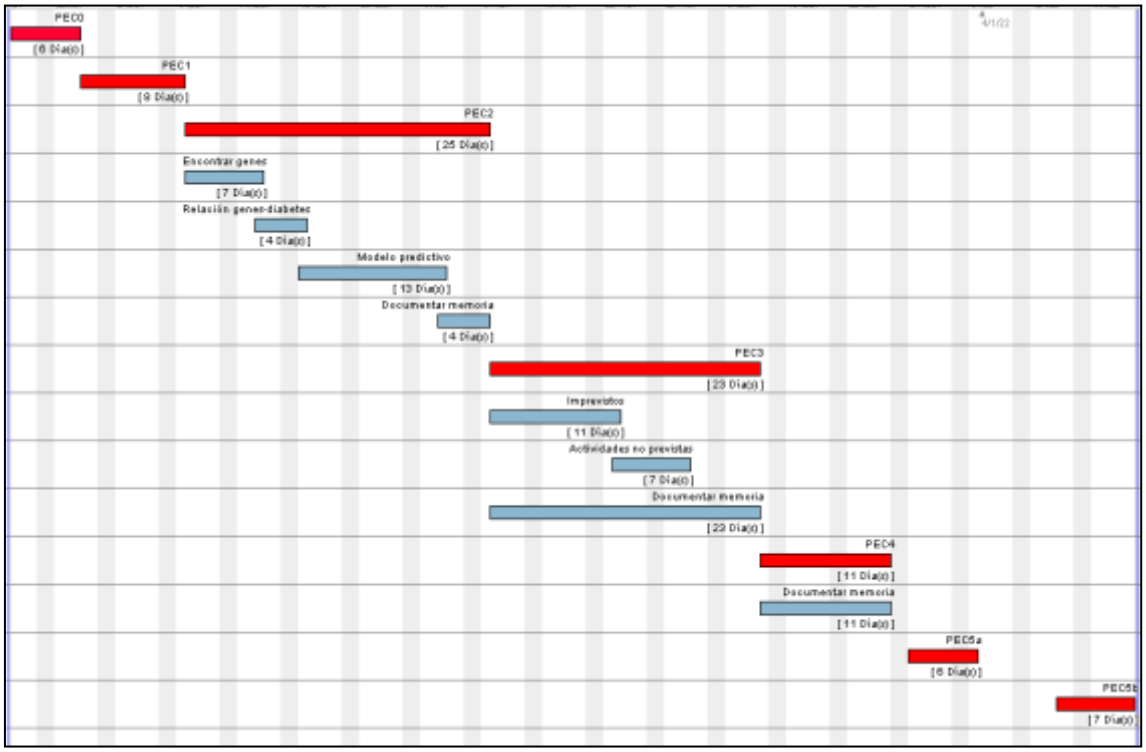

Figura 2. Planificación temporal de las tareas

Proceso	Fecha inicio	Fecha fin	Hito
Definición de los contenidos del trabajo	**15/09/2021**	**22/09/2021**	**Entrega PEC0**
Plan de trabajo	**23/09/2021**	**04/10/2021**	**Entrega PEC1**
Fase 1 del desarrollo del trabajo	**05/10/2021**	**08/11/2021**	**Entrega PEC2**
Encontrar los genes expresados de forma diferencial a través del análisis de datos ómicos mediante Bioconductor. Proceso clásico de análisis de datos de microarrays.	05/10/2021	13/10/2021	
Anotación del análisis de significación biológica a través de las bases de datos Gene Ontology y Reactome.	13/10/2021	18/10/2021	
Realizar diversos modelos predictivos con las diferentes técnicas de IA	18/10/2021	03/11/2021	
Documentar en la memoria	03/11/2021	08/11/2021	
Fase 2 del desarrollo del trabajo	**09/11/2021**	**09/12/2021**	**Entrega PEC3**
Ajuste de imprevistos relacionados con la Fase 1	09/11/2021	23/11/2021	
Actividades no previstas y realizadas	23/11/2021	09/12/2021	
Documentar en la memoria	09/11/2021	09/12/2021	
Cierre de la memoria	**10/12/2021**	**24/12/2021**	**Entrega PEC4**
Elaboración de la presentación	**27/12/2021**	**03/01/2022**	**Entrega PEC5a**
Defensa pública	**13/01/2022**	**21/01/2022**	**Entrega PEC5b**

Figura 3. Descripción de las tareas a realizar en cada PEC.

2.5 Breve sumario de contribuciones y productos obtenidos

La principal contribución es la validación del binomio entre los genes expresado de forma diferencial a través de un análisis de datos de microarrays y la posterior comprobación de la efectividad de estos datos para realizar predicciones a través de herramientas de ML. En este caso, se ha descubierto que con 60 genes expresados de forma diferencial, se puede llegar a predecir con una exactitud superior al 90% la clase (IS, IR o DB) a la que podría pertenecer el individuo.

Por otra parte, a continuación se describen los archivos obtenidos tras la ejecución del archivo R Markdown que contiene la parte de ejecución de ADO y ML:

- Ficheros asociados a la parte de ADO:
    - Resultados del control de calidad realizado con arrayQualityMetrics de los datos crudos.
    - Resultados del control de calidad realizado con arrayQualityMetrics de los datos normalizados mediante la función “rma”.
    - Fichero en formato .csv ordenado por p-valor ajustado de las comparaciones entre los grupos DB vs IR.
    - Fichero en formato .csv ordenado por p-valor ajustado de las comparaciones entre los grupos DB vs IS.
    - Fichero en formato .csv ordenado por p-valor ajustado de las comparaciones entre los grupos IR vs IS.
    - Fichero en formato .csv que contiene todos los genes expresado de forma diferencial
    - Fichero en formato .csv que contiene los 60 genes que se presentan un p-valor más ajustado y que es el fichero que se va a utilizar como fichero de entrada en la parte de Machine Learning.
    - Ficheros en formato .html asociados a la parte de análisis de significación biológica, más concretamente a la parte de Gene Ontology (paquete GOstats).
    - Ficheros en formato .pdf asociados a la parte de análisis de significación biológica, más concretamente a la base de datos “REACTOME pathway database” (paquete ReactomePA).
- Ficheros asociados a la parte de ML:
    - Ficheros en formato .csv de cada una de las técnicas de ML utilizadas y que contiene los resultados de la exactitud de la matriz de confusión 3*3, los resultados de la exactitud de la matriz de confusión 2*2, la sensibilidad, especificidad, vpp y vpn de cada una de las 1.000 ejecuciones.
    - Ficheros en formato “.txt” que contiene para cada una de las técnicas utilizadas de ML la matriz de confusión 3*3 acumulada con los datos de cada una de las 1.000 ejecuciones e ídem para la matriz de confusión 2*2.
    - Ficheros en formato “.txt” que contiene para cada una de las técnicas utilizadas de ML la matriz de confusión 3*3 y la matriz de confusión 2*2 de cada una de las 1.000 ejecuciones.
    - Tabla resumen con las medidas de rendimiento (exactitud, sensibilidad, especificidad, vpp y vpn) obtenidas tras las 1.000 ejecuciones.
- Fichero en formato .html que contiene todo el proceso de la ejecución de ADO y ML.

2.6 Breve descripción de los otros capítulos de la memoria

En el capítulo 1, se describe de forma sucinta todo el proceso realizado en este análisis, tanto del proceso de ADO como ML.

En el capítulo 2, se describe la justificación del trabajo, los datos asociados a la diabetes, los objetivos, tanto los generales como los específicos, un breve resumen de la metodología empleada, los productos obtenidos (resultados, ficheros, gráficos, etc.).

En el capítulo 3, se describe el estado del arte, es decir, cual es la situación actual de la diabetes y lo que aporta de nuevo este análisis.

En el capítulo 4, se describe a fondo la metodología empleada en las dos partes del análisis. En primer lugar se describen los pasos utilizados para obtener los genes expresados de forma diferencial y el análisis de significación biológica a través de dos plataformas distintas (GOstats y ReactomePA) a través de la plataforma de Bioconductor. En la segunda parte del análisis, se describen las técnicas utilizadas de ML para predecir la variable objetivo y los respectivos parámetros utilizados.

En el capítulo 5, se describen los resultados de las dos partes del análisis. En el capítulo 6 se describe la discusión y en el capítulo 7, las conclusiones.

3 Estado del arte

La Diabetes mellitus es una enfermedad crónica que se caracteriza por presentar un conjunto de trastornos metabólicos relativos a la aparición de hiperglucemia crónica, así como alteraciones en el metabolismo de los hidratos de carbono, las grasas y las proteínas, debido a la existencia de problemas en la secreción y/o acción de la insulina[6]. Esto se debe a que el páncreas no produce insulina en una cantidad suficiente o bien a que el propio organismo no es capaz de utilizar de forma adecuada la insulina que genera, y consecuentemente la glucosa no es asimilada por las células y permanece en la sangre, donde se produce un incremento de su concentración[7].

Se diferencian las siguientes tipologías de la DM:

- Diabetes mellitus tipo 1 (DM1): es la causante de la destrucción de las células beta del páncreas, que origina una carencia total de insulina. La DM1 representa únicamente entre el 5% y el 10% de las personas que tienen diabetes, pero se presenta de forma mayoritaria en los tramos de edad más jóvenes de la población[6]. La DM1 aparece de forma súbita y una vez diagnosticada el paciente suele requerir tratamiento con insulina para regular la glucemia.

  Los síntomas de la DM1 son la diuresis y la sed en exceso, la sensación de hambre de forma constante, el adelgazamiento, la aparición de problemas en la vista y el cansancio[7].
- Diabetes mellitus tipo 2 (DM2): se origina cuando se genera una resistencia a la acción de la insulina por parte del organismo humano junto con una carencia progresiva de producción de la misma. La DM2 se produce de forma gradual, afectando en torno, al 85% y 95% de la población diabética[6]. En mayor medida, la DM2 afecta a la población adulta, en tramos superiores a los 40 años, aunque cada vez se observan más casos en pacientes jóvenes, gran parte de los cuales están asociados a la obesidad. Tras el diagnóstico, el paciente no requiere un tratamiento continuo con insulina, pero puede ser necesaria su administración con la evolución de la enfermedad.

  Los síntomas de la DM2 pueden ser similares a los de la DM1 pero de menor intensidad, y en ocasiones, no se producen síntomas. Todo ello puede generar que la enfermedad no sea diagnosticada hasta la aparición de ciertas complicaciones varios años después de su inicio.

  La prediabetes implica la presencia de niveles elevados de glucosa en sangre, sin llegar a padecer la DM2, constituyendo un riesgo elevado para el desarrollo de la misma si no se adoptan medidas preventivas[7].
- Diabetes mellitus gestacional (DMG): se caracteriza por la aparición durante el embarazo de una concentración elevada de glucosa en sangre, que no existía previamente a la gestación. Este tipo de diabetes se diagnostica mediante la realización de un cribado inicial, el test de O'Sullivan, que, en el caso de resultar positivo, se realiza una sobrecarga oral de glucosa (SOG) para confirmar el diagnóstico. El test de cribado se realiza durante el primer trimestre del embarazo en los casos de alto riesgo y durante el segundo trimestre, en las semanas 24 a28 de gestación, a todas aquellas mujeres que no hayan sido diagnosticadas con anterioridad[7].
- Otros tipos específicos de diabetes mellitus: son de baja frecuencia, incluyéndose entre otras, DM producidas por patologías del páncreas, de

carácter genético o diabetes motivada por la exposición a determinados fármacos[7].

La Diabetes Mellitus tiene una elevada prevalencia e incidencia en la población, y puede ocasionar importantes problemas de salud, dando lugar a serias complicaciones como son la enfermedad cardiovascular, el ictus, la ceguera, la amputación de miembros inferiores, entre otros. En el caso de la diabetes gestacional, puede ocasionar problemas en el embarazo, tanto para la madre como para el feto o el recién nacido.

Algunas de las complicaciones que produce la diabetes se pueden evitar o demorar mediante el desarrollo de acciones preventivas y un buen control de la misma. En este sentido, existen diversos factores de riesgo que predisponen a las personas a desarrollar la enfermedad, y específicamente a las mujeres la DMG. Algunos de estos factores de riesgo no se pueden modificar, pero sobre otros es posible desarrollar acciones de prevención que eviten o retrasen su impacto en el desarrollo de la patología.

Entre los factores de riesgo no modificables se encuentran los factores genéticos y metabólicos y la edad. En este sentido, la prevalencia de la DM2 se incrementa a partir de la mediana edad y de forma más acusada en la población más envejecida[7].

En la actualidad, la prevalencia de la diabetes, una enfermedad potencialmente mortal, está aumentando en todo el mundo, alrededor de 463 millones de adultos de entre 20 y 79 años tienen diabetes. Esto representa el 9,3% de la población mundial en este grupo de edad. Se prevé que la cantidad total aumente a 578 millones (10,2%) para 2030 y a 700 millones (10,9%) para 2045. El número calculado de adultos de entre 20 y 79 años con tolerancia anormal a la glucosa (TAG) es de 374 millones (7,5% de la población mundial en este grupo de edad). Se prevé que esta cifra aumente a 454 millones (8,0%) para 2030 y a 548 millones (8,6%) para 2045. Se calcula que 1,1 millones de niños y adolescentes (que no superan los 20 años de edad) tienen diabetes tipo 1 (no se disponen datos en relación a la cantidad de niños y adolescentes que tienen diabetes tipo 2). Se calcula que el gasto anual en salud a nivel mundial destinado a la diabetes será de 760 mil millones de USD. Se pronostica que para los años 2030 y 2045, el gasto alcanzará los 825 mil y 845 mil millones de USD, respectivamente[1].

En Europa, se estima que el número de adultos de entre 20 y 79 años con diabetes es de 59,3 millones, lo que representa el 8,9% de la población regional en este grupo de edad. Esto incluye a 24,2 millones de adultos con diabetes no diagnosticada. Se estima que otros 36,6 millones de adultos de entre 20 y 79 años, el 5,5% de la población regional en este grupo de edad, tiene TAG en 2019. En 2030, se pronostica que habrá 66 millones de adultos con diabetes y 39,7 millones con TAG en la región. Las predicciones para 2045 sugieren que esta cifra aumentará hasta 68,1 millones de personas con diabetes y 40,3 millones con TAG. En gran medida, la alta prevalencia de la diabetes tipo 2 y la TAG son una consecuencia del envejecimiento de la población en la región. En relación a los jóvenes, Europa registra el mayor número de niños y adolescentes (entre 0 y 19 años) con diabetes tipo 1, con 296.500 afectados. La región también registra uno de los índices de incidencia más altos de diabetes tipo 1 en niños y adolescentes con una estimación de 31.100 nuevos casos por año. En la región europea, se estima que casi 465.900 muertes en adultos de entre 20 y 79 años son atribuibles a la diabetes y sus complicaciones en 2019 (8,5% de la mortalidad por todas las causas). En 2019, se estimó que el gasto sanitario total relacionado con la diabetes en la región EUR era de 161,4 mil millones de USD. En los adultos de entre 20 y 79 años, se proyecta que el gasto sanitario relacionado con la diabetes alcance

los 168,5 mil millones de USD en 2030 y 159,6 mil millones de USD en 2045. Aproximadamente 32.3 millones de adultos fueron diagnosticados de diabetes en la Unión Europea en 2019, frente a un estimado de 16.8 millones de adultos en 2020. Otros 24.2 millones de personas en Europa se estima que tienen diabetes pero no se diagnosticaron en 2019[1]. En España, unos 3.6 millones de adultos de entre 20 y 79 años presentaban diabetes en 2019, lo que representa el 10.5% de la población. Por otra parte, la prevalencia declarada de la diabetes en Canarias en la población de 15 y más años, es la más alta de España, habiendo mantenido una tendencia creciente en los últimos años. En el año 2017, la prevalencia de la diabetes en Canarias asciende casi al 11%, 3.19 puntos porcentuales superior a la media española (7.82%.)[7].

Los cambios en el estilo de vida, el consumo de una dieta alta en calorías y la falta de ejercicio se asocian con la resistencia a la insulina y puesto que la incidencia de estos trastornos ha aumentado drásticamente parece plausible que el exceso de nutrientes y la obesidad ejercen de factores causales de los estadios pre-diabéticos y de la diabetes. Las alteraciones extracelulares, como el exceso de nutrientes, la hiperinsulinemia, los glucocorticoides o la inflamación, desencadenan estrés intracelular en los tejidos metabólicos clave y estos son, fundamentalmente, el músculo y el tejido adiposo.

La resistencia a la insulina en el músculo esquelético se manifiesta por una disminución de la captación de glucosa estimulada por la insulina y es el resultado de una señalización alterada de la insulina y múltiples defectos intracelulares pos-receptores que incluyen transporte alterado de la glucosa, fosforilación de la glucosa y reducción de la oxidación de la glucosa y la síntesis de glucógeno.

Los estudios genéticos de asociación o prospectivos llevados a cabo en los años anteriores al advenimiento de las técnicas de secuenciación masiva o NGS se basaban en estrategias a apriorísticas en la que se presumía la participación del gen o de los genes a evaluar. Las técnicas NGS no requieren de hipótesis previas y permiten la evaluación por asociación de variantes génicas en el caso del genoma completo o del exoma y menos frecuentemente en el ARN o de transcritos diferencialmente expresados en el ARN.

La diabetes tipo 2 (DM2), que representa el 90-95% de todas las diabetes, se caracteriza esencialmente por una disfunción de las células β pancreáticas y resistencia a la insulina, por lo que se requiere un aumento de la secreción de insulina para compensar[2]. El descubrimiento de la insulina en 1921 fue un Big Bang del que ha surgido un vasto y creciente universo de investigación sobre la acción y la resistencia de la insulina[3]. La insulina es una potente hormona pleiotrópica que afecta a los procesos como el crecimiento celular, la diferenciación, la apoptosis, el flujo de iones, el gasto energético y el metabolismo de carbohidratos, lípidos y proteínas[13]. Estas acciones diversas se inician por la unión específica a receptores de alta afinidad en la membrana plasmática de las células diana[14,15], que luego activan tanto una vía de señalización metabólica a través de la quinasa PI-3 como una vía mitogénica a través de la cascada Ras/MAPK. La señal mediada por insulina se ha estudiado extensamente en lo que se refiere a eventos tempranos en la traducción de genes. Sin embargo, se carece de una comprensión de los eventos más distales en la señalización de la insulina que involucran varios sistemas efectores y de los efectos integrados sobre la expresión génica que subyacen a las acciones múltiples de la hormona. Ahora es posible una evaluación completa de la expresión diferencial en respuesta a la insulina utilizando la tecnología de microarrays. Este conocimiento podría mejorar la comprensión de la acción de la insulina y cómo se integran las respuestas para mediar en el espectro de efectos hormonales[13].

El músculo esquelético es el sitio principal para la eliminación de la glucosa insulinodependiente en los seres humanos[16,17]. La insulina estimula la captación y el uso de glucosa en las vías oxidativas y de almacenamiento. Aproximadamente el 80% de la captación de glucosa que responde a la insulina afecta a los músculos esqueléticos y este tejido es el sitio principal de almacenamiento de glucógeno, oxidación de lípidos, recambio de proteínas y termogénesis. La resistencia a la insulina que involucra al músculo esquelético es fundamental en la patogénesis de las enfermedades humanas, incluido el síndrome metabólico y la diabetes tipo 2, que causan una gran y creciente carga para la salud pública. Los mecanismos moleculares por los que la insulina regula el metabolismo muscular y los defectos subyacentes que causan la resistencia a la insulina no se han dilucidado por completo[13].

Debido a ello, se pretende analizar los genes relacionados con las 3 posibilidades respecto del binomio insulina-diabetes:

- Insulina-sensibles (IS), la cual sería la situación perfecta. Es decir, cuando el páncreas solo se necesita segregar un poco de insulina para hacer descender los niveles de glucosa en sangre.
- Insulina-resistentes (IR), es decir, que cada vez se va necesitando más cantidad de insulina para que ésta haga su función correctamente ya que las células no hacen caso a las señales y no quieren absorber glucosa. Se puede considerar un estadio pre-diabético. Tener la insulina alta hace aumentar el almacenamiento de grasa, lo cual a la vez te hace más resistente a la insulina. Por eso la resistencia a la insulina es un factor de riesgo para desarrollar obesidad, enfermedades cardiovasculares y diabetes tipo 2.
- Diabéticos (DB), sería el peor estadio. La diabetes es una enfermedad en la que los niveles de glucosa de la sangre están muy altos. La glucosa proviene de los alimentos que consume. La insulina es una hormona que ayuda a que la glucosa entre a las células para suministrarles energía. En la diabetes tipo 1, el cuerpo no produce insulina. En la diabetes tipo 2, la más común, el cuerpo no produce o no usa la insulina de manera adecuada y sin suficiente insulina, la glucosa permanece en la sangre.

Realizando las comparaciones dos a dos entre estás 3 categorías (DBvsIR, DBsvIS y IRvsIS), se podría encontrar los genes que caracterizan cada una de las 3 situaciones. Una vez encontrados los genes que se expresan de forma diferencial, se pretende dar un enfoque práctico y por ello se pretende elaborar modelos predictivos mediante técnicas de Machine Learning (ML) con la finalidad de poder clasificar a los pacientes con la mayor antelación y validez posible sobre su posible estadio respecto de la insulina y diabetes.

En relación al análisis propiamente dicho, en relación al ADO, se va a utilizar el proyecto Bioconductor[8], que hoy en día es una de las principales herramientas para el análisis de datos de microarrays u otros datos ómicos. En relación al análisis de los modelos predictivos, se van a utilizar técnicas de ML como SPV, ANN, RF, KNN y técnicas de Deep Learning como MLP.

En cuanto a las hipótesis planteadas en este análisis, son completamente distintas a las hipótesis relacionadas con los dos artículos publicados en relación a estos datos[13,18].

En el primer artículo (The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle) se utilizan los datos de los pacientes que son sensibles a la insulina antes y después de una estimulación de insulina a través de la

aplicación de pinzas hiperinsulinémicas euglucémicas (este método se considera el método de referencia para evaluar la sensibilidad a la insulina.

En el segundo artículo, (Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning) se utilizan los datos de los pacientes que son sensibles a la insulina y los pacientes que son resistentes a la insulina (individuos normoglucémicos).

Citation(s)	Wu X, Wang J, Cui X, Maianu L et al. The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle. *Endocrine* 2007 Feb;31(1):5-17. PMID: 17709892 Wu X, Patki A, Lara-Castro C, Cui X et al. Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning. *J Appl Physiol (1985)* 2011 Mar;110(3):746-55. PMID: 21109598

Figura 4. Artículos publicados con los datos utilizados

Dado que en los dos artículos comentados previamente se realizan las comparaciones de los tres grupos planteados (IS, IR y DB), resulta interesante utilizar los datos de los 3 grupos disponibles para realizar las comparaciones dos a dos y de esta forma poder observar los genes que se expresan de forma diferencial entre los grupos. Por último, y para dar un enfoque más práctico a los genes diferencialmente expresados, estudiamos la capacidad predictiva de estos genes con la finalidad de poder clasificar a los pacientes con la mayor validez posible.

4 Metodología

En este apartado se describe la metodología empleada en la parte de ADO como en la parte de ML.

4.1 Análisis de datos de microarrays

La biología molecular ha estado interesada desde sus comienzos en poder determinar el nivel de expresión de los genes integrados en el genoma humano. Especial interés ha tenido siempre sobre las otras biomoléculas el estudio de los niveles del ARN transcrito. Hace unos años se acuñó el término de transcriptómica para referirse al estudio del ARN en cada una de sus formas. La transcriptómica ha contribuido a tener una mejor comprensión de la patogénesis de las enfermedades[5,4].

En el año 2009 Wang y colaboradores[19] definieron el transcriptoma como el conjunto completo de transcritos en una célula y su cantidad en un momento determinado del desarrollo o en una determinada condición fisiológica. El interés de la transcriptómica no solamente se ha centrado en el desarrollo de nuevas tecnologías que mejoran su estudio, sino también en el desarrollo de nuevos métodos de extraer la gran cantidad de información que se genera con estas nuevas técnicas[5,4].

Una de las técnicas que revolucionó el estudio del transcriptoma fueron los microarrays. Un microarray es un artefacto para la realización de experimentos que permite estudiar simultáneamente múltiples unidades, que representan los genes, proteínas o metabolitos, sobre un sustrato sólido de cristal, plástico o sílice, y expuestos a la acción de las moléculas diana cuya expresión se desea analizar.

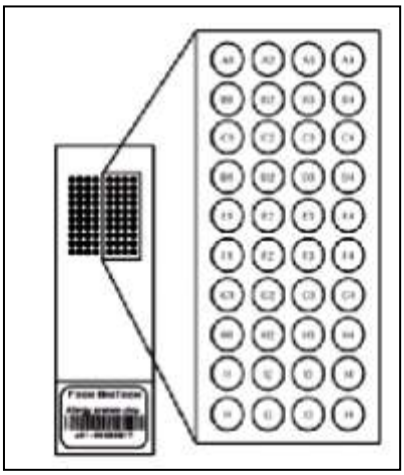

Figura 5. Imagen de un microarray

La utilización de los microarrays en la última década ha generado inmensas cantidades de datos útiles para el estudio y el desarrollo de enfermedades, dando a conocer múltiples mapas de expresión génica, encontrar biomarcadores o construir firmas génicas para determinadas enfermedades. Lo que caracteriza a los nuevos métodos utilizados para estudiar el transcriptoma no es lo que pueden medir, sino la cantidad de mediciones simultáneas que pueden realizar. Mientras que hasta hace apenas una década se estudiaban los genes uno a uno en profundidad, a partir del uso de estas nuevas tecnologías se pueden estudiar muchísimos genes a la vez, pero

en contrapartida con mucho menos detalle y más ruido. Los microarrays han sido cruciales para concebir una nueva manera de estudiar el transcriptoma, especialmente en el campo de la expresión génica. Después de ellos han venido otras técnicas que tienen en común con ellos el “alto rendimiento” es decir la capacidad para medir muchas variables, cientos o miles a la vez. Entre estas técnicas podemos destacar los arrays de SNPs, los microRNAs, la metilación y especialmente la secuenciación de nueva generación. La tecnología de los microarrays se ha aplicado a una inmensa variedad de problemas, en particular, en este estudio se van a utilizar para el estudio de genes que se expresan diferencialmente entre varios tipos de pacientes (IS, IR o DB). En términos generales los microarrays funcionan mediante la hibridación de una sonda específica (“probe”) y una molécula diana (“target”). La hibridación que ha tenido lugar se detecta mediante fluorescencia y se visualiza con la ayuda de un escáner. Los niveles de fluorescencia detectados reflejan la cantidad de moléculas diana presentes en la muestra problema. Existen diferentes tipos de microarrays según la naturaleza del “target” que se está estudiando, aunque en este análisis se utilizan microarrays de ARN o de expresión. Los microarrays se pueden clasificar por el número de muestras que se hibridan simultáneamente. Entonces, tenemos los microarrays de dos colores, los cuales están basados en la hibridación competitiva de dos muestras, cada una de las cuales ha sido marcada con un marcador fluorescente diferente (normalmente Cy3 y Cy5, verde y rojo respectivamente). Y por otra parte, están los microarrays de un color o arrays de oligonucleótidos[5,4].

Por otra parte, en este análisis se hibridan muestras de un solo color, lo que se llama microarrays de un color o arrays de oligonucleótidos. En estos arrays las muestras están marcadas únicamente con un marcador fluorescente. En cada array solamente se hibrida una muestra, por lo que no se da la hibridación competitiva como pasaba en los arrays de dos colores. En los arrays de un color en cambio tan solo se mide cuánto se expresa un gen en una escala que carece de sentido biológico. El valor que se obtiene después de iluminar el array con el láser es una medida numérica que se obtiene directamente del escáner, es decir, los microarrays permiten cuantificar la expresión de los genes a través de la intensidad de la fluorescencia que es capturada por los escáneres. Por último comentar, que el análisis de datos de microarrays, es decir, la parte relacionada con ADO, se ha desarrollado mediante la plataforma Bioconductor. De forma muy simple, se podría decir que un conjunto de datos de microarrays es una matriz de valores continuos que representan las expresiones de un conjunto de genes (un gen por fila), en una variedad de condiciones o muestras (una muestra por columna). A continuación se va a describir un flujo de trabajo, una serie de pasos ordenados, desde los datos sin procesar, es decir, las imágenes digitalizadas producidas por el sistema de hibridación, hasta una o más listas de genes que pueden usarse para ayudar a responder una determinada pregunta biológica. Un resumen del proceso se describe en la siguiente figura[5,4]:

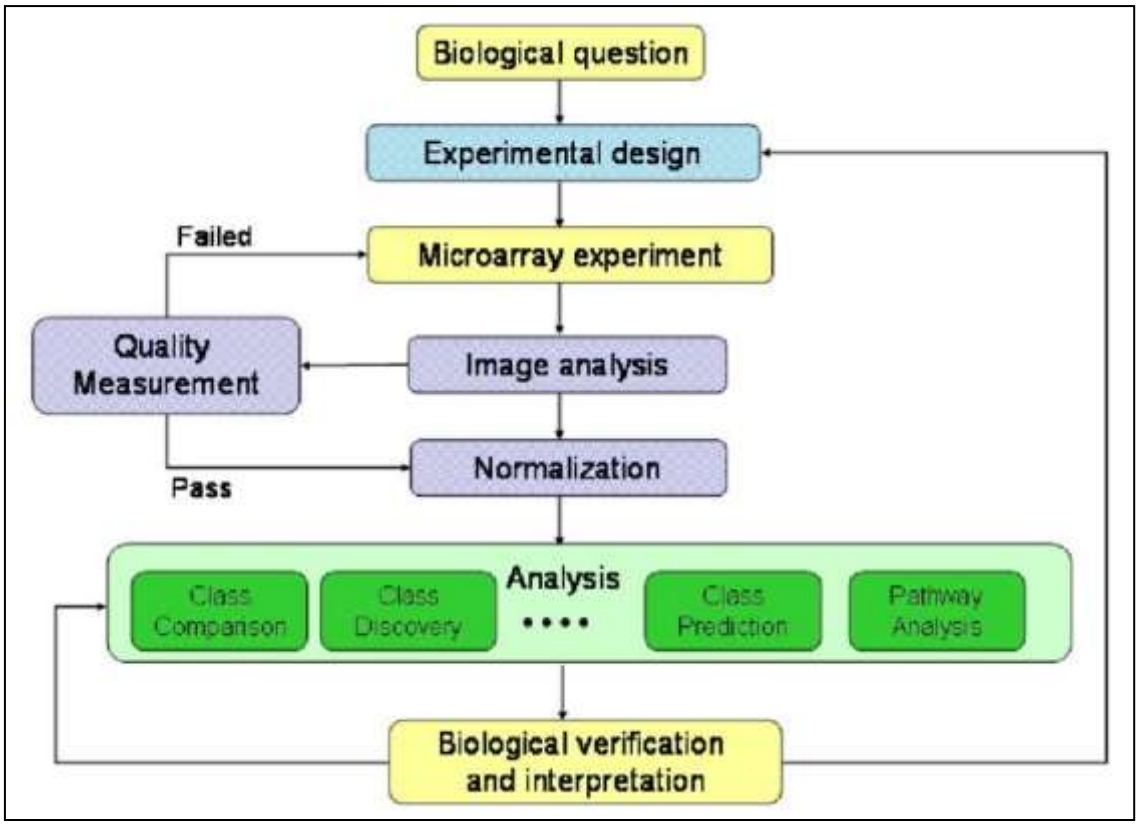

Figura 6. Proceso de datos de microarrays[5]

Tal como ilustra la figura superior, el análisis de microarrays puede ser fácilmente visualizado como un proceso que empieza por una pregunta biológica y concluye con una interpretación de los resultados de los análisis que, de alguna forma, se espera que nos aclare algo sobre la respuesta de la pregunta inicial.

En este estudio, el análisis de datos de microarrays fue llevado a cabo usando el chip Hu95A de biopsias del músculo esquelético humano de Affymetrix.

4.2 Ficheros de entrada

El análisis se basará en los datos de un estudio depositado en Gene Expresion Omnibus (GEO) con identificador "GSE22309" y cuyo título es "*Expression data from human skeletal muscle*". Estos datos han sido depositados en GEO siguiendo los estándares MIAME[20] (Minimum information about a microarray experiment). Se puede acceder a los datos a través del siguiente enlace:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22309

donde se puede encontrar el enlace para la descarga al final de la página.

Antes de nada, se debe aclarar que en este estudio únicamente se analizan la información de los 55 pacientes después de haber realizado la técnica de la aplicación de pinzas hiperinsulinémicas euglucémicas.

Platforms (1)	GPL91 [HG_U95A] Affymetrix Human Genome U95A Array
Samples (110)	GSM555237 a10a.000606jd.2.SkMNor907.CEL-IS basal
Less...	GSM555238 a10a.000606jd.2.SkMNor908.CEL-IS stimulate
	GSM555239 a10a.000606jd.2.SkMNor911.CEL-IS basal
	GSM555240 Run1.a10a.000606jd.2.SkMNor912.CEL-IS stimulate
	GSM555241 Run1.a10a.000607jd.2.SkMNor915.CEL-IS basal
	GSM555242 Run1.a10a.000607jd.2.SkMNor916.CEL-IS stimulate
	GSM555243 Run1.a10a.000607jd.2.SkMNor917.CEL-IS basal
	GSM555244 Run1.a10a.000607jd.2.SkMNor918.CEL-IS stimulate

Figura 7. Archivos depositados en GEO. Se han utilizado los archivos "stimulate".

La información basada en la imagen del microarray ha quedado registrada en 55 archivos (20 archivos asociados a IS, 20 a IR y 15 a diabéticos) tipo ".cel" (Cell Intensity File). A partir de las intensidades de los archivos ".CEL" se genera la matriz de expresión, que contiene una columna por chip con los valores de intensidad absoluta, y una fila por grupo de sondas.

Por lo tanto, se leen estos 55 archivos tipo ".cel" y se unifica con el fichero "targets.csv" que contiene las características de los registros. El objeto conjunto es del tipo "ExpressionSet", diseñado para combinar distintos tipos de información.

4.3 Tipo de Microarray (plataforma)

Para trabajar con microarrays, necesitamos conocer de qué tipo son, puesto que ello nos permite saber qué paquete de anotaciones necesitamos. Esta información se encuentra también en la página de GEO correspondiente a los datos:

Platforms (1)	GPL91 [HG_U95A] Affymetrix Human Genome U95A Array

Figura 8. Tipo de Microarray

Esto nos indica que se trata de arrays del modelo Hu95A de Affymetrix, cuyas anotaciones se encuentran en el paquete hgu95av2.db de Bioconductor.

4.4 Exploración y control de calidad

Los gráficos son útiles para comprobar la calidad de los datos de microarrays, obtener información sobre cómo se deben preprocesar los datos y comprobar, finalmente, que el preprocesado se haya realizado correctamente. Siguiendo el esquema presentado en la figura 4 se presentan a continuación los distintos gráficos utilizados con una breve descripción de lo que representa cada uno y cómo interpretarlos adecuadamente.

4.4.1 Control de calidad con gráficos estadísticos generales

4.4.1.1 Gráficos de densidad

Los gráficos de densidad muestran una idea de si las distribuciones de los distintos arrays son similares en forma y posición[5,4].

4.4.1.2 Diagrama de cajas o boxplots

Los diagramas de caja o boxplots, basados en los cuartiles de los valores, dan una idea de la distribución de las intensidades[5,4].

4.4.1.3 Gráficos de componentes principales

El análisis de componentes principales puede servir para detectar si las muestras se agrupan de forma "natural" es decir con otras muestras provenientes del mismo grupo o si no hay correspondencia clara entre grupos experimentales y proximidad en este gráfico. Cuando esto sucede no significa necesariamente que haya un problema pero puede ser indicativo de efectos técnicos, como el conocido efecto "batch", que podría ser necesario corregir[5,4].

4.4.1.4 Cluster jerárquico (Dendograma)

Otra forma de ver si las muestras se agrupan según los grupos experimentales, o mediante otros criterios es usando un clúster jerárquico que realiza una agrupación básica de las muestras por grado de similaridad según la distancia que se utilice. Si las muestras se agrupan según las condiciones experimentales es una buena señal pero si no es así, puede deberse a la presencia de otra fuente de variación[5,4].

4.4.2 Gráficos de diagnóstico para microarrays de un color

4.4.2.1 Gráfico "M-A"

En los chips de Affymetrics, en que sólo hay un canal en cada array, en el eje-y (abscisas) se plasma el valor M (el log ratio) a partir de la comparación entre pares de valores, ya sea los arrays dos a dos o bien cada array respecto un valor de referencia que puede ser la mediana, punto a punto, de todos los arrays. En el eje-x se plasma el valor A como la media de la intensidad, i.e:

$$M = log_2(I_1) - log_2(I_2)\ ;\ A = \frac{1}{2}(log_2(I_1) + log_2(I_2))$$

donde I_1 es la intensidad del array de estudio e I_2 es la intensidad media de arrays. Se espera que la distribución en el gráfico se concentre a lo largo del eje M=0[5,4]. Este gráfico es similar a lo que en estadística se llama el gráfico de Bland-Altman.

4.4.3 Control de calidad con el paquete ArrayQualityMetrics

Otra forma de observar la calidad de los datos es a través del paquete ArrayQualityMetrics[21], que también realiza gráficos como los anteriormente comentados, como el ACP, MA, gráficos de densidad, etc. En el archivo con extensión ".html" creado con este paquete, merece la pena observar la figura denominada "Array metadata and outlier detection overview" y observar si alguna de las muestras está marcada con tres "x" en los criterios de calidad, lo que indicaría posibles problemas de calidad[5,4].

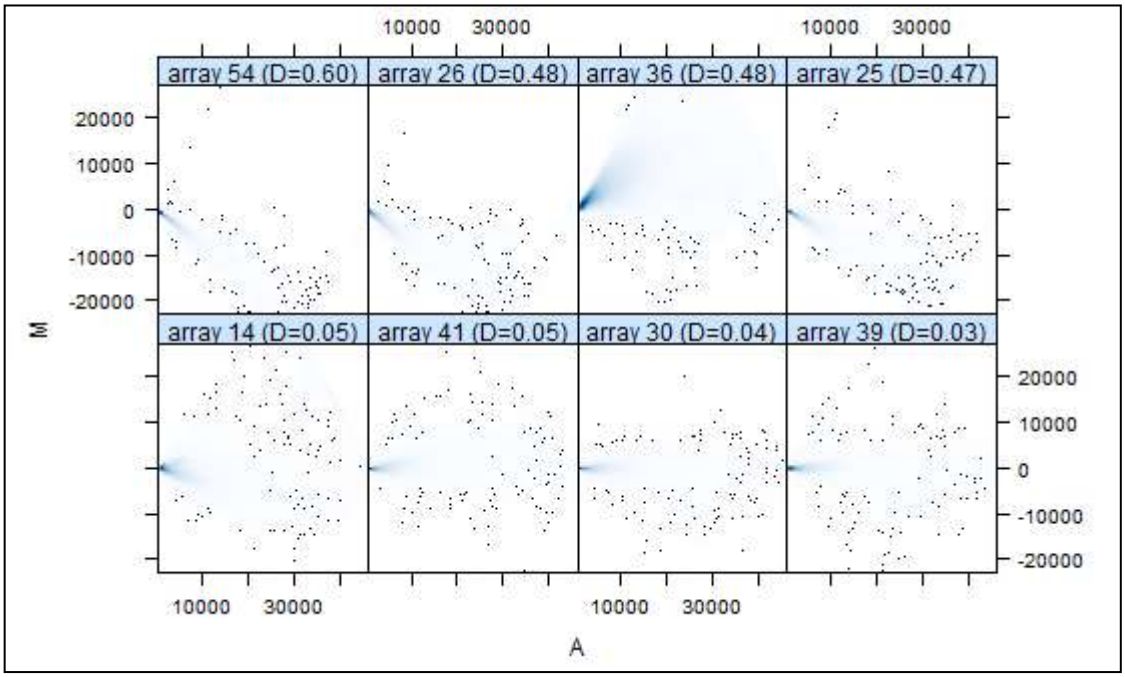

Figura 9. MA plot obtenido con el paquete ArrayQualityMetrics.

4.5 Análisis de efectos Batch

Dado que una posible causa de efectos Batch es la fecha en la que se procesan las muestras, se muestra la fecha de hibridación de los archivos .CEL mediante la función "*get.celfile.dates*" del paquete affyio[5,4].

4.6 Normalización de microarrays de Affymetrix

4.6.1 El método RMA (Robust Multi-Array Average)

Antes de comenzar con el análisis de expresión diferencial, es necesario hacer que los arrays sean comparables entre ellos y tratar de reducir, y si es posible eliminar, toda la variabilidad en las muestras que no sean producidas por razones biológicas El proceso de normalización intenta asegurar que las diferencias de intensidad presentes en los arrays sean por la expresión diferencial de los genes, en lugar de sesgos artificiales debidos a problemas técnicos. El método más utilizado para la normalización de arrays es el "Robust Multichip Analysis" (RMA)[22]. Es un método basado en la modelización de las intensidades de las sondas que se basa en los distintos valores de la misma sonda entre todos los arrays disponibles. Esquemáticamente los pasos que realiza este método son:

1. Ajusta el ruido de fondo (background) basándose solo en los valores PM y utilizando un modelo estadístico complejo en el que combina la modelización de la señal mediante una distribución exponencial con la del ruido mediante una distribución normal.
2. Toma logaritmos base 2 de cada intensidad ajustada por el background.
3. Realiza una normalización por cuantiles de los valores del paso 2 consistente en sustituir cada valor individual por el que tendría la misma posición en la distribución empírica estimada sobre todas las muestras, es decir los promedios de las distribuciones de los valores ordenados de cada array.
4. Estima las intensidades de cada gen separadamente para cada conjunto de sondas. Para ello realiza una técnica similar a una regresión robusta denominada "pulido de medianas" (median polish) sobre una matriz de datos que tiene los arrays en filas y los grupos de sondas en columnas.

Como resultado final de todos los pasos anteriores se obtiene la matriz con los datos sumarizados y normalizados. Este método se ha convertido en el estándar "de facto" actualmente en Bioconductor. Tras realizar el proceso de normalización de los datos, se realizan las representaciones gráficas comentadas previamente de estos nuevos datos y son con los que se trabaja de aquí en adelante[5,4].

4.7 Detección de los genes que presentan mayor variabilidad

Si un gen se expresa de manera diferencial, se espera que exista una cierta diferencia entre los grupos y, por lo tanto, la varianza general del gen será mayor que la de aquellos que no tienen expresión diferencial. Mostrar la variabilidad general de todos los genes es útil para decidir qué porcentaje de genes muestra una variabilidad que puede atribuirse a otras causas distintas a la variación aleatoria. Por lo tanto, se realiza una representación gráfica de la desviación típica de cada uno de los genes ordenados de menor a mayor valor. Aquellos genes con mayor variabilidad, son los que se espera que se expresen de forma diferencial[5,4].

4.8 Filtraje no específico

El objetivo del filtraje es eliminar aquellos genes o aquellos spots cuyas imágenes o señales sean erróneas por diferentes motivos, con el fin de reducir el ruido de fondo o

eliminar aquellos genes que varíen poco entre condiciones. En este caso, los procesos de filtraje utilizados han sido:

1. Eliminación de genes que no presenten una variación significativa en su señal, entre distintas condiciones experimentales (Filtraje por variabilidad).
2. Eliminación de genes que no se dispone de su anotación para ellos[5,4].

Es decir, se ha aplicado un filtraje "estándar" que retiene el 50% de los genes con mayor variabilidad de entre aquellos que están correctamente anotados (la base de datos utilizada ha sido hgu95av2.db), es decir se han eliminado aquellos que no tienen identificador en la base de datos Entrez y el 50% de los genes con menor variabilidad. El resultado del filtraje es una lista con varios objetos, que informan de lo que se ha descartado y un objeto expressionSet que, en lugar de 12.626 "*features*" tiene 4.380 que son los que están anotados y tienen mayor variabilidad. La selección de genes diferencialmente expresados, se llevará a cabo sobre esta lista.

4.9 Selección de genes diferencialmente expresados

4.9.1 Introducción

El motivo más habitual para el que se suelen utilizar microarrays es la búsqueda de genes cuya expresión cambia entre dos o más condiciones experimentales, por ejemplo a consecuencia de un tratamiento, una enfermedad u otras causas (distintos tiempos, distintas líneas celulares, etc.). El problema consiste en identificar estos genes y suele denominarse selección de genes diferencialmente expresados ("DEG") o bien comparación de clases. El problema de seleccionar genes diferencialmente expresados se traduce de manera casi inmediata al problema estadístico de comparar variables y, en años recientes, se han desarrollado un gran número de métodos estadísticos para resolverlo. La mayoría son extensiones de los métodos estadísticos clásicos, pruebas-t o análisis de la varianza, adaptados en uno u otro sentido para tener en cuenta las peculiaridades de los microarrays[5,4].

4.9.2 Modelos lineales para la selección de genes

La selección de genes diferencialmente expresados puede basarse en distintas aproximaciones, desde la t de Student al programa SAM pasando por multitud de variantes. En este ejemplo, dado que se realizaran tres comparaciones que luego deseamos comparar entre ellas, se aplicará la aproximación presentada por Smyth[23] basado en la utilización del modelo lineal general combinada con un método para obtener una estimación mejorada de la varianza. El procedimiento es el siguiente:

- Se plantea el problema como un modelo lineal con una componente bayesiana ya que se supone que los mismos parámetros a estimar son variables (no constantes) con distribuciones "prior" que se estimaran a partir de la información de todos los genes.
- A continuación se obtienen las estimaciones de los parámetros del modelo. La aproximación utilizada garantiza que estos estimadores tienen un comportamiento robusto incluso para pequeño número de arrays.

- Finalmente se calcula un “odd-ratio” que viene a ser la probabilidad de que un gen esté diferencialmente expresado frente a la de que no lo esté y se asocia este valor denominado estadístico B con un estadístico t moderado y su p-valor.

$$B = \log \frac{P\ [Afectado|M_{ij}]}{P\ [No\ Afectado|M_{ij}]}$$

gen = i (i = 1, ..., N), réplica = j (j = 1,..., n)

El hecho de trabajar con logaritmos permite poner el punto de corte en el cero. A mayor valor positivo más probable es que el gen esté diferencialmente expresado. A mayor valor negativo, más probable es que no lo esté[5,4].

4.9.3 Matriz de diseño y de contrastes

El primer paso para el análisis basado en modelos lineales es crear la matriz de diseño. Básicamente, la matriz de diseño es una tabla que describe la asignación de cada muestra a un grupo. Tiene tantas filas como muestras y tantas columnas como grupos (si solo se considera un factor). Cada fila contiene un uno en la columna del grupo al que pertenece la muestra y un cero en las restantes[5,4]. En este estudio, la matriz de diseño está formada por 55 filas y 3 columnas (DB, IR e IE).

Por otra parte, la matriz de contrastes se utiliza para describir las comparaciones entre grupos (DB vs IR; DB vs IS y IR vs IS). Consta de tantas columnas como comparaciones y tantas filas como grupos (es decir como columnas de la matriz de diseño). Una comparación entre grupos –llamada “contraste”– se representa con un “1” y un “-1” en las filas de los grupos a comparar y ceros en las restantes. Si varios grupos intervinieran en la comparación se tendría tantos coeficientes como grupos con la única restricción de que su suma sería cero. Las comparaciones a realizar son las diferencias, dos a dos entre los 3 grupos posibles[5,4]:

- DB – IR
- DB – IS
- IR – IS

4.9.4 Estimación del modelo y selección de genes

Una vez definida la matriz de diseño y los contrastes se puede pasar a estimar el modelo, estimar los contrastes y realizar las pruebas de significación que nos indiquen, para cada gen y cada comparación, si puede considerarse diferencialmente expresado. El análisis proporciona los estadísticos de test habituales como Fold–change, t-moderados o p-valores ajustados que se utilizan para ordenar los genes de más a menos diferencialmente expresados[5,4].

A fin de controlar el porcentaje de falsos positivos que puedan resultar del alto número de contrastes realizados simultáneamente, los p–valores se ajustan de forma que tengamos control sobre la tasa de falsos positivos utilizando el método de Benjamini y Hochberg[24].

4.9.5 Volcano plots

Dado que se han calculado los valores de significación (p–valores) de los genes, resulta interesante comparar el tamaño del cambio del nivel de significación estadístico. El "volcano plot" es una representación gráfica que permite ordenar los genes a lo largo de dos dimensiones, la biológica, representada por el "fold change" y la estadística representada por el logaritmo negativo del p–valor. En la escala horizontal se representa el cambio entre los dos grupos (en escala logarítmica, de manera que la regulación positiva o negativa se representa de forma simétrica). En la escala vertical se representa el p–valor del test en una escala logarítmica negativa, de forma que los p–valores más pequeños aparecen mayores. Así pues puede considerarse que el primer eje indica impacto biológico del cambio (a más efecto biológico mayor "*fold-change*") y el segundo muestra la evidencia estadística, o la fiabilidad del cambio[5,4].

4.9.6 Comparaciones múltiples

Cuando se realizan varias comparaciones a la vez puede resultar importante ver qué genes cambian simultáneamente en más de una comparación. Si el número de comparaciones es alto también puede ser necesario realizar un ajuste de p-valores entre las comparaciones, distinto del realizado entre genes. En este análisis, no se ajustaran los p-valores entre comparaciones. Tan solo se seleccionaran los genes que cambian en una o más condiciones. El resultado del análisis es una tabla que para cada gen y cada comparación contiene un 1 (si el gen esta sobre-expresado), un 0 (si no hay cambio significativo) o un -1 (si está infra-expresado).

Por otra parte, se ha realizado un diagrama de Venn para observar cuantos genes se expresan de forma diferencial entre las tres comparaciones sin diferenciar entre genes sobre o infra regulados[5,4].

4.9.7 Visualización de los perfiles de expresión

Tras seleccionar los genes diferencialmente expresados podemos visualizar las expresiones de cada gen agrupándolas para destacar los genes que se encuentran up o down regulados simultáneamente constituyendo perfiles de expresión. En este caso, se realiza a través del uso de "mapas de calor" o "*heatmaps*". En este apartado, se realizan 3 *heatmaps*:

- Se toman todos aquellos genes que han resultado diferencialmente expresados en alguna de las tres comparaciones.
- Se toman los 60 genes que se van a utilizar como conjunto de datos de entrada de la parte de ML.
- Se toman los 4 genes que se expresan de forma diferencial en las 3 comparaciones realizadas.

4.10 Análisis de significación biológica

Una vez obtenidas las listas de genes diferencialmente expresados, se pueden llevar a cabo todo tipo de análisis sobre ellas, generalmente encaminados a facilitar la interpretación biológica de los resultados. Entre estas exploraciones se encuentra la anotación de las listas de genes en diversas bases de datos.

Por una parte, se ha realizado el análisis básico de enriquecimiento (Gene Enrichment Analysis) tal como se describe en los trabajos de Falcon y Gentleman[10] implementados en el paquete GOstats de Bioconductor. El análisis se realiza sobre la base de datos anotaciones "Gene Ontology". Los análisis de este tipo necesitan un número mínimo de genes para resultar fiables por lo que se incluirán en todos los genes con p–valores ajustados inferiores a 0.05, sin filtrar por mínimo "*fold–change*".

Por otra parte, se ha realizado un análisis de conjuntos de genes (Gene set analysis) para la interpretación biológica donde se accede a la base de anotaciones [11]. En este caso, se seleccionan todos los genes expresados de forma diferencial en la comparación DBvsIR y aquellos que presentan un p-valor ajustado < 0.05 en las otras dos comparaciones.

4.11 Técnicas de Inteligencia Artificial

A continuación, se describen las técnicas de Machine Learning utilizadas para elaborar los modelos predictivos donde las variables independientes son los 60 genes expresados de forma diferencial entre las 3 comparaciones comentadas previamente y la variable objetivo es la situación de los pacientes respecto de la insulina y diabetes (IS, IR o DB).

4.11.1 Multilayer Perceptron con Keras

La técnica Multilayer Perceptron (MLP) es un tipo especial de red totalmente conectada con múltiples neuronas individuales. El modelo secuencial definido es una cadena lineal de capas. El modelo necesita saber qué forma de entrada debe esperar. Por esta razón, la primera capa en un modelo secuencial necesita recibir información sobre su forma de entrada. Por ello, la capa de entrada tiene el mismo número de entradas que el total de las variables predictivas, en este caso 60. La capa intermedia busca características asociadas a los datos. En este caso, se ha definido una capa intermedia con 64 nodos. La capa de salida tiene el mismo número de salidas que las categorías a predecir, en este caso 3. La función de activación de la última capa es "softmax", que convierte un vector de valores en una distribución de probabilidad[25].

4.11.2 k-nearest neighbors (KNN)

El algoritmo de KNN aplica el enfoque de los vecinos más próximos para clasificar un nuevo individuo. El algoritmo KNN comienza con un conjunto de datos de entrenamiento donde se conoce sus categorías respecto de la variable objetivo. Por otra parte, se utiliza el conjunto de datos de prueba que contiene ejemplos no

etiquetados del mismo tipo que los datos de entrenamiento. Para cada registro en el conjunto de datos de prueba, KNN identifica k registros en los datos de entrenamiento que son los "más cercanos" en similitud (se ha utilizado la distancia Euclídea), donde k es un valor positivo especificado de antemano y que en este caso se ha tomado como la raíz cuadrada del tamaño del conjunto de datos de entrenamiento. Al registro no etiquetado se le asigna la clase de la mayoría de los k vecinos más cercanos (6 vecinos en nuestro caso). Es un algoritmo simple y efectivo, donde no se realizan suposiciones sobre la distribución de los datos[26].

4.11.3 Redes Neuronales

Las redes neuronales artificiales se inspira en las redes neuronales como las que se tiene en el cerebro. Así como un cerebro usa una red de células interconectadas llamadas neuronas para crear un procesador paralelo masivo, la ANN utiliza una red de neuronas o nodos artificiales para resolver problemas de aprendizaje. Las neuronas se sustituyen por nodos que reciben y envían señales (información). Se crea una red con diferentes capas interconectadas para procesar la información. Cada capa está formada por un grupo de nodos que transmite la información a los otros nodos de las capas siguientes. Aunque existen numerosas variantes de redes neuronales, cada una se puede definir en términos de las siguientes características:

- La topología: Esto corresponde al número de capas y de nodos. Además de la dirección en que se la información pasa de un nodo al siguiente, dentro de capas o entre capas. La topología determina la complejidad de las tareas que la red puede aprender. Generalmente, las redes más grandes y complejas son capaces de identificar patrones más sutiles y límites de decisión complejos.
- La función de activación: Función que recibe un conjunto de entradas e integra la señales para transmitir la información a otro nodo/capa. Es el mecanismo por el cual la neurona artificial procesa información y la pasa a través de la red.
- El algoritmo de entrenamiento: Establece la importancia de cada conexión para transmitir o no la señal a los nodos correspondientes. El más usado es el algoritmo "backpropagation". El nombre indica que para corregir los errores de predicción va hacia atrás de la red corrigiendo los pesos de los nodos. Debido a que la red no contiene conocimiento a priori (existente), típicamente los pesos se establecen aleatoriamente antes de comenzar. Luego, el algoritmo recorre el proceso hasta que el criterio de parada es alcanzado. El recorrido incluye:
    - Una fase hacia adelante: En la que las neuronas se activan en secuencia desde la capa de entrada a la capa de salida, aplicando los pesos de cada neurona y función de activación en el camino. Al llegar a la capa final, se produce una señal de salida.
    - Una fase hacia atrás: En la que la señal de salida de la red resultante de la fase hacia adelante se compara con el valor real en los datos de entrenamiento. La diferencia entre la señal de salida de la red y los resultados del valor real, es el error que se propaga hacia atrás en la red para modificar el pesos de conexión entre neuronas y reducir errores futuros[26].

En este análisis, se han utilizados dos modelos de redes neuronales. En el primero de ellos, se han utilizado dos capas ocultas, la primera con 30 nodos y la segunda con 20 nodos. En el segundo, las variables de entrada han sido las 7 variables cuyo autovalor ha sido superior a 1 en el análisis de componentes principales realizado con los 60 genes obtenidos en el proceso de ADO. En este segundo modelo, también se han

utilizado dos capas ocultas con 5 y 3 nodos respectivamente. En ambos casos, la función de activación ha sido la función logística (es la principal función de activación y es muy importante ya que se puede derivar) y el algoritmo de entrenamiento, ha sido el "backpropagation"[26].

4.11.4 Support Vector Machine

Una máquina de vectores de soporte (SVM) se puede imaginar como una superficie que define un límite entre varios puntos de datos que representan ejemplos trazados en espacio multidimensional de acuerdo con sus características. El objetivo de un SVM es crear un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta para crear particiones bastante homogéneas a cada lado.

Cuando los datos no son separables de forma lineal, es necesario el uso de kernels, o funciones de similitud y especificar un parámetro C para minimizar la función de coste. La elección de este parámetro es a base de ensayo/error, pero se buscan valores que no sean extremos en la búsqueda del equilibrio sesgo/varianza. Los kernels más populares son el lineal y el gaussiano.

En este análisis, se ha aplicado dos veces la técnica de SVM. La primera de ellas ha sido lineal y la segunda el gaussiano. En ambos casos, el parámetro C ha tomado el valor 1[26].

4.11.5 Random Forest

Este algoritmo sólo usa una parte de las variables del conjunto de datos, que se eligen de forma aleatoria. Así se pueden manejar conjunto de datos extremadamente grandes, y el problema de "*curse of dimensionality*" o maldición de la dimensión, se evita y se consigue tasas de error similares a las de otros algoritmos. Su técnica es estimar a partir de varios árboles de decisión y varios subconjuntos de entrenamiento intentar reducir el sobredimensionamiento y mejorar la precisión de los resultados. Es decir, se seleccionan individuos al azar con reemplazamiento, formando así diferentes conjuntos de datos. Posteriormente se crea un árbol de decisión con cada conjunto de datos, de forma que se obtienen diferentes árboles. Al crear el árbol se elijen las variables al azar en cada nodo del árbol, y así sin poda se deja crecer el árbol. Posteriormente, se predicen los nuevos datos usando el voto mayoritario, clasificándose como positivo si la mayoría de los árboles predicen la observación como positiva[26]. En este análisis, el bosque aleatorio incluyó 500 árboles y probó 7 variables en cada división.

5 Resultados

5.1 Resultados asociados al proceso de Análisis de datos ómicos

En este apartado, se muestran de forma secuencial los resultados obtenidos en el proceso. El conjunto de datos está formado por 409.600 genes de 55 pacientes. En primer lugar, se muestran los diagramas de cajas y bigotes de la intensidad de los valores, el gráfico de densidad de la señal de distribución de los datos, el gráfico de dos dimensiones del análisis de componentes principales de los datos crudos y una tabla con tres columnas indicando la calidad de los datos. En la figura inferior se puede observar que hay 7 datos asociados al grupo de IR (IR_11, IR_12, IR_13, IR_14, IR_15, IR_16 y IR_17) y un dato asociado al grupo DB (DB_4) que podrían presentar algún efecto Batch.

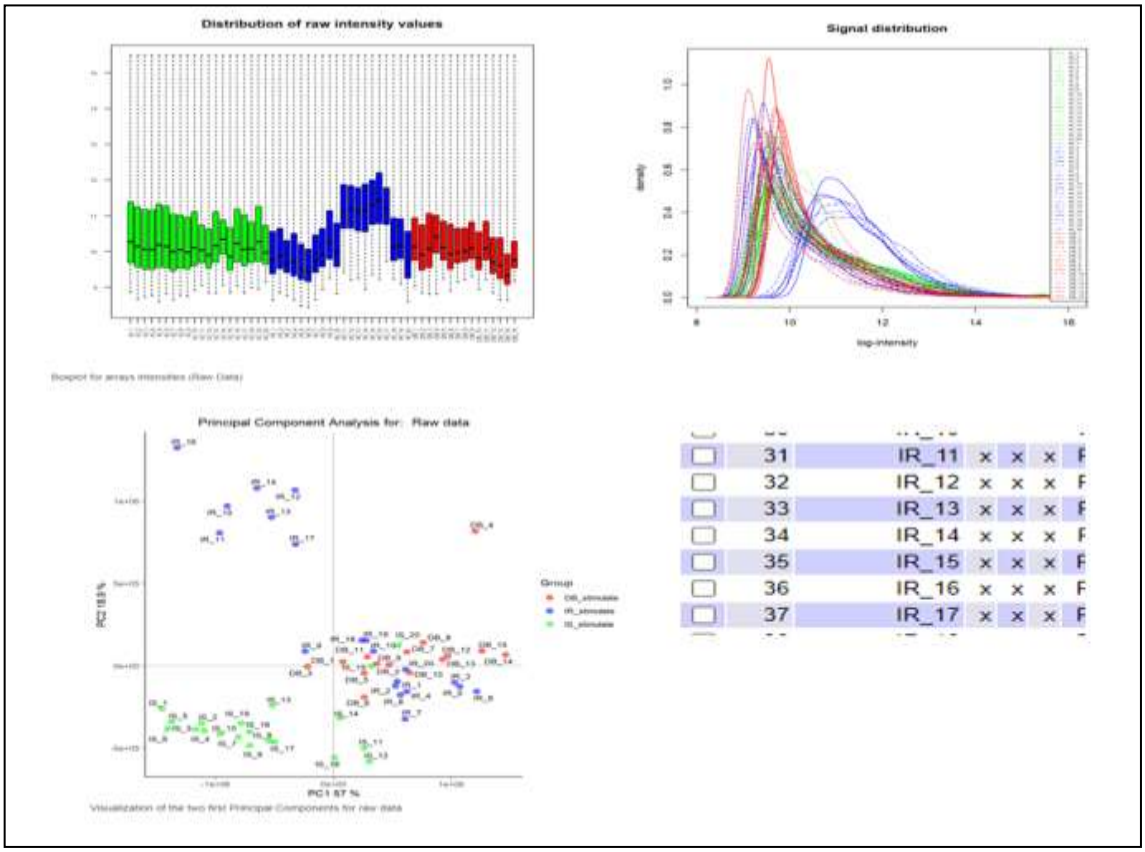

Figura 10. Control de calidad de los datos crudos

Los gráficos sugieren que puede haber algún factor que se superponga a las diferencias entre los grupos. Dado que no se dispone de información sobre otras covariables es difícil decidir si esto es así. Una causa habitual del efecto batch es la fecha en que se procesan las muestras. Por ello, en la tabla inferior se muestran las fechas de hibridación aunque de los 10 primeros registros de los pacientes IS no se dispone de la fecha.

HybDates/	DB_stimulate	IR_stimulate	IS_stimulate
2000-12-06	0	0	8
2000-12-20	0	8	0
2001-01-10	0	9	0
2001-01-11	5	3	0
2001-01-12	6	0	2
2001-02-06	4	0	0

Figura 11. Fechas de hibridación

Se puede observar que hay 4 fechas (2000-12-06, 2000-12-20, 2001-01-10 y 2001-02-06) en las que únicamente se hibridaron una de las categorías (IS, IR, IR y DB respectivamente). En cualquier caso, continuamos con todos los datos.

Tras realizar el proceso de normalización de los datos con la función "*rma*", se seleccionan 12.626 genes de los 55 pacientes. Se vuelven a mostrar las mismas imágenes pero de los datos normalizados, donde se puede observar que los datos presentan aspecto similar y los datos que presentaban tres cruces en la imagen anterior, ahora presentan una o ninguna. Por lo que se puede afirmar que la normalización ha funcionado bien.

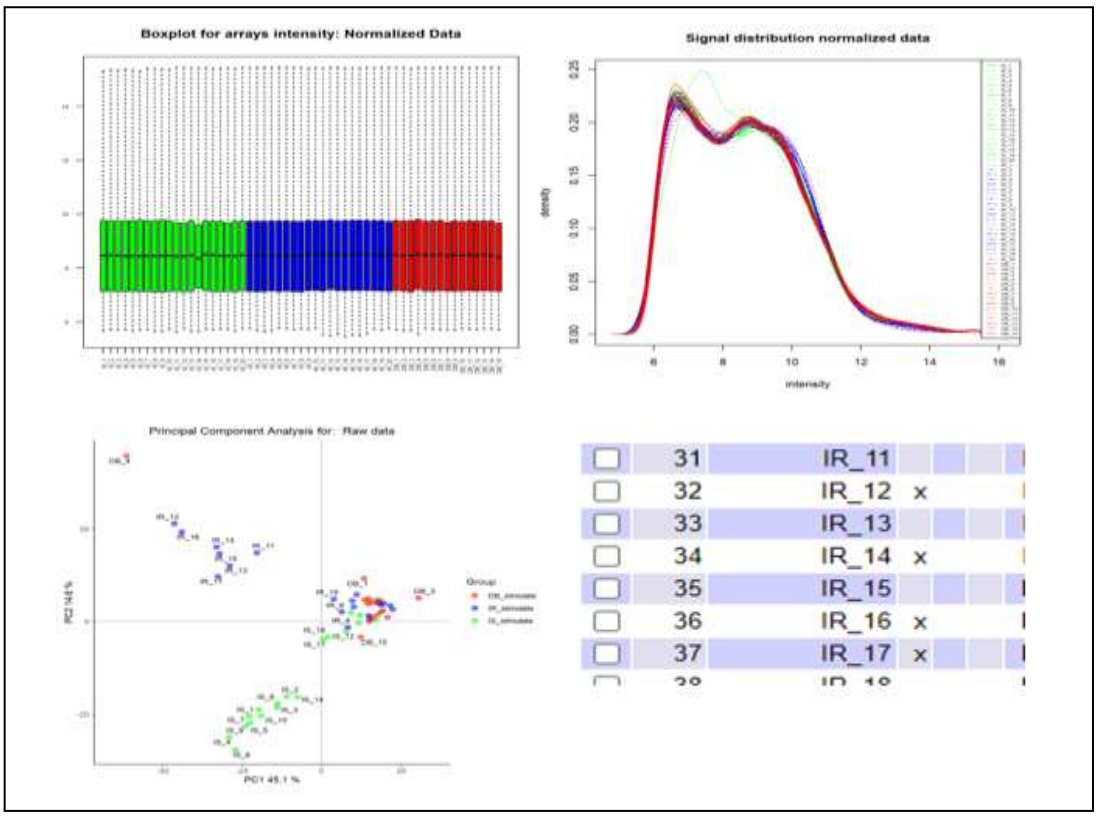

Figura 12. Control de calidad de los datos normalizados

A continuación se muestra una figura donde se analiza la distribución de la variabilidad de los genes. Esta figura presenta el aspecto habitual, sin que haya nada reseñable.

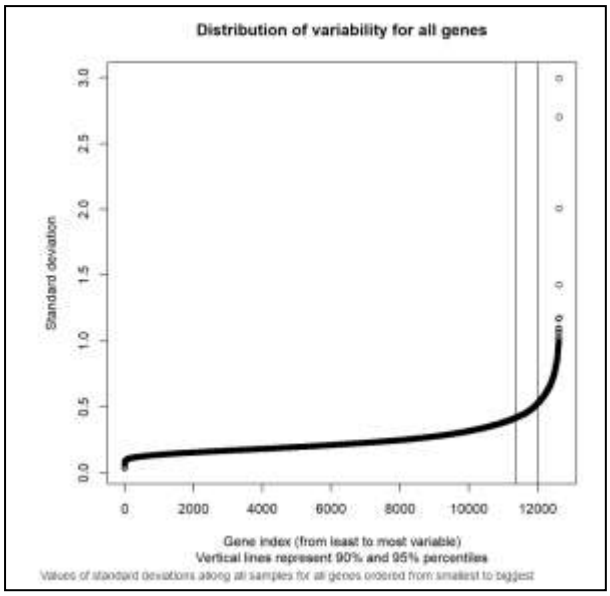

Figura 13. Distribución de variabilidad de todos los genes

Tras realizar el filtraje de los datos, donde se han seleccionado el 50% de los genes que presentan mayor variabilidad y que además, estén correctamente anotados en la base de datos de Entrez, se seleccionan 4.380 genes de los 55 pacientes. Tras crear la matriz de diseño y la matriz de contraste se realizar la selección de genes expresados de forma diferencial en las comparaciones dos a dos.

A continuación, para cada una de las 3 comparaciones, se muestran las tablas con los 20 genes que presentan un p-valor ajustado más pequeño ordenados de menor a mayor. Para cada una de las comparaciones, se muestran los siguientes estadísticos:

- logFC: Diferencia de medias entre grupos.
- AveExpr: Expresión media de todos los genes de la comparación.
- t : Estadístico t.
- P.Value: Test p-value.
- adj.P.Val: p-valor ajustado según Benjamini[24]
- B: B-statistic: Logaritmo del ratio de los genes expresados diferencialmente vs no diferenciados.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
34478_at	RAB11B	-0.39615719	8.81142173	-5.78620267	3.25E-07	0.00142256	6.38286312
33457_at	TASOR	0.35127291	7.56423507	5.03596547	5.13E-06	0.00591424	3.90093662
39945_at	FAP	0.50047665	7.00698386	5.02332572	5.37E-06	0.00591424	3.85997648
38207_at	**NEAT1**	**0.56517456**	**7.85489804**	**5.02168818**	**5.40E-06**	**0.00591424**	**3.85467239**
38038_at	LUM	0.57710895	7.46689885	4.94839025	7.03E-06	0.00615724	3.61785581
38249_at	VGLL1	-0.3478527	8.0016481	-4.63573234	2.13E-05	0.01212756	2.62207645
40007_at	IKZF1	-0.33253213	7.92532503	-4.63359027	2.15E-05	0.01212756	2.61534147
38099_r_at	ACSL4	0.3326218	6.08368208	4.62162673	2.24E-05	0.01212756	2.5777497
32656_at	MPDZ	0.31083766	6.25503338	4.58246148	2.57E-05	0.01212756	2.45496474
31495_at	XCL2	-0.35891966	9.74604429	-4.56056313	2.77E-05	0.01212756	2.38650186
41278_at	ACTL6A	0.44840212	8.23546516	4.49634679	3.46E-05	0.01262836	2.1865393
31973_at	CACNA1G	-0.31810199	9.38088579	-4.48699218	3.57E-05	0.01262836	2.15751187

38790_at	EPHX1	-0.36348821	8.68605524	-4.4616553	3.90E-05	0.01262836	2.07902375
39052_at	KRT14	-0.60357868	9.43024458	-4.4517832	4.04E-05	0.01262836	2.04849497
39923_at	**ARHGAP12**	**0.36952205**	**7.43173298**	**4.41645443**	**4.56E-05**	**0.0128861**	**1.93948902**
39507_at	OGT	0.51465793	8.36637773	4.40711172	4.71E-05	0.0128861	1.91072723
39356_at	NEDD4L	0.30772451	7.74164902	4.32798022	6.17E-05	0.01589945	1.66823249
36660_at	RAB11A	0.38576952	6.92484233	4.29982043	6.79E-05	0.0165268	1.58242888
40591_at	CDC27	0.28879418	6.49990407	4.24252246	8.25E-05	0.0180953	1.40865934
40750_at	PDE4A	-0.35681298	9.57443378	-4.2419872	8.26E-05	0.0180953	1.40704129

Tabla 1. Comparación entre DB vr IR, 20 genes con menor p-valor ordenados de menor a mayor p-valor

Los genes marcados en negrita, corresponden a los dos genes que también aparecen diferencialmente expresados en la comparación DBvsIS.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
38207_at	**NEAT1**	**0.82128422**	**7.85489804**	**7.29727334**	**1.04E-09**	**4.54E-06**	**11.914415**
34352_at	PCBD1	-0.49538474	9.80168755	-7.01799913	3.02E-09	6.62E-06	10.9132117
41064_at	PCGF1	-0.34511336	9.65590303	-6.34689292	3.92E-08	5.72E-05	8.5130661
35670_at	ATP1A3	-0.39997938	8.73573379	-6.22935656	6.12E-08	6.70E-05	8.09505087
38905_at	PRKAR2A	0.4065841	8.38434579	6.0852039	1.06E-07	7.66E-05	7.58392157
35124_at	ALOX12	-0.36012148	9.70503234	-6.0468765	1.22E-07	7.66E-05	7.44834713
38751_i_at	**ATP5ME**	**-0.58396684**	**14.0804361**	**-6.02372853**	**1.33E-07**	**7.66E-05**	**7.36653807**
36237_at	SLC22A6	-0.42877806	10.9283218	-6.01046937	1.40E-07	7.66E-05	7.31970282
33932_at	GSPT1	0.38876295	8.24454878	5.94077341	1.82E-07	8.85E-05	7.07382716
40459_at	**ACOX1**	**0.41510904**	**7.39343164**	**5.90669658**	**2.07E-07**	**9.05E-05**	**6.95380812**
40311_at	TFR2	-0.34738232	9.68716334	-5.88159975	2.27E-07	9.05E-05	6.86550376
34990_at	SETBP1	0.47431002	8.29623907	5.82230364	2.84E-07	0.00010026	6.65717286
37139_at	EDA	-0.39415413	8.94452988	-5.80960064	2.98E-07	0.00010026	6.61259971
38076_at	ATP5MC1	-0.50362152	12.7686502	-5.72586795	4.07E-07	0.00012292	6.31932683
32508_at	PRRC2C	0.81969335	8.77836752	5.71676875	4.21E-07	0.00012292	6.28751467
32734_at	PPP2R5E	0.43354826	7.16512269	5.68834824	4.68E-07	0.00012524	6.18822794
34811_at	ATP5MC3	-0.78219826	12.0329648	-5.67816404	4.86E-07	0.00012524	6.15267778
36315_i_at	**EXOC6B**	**-0.60562193**	**8.16134248**	**-5.607913**	**6.31E-07**	**0.00013407**	**5.90787069**
37254_at	ZNF133	-0.40444	8.87517429	-5.60419349	6.40E-07	0.00013407	5.89493005
242_at	MAP4	0.81042194	9.75721204	5.59988539	6.50E-07	0.00013407	5.87994429
39923_at	**ARHGAP12**	**0.46531087**	**7.43173298**	**5.56130353**	**7.50E-07**	**0.00013407**	**5.74586748**
38110_at	SDCBP	0.71914949	8.27191719	5.55987792	7.54E-07	0.00013407	5.74091786

Tabla 2. Comparación entre DB vr IS, 24 genes con menor p-valor ordenados de menor a mayor p-valor

Como los genes NEAT1 y ARHGAP12 aparecen diferencialmente expresados en la comparación DBvsIR, se introducen 2 genes más, ya que posteriormente se van a eliminar duplicados.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
41316_s_at	SAFB	0.5993214	9.52068765	7.2305267	1.34E-09	5.87E-06	11.6557926
36988_at	TNFAIP1	-0.36085963	8.78733814	-6.47674609	2.39E-08	4.84E-05	8.96488623

36315_i_at	**EXOC6B**	**-0.63896567**	**8.16134248**	**-6.39073181**	**3.32E-08**	**4.84E-05**	**8.65902876**
33329_at	NFIC	0.34843629	9.80583674	6.22567853	6.21E-08	5.76E-05	8.0734709
33372_at	RAB31	-0.46668757	8.35018991	-6.15905347	7.99E-08	5.76E-05	7.83769565
35892_at	CR1	-0.50185741	6.91552933	-6.13894708	8.62E-08	5.76E-05	7.76661691
38751_i_at	**ATP5ME**	**-0.5494132**	**14.0804361**	**-6.1213849**	**9.21E-08**	**5.76E-05**	**7.70456182**
39143_at	NFATC1	-0.40557532	8.20026715	-6.02815812	1.31E-07	7.17E-05	7.37563588
36540_at	RHOBTB2	-0.46508546	8.88426853	-5.86568154	2.41E-07	0.00011737	6.80455289
36151_at	PLD3	0.35770304	8.22117722	5.81465936	2.92E-07	0.00012453	6.62585189
32644_at	NUP188	-0.44915855	9.36471426	-5.79629861	3.13E-07	0.00012453	6.56162507
31527_at	RPS2	-0.51562039	13.3625436	-5.67056581	5.00E-07	0.00018252	6.12301862
32545_r_at	RSU1	-0.51478708	9.34344639	-5.55322916	7.73E-07	0.00022702	5.71576851
41091_at	BPTF	0.68294266	8.45334067	5.5499616	7.83E-07	0.00022702	5.70445817
31717_at	PIN1P1	-0.39652909	7.81327394	-5.53495861	8.27E-07	0.00022702	5.65254892
34656_at	MPP2	-0.30701246	8.22746185	-5.53433271	8.29E-07	0.00022702	5.65038414
36516_at	ZNF473	-0.34705087	9.3784589	-5.51009936	9.07E-07	0.00023371	5.56661867
39969_at	H4C3	-0.65127918	9.8315898	-5.4667044	1.06E-06	0.0002429	5.4168641
39106_at	APOA1	-0.34661177	7.54843801	-5.45001871	1.13E-06	0.0002429	5.35936801
33742_f_at	ATP6V1H	-0.9349265	8.83903822	-5.44302276	1.16E-06	0.0002429	5.33527556
40459_at	**ACOX1**	**0.35410943**	**7.39343164**	**5.44243626**	**1.16E-06**	**0.0002429**	**5.33325618**
39386_at	MAD2L1BP	-0.40898111	8.165344	-5.34082153	1.69E-06	0.00032918	4.98432066
33778_at	TBC1D22A	-0.28749097	10.3891952	-5.33503493	1.73E-06	0.00032918	4.96450752

Tabla 3. Comparación entre IR vr IS, 20 genes con menor p-valor ordenados de menor a mayor p-valor

Como los genes EXOC6B, ATP5ME y ACOX1 aparecen diferencialmente expresados en la comparación DBvsIS, se introducen 3 genes más, ya que posteriormente se van a eliminar duplicados.

Por lo tanto, los 60 genes utilizados para el proceso de predicción se muestran en la siguiente tabla:

N	DBvsIR	DBvsIS	IRvsIS
1	RAB11B	PCBD1	SAFB
2	TASOR	PCGF1	TNFAIP1
3	FAP	ATP1A3	NFIC
4	NEAT1	PRKAR2A	RAB31
5	LUM	ALOX12	CR1
6	VGLL1	ATP5ME	NFATC1
7	IKZF1	SLC22A6	RHOBTB2
8	ACSL4	GSPT1	PLD3
9	MPDZ	ACOX1	NUP188
10	XCL2	TFR2	RPS2
11	ACTL6A	SETBP1	RSU1
12	CACNA1G	EDA	BPTF
13	EPHX1	ATP5MC1	PIN1P1
14	KRT14	PRRC2C	MPP2
15	ARHGAP12	PPP2R5E	ZNF473

16	OGT	ATP5MC3	H4C3
17	NEDD4L	EXOC6B	APOA1
18	RAB11A	ZNF133	ATP6V1H
19	CDC27	MAP4	MAD2L1BP
20	PDE4A	SDCBP	TBC1D22A

Tabla 4. Genes utilizados como variables predictoras en el modelo de ML

Continuando con el proceso, se muestran los "volcano plots" para cada una de las comparaciones. Se muestran los nombres de los 5 genes con el p-valor ajustado más pequeño. Se puede observar que genes están sobre-expresados o infra-expresados.

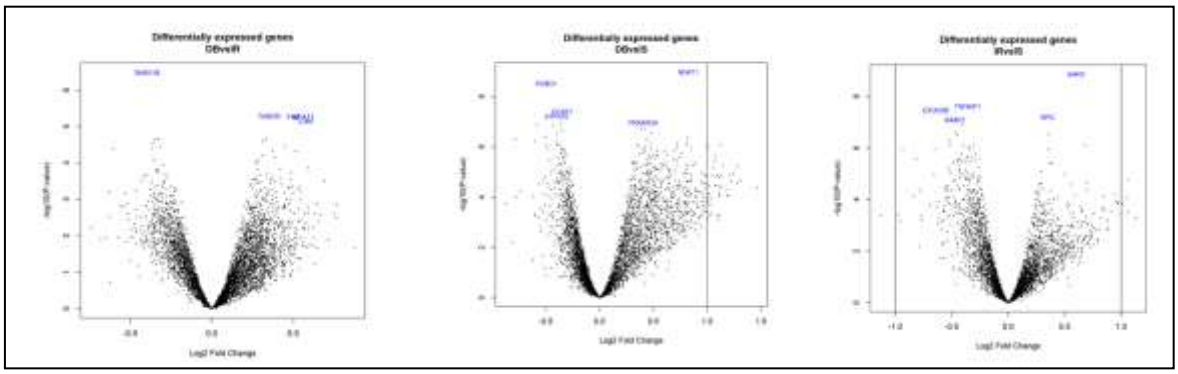

Figura 14. Volcano plot de las distintas comparaciones

En el apartado de las comparaciones múltiples:

- Se puede observar que hay 294 genes expresado de forma diferencial entre DB y IR:
  - Hay 139 genes poco expresados entre DB y IR
  - Hay 155 genes sobreexpresados entre DB y IR
- Se puede observar que hay 1843 genes expresado de forma diferencial entre DB y IS:
  - Hay 731 genes poco expresados entre DB y IS
  - Hay 1112 genes sobreexpresados entre DB y IS
- Se puede observar que hay 863 genes expresado de forma diferencial entre IR y IS:
  - Hay 511 genes poco expresados entre IR y IS
  - Hay 352 genes sobreexpresados entre IR y IS

	DBvsIR	DBvsIS	IRvsIS
Down	139	731	511
NotSig	4086	2537	3517
Up	155	1112	352

Figura 15. Nº de genes infra o sobre regulados

En el diagrama de Venn inferior se puede observar cuantos genes se expresan de forma diferencial entre las tres comparaciones sin diferenciar entre genes up o down

regulados. Se puede observar que hay 4 genes que se expresan de forma diferencial en las 3 comparaciones.

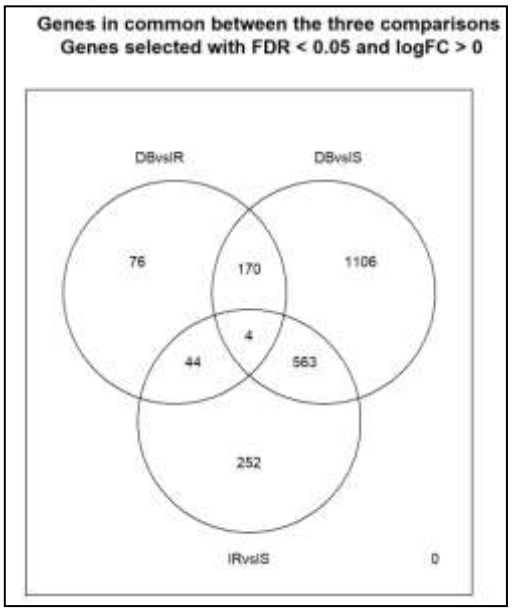

Figura 16. Diagrama de Venn

A través de un "heatmap" o mapa de calor, se puede visualizar las expresiones de cada gen agrupándolas para destacar los genes que se encuentran "*up*" o "*down*" regulados simultáneamente constituyendo perfiles de expresión. En la imagen inferior se muestran respectivamente, el "heatmap" con los 2.215 genes que se expresan de forma diferencial en alguna de las comparaciones (izquierda), el heatmap con los 60 genes utilizados en el proceso de ML (centro) y el heatmap de los 4 genes que se expresan de forma diferencial en las 3 comparaciones.

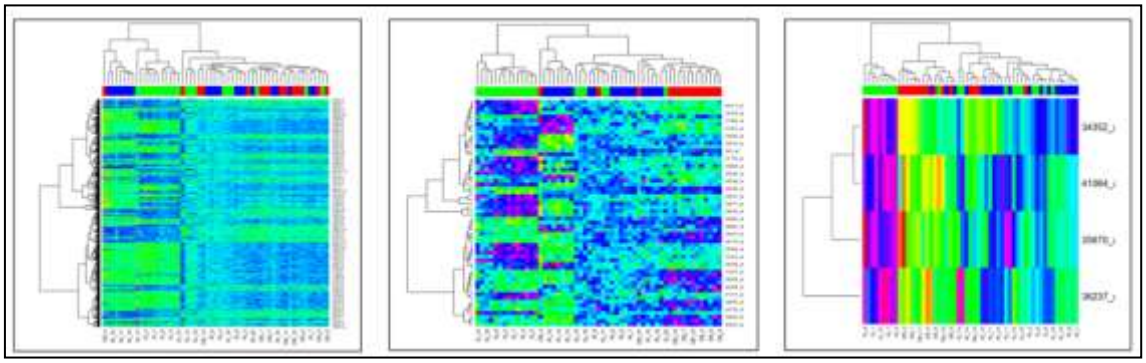

Figura 17. Heatmaps de genes expresados de forma diferencial agrupados por similariada

En los "*heatmaps*", se puede ver que el modelo del centro, con los 60 genes, es el que mejor discrimina entre las categorías IS, IR y DB. A la izquierda, se puede observar que con estos 4 genes, no es suficiente para elaborar modelos predictivos válidos a pesar de expresarse de forma diferencial en las 3 comparaciones.

En relación a los resultados de análisis de significación biológica, se describen los resultados respecto de dos bases de datos.

En relación a la base de datos GO, se han realizados las búsquedas tanto para los genes sobre-expresados como infra-expresados, por lo que se han obtenido 6 archivos en formato html. En la tabla inferior se enlazan los 3 primeros enlaces de cada resultado:

Comparación	Proceso
DBvsIR Over-representation	spleen development embryonic heart tube left/right pattern formation left/right pattern formation
DBvsIR under-representation	mRNA catabolic process RNA catabolic process translation
DBvsIS Over-representation	mRNA metabolic process regulation of translation posttranscriptional regulation of gene expression
DBvsIS under-representation	extracellular matrix organization extracellular structure organization external encapsulating structure organization
IRvsIS Over-representation	translational initiation SRP-dependent cotranslational protein targeting to membrane nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
IRvsIS Under-representation	external encapsulating structure organization urogenital system development extracellular matrix organization

Tabla 5. Procesos obtenidos de la base de datos GO partiendo de los genes expresados de forma diferencial

En relación a la base de datos de procesos de REACTOME, no se encuentran relaciones entre los genes expresados de forma diferencial entre los grupos DBvsIR, esto puede ser debido a que únicamente hay 294 genes expresados de forma diferencial. En relación a las otras dos comparaciones, únicamente se tienen en cuenta los genes cuyo p-valor ajustado es menor de 0.05.

En relación a la comparación entre DB e IS, se describen los 5 principales procesos relacionados con los genes expresados de forma diferencial:

- Signaling by ROBO receptors
- Regulation of expression of SLITs and ROBOs
- Eukaryotic Translation Initiation
- Cap−dependent Translation Initiation
- L13a−mediated translational silencing of Ceruloplasmin expression

En relación a la comparación entre IR e IS, se describen los 5 principales procesos relacionados con los genes expresados de forma diferencial:

- Eukaryotic Translation Initiation
- Cap−dependent Translation Initiation
- GTP hydrolysis and joining of the 60S ribosomal subunit
- L13a−mediated translational silencing of Ceruloplasmin expression
- Regulation of expression of SLITs and ROBOs

En la figura inferior, se detalla de forma gráfica las redes producidas por los genes expresados de forma diferencial:

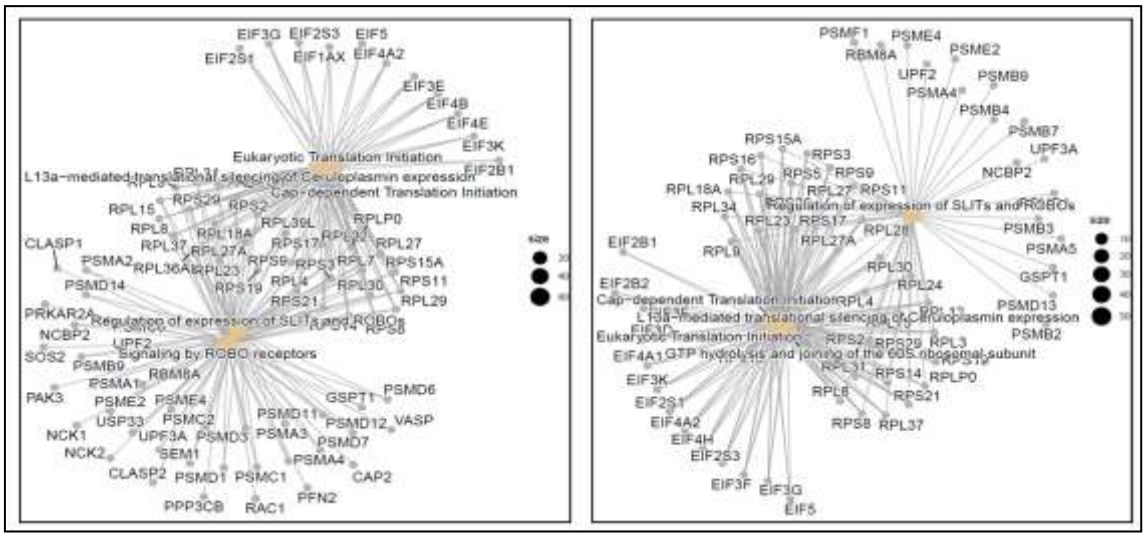

Figura 18. A la izda., red producida por los genes expresados de forma diferencial entre DBvsIS y a la derecha IRvsIS

5.2 Resultados asociados al proceso de Machine Learning

A continuación, pasan a detallarse los resultados del proceso de ML.

En primer lugar, se ha realizado una descripción gráfica de las variables predictivas (genes), un diagrama de cajas y bigotes y el gráfico de dos dimensiones asociado a un ACP realizado sobre las 60 variables.

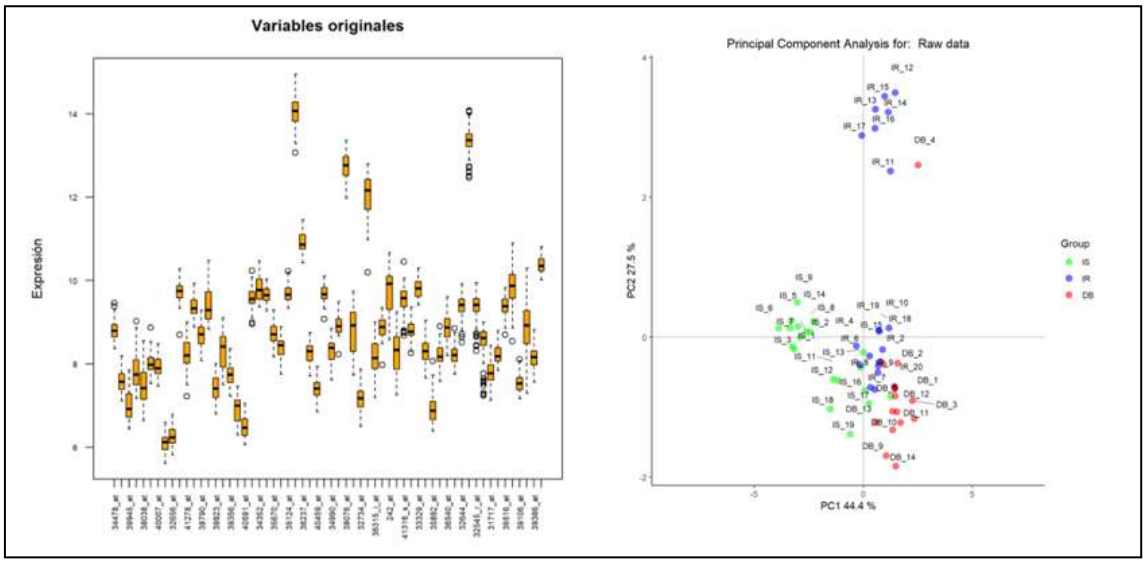

Figura 19. Diagrama de cajas y bigotes (izda.) y ACP (derecha).

En la figura de la izquierda se puede observar que va a ser necesaria una normalización de las variables, mientras que en la figura de la derecha, se puede apreciar que los grupos están muy bien clasificados, a excepción de los datos que aparecen en la parte superior de la imagen.

Una vez dividido el conjunto de datos de entrada, en datos de entrenamiento (70% de los datos) y datos de validación (30% de los datos), se normaliza el conjunto de datos de entrenamiento mediante la siguiente fórmula:

$$Z = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Y posteriormente, se normaliza el conjunto de datos de validación con los valores máximos y mínimos obtenidos en el conjunto de datos de entrenamiento para cada una de las variables. A continuación se muestra una figura de los datos de entrenamiento y validación normalizados mediante el siguiente procedimiento:

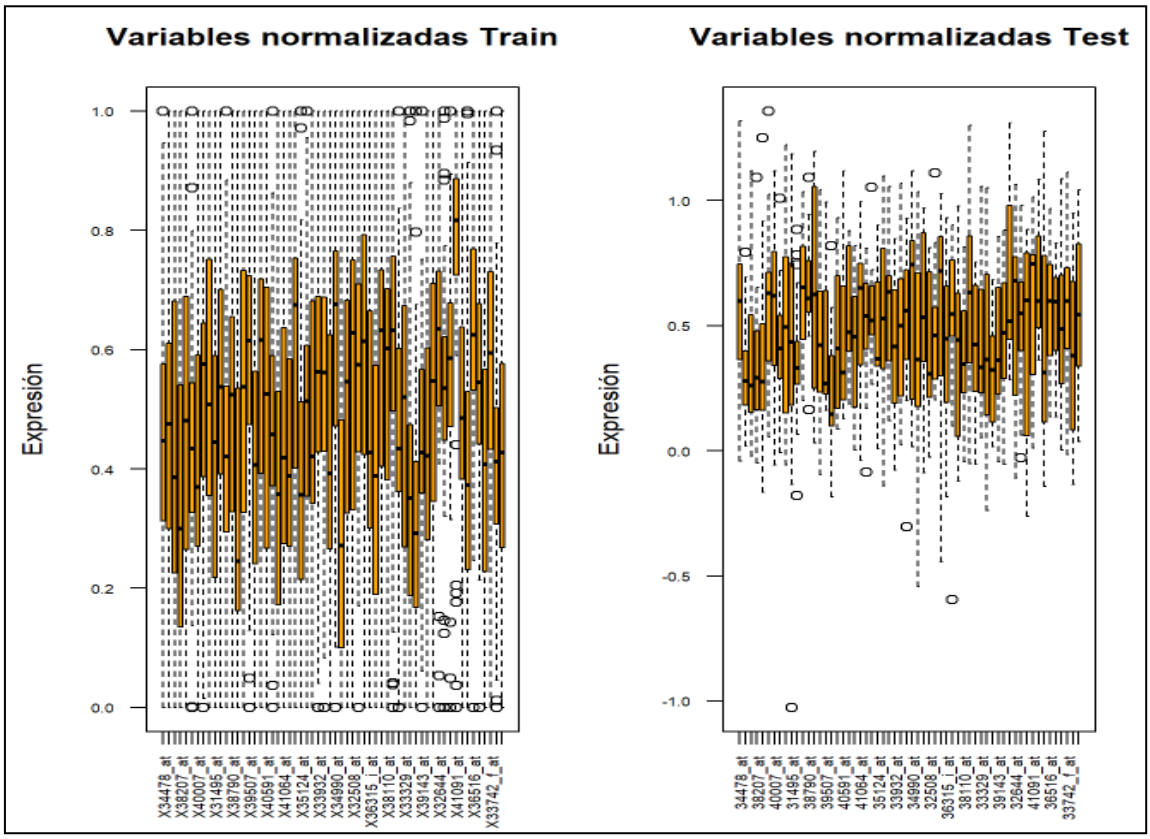

Figura 20. Datos de entrenamiento y validación normalizados

Una vez normalizado el conjunto de datos de entrenamiento y validación, se procesan las 7 técnicas utilizadas:

- Multilayer perceptron (MLP)
- K-nearest neighbors (KNN)
- Redes neuronales (ANN)
- Redes neuronales utilizando las variables obtenidas en el ACP (ANN-ACP)
- Support Vector Machine gaussiano (SVM-gaussiano)
- Support Vector Machine lineal (SVM-lineal)
- Random Forest (RF)

Después de ejecutar las técnicas, con los datos de validación, se construye una matriz de confusión 3*3 para cada técnica donde se calcula la exactitud (accuracy). Posteriormente, en esta matriz de confusión, se unifican las categorías IR y DB (teniendo en cuenta la similitud que presentan) y se construye una nueva matriz de confusión 2*2 donde se calcula la exactitud, sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo.

Con la intención de obtener unos resultados robustos y no debidos al azar, el proceso de partición de datos de entrenamiento, validación y ejecución de las distintas técnicas, se repite 1.000 veces, donde en cada una de ellas, se van almacenando los resultados de cada ejecución.

Una vez realizadas las 1.000 ejecuciones, los resultados finales se han calculado de dos formas distintas, siendo muy similares en ambos casos:

En primer lugar, se muestran los resultados obtenidos a través de una matriz de confusión con 17.000 (17x1.000) registros que se ha ido acumulando tras cada una de las 1.000 ejecuciones. La exactitud de 2 grupos se refiere a la exactitud obtenida tras unificar en la matriz de confusión 3x3 las categorías IR y DB, y posteriormente, en esta matriz de 2x2 (IS e IR-DB), se calculan la sensibilidad, especificidad, vpp y vpn:

Técnica	Exactitud 3G	Exactitud 2G	Sens	Espec	VPP	VPN
MLP	95.42	96.31	97.65	93.94	96.58	95.8
KNN	85.51	90.65	98.49	76.88	88.2	96.68
ANN	88.34	92.96	96.04	87.57	93.13	92.65
ANN-PCA	89.01	91.99	93.32	89.65	94.05	88.44
SVM-radial	89.55	93.06	99.54	81.69	90.51	99.02
SVM-lineal	90.99	94.53	97.09	90.04	94.47	94.64
RF	80.97	89.57	95.14	79.8	89.2	90.35

Tabla 6. Resultados obtenidos a través de una matriz de confusión con datos acumulados unificando IR y DB como categoría positiva.

En segundo lugar, se muestran los resultados obtenidos como las medias de los resultados de cada una de las 1.000 ejecuciones. Respecto a la matriz 2x2, la unión de las categorías IR y DB se establece como categoría positiva :

Técnica	Exactitud 3G	Exactitud 2G	Sens	Espec	VPP	VPN
MLP	95.42	96.31	97.51	93.87	96.61	96.37
KNN	85.51	90.65	98.6	77.77	88.6	97.08
ANN	88.34	92.96	96.18	87.85	93.28	93.17
ANN-PCA	89.01	91.99	93.5	90.08	94.26	89.11
SVM-radial	89.55	93.06	99.58	82.5	90.64	99.06
SVM-lineal	90.99	94.53	97.16	90.29	94.55	94.81
RF	80.97	89.57	95.44	80.45	89.5	90.56

Tabla 7. Resultados obtenidos como media de los resultados de cada una de las 1.000 ejecuciones

Se puede observar que los resultados obtenidos calculados de las dos formas, son iguales en la exactitud y muy similares en la sensibilidad, especificidad, vpp y vpn. La exactitud (3 grupos) es superior al 80% en todos los casos, siendo superior al 90% con las técnicas MLP y SVM-lineal. Si nos fijamos en la exactitud unificando las categorías IR y DB (2 grupos), ésta alcanza el 90% en todos los casos excepto en RF, que se queda rozando el 90% (89.57%). Se puede observar el buen comportamiento de todas las técnicas y en particular de MLP.

En relación a la sensibilidad, es superior al 90% en todos los casos, siendo superior del 99.50% con la técnica SVM-radial. En relación a la especificidad, los resultados varían más, la técnica MLP es la que presenta un valor superior con un valor cercano al 94% y la técnica KNN es la que presenta el valor más bajo con un valor entorno al 77%. En relación al VPP, la técnica que presenta mayor grado de acierto vuelve a ser MLP, con un valor entorno al 96.6% y la menor vuelve a ser KNN con un valor entorno al 88%. Respecto al VPN, todas las técnicas presentan valores cercanos al 90% (ANN-PCA) o superiores, alcanzando casi la predicción perfecta la técnica SVM-radial con un 99.00% de predicción. Hay que recordar, que a igualdad de prevalencia, el vpp está afectado por la especificidad, cuanta más alta es la especificidad, más alta va a ser el vpp (suponiendo constante la prevalencia). Y lo mismo ocurre con el vpn, cuando más alta sea la sensibilidad, más alta va a ser el vpn (suponiendo constante la prevalencia). Por ello, la técnica que presenta especificidad más alta es MLP y por lo tanto, la que presenta mayores valores de vpp también es MLP. Lo mismo ocurre con vpn, la técnica que presenta la sensibilidad más alta es SVM-radial y por lo tanto esta misma es la que presenta valores más altos de vpn.

6 Discusión

Este análisis está compuesto por dos hipótesis de trabajo complementarias y se deseaba saber si con un número determinado de genes expresados de forma diferencial en la comparativa entre tres grupos (IS, IR y DB) en una análisis de microarrays iban a ser válidos posteriormente para predecir la variable objetivo compuesta por esos tres grupos, utilizando como variables predictivas dicho número determinado de genes expresados de forma diferencial. Una vez analizados los resultados, se puede afirmar que se han encontrado genes expresados de forma diferencial entre los 3 grupos, también se ha podido comprobar su significación biológica, aunque en relación a la comparativa DBvsIR no se ha obtenido ningún resultado en la base de datos ReactomePA, aunque sí en la base de datos GO. Posteriormente, tras la selección de los 20 genes expresados de forma diferencial entre las 3 comparaciones con el p-valor ajustado más bajo, se ha realizado unos modelos predictivos con la finalidad de comprobar si con estos 20 genes es suficiente para predecir con un alto grado de validez la variable objetivo (IS, IR y DB). A la vista de los resultados, se puede afirmar que estos 60 genes sirven para predecir con una gran validez la variable objetivo. Sin embargo, hay que ser cautelosos con los resultados obtenidos, ya que, además del problema típico del efecto Batch en el análisis de datos de microarrays, hay que tener en cuenta que el tamaño muestra está formado por 55 individuos, ello hace que el estudio tenga poca potencia y por lo tanto haya más falsos negativos en caso de utilizar un tamaño muestral superior.

Por otra parte, merece la pena comentar y una vez visto los gráficos y resultados, que el grupo IR es más similar al grupo DB que al grupo IS. Esto es curioso ya que en uno de los dos estudios publicados con estos datos (Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning)[18], trabaja con los pacientes IS e IR y los cataloga como individuos sanos.

También merece la pena comentar que se hicieron unos análisis preliminares predictivos con los 4 genes que se expresaban de forma diferencial en las 3 comparaciones, pero como se puede observar en la figura 15 relacionada con el heatmap de estos 4 genes, estos 4 genes no son suficientes para realizar predicciones válidas y por ello ni tan siquiera se han incluido estos resultados en el análisis.

7 Conclusiones

7.1 Conclusiones

En este análisis se planteaba el funcionamiento del binomio entre los genes expresados de forma diferencial a través de un análisis de datos de microarrays y la posterior comprobación de la efectividad de estos datos para realizar predicciones a través de herramientas de ML.

Por otra parte, a la vista de los resultados se puede concluir que un número determinado de genes que se expresan de forma diferencial sirven para realizar modelos predictivos con una alta validez, esto sería interesante poder llevarlo a la práctica por el beneficio de los pacientes o personas en riesgo de ser diabéticas.

Se han cumplido los objetivos propuestos, encontrar genes expresados de forma diferencial y sobretodo, observar el alto poder predictivo obtenido con estos genes, obteniendo una exactitud del 95% utilizando la técnica de MLP, por lo tanto, se puede decir que los objetivos han sido alcanzados.

7.2 Líneas de futuro

Para futuras líneas de trabajo, habría que confirmar con técnicas biológicas más avanzadas como técnicas de ultrasecuenciación, si se confirman los genes que se han expresados de forma diferencial en este análisis. También sería bastante útil intentar reproducir este análisis con un tamaño muestral mayor e intentando minimizar los efectos Batch, como puede ser la fecha de hibridación de los arrays.

En este análisis, se han utilizado técnicas de ML y una técnica de Deep Learning como ha sido MLP, pero sería interesante realizar los cálculos con otros técnica predictivas como redes convolucionales, mapas autoorganizados (a pesar que los mapas autoorganizados son técnicas no supervisadas, en el paquete kohonen[27] de R se puede realizar predicciones) o realizar los cálculos con otros parámetros a los utilizamos con la finalidad de aumentar la validez de las predicciones.

Por último, en este estudio se han seleccionado 20 genes de cada una de las 3 comparaciones realizadas debido a que el tamaño muestral era de 55 pacientes. Sería interesante comprobar si aumentando o disminuyendo el número de variables, sería posible aumentar los valores predictivos.

Una vez validado los pasos anteriores, sería estupendo crear una herramienta on-line, donde se puedan introducir los valores de un número determinado genes y que se pueda obtener la predicción en relación a la clase de IS, IR o DB.

7.3 Seguimiento de la planificación

Se ha seguido la planificación asociada a las PECs. Por lo tanto, se considera que la metodología aplicada ha sido la correcta. No ha habido que realizar cambios significativos ya que los resultados obtenidos han sido satisfactorios, se han encontrado genes expresados de forma diferencial y la validez obtenida en la parte predictiva ha sido superior al 90%.

8 Glosario

Ordenado alfabéticamente:

ACP: Análisis de componentes principales
ADO: Análisis de datos ómicos
ANN: Artificial neural network (redes neuronales artificiales)
ARN: Ácido ribonucleico
CEL: Cell Intensity File. Archivos que contienen la información de los arrys
DB: Diabéticos
DM: Diabetes Mellitus
DMG: Diabetes mellitus gestacional
Espec: Especificidad
GEO: Gene Expresion Omnibus
GO: Gene Ontology (base de datos Gene Ontology)
IR: Insulina-resistentes
IS: Insulina-sensibles
KNN: k-nearest neighbors (k vecinos más próximos)
MIAME: Minimum information about a microarray experiment
ML: Machine Learning
MLP: Multilayer perceptron
PEC: Prueba de evaluación continua
ReactomePA: Reactome Pathway Database (base de datos Reactome)
RMA: Robust Multi-Array Average
RF: Random forest
Sens: Sensibilidad
SOG: Sobrecarga oral de glucosa
SVM-gaussiano: Support vector machine gaussiano
SVM-lineal: Support vector machine lineal
TAG: Tolerancia anormal a la glucosa
TFG: Trabajo fin de grado
VPN: Valor predictivo negativo
VPP: Valor predictivo positivo

9 Bibliografía

1. Williams R, Colagiuri AR, Aschner Montoya B, others. Atlas de la Diabetes de la FID. Fed Int Diabetes Suvi Karuranga, Belma Malanda, Pouya Saeedi, Paraskevi Salpea. 2019;
2. Gheibi S, Singh T, da Cunha JPMCM, Fex M, Mulder H, others, et al. Insulin/glucose-responsive cells derived from induced pluripotent stem cells: disease modeling and treatment of diabetes. Cells. 2007;9(1):2465.
3. Petersen MC, Shulman GI. Mechanisms of Insulin Action and Insulin Resistance. Physiol Rev. 2018;98:2133-223.
4. Sanz RG, Sánchez-Pla A. Statistical Analysis of Microarray Data. En: Microarray Bioinformatics. Springer; 2019. p. 87-121.
5. Sánchez-Pla, Alex; Gonzalo Sanz R. Análisis de datos ómicos.
6. Mediavilla Bravo JJ, Alonso Fernández M, Moreno Moreno A, Carramiñana Barrera F, SO S, Holland JH (John H, et al. Guías clínicas diabetes mellitus. Diabetes. 2019;9(2):5-32.
7. Salud SC de la. Estrategia de Abordaje de la Diabetes Mellitus en Canarias. 2021.
8. R Core Team, Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, others. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115-21.
9. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021.
10. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257-8.
11. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12(2):477-9.
12. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-10.
13. Wu X, Wang J, Cui X, Maianu L, Rhees B, Rosinski J, So WV, Willi SM, Osier M V, Hill HS, others. The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle. Endocrine. 2007;31(1):5-17.
14. Virkamäki A, Ueki K, Kahn CR. Protein–protein interaction in insulin signaling and the molecular mechanisms of insulin resistance. J Clin Invest. 1999;103(7):931-43.
15. Pawson T, Scott JD. Signaling through scaffold, anchoring, and adaptor proteins. Science (80-). 1997;278(5346):2075-80.
16. DeFronzo RA. The triumvirate: $β$-cell, muscle, liver: a collusion responsible for NIDDM. Diabetes. 1988;37(6):667-87.
17. DeFronzo RA, Jacot E, Jequier E, Maeder E, Wahren J, Felber JP. The effect of insulin on the disposal of intravenous glucose: results from indirect calorimetry and hepatic and femoral venous catheterization.

Diabetes. 1981;30(12):1000-7.

18. Wu X, Patki A, Lara-Castro C, Cui X, Zhang K, Walton RG, Osier M V, Gadbury GL, Allison DB, Martin M, others. Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning. J Appl Physiol. 2011;110(3):746-55.
19. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.
20. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, others. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29(4):365-71.
21. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25(3):415-6.
22. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Vol. 4, Biostatistics. 2003.
23. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1).
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289-300.
25. Kalinowski T, Allaire J, Chollet F. Guide to the Sequential Model [Internet]. Disponible en: https://keras.rstudio.com/articles/sequential_model.html
26. Lantz B. Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd; 2019.
27. Wehrens R, Buydens LMC. Self- and super-organizing maps in R: The kohonen package. J Stat Softw. 2007;21(5):1-19.

10 Anexos

En primer lugar, se muestran los primeros registros de los procesos obtenidos con los genes expresados de forma diferencial en las 3 comparaciones, tanto para los genes sobre-expresados como infra-expresados en la base de datos de GO.

Gene to GO BP test for over-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0048536	0.000	9.491	1	6	15	spleen development
GO:0060971	0.000	Inf	0	3	3	embryonic heart tube left/right pattern formation
GO:0060972	0.001	42.336	0	3	4	left/right pattern formation
GO:0046621	0.002	7.091	1	5	15	negative regulation of organ growth
GO:0046661	0.003	3.034	4	11	63	male sex differentiation
GO:1901135	0.003	1.738	23	36	342	carbohydrate derivative metabolic process
GO:1901137	0.004	1.993	13	23	191	carbohydrate derivative biosynthetic process
GO:0007389	0.004	2.403	7	15	105	pattern specification process
GO:0055022	0.004	8.081	1	4	11	negative regulation of cardiac muscle tissue growth
GO:0061117	0.004	8.081	1	4	11	negative regulation of heart growth
GO:0019376	0.004	Inf	0	2	2	galactolipid catabolic process
GO:0060075	0.004	Inf	0	2	2	regulation of resting membrane potential
GO:0070986	0.004	Inf	0	2	2	left/right axis specification
GO:0090234	0.004	Inf	0	2	2	regulation of kinetochore assembly
GO:1901374	0.004	Inf	0	2	2	acetate ester transport
GO:2000563	0.004	Inf	0	2	2	positive regulation of CD4-positive, alpha-beta T cell proliferation
GO:0007350	0.005	14.105	0	3	6	blastoderm segmentation
GO:0060044	0.005	14.105	0	3	6	negative regulation of cardiac muscle cell proliferation
GO:0030004	0.005	3.830	2	7	33	cellular monovalent inorganic cation homeostasis
GO:0007548	0.006	2.368	7	14	99	sex differentiation

Figura 21. Análisis de enriquecimiento de los genes sobre-expresados en la comparación DBvsIR en GO

Gene to GO BP test for under-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006402	0.000	0.146	12	2	186	mRNA catabolic process
GO:0006401	0.001	0.206	13	3	198	RNA catabolic process
GO:0006412	0.002	0.338	16	6	244	translation
GO:0006518	0.003	0.407	20	9	304	peptide metabolic process
GO:0009057	0.003	0.533	34	20	513	macromolecule catabolic process
GO:0016071	0.003	0.463	24	12	357	mRNA metabolic process
GO:0043043	0.004	0.380	17	7	254	peptide biosynthetic process
GO:1903047	0.004	0.442	21	10	312	mitotic cell cycle process
GO:0044265	0.005	0.529	30	17	443	cellular macromolecule catabolic process
GO:0000184	0.005	0.000	5	0	75	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
GO:0044270	0.007	0.429	17	8	259	cellular nitrogen compound catabolic process
GO:0046700	0.007	0.429	17	8	259	heterocycle catabolic process
GO:0031329	0.008	0.497	22	12	335	regulation of cellular catabolic process
GO:0000956	0.009	0.141	7	1	98	nuclear-transcribed mRNA catabolic process
GO:0030258	0.010	0.144	6	1	96	lipid modification
GO:0031331	0.012	0.294	9	3	142	positive regulation of cellular catabolic process
GO:0034655	0.013	0.454	16	8	246	nucleobase-containing compound catabolic process
GO:1903311	0.014	0.346	11	4	161	regulation of mRNA metabolic process
GO:0009896	0.014	0.349	11	4	160	positive regulation of catabolic process
GO:0043604	0.015	0.514	20	11	299	amide biosynthetic process

Figura 22. Análisis de enriquecimiento de los genes infra-expresados en la comparación DBvsIR en GO

Gene to GO BP test for over-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0016071	0.000	1.540	150	185	357	mRNA metabolic process
GO:0006417	0.000	1.953	60	83	143	regulation of translation
GO:0010608	0.000	1.649	109	139	259	posttranscriptional regulation of gene expression
GO:0034248	0.000	1.846	70	94	166	regulation of cellular amide metabolic process
GO:0006302	0.000	2.568	29	44	68	double-strand break repair
GO:0006996	0.000	1.287	507	560	1206	organelle organization
GO:0006281	0.000	1.794	67	89	159	DNA repair
GO:0043087	0.000	1.841	60	81	143	regulation of GTPase activity
GO:0006282	0.000	4.005	13	23	31	regulation of DNA repair
GO:0006886	0.001	1.420	169	200	401	intracellular protein transport
GO:0017148	0.001	2.572	23	35	54	negative regulation of translation
GO:0007052	0.001	2.888	18	29	43	mitotic spindle organization
GO:0046907	0.001	1.338	246	282	585	intracellular transport
GO:0044248	0.001	1.290	339	380	807	cellular catabolic process
GO:2001022	0.001	3.652	12	21	29	positive regulation of response to DNA damage stimulus
GO:0120031	0.001	1.688	65	84	154	plasma membrane bounded cell projection assembly
GO:0034249	0.001	2.270	26	39	63	negative regulation of cellular amide metabolic process
GO:0051276	0.001	1.395	164	193	390	chromosome organization
GO:0045184	0.001	1.304	274	310	652	establishment of protein localization
GO:0033044	0.001	1.890	42	58	101	regulation of chromosome organization

Figura 23. Análisis de enriquecimiento de los genes sobre-expresados en la comparación DBvsIS en GO

Gene to GO BP test for under-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0030198	0.000	0.364	51	26	122	extracellular matrix organization
GO:0043062	0.000	0.364	51	26	122	extracellular structure organization
GO:0045229	0.000	0.374	52	27	124	external encapsulating structure organization
GO:0048513	0.000	0.736	468	408	1114	animal organ development
GO:0032501	0.000	0.793	906	847	2156	multicellular organismal process
GO:0022610	0.000	0.693	204	167	486	biological adhesion
GO:1902105	0.000	0.440	41	24	98	regulation of leukocyte differentiation
GO:0007155	0.000	0.701	203	167	483	cell adhesion
GO:0048646	0.000	0.675	158	126	375	anatomical structure formation involved in morphogenesis
GO:1903706	0.000	0.530	60	40	142	regulation of hemopoiesis
GO:0048731	0.000	0.802	638	586	1517	system development
GO:0071294	0.000	0.000	6	0	14	cellular response to zinc ion
GO:0007275	0.001	0.808	690	638	1641	multicellular organism development
GO:0031638	0.001	0.130	10	2	23	zymogen activation
GO:0016485	0.001	0.393	28	13	67	protein processing
GO:1904029	0.001	0.264	16	6	37	regulation of cyclin-dependent protein kinase activity
GO:0040008	0.001	0.635	98	75	234	regulation of growth
GO:0030808	0.001	0.000	5	0	13	regulation of nucleotide biosynthetic process
GO:1900371	0.001	0.000	5	0	13	regulation of purine nucleotide biosynthetic process
GO:0002237	0.001	0.530	51	34	121	response to molecule of bacterial origin

Figura 24. Análisis de enriquecimiento de los genes infra-expresados en la comparación DBvsIS en GO

Gene to GO BP test for over-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006413	0.000	4.283	21	33	106	translational initiation
GO:0006614	0.000	5.469	12	33	62	SRP-dependent cotranslational protein targeting to membrane
GO:0000184	0.000	4.581	15	39	75	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
GO:0006613	0.000	5.088	13	35	64	cotranslational protein targeting to membrane
GO:0019080	0.000	3.831	19	46	97	viral gene expression
GO:0006612	0.000	3.794	18	43	91	protein targeting to membrane
GO:0019083	0.000	3.794	18	43	91	viral transcription
GO:0045047	0.000	4.466	13	35	68	protein targeting to ER
GO:0072599	0.000	4.209	14	35	70	establishment of protein localization to endoplasmic reticulum
GO:0070972	0.000	3.639	16	38	82	protein localization to endoplasmic reticulum
GO:0006401	0.000	2.341	39	70	198	RNA catabolic process
GO:0006412	0.000	2.178	48	82	244	translation
GO:0043043	0.000	2.127	50	84	254	peptide biosynthetic process
GO:0006605	0.000	2.451	33	61	167	protein targeting
GO:0000956	0.000	3.030	19	41	98	nuclear-transcribed mRNA catabolic process
GO:0006402	0.000	2.289	37	63	186	mRNA catabolic process
GO:0016071	0.000	1.854	70	107	357	mRNA metabolic process
GO:0006518	0.000	1.929	60	94	304	peptide metabolic process
GO:0090150	0.000	2.466	28	53	144	establishment of protein localization to membrane
GO:0016032	0.000	1.728	86	124	436	viral process

Figura 25. Análisis de enriquecimiento de los genes sobre-expresados en la comparación IRvsIS en GO

Gene to GO BP test for under-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0045229	0.000	0.311	24	9	124	external encapsulating structure organization
GO:0001655	0.000	0.279	21	7	107	urogenital system development
GO:0030198	0.000	0.317	24	9	122	extracellular matrix organization
GO:0043062	0.000	0.317	24	9	122	extracellular structure organization
GO:0072001	0.000	0.276	18	6	93	renal system development
GO:0001822	0.001	0.289	18	6	89	kidney development
GO:0060485	0.001	0.326	18	7	93	mesenchyme development
GO:0007423	0.001	0.467	32	17	162	sensory organ development
GO:0048856	0.001	0.780	348	308	1763	anatomical structure development
GO:0001501	0.001	0.462	30	16	154	skeletal system development
GO:0032501	0.001	0.790	425	386	2156	multicellular organismal process
GO:0007610	0.002	0.512	36	21	184	behavior
GO:0002088	0.002	0.000	6	0	29	lens development in camera-type eye
GO:0032502	0.002	0.793	375	337	1901	developmental process
GO:0035295	0.002	0.635	68	48	346	tube development
GO:0007218	0.002	0.000	6	0	28	neuropeptide signaling pathway
GO:0031401	0.002	0.654	76	55	386	positive regulation of protein modification process
GO:0048705	0.002	0.268	13	4	64	skeletal system morphogenesis
GO:0040008	0.003	0.584	46	30	234	regulation of growth

Figura 26. Análisis de enriquecimiento de los genes infra-expresados en la comparación IRvsIS en GO

A continuación, se muestran algunas figuras asociadas a las técnicas de Machine Learning utilizadas.

En primer lugar, se muestra un resumen del modelo creado con la técnica de MLP. Los datos iniciales estaban compuestos por 60 variables y la primera capa tenía 64 nodos, más 64 parámetros del concepto “bias”, en total se estimaban 3904 parámetros en la primera capa (60*64 +64 = 3.904). En la segunda capa oculta también hay 64 nodos, más los 64 parámetros del concepto “bias”, pues son 4.160 parámetros a estimar (64*64+64). Por último, en relación a la capa de salida, compuesta por 3 variables se estiman 195 parámetros (3*64+3 = 195).

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_2 (Dense)              (None, 64)                3904
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                4160
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 3)                 195
=================================================================
Total params: 8,259
Trainable params: 8,259
Non-trainable params: 0
_________________________________________________________________
```

Figura 27. Resumen del modelo MLP creado

A continuación, se muestra el modelo de red neuronal creado con 30 nodos en la primera capa y 20 en la segunda:

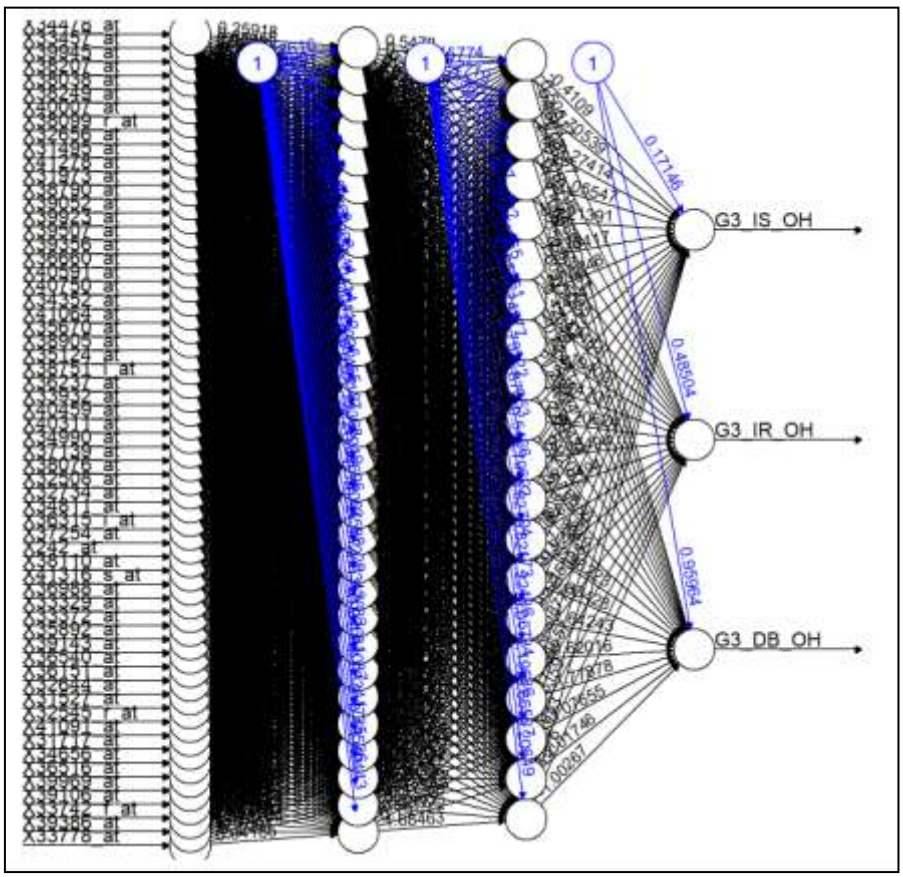

Figura 28. Resumen del modelo ANN creado con las 60 genes

La siguiente figura muestra el modelo de red neuronal utilizada, pero donde las variables de entrada han sido las 7 variables que presentaban un autovalor más alto obtenidas de realizar un ACP sobre el conjunto de los 60 genes:

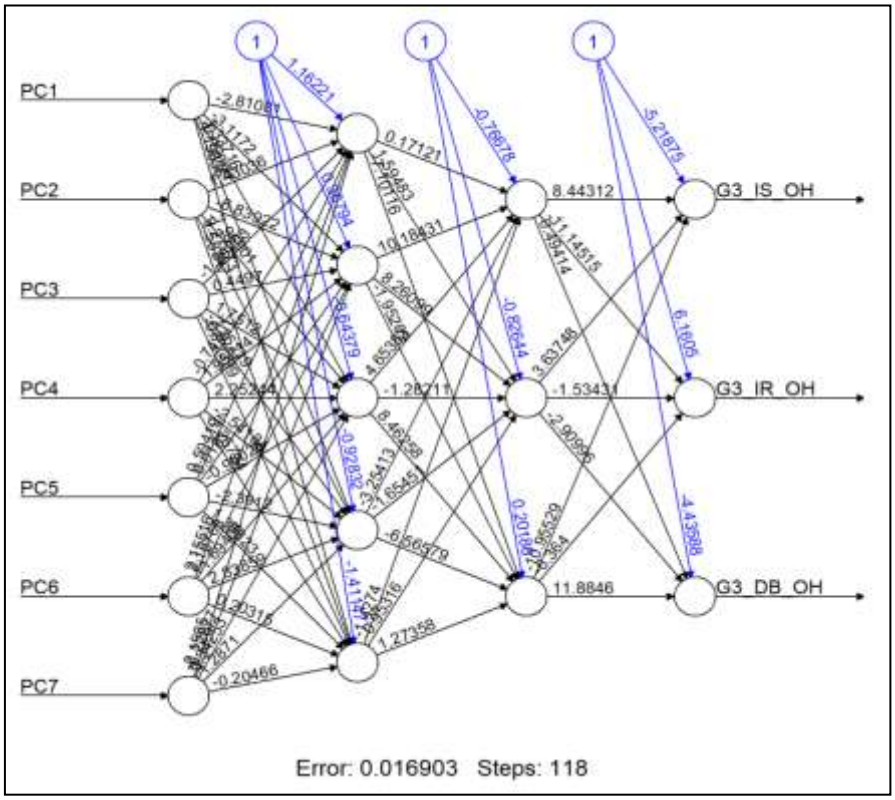

Figura 29. Resumen del modelo ANN creado con las 7 variables del ACP

En relación a la técnica RF, se muestran dos figuras. En la figura de la izquierda se indica que el bosque aleatorio incluyó 500 árboles. La línea negra representa el Out-of-bag (OOB), es una estimación de error para los casos que no se utilizaron durante la construcción del árbol. La línea roja es el error al intentar predecir la variable objetivo y la línea verde es el error en la predicción de la localización. En la figura de la derecha, se puede observar la importancia de las variables en el modelo predictivo en orden descendente.

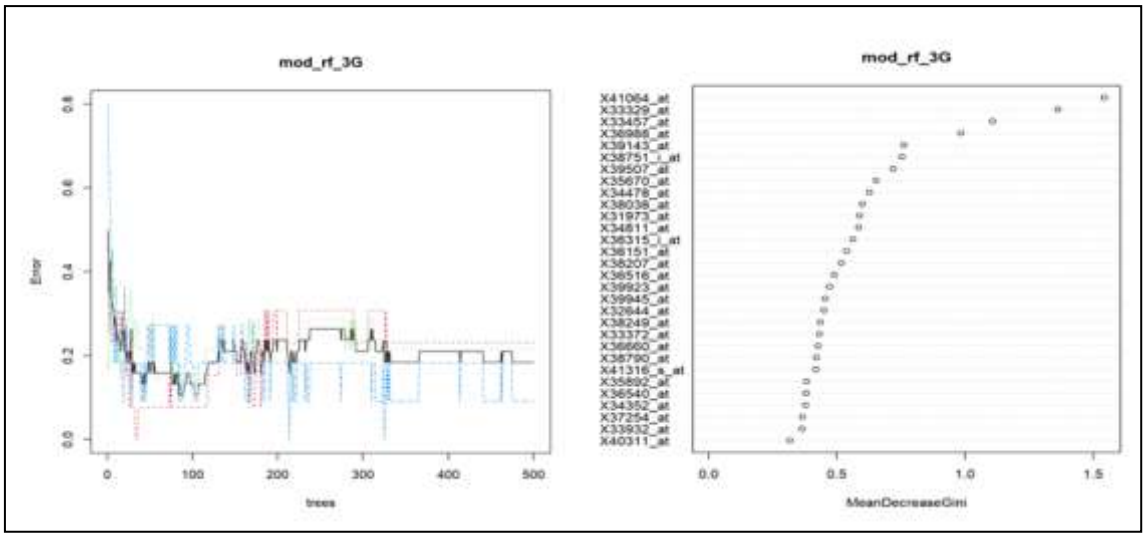

Figura 30. Resumen del modelo RF creado

Por último, se muestran las matrices de confusión acumuladas de dimensiones 3*3 y 2*2 para cada técnica después de las 1.000 ejecuciones y que han servido para calcular los resultados de la tabla 5.

```
                 real_MLP_3G
prediction_MLP_3G   IS   IR   DB
               IS 5798  105  149
               IR    7 5956   77
               DB  367   74 4467
        IS IR-DB
IS    5798   254
IR-DB  374 10574
```

```
prediction_KNN_3G   IS   IR   DB
               IS 4745   94   69
               IR 1100 5675  507
               DB  327  366 4117
        IS IR-DB
IS    4745   163
IR-DB 1427 10665
```

```
prediction_ANN_3G   IS   IR   DB
               IS 5405  249  180
               IR  435 5425  325
               DB  332  461 4188
        IS IR-DB
IS    5405   429
IR-DB  767 10399
```

```
prediction_ANN_3G_ACP   IS   IR   DB
                  IS 5533  414  309
                  IR  276 5453  239
                  DB  363  268 4145
        IS IR-DB
IS    5533   723
IR-DB  639 10105
```

```
prediction_SVM_3G_KG   IS   IR   DB
                  IS 5042   25   25
                  IR  803 5875  362
                  DB  327  235 4306
        IS IR-DB
IS    5042    50
IR-DB 1130 10778
```

```
prediction_SVM_3G_KL   IS   IR   DB
                  IS 5557  118  197
                  IR  310 5667  251
                  DB  305  350 4245
        IS IR-DB
IS    5557   315
IR-DB  615 10513
```

```
prediction_RF_3G   IS   IR   DB
              IS 4925  337  189
              IR  862 5210  874
              DB  385  588 3630
        IS IR-DB
IS    4925   526
IR-DB 1247 10302
```

Figura 31. Matrices de confusión acumuladas

Partiendo de los datos de la figura superior, se calcula la matriz de confusión 2x2 unificando las categorías IS e IR, como clase negativa y DB como clase positiva:

Técnica	Exactitud 3G	Exactitud 2G	Sens	Espec	VPP	VPN
MLP	95.42	96.08	95.18	96.42	91.01	98.13

KNN	85.51	92.54	87.73	94.37	85.59	95.27
ANN	88.34	92.36	89.24	93.56	84.08	95.80
ANN-PCA	89.01	93.06	88.32	94.87	86.79	95.52
SVM-radial	89.55	94.42	91.75	95.43	88.46	96.81
SVM-lineal	90.99	93.51	90.45	94.68	86.63	96.30
RF	80.97	88.02	77.35	92.09	78.86	91.43

Tabla 8. Resultados obtenidos a través de una matriz de confusión con datos acumulados unificando IS e IR como categoría negativa.

Se puede observar que los resultados no difieren mucho de los mostrados en la tabla 6, por lo que se podría concluir que las personas IR están en una situación intermedia entre las personas IS y las personas DB.